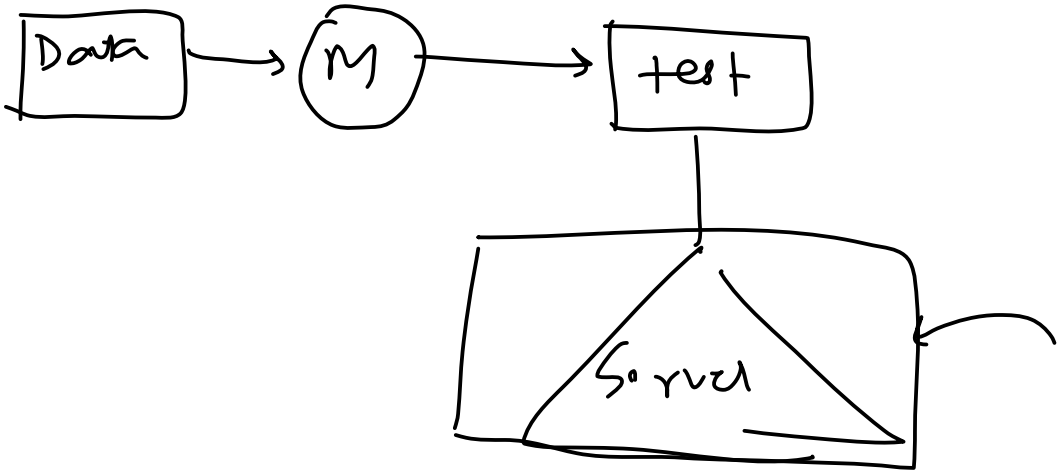
1. Batch Vs Online ML

Wednesday, March 17, 2021 5:30 PM

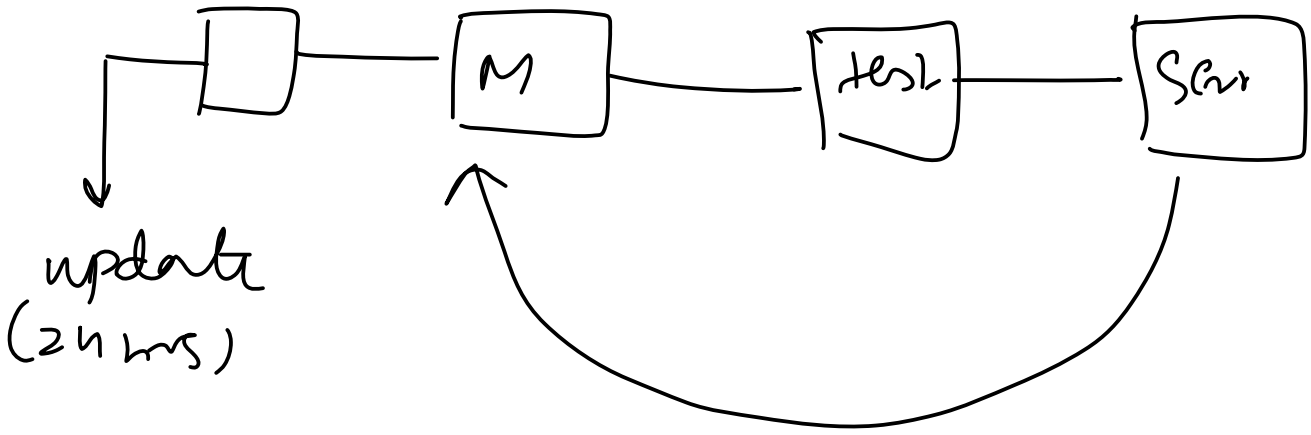
2. Batch/Offline ML

Wednesday, March 17, 2021 5:31 PM



3. The problem with Batch Learning

Wednesday, March 17, 2021 5:47 PM



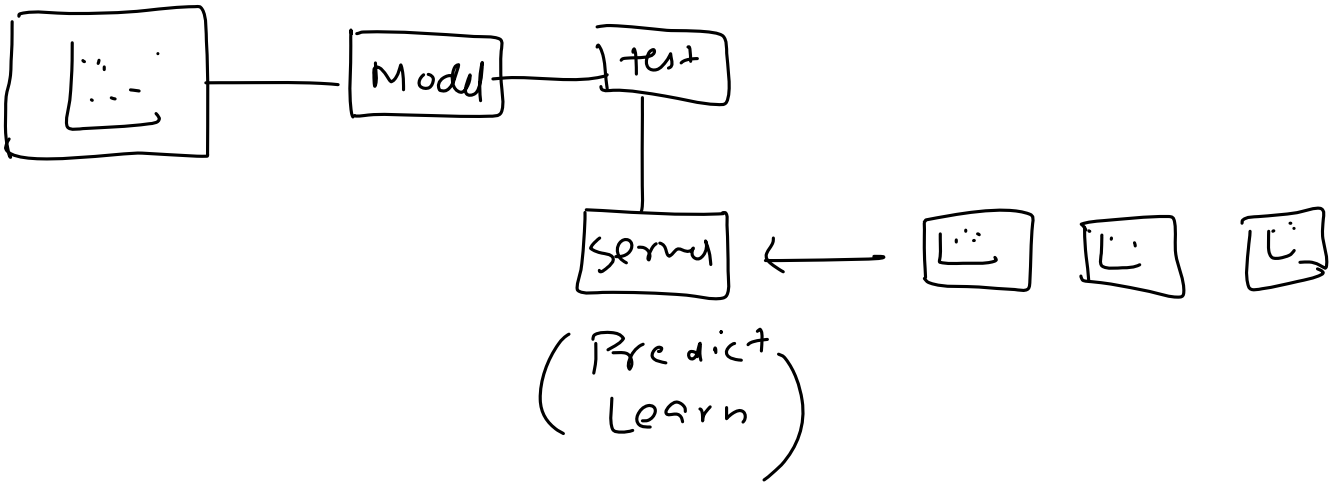
4. Disadvantages of Batch ML

Wednesday, March 17, 2021 5:32 PM

1. Lots of Data
2. Hardware Limitation
3. Availability

1. Online Machine Learning

Thursday, March 18, 2021 4:27 PM



2. When to use?

Thursday, March 18, 2021 4:33 PM

1. Where there is a concept drift
2. Cost Effective
3. Faster solution

3. How to implement?

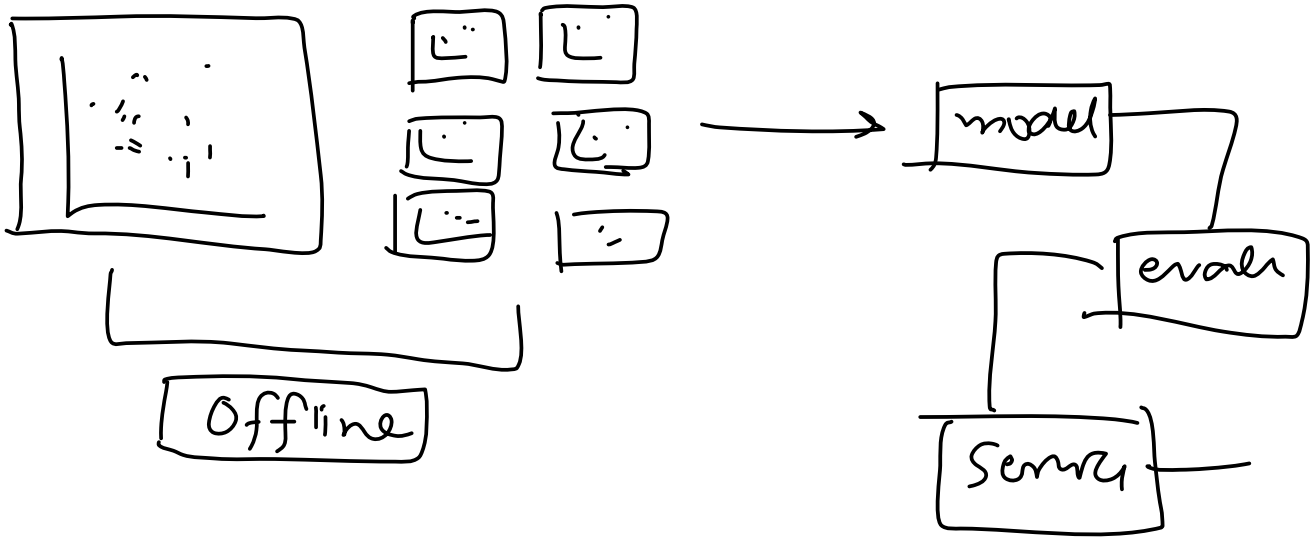
Thursday, March 18, 2021 4:28 PM

4. Learning Rate

Thursday, March 18, 2021 4:28 PM

5. Out of Core Learning

Thursday, March 18, 2021 4:28 PM



6. Disadvantage

Thursday, March 18, 2021 4:29 PM

1. Tricky to use
2. Risky

7. Batch Vs Online Learning

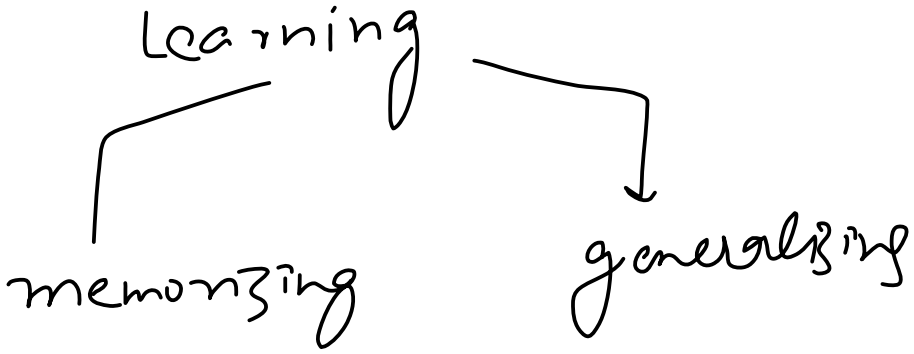
Thursday, March 18, 2021 4:29 PM

Offline Learning	Features	Online Learning
Less complex as model is constant	**Complexity**	Dynamic complexity as the model keeps evolving over time
Fewer computations, single time batch-based training	**Computational Power**	Continuous data ingestions result in consequent model refinement computations
Easier to implement	**Use in Production**	Difficult to implement and manage
Image Classification or anything related to Machine Learning - where data patterns remains constant without sudden concept drifts	**Applications**	Used in finance, economics, heath where new data patterns are constantly emerging
Industry proven tools. E.g. Sci-kit, TensorFlow, Pytorch, Keras, Spark Mlib	**Tools**	Active research/New project tools: E.g. MOA, SAMOA, scikit-multiflow, streamDM

Image courtesy - https://www.iunera.com/kraken/fabric/simple-introduction-to-online-learning-in-machine-learning/

1. Instance Vs Model Based Learning

Friday, March 19, 2021 4:05 PM

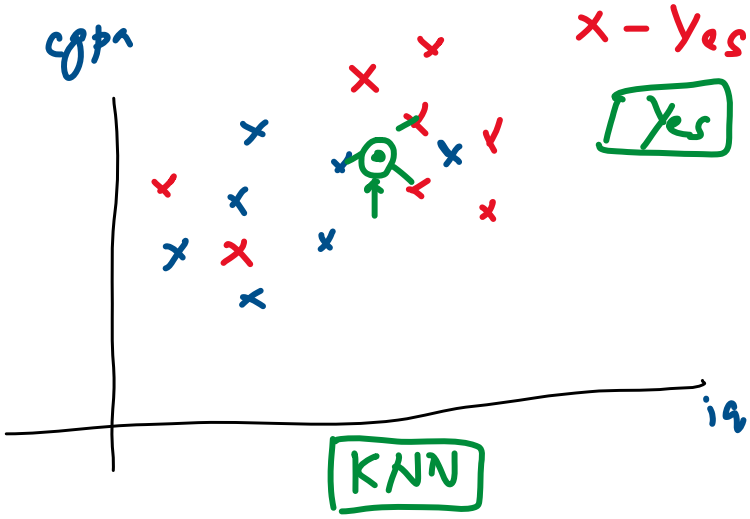


2. Instance Based

Friday, March 19, 2021 4:06 PM

iq	cgpa	placement Y/N
80	8	Y
70	7	N

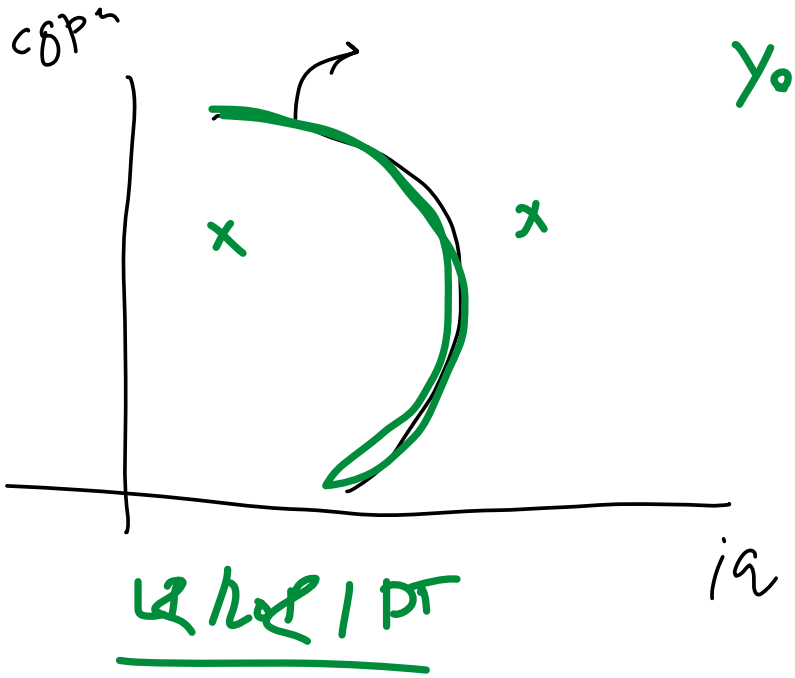
7.5, 103



3. Model Based

Friday, March 19, 2021 4:06 PM

iq | cgpa | placement



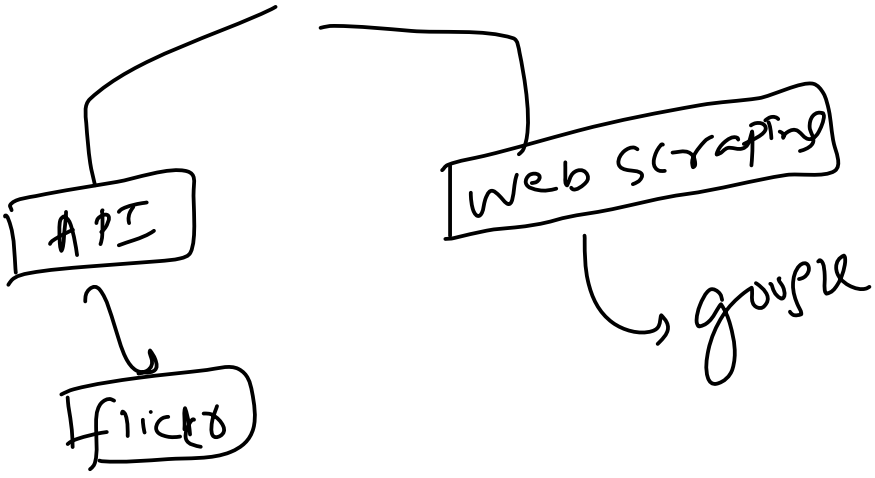
4. Differences

Friday, March 19, 2021 4:06 PM

Usual/Conventional Machine Learning	**Instance Based Learning**
Prepare the data for model training	Prepare the data for model training. No difference here
Train model from training data to estimate model parameters i.e. discover patterns	Do not train model. Pattern discovery postponed until scoring query received
Store the model in suitable form	There is no model to store
Generalize the rules in form of model, even before scoring instance is seen	No generalization before scoring. Only generalize for each scoring instance individually as and when seen
Predict for unseen scoring instance using model	Predict for unseen scoring instance using training data directly
Can throw away input/training data after model training	Input/training data must be kept since each query uses part or full set of training observations
Requires a known model form	May not have explicit model form
Storing models generally requires less storage	Storing training data generally requires more storage

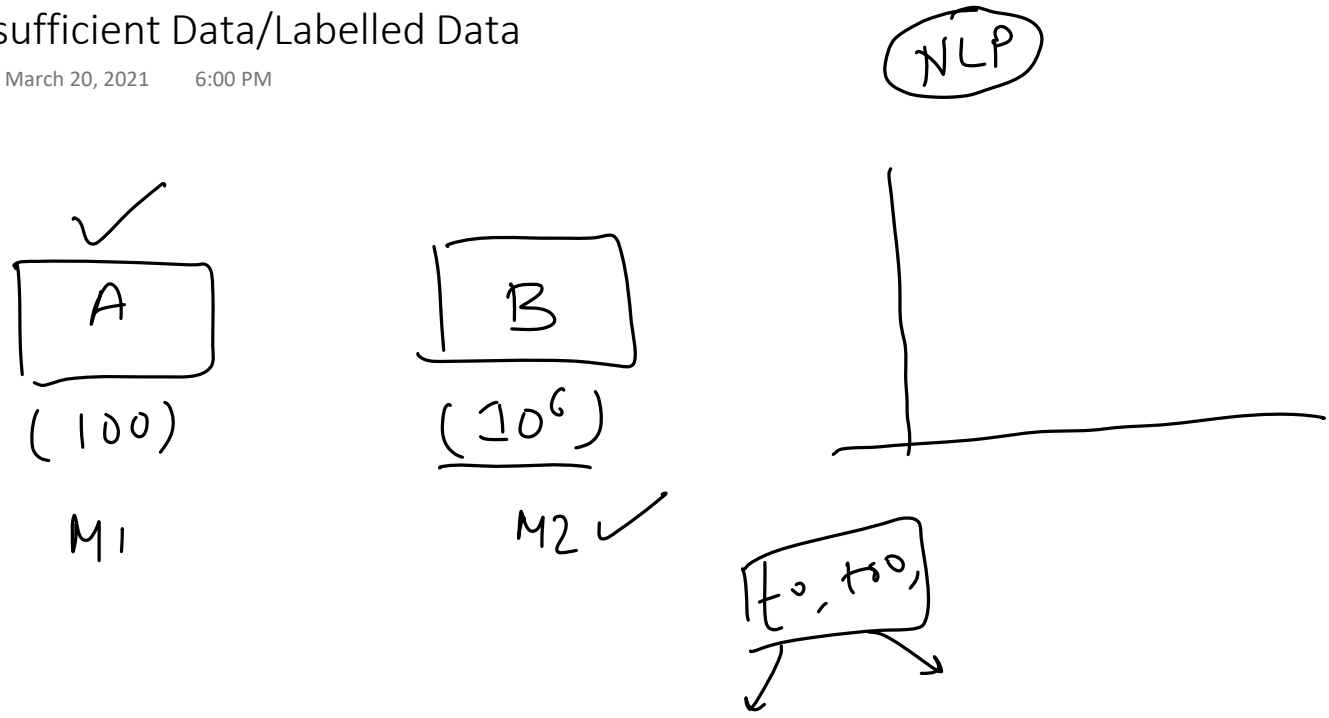
1. Data Collection

Saturday, March 20, 2021 5:59 PM



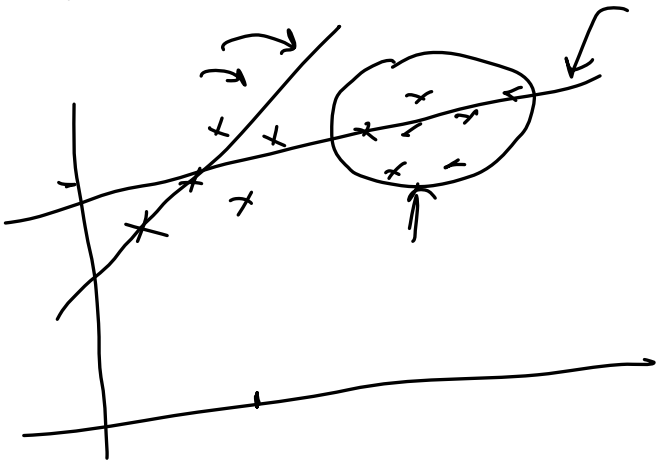
2. Insufficient Data/Labelled Data

Saturday, March 20, 2021 6:00 PM



3. Non Representative Data

Saturday, March 20, 2021 6:00 PM



Sampling Noise

Sampling bias

4. Poor Quality Data

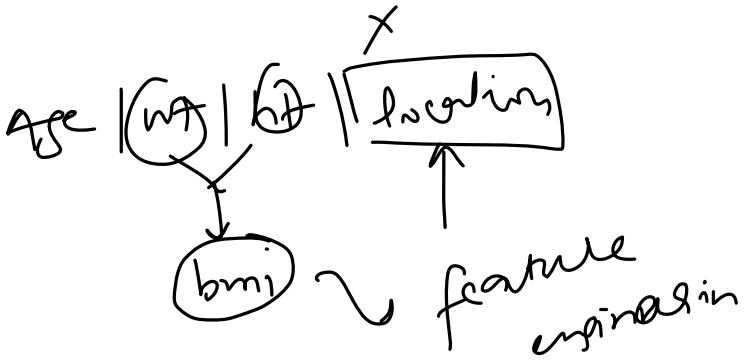
Saturday, March 20, 2021 6:00 PM

60 %

5. Irrelevant Features

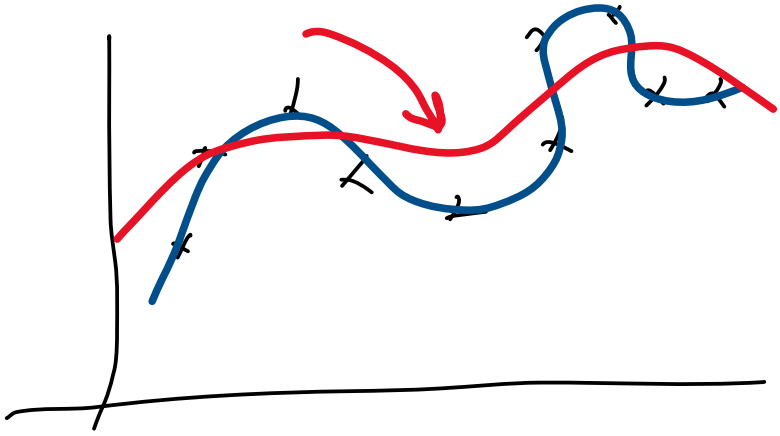
Saturday, March 20, 2021 6:00 PM

Garbage In
Garbage out



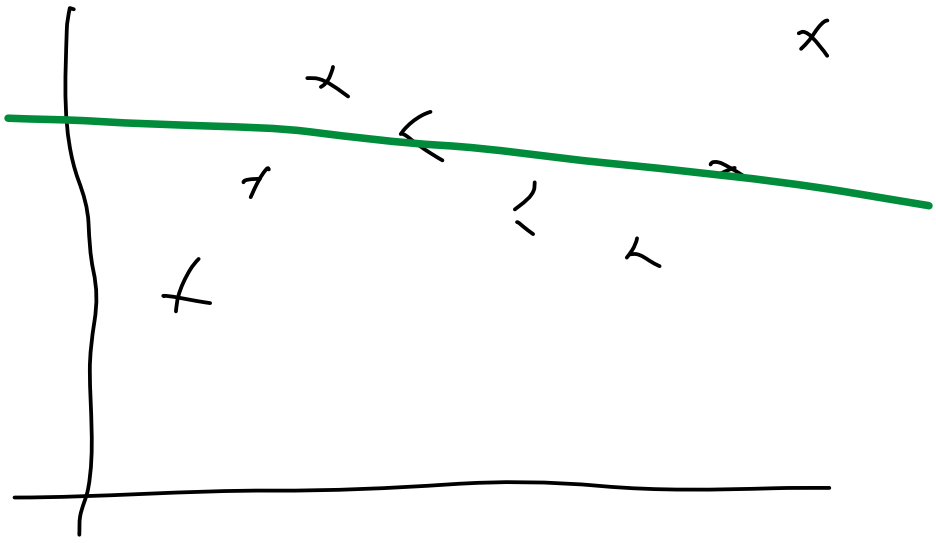
6. Overfitting

Saturday, March 20, 2021 6:01 PM



7. Underfitting

Saturday, March 20, 2021 6:01 PM

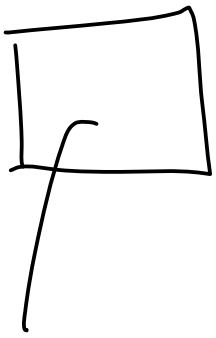


8. Software Integration

Saturday, March 20, 2021 6:01 PM

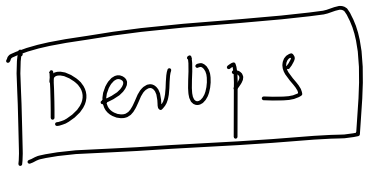
9. Offline Learning/ Deployment

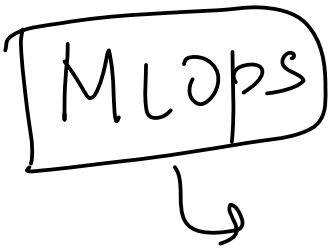
Saturday, March 20, 2021 6:01 PM



10. Cost Involved

Saturday, March 20, 2021 6:01 PM





1. Retail - Amazon/Big Bazaar

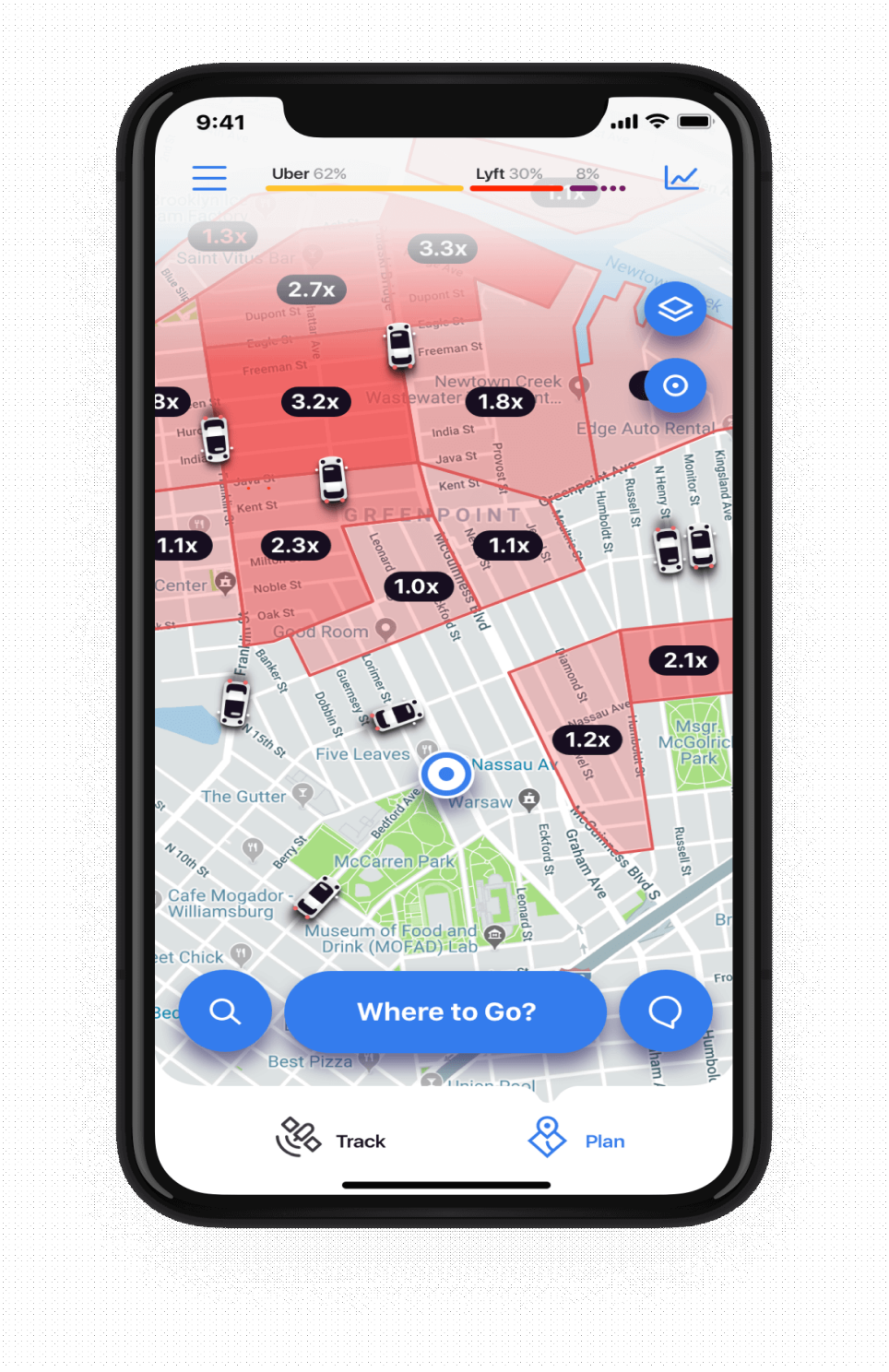
Monday, March 22, 2021 6:07 PM

2. Banking and Finance

Monday, March 22, 2021 6:07 PM

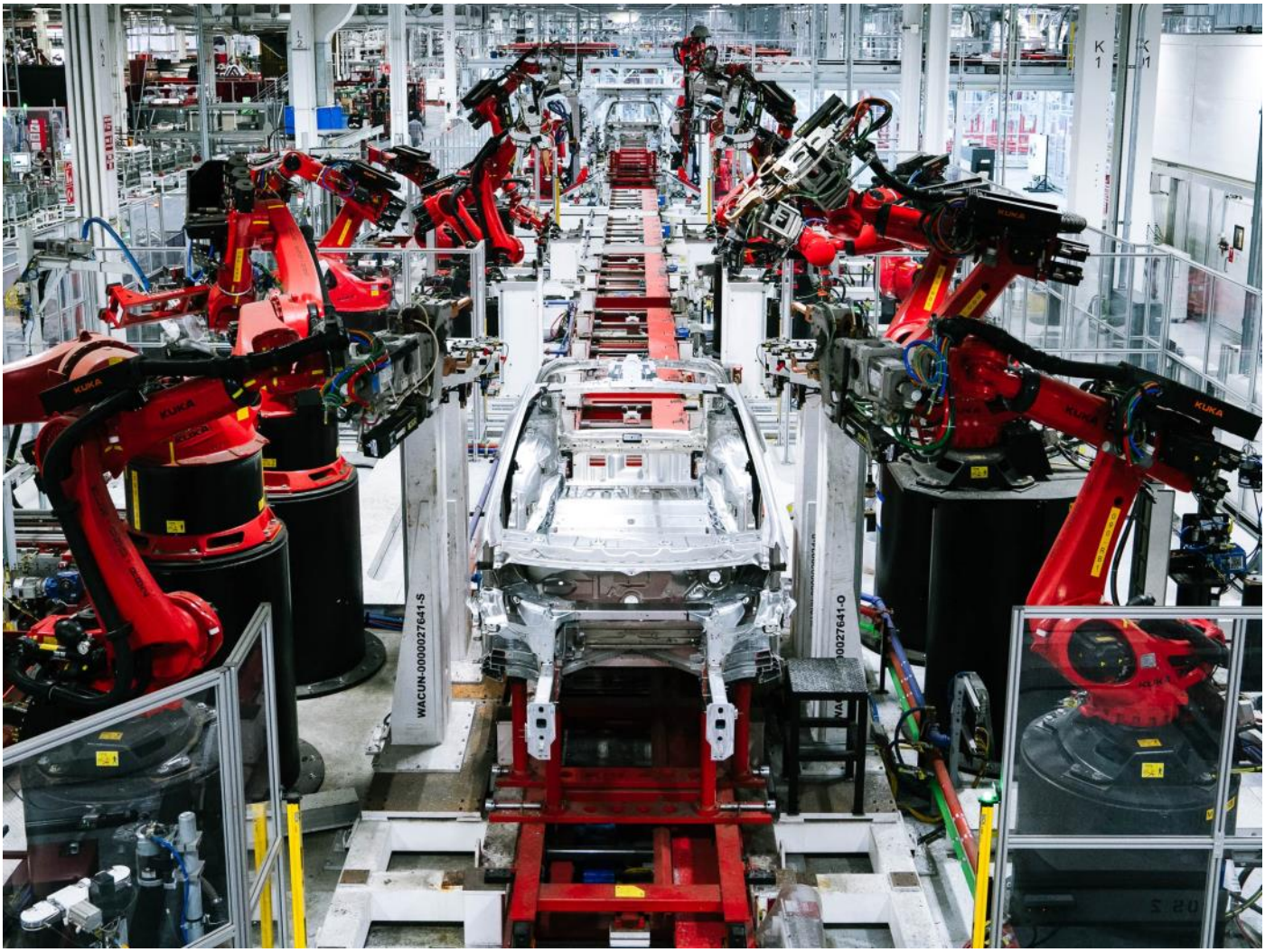
3. Transport - OLA

Monday, March 22, 2021 6:07 PM



4. Manufacturing - Tesla

Monday, March 22, 2021 6:08 PM



5. Consumer Internet - Twitter

Monday, March 22, 2021 6:08 PM

Machine Learning Development Life Cycle(MLDLC/MLDC)

Tuesday, March 23, 2021 12:09 PM

SDLC

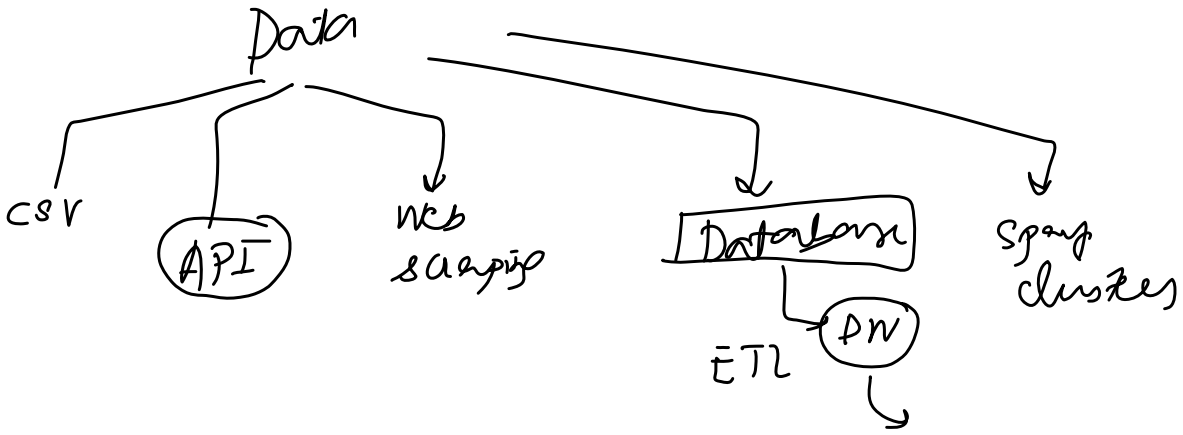
ML DLC

1. Frame the Problem

Tuesday, March 23, 2021 12:10 PM

2. Gathering Data

Tuesday, March 23, 2021 12:11 PM



3. Data Preprocessing

Tuesday, March 23, 2021 12:11 PM

→ Remove duplicates
→ Remove missing val
→ Outliers
→ Scale

4. Exploratory Data Analysis

Tuesday, March 23, 2021 12:11 PM

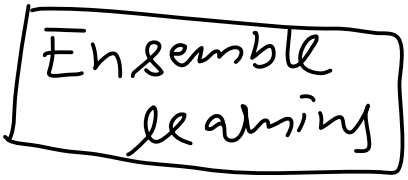
Viz s
Univariate / Biavariate
Outlier detection
Imbalance →

5. Feature Engineering and Selection

Tuesday, March 23, 2021 12:12 PM

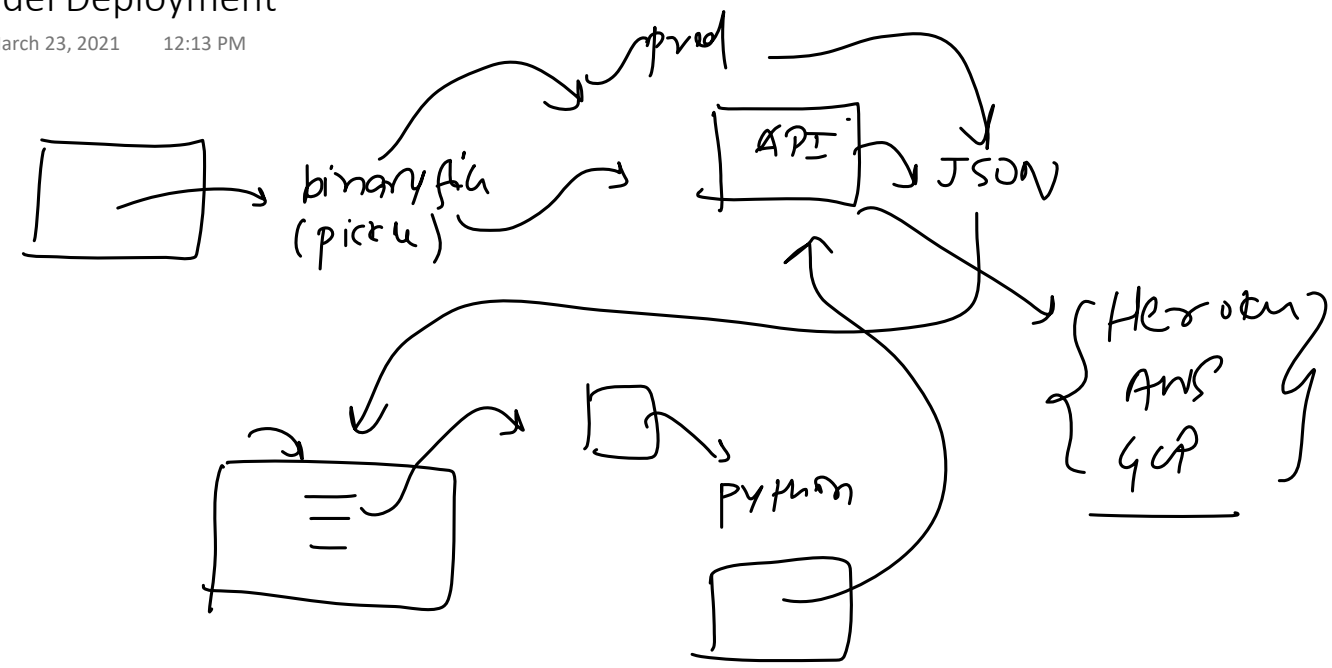
6. Model Training,Evalation and Selection

Tuesday, March 23, 2021 12:12 PM



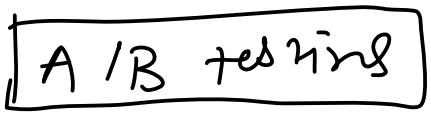
7. Model Deployment

Tuesday, March 23, 2021 12:13 PM



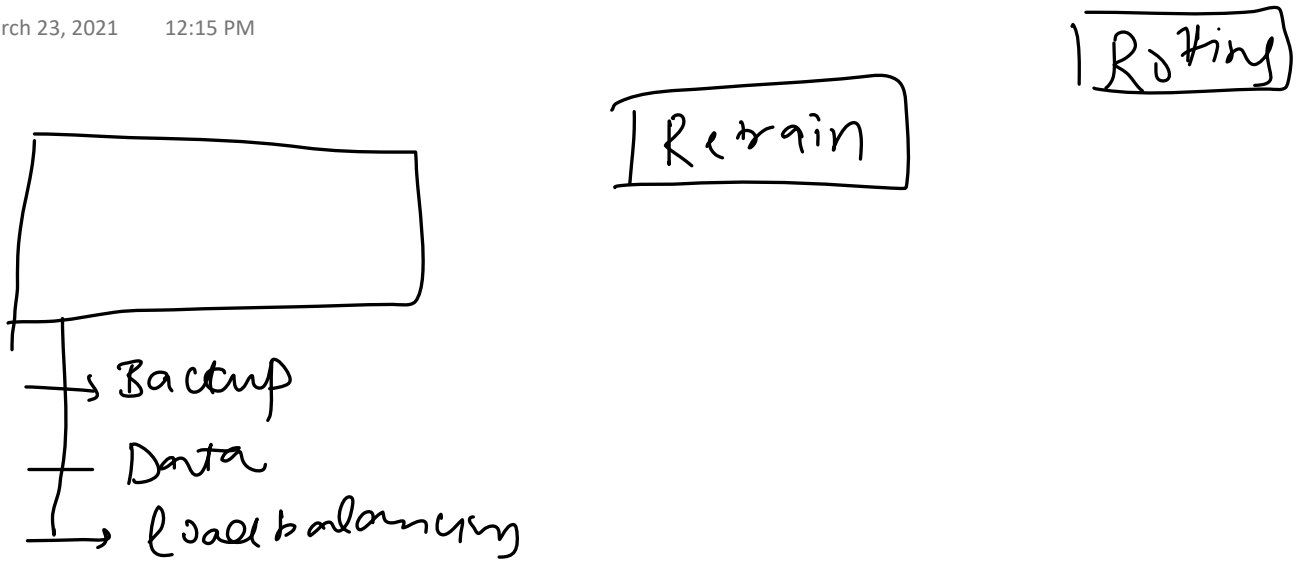
8. Testing

Tuesday, March 23, 2021 12:14 PM



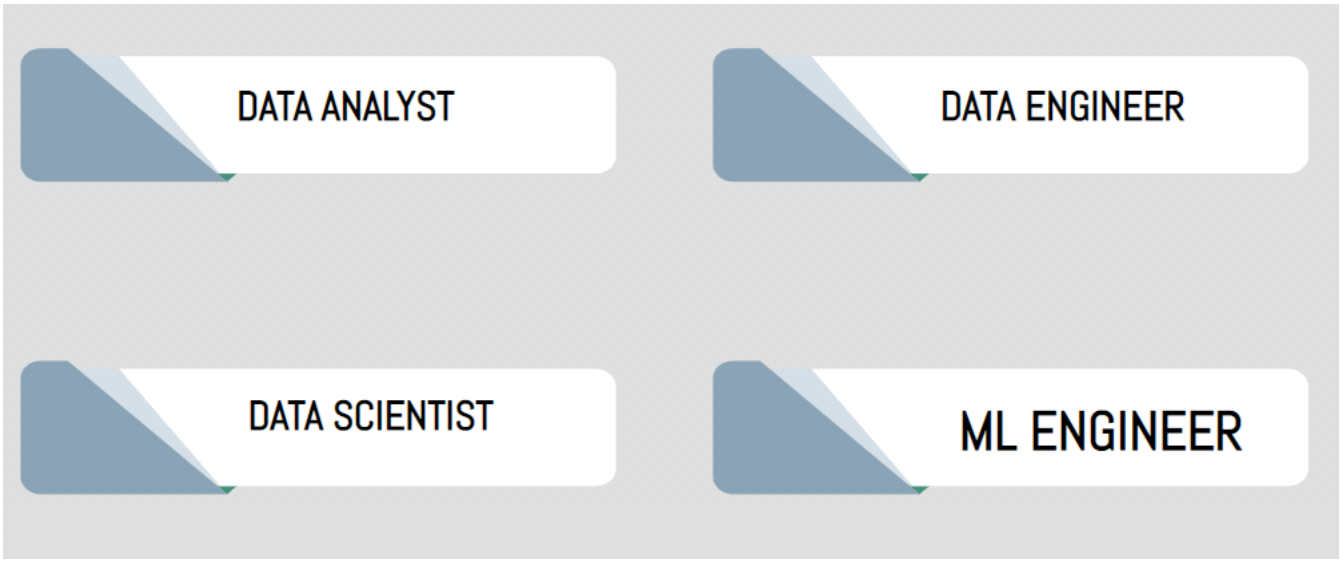
9. Optimize

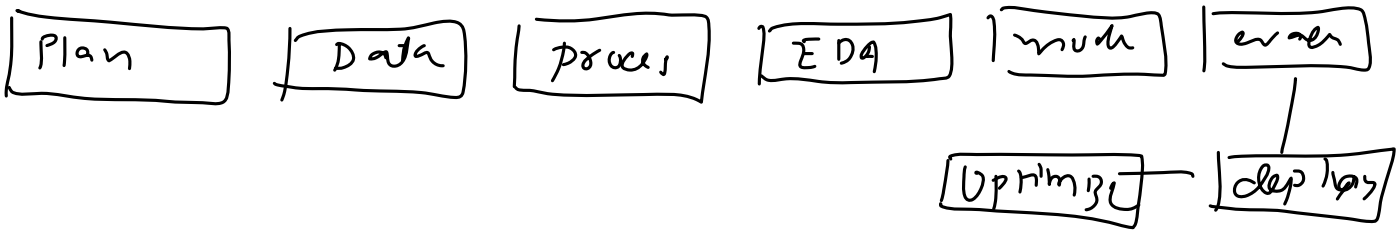
Tuesday, March 23, 2021 12:15 PM



1. Various Data Based Job Roles

Wednesday, March 24, 2021 1:25 PM



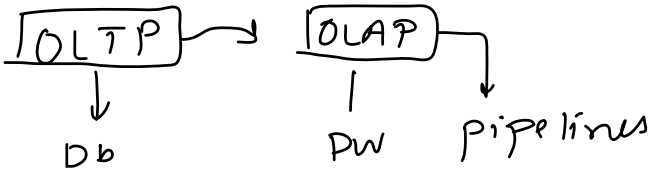


1. Data Engineer

Wednesday, March 24, 2021 1:25 PM

Job Roles

- Scrape Data from the given sources.
- Move/Store the data in optimal servers/warehouses.
- Build data pipelines/APIs for easy access to the data.
- Handle databases/data warehouses.



Skills Required

- Strong grasp of algorithms and data structures
- Programming Languages (Java/R/Python/Scala) and script writing
- Advanced DBMS's
- BIG DATA Tools (Apache Spark, Hadoop, Apache Kafka, Apache Hive)
- Cloud Platforms (Amazon Web Services, Google Cloud Platform)
- Distributed Systems
- Data Pipelines

2. Data Analyst

Wednesday, March 24, 2021 1:26 PM

Responsibilities of a Data Analyst

- *Cleaning and organizing Raw data.*
- *Analyzing data to derive insights.*
- *Creating data visualizations.*
- *Producing and maintaining reports.*
- *Collaborating with teams/colleagues based on the insight gained.*
- *Optimizing data collection procedures*

Skills

- *Statistical Programming*
- *Programming Languages (R/SAS/Python)*
- *Creative and Analytical Thinking*
- *Business Acumen — Medium to High preferred*
- *Strong Communication Skills.*
- *Data Mining, Cleaning, and Munging*
- *Data Visualization*
- *Data Story Telling*
- *SQL*
- *Advanced Microsoft Excel*

3. Data Scientist

Wednesday, March 24, 2021 1:26 PM

“A data scientist is someone who is better at statistics than any software engineer and better at software engineering than any statistician”.

4. ML Engineer

Wednesday, March 24, 2021 1:26 PM

Responsibilities

- Deploying machine learning models to production ready environment
- Scaling and optimizing the model for production
- Monitoring and maintenance of deployed models

Skills

- Mathematics
- Programming Languages (R/Python/Java/Scala mainly)
- Distributed Systems
- Data model and evaluation
- Machine Learning models
- Software Engineering & Systems design

5. Comparison

Wednesday, March 24, 2021 1:26 PM

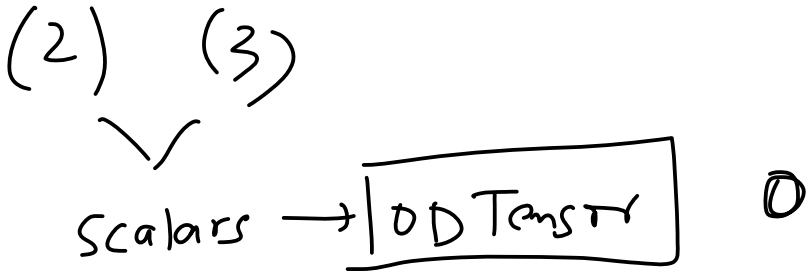
	ANALYTICAL SKILLS	BUSINESS ACUMEN	DATA STORYTELLING	SOFT SKILLS	SOFTWARE SKILLS
DATA ANALYST	HIGH	MEDIUM TO HIGH	HIGH	MEDIUM TO HIGH	MEDIUM
DATA ENGINEER	MEDIUM	LOW	LOW	MEDIUM	HIGH
DATA SCIENTIST	HIGH	HIGH	HIGH	HIGH	MEDIUM
ML ENGINEER	MEDIUM TO HIGH	MEDIUM	LOW	HIGH	HIGH

1. What are Tensors

Thursday, March 25, 2021 4:44 PM

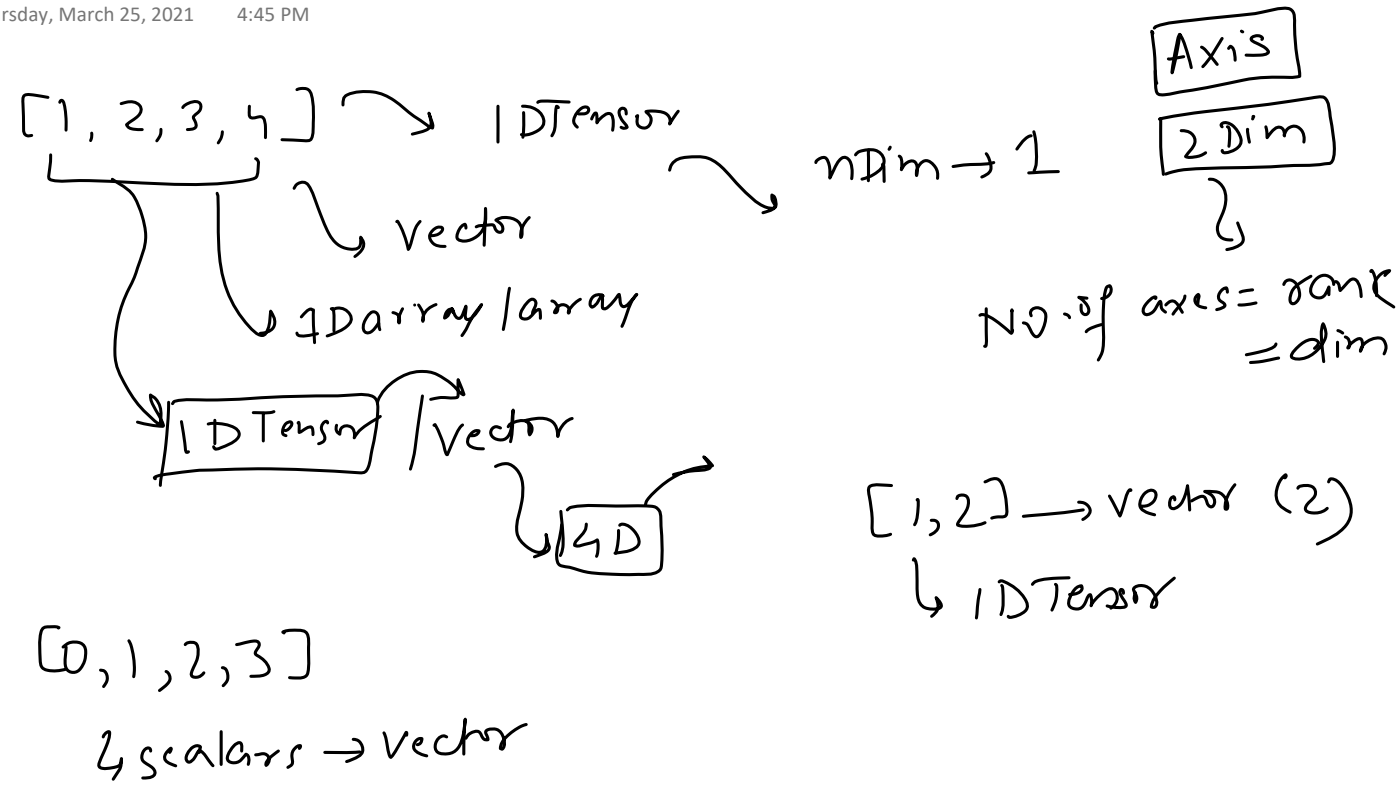
2. 0D Tensor/Scalar

Thursday, March 25, 2021 4:44 PM



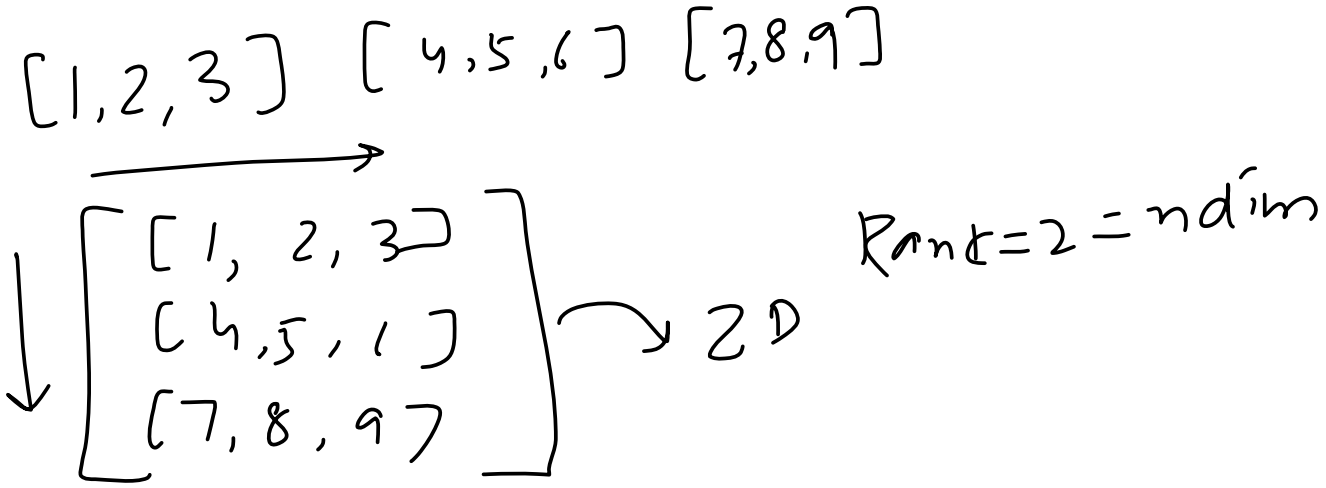
3. 1D Tensor/Vector

Thursday, March 25, 2021 4:45 PM



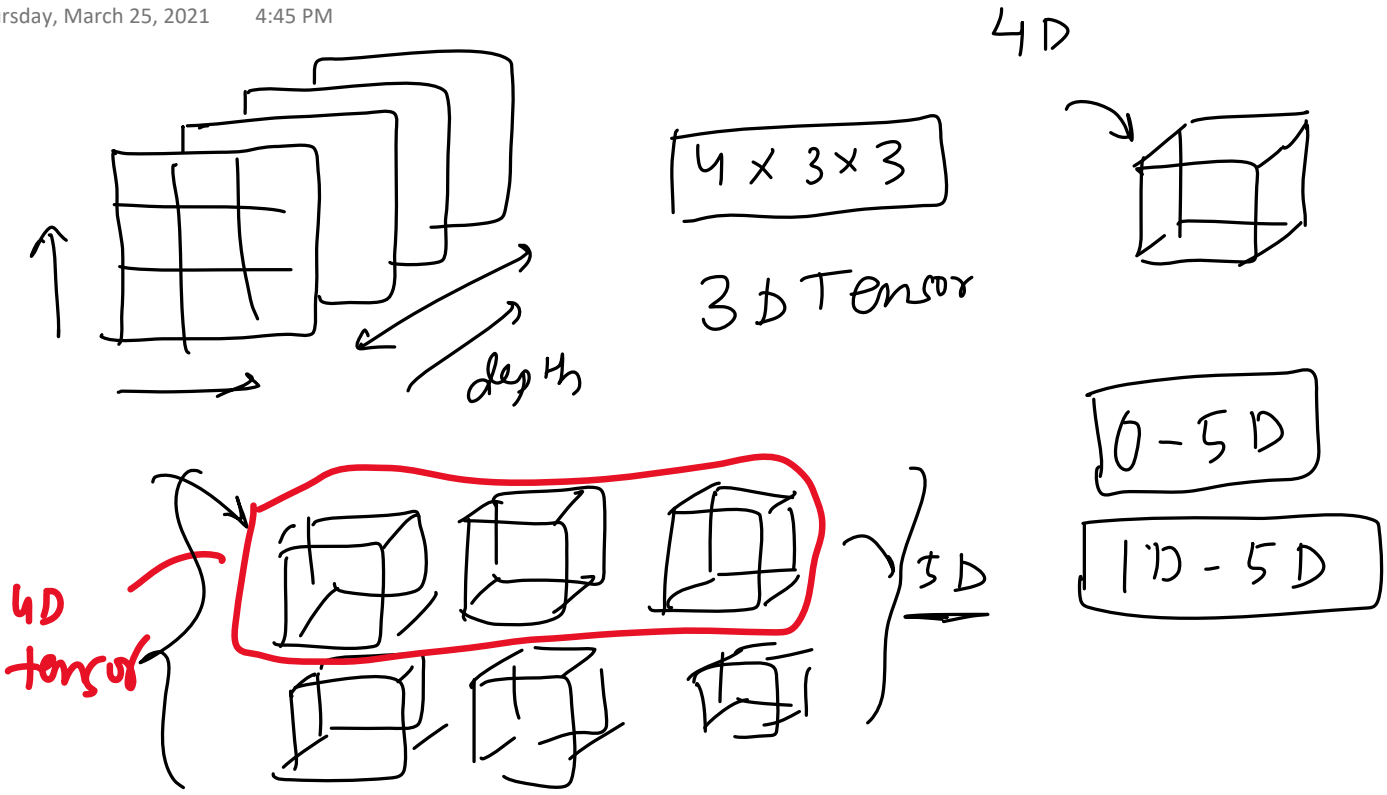
4. 2D Tensor/Matrices

Thursday, March 25, 2021 4:45 PM



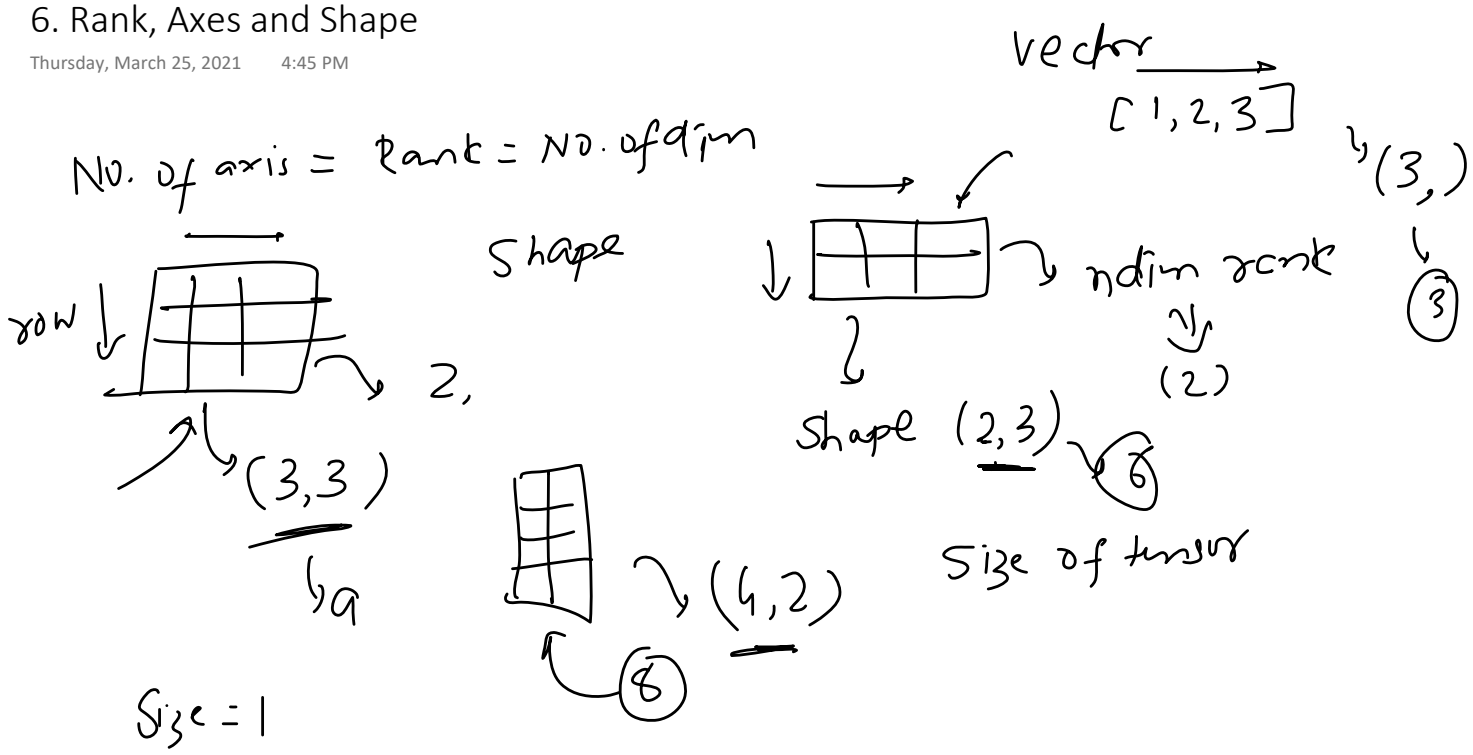
5. ND Tensors

Thursday, March 25, 2021 4:45 PM



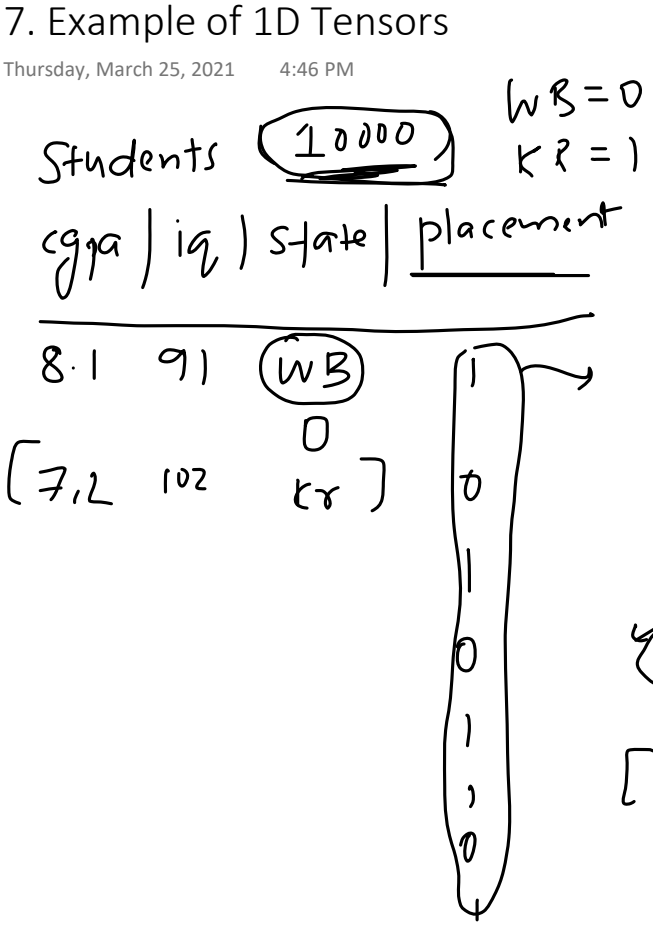
6. Rank, Axes and Shape

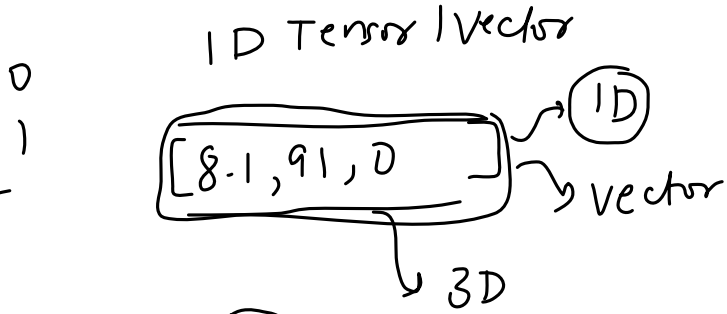
Thursday, March 25, 2021 4:45 PM

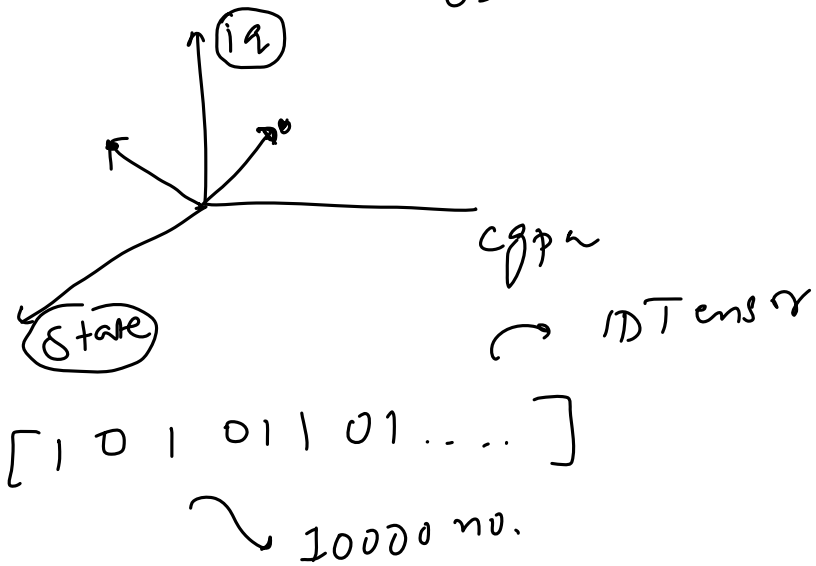


7. Example of 1D Tensors

Thursday, March 25, 2021 4:46 PM

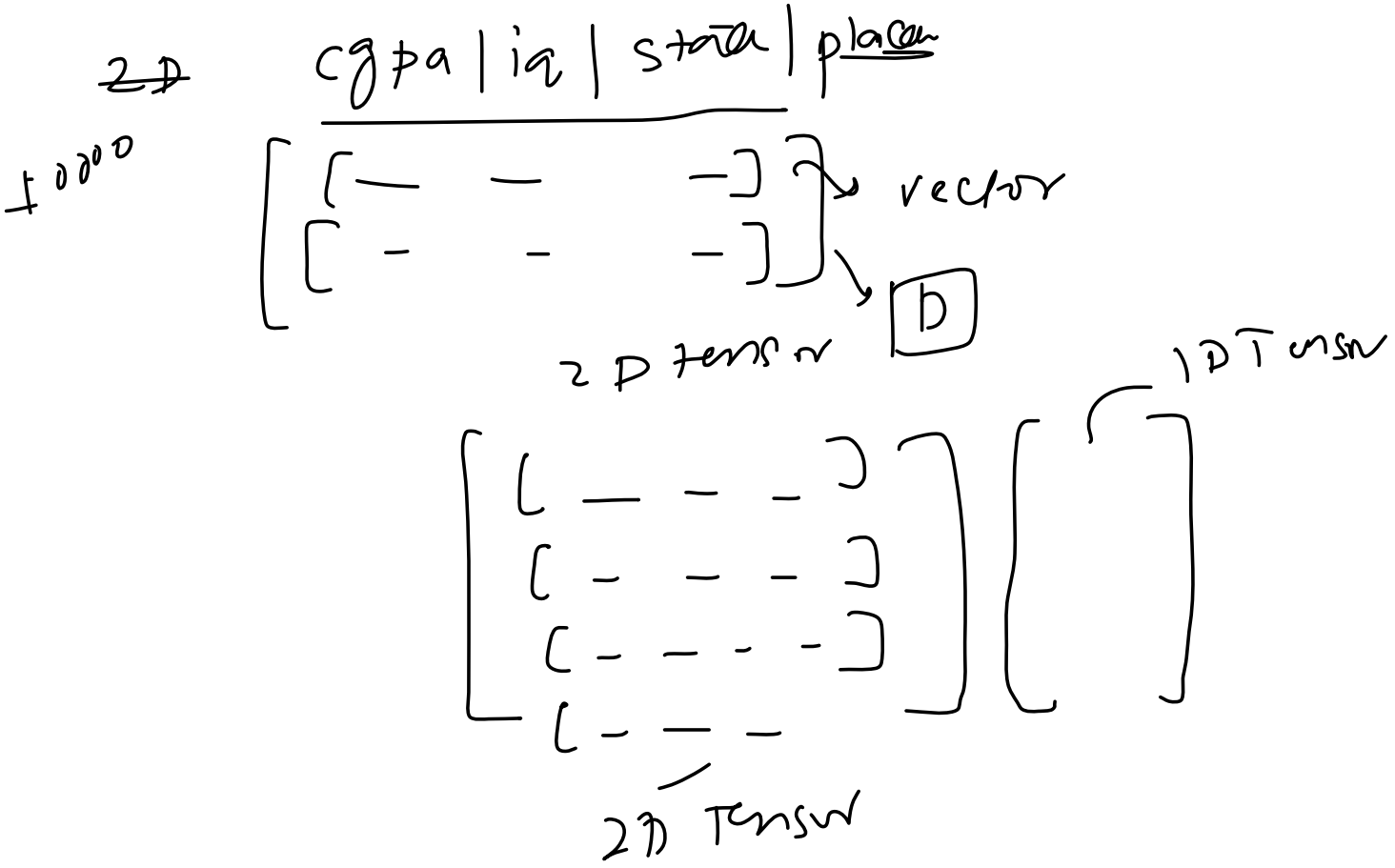






8. Example of 2D Tensors

Thursday, March 25, 2021 4:46 PM



9. Example of 3D Tensors

Thursday, March 25, 2021 4:46 PM

NLP

Hi Nitish

Hi Rahul

Hi Ankit

Hi	Nitish	Rahul	Ankit
1	0	0	0
0	1	0	0
0	0	1	0
0	0	0	1

[[[1,0,0,0],[0,1,0,0]]

(3,2,4)

[[1,0,0,0],[0,0,1,0]] → 3D Tensor

[[1,0,0,0],[0,0,0,1]]]

Timeseries Data

Highest | Lowest

10 years

(365,2)

Day 1 — —

Day 2 — — → 2D → 10 →

Day 365 — —

365

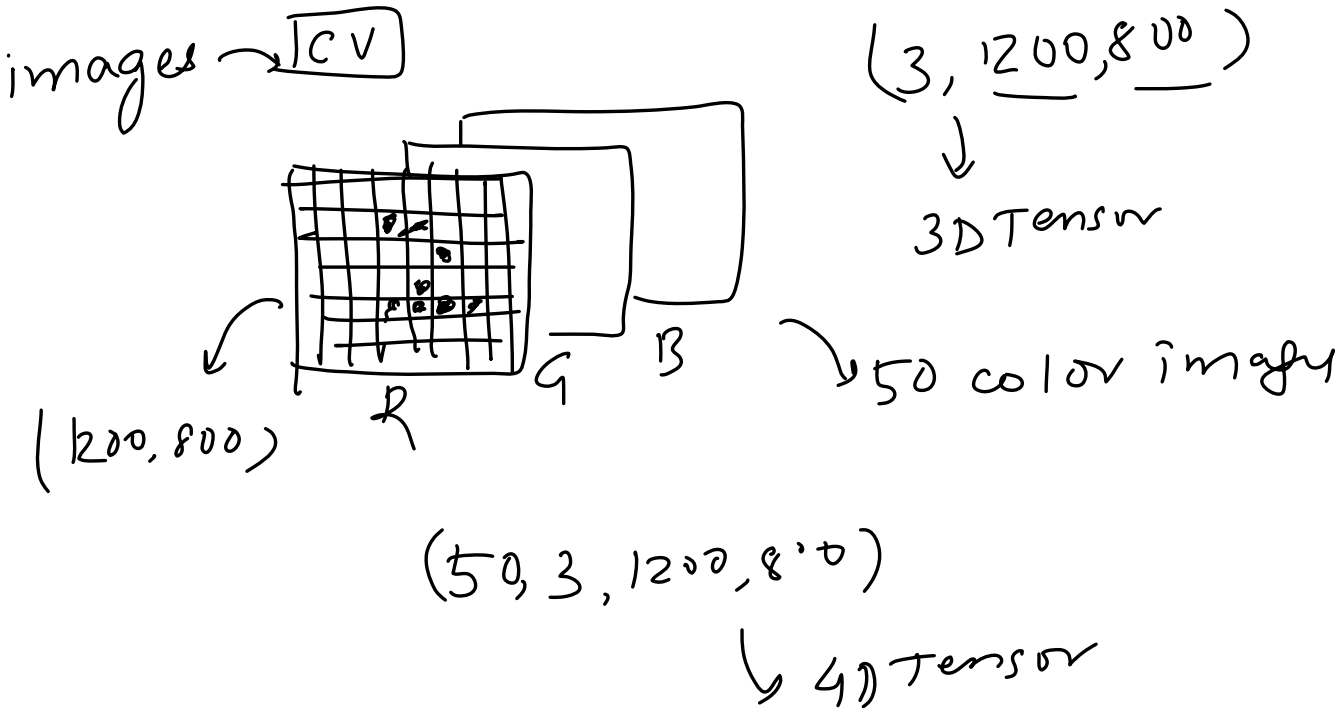
(10, 365, 2)

time axis

→ 3D Tensor

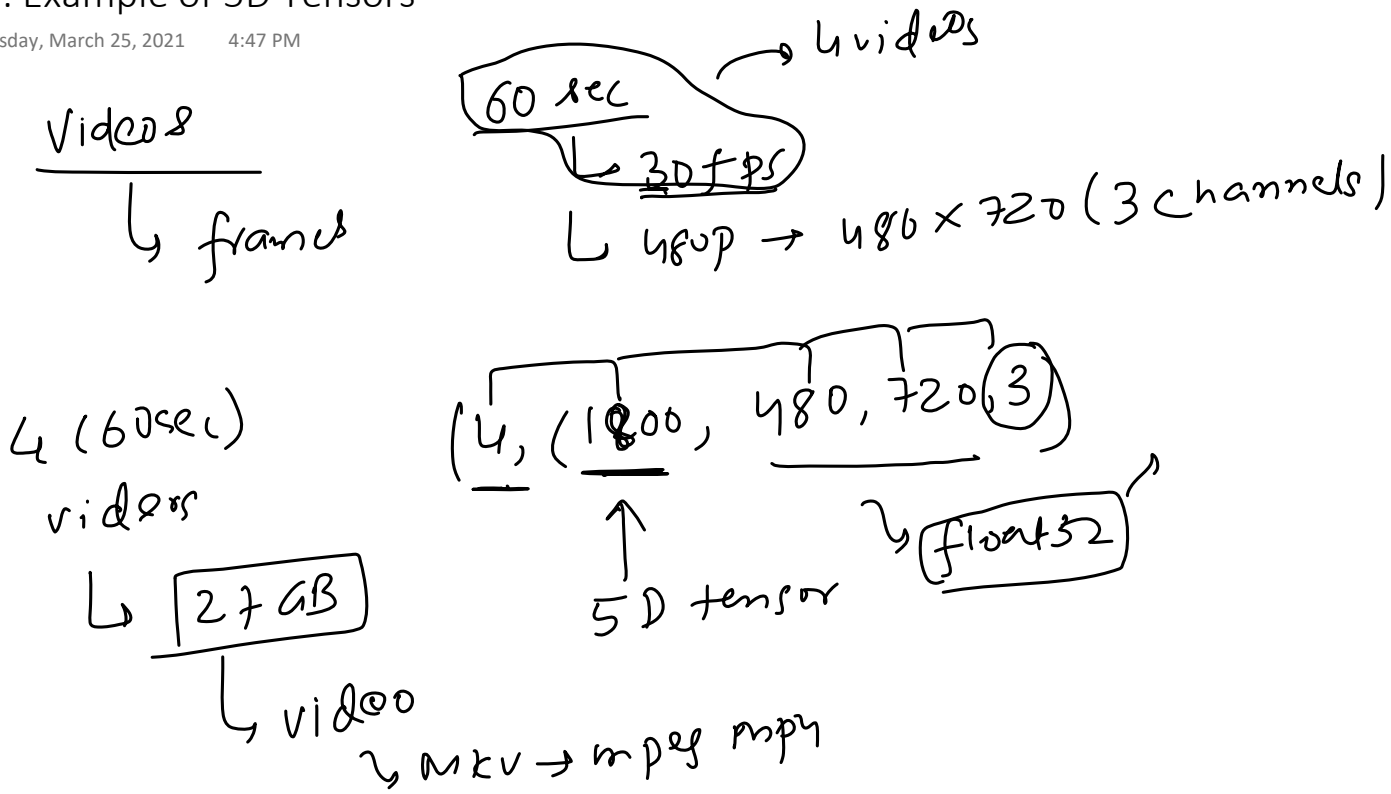
10. Example of 4D Tensors

Thursday, March 25, 2021 4:47 PM



12. Example of 5D Tensors

Thursday, March 25, 2021 4:47 PM



1. Installing Anaconda

Friday, March 26, 2021 5:40 PM

2. Jupyter Notebook Intro

Friday, March 26, 2021 5:40 PM

3. Virtual Env

Friday, March 26, 2021 5:40 PM

4. Using Kaggle

Friday, March 26, 2021 5:41 PM

5. Using Google Colab

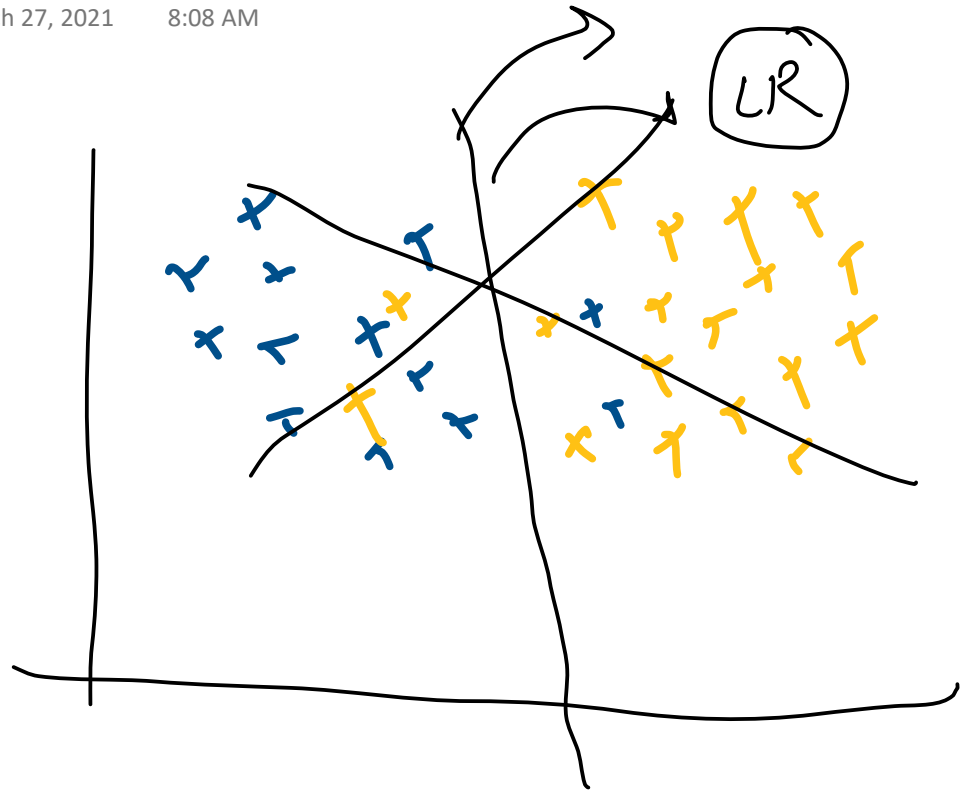
Friday, March 26, 2021 5:41 PM

6. Running Kaggle Data on Google Colab

Friday, March 26, 2021 5:41 PM

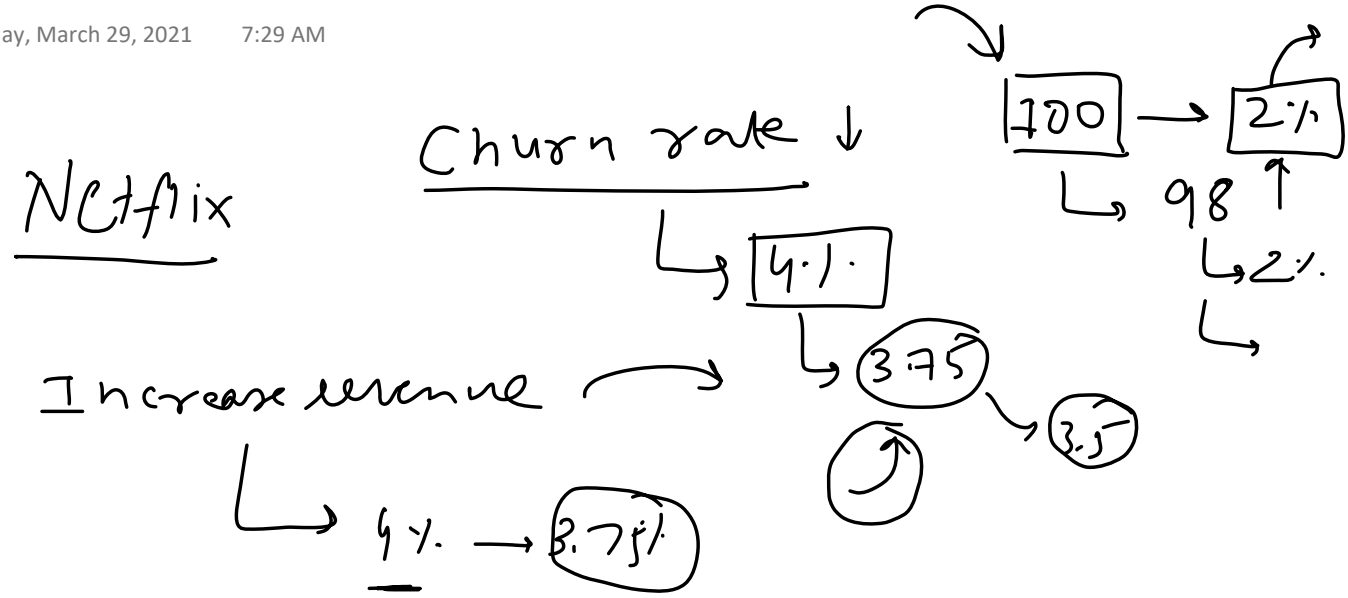
End to End Example

Saturday, March 27, 2021 8:08 AM



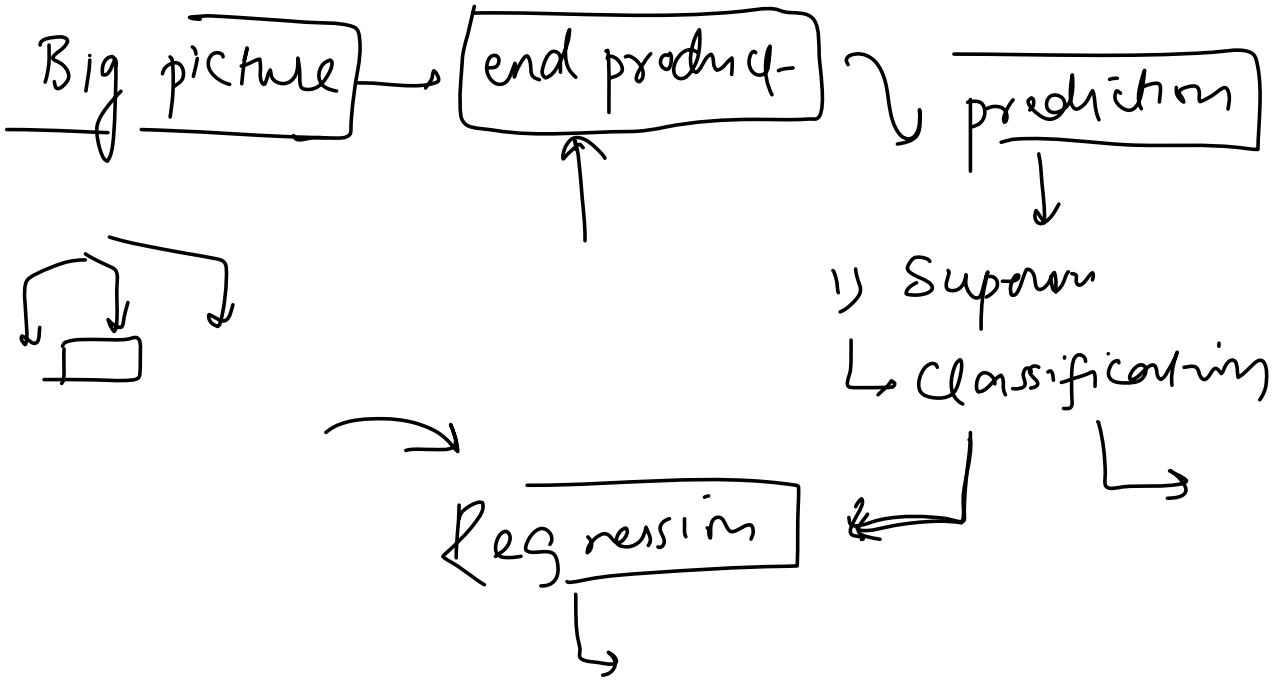
1. Business Problem to ML Problem

Monday, March 29, 2021 7:29 AM



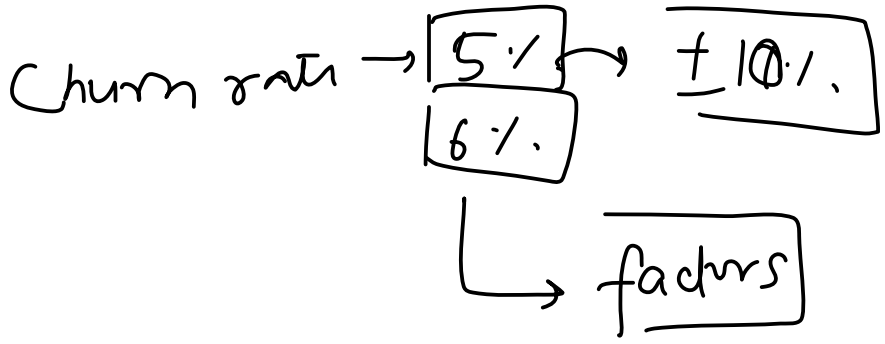
2. Type of Problem

Monday, March 29, 2021 7:30 AM



3. Current Solution

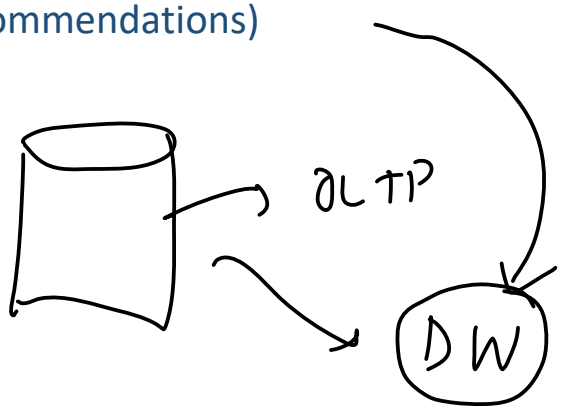
Monday, March 29, 2021 7:30 AM



4. Getting Data

Monday, March 29, 2021 7:30 AM

1. Watch time
2. Search but did not find
3. Content left in the middle
4. Clicked on recommendations(order of recommendations)

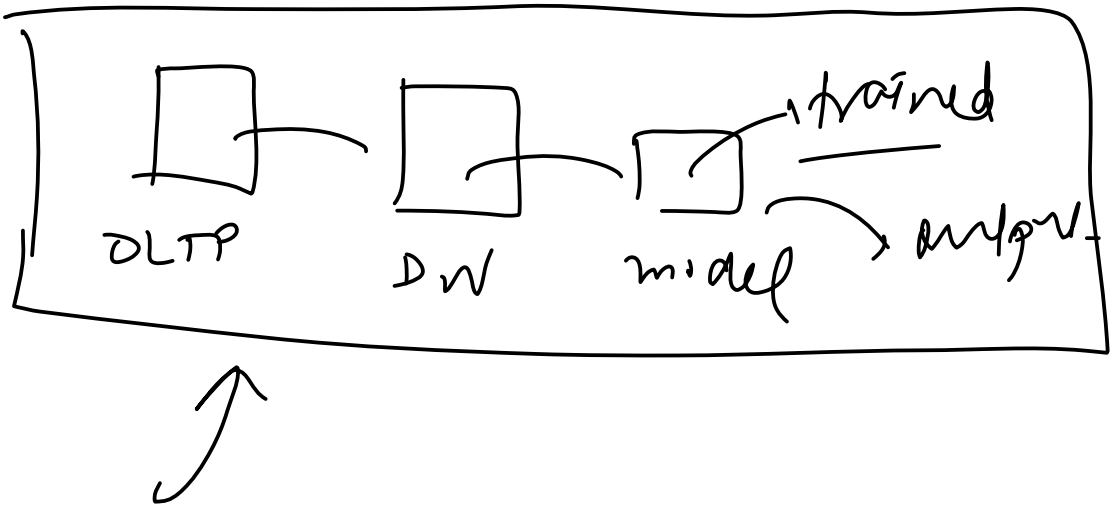


5. Metrics to measure

Monday, March 29, 2021 7:30 AM

6. Online Vs Batch?

Monday, March 29, 2021 7:31 AM

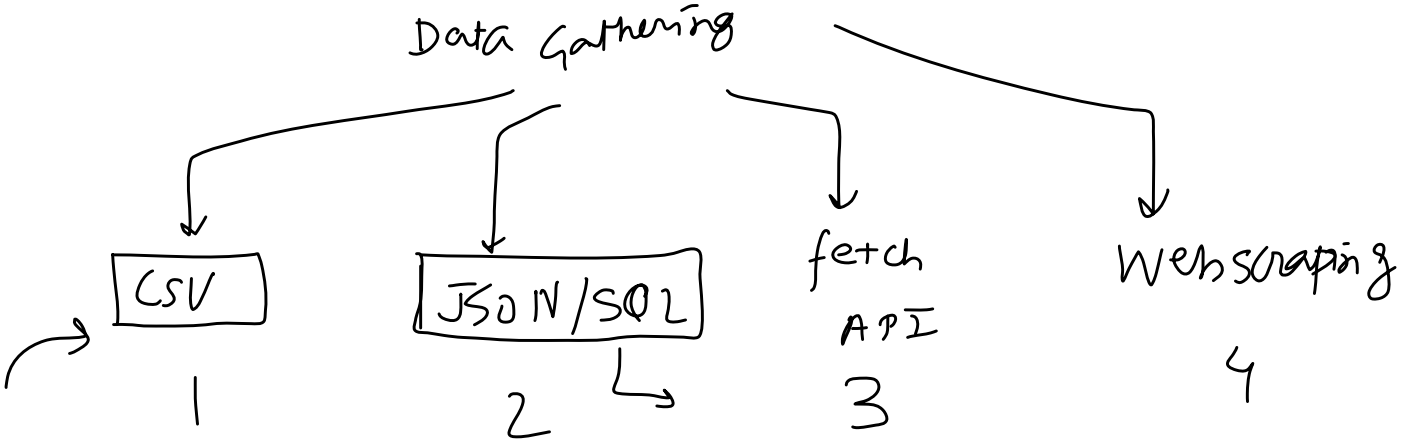


7. Check Assumptions

Monday, March 29, 2021 7:31 AM

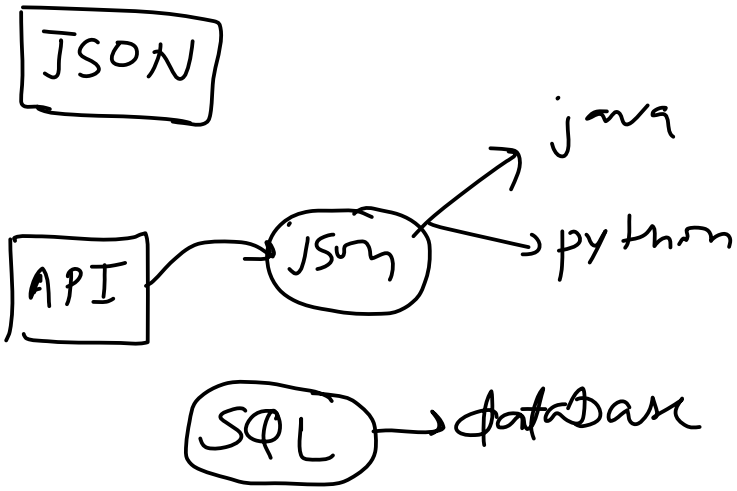
Gathering Data

Tuesday, March 30, 2021 5:35 PM



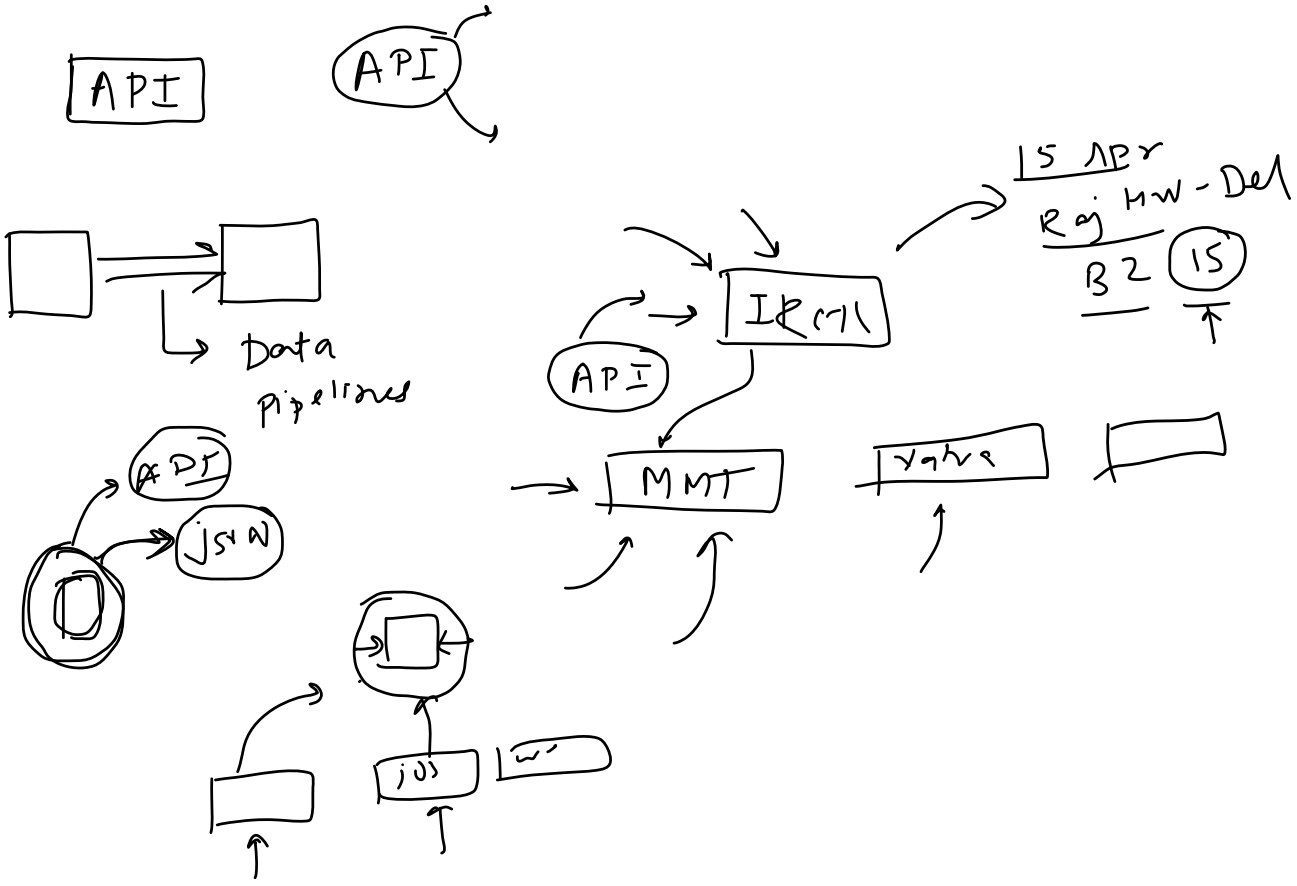
JSON/SQL

Wednesday, March 31, 2021 8:10 AM



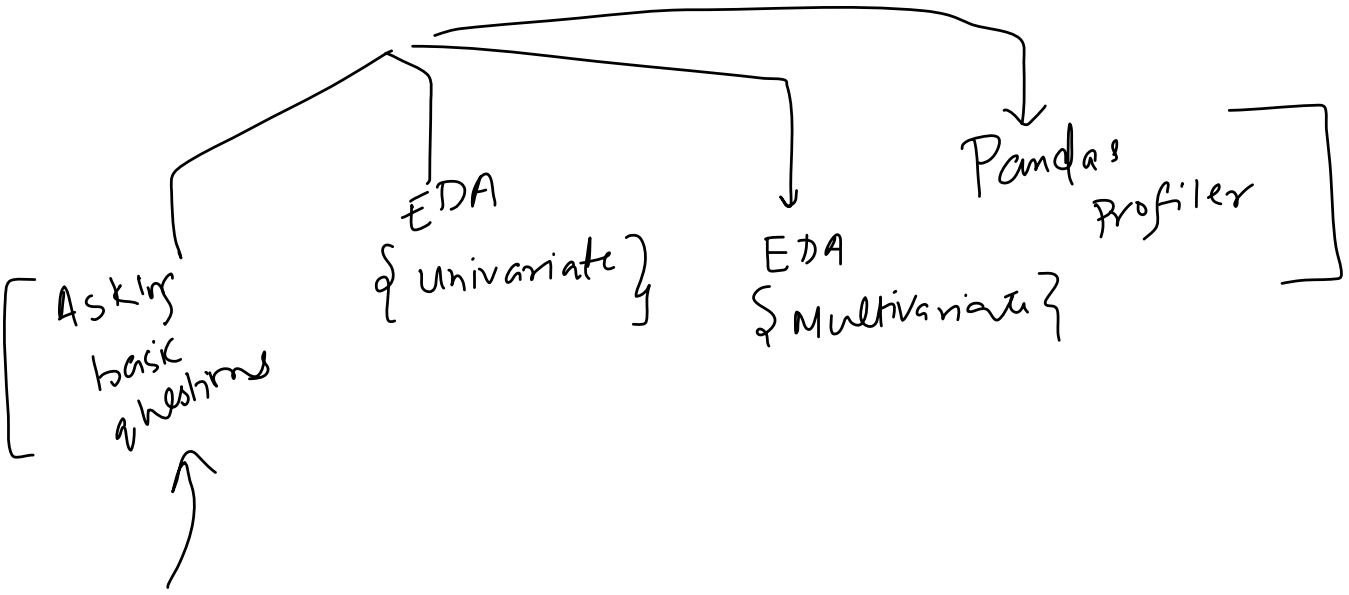
What is an API

Thursday, April 1, 2021 6:55 AM



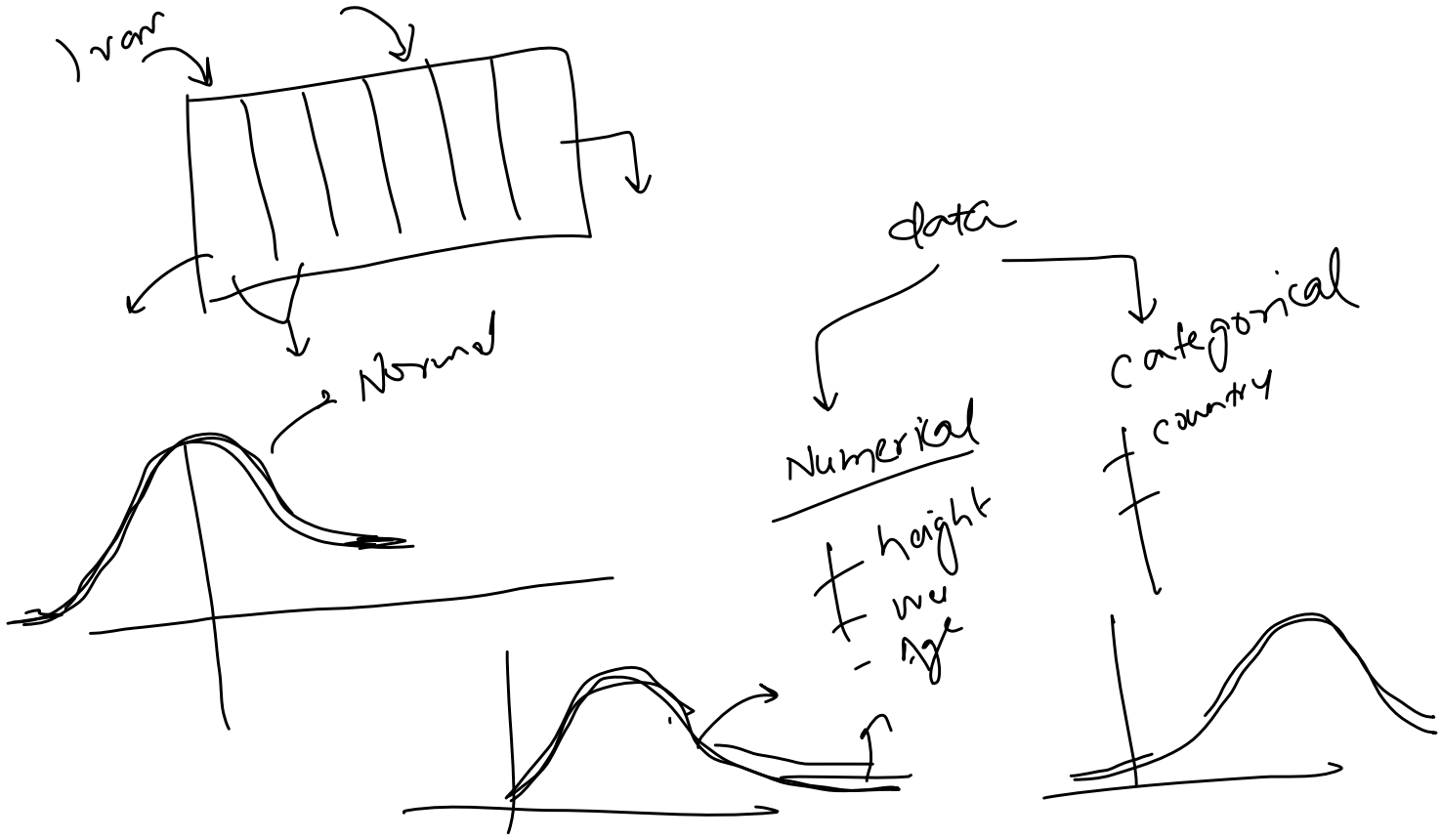
Understanding Your Data

Saturday, April 3, 2021 9:07 AM



Univariate Analysis

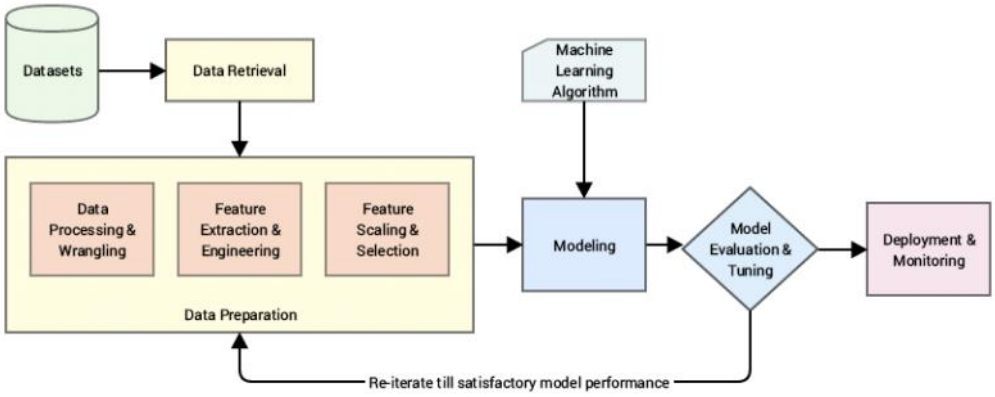
Sunday, April 4, 2021 5:39 PM

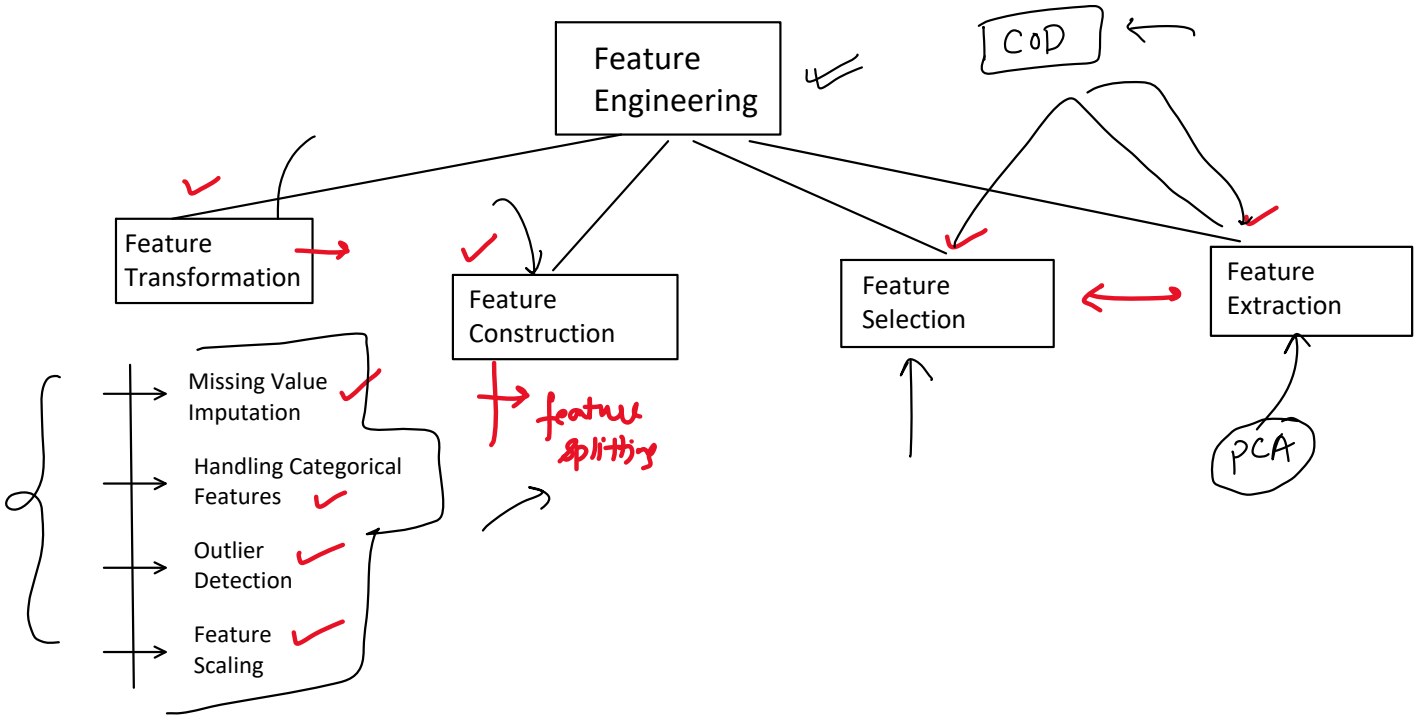


Feature Engineering

Thursday, April 8, 2021 6:02 PM

Feature engineering is the process of using domain knowledge to extract features from raw data. These features can be used to improve the performance of machine learning algorithms.





1.1 Missing Values Imputation

Thursday, April 8, 2021 7:05 PM

Average_Age = 26.0

ID	City	Age	Married ?
1	Lisbon	25	0
2	Berlin	25	1
3	Lisbon	30	1
4	Lisbon	30	1
5	Berlin	18	0
6	Lisbon	NaN	0
7	Berlin	30	1
8	Berlin	NaN	0
9	Berlin	25	1
10	Madrid	25	1

ID	City	Age	Married ?
1	Lisbon	25	0
2	Berlin	25	1
3	Lisbon	30	1
4	Lisbon	30	1
5	Berlin	18	0
6	Lisbon	26	0
7	Berlin	30	1
8	Berlin	26	0
9	Berlin	25	1
10	Madrid	25	1

1.2 Handling Categorical Values

Thursday, April 8, 2021 7:06 PM

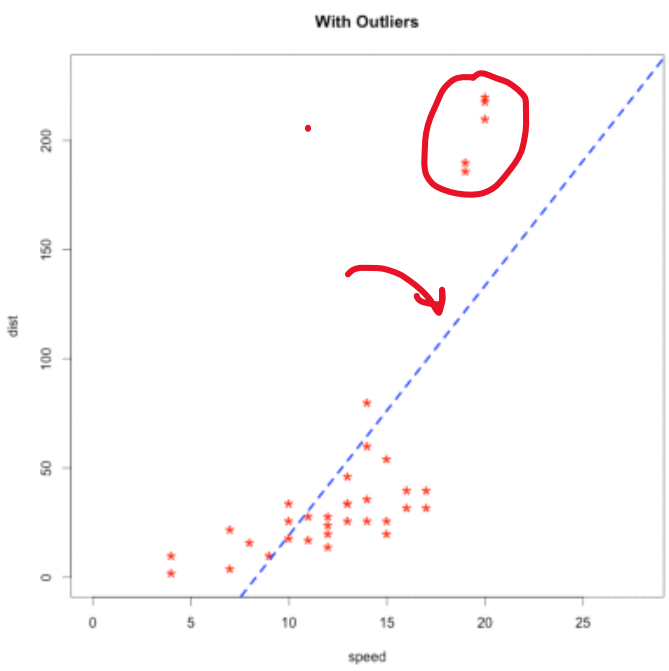
Index	Animal
0	Dog
1	Cat
2	Sheep
3	Horse
4	Lion

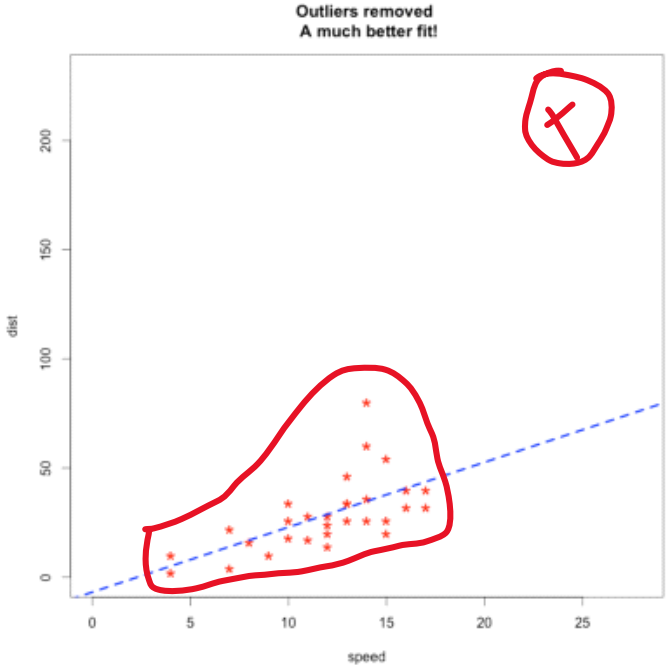
One-Hot code

Index	Dog	Cat	Sheep	Lion	Horse
0	1	0	0	0	0
1	0	1	0	0	0
2	0	0	1	0	0
3	0	0	0	0	1
4	0	0	0	1	0

1.3 Outlier Detection

Thursday, April 8, 2021 7:07 PM

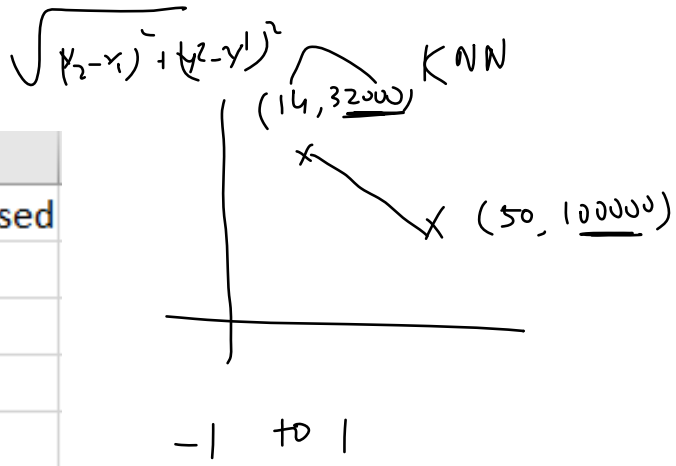




1.4 Feature Scaling

Thursday, April 8, 2021 7:08 PM

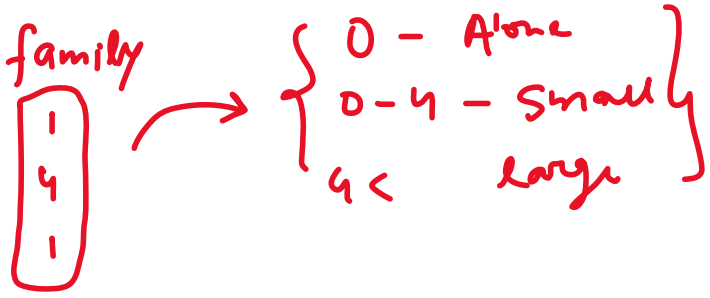
	A	B	C	D
1	Country	Age	Salary	Purchased
2	France	44	72000	No
3	Spain	27	48000	Yes
4	Germany	30	54000	No
5	Spain	38	61000	No
6	Germany	40		Yes
7	France	35	58000	Yes
8	Spain		52000	No
9	France	48	79000	Yes
10	Germany	50	83000	No
11	France	37	67000	Yes



2. Feature Construction

Thursday, April 8, 2021 7:09 PM

P. Id	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
1	0	3	Braund, Mr. Owen Harris	male	22	1	0	A/5 21171	7.25		S
2	1	1	Cumings, Mrs. John Bradley (Florence Brigg	female	38	1	0	PC 17599	71.2833	C85	C
3	1	3	Heikkinen, Miss. Laina	female	26	0	0	STON/O2.	7.925		S
4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35	1	0	113803	53.1	C123	S
5	0	3	Allen, Mr. William Henry	male	35	0	0	373450	8.05		S
6	0	3	Moran, Mr. James	male		0	0	330877	8.4583		Q
7	0	1	McCarthy, Mr. Timothy J	male	54	0	0	17463	51.8625	E46	S
8	0	3	Palsson, Master. Gosta Leonard	male	2	3	1	349909	21.075		S
9	1	3	Johnson, Mrs. Oscar W (Elisabeth Vilhelmin	female	27	0	2	347742	11.1333		S
10	1	2	Nasser, Mrs. Nicholas (Adele Achem)	female	14	1	0	237736	30.0708		C



3. Feature Selection

Thursday, April 8, 2021 7:11 PM

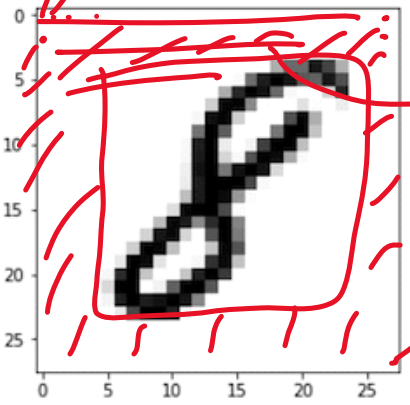
```
In [5]: train_df = pd.read_csv(f'{PATH}train.csv', header = 0)
```

```
In [19]: train_df
```

Out[19]:

	label	pixel0	pixel1	pixel2	pixel3	pixel4	pixel5	pixel6	pixel7	pixel8	...	pixel774
0	1	0	0	0	0	0	0	0	0	0	...	0
1	0	0	0	0	0	0	0	0	0	0	...	0
2	1	0	0	0	0	0	0	0	0	0	...	0
3	4	0	0	0	0	0	0	0	0	0	...	0

784 features



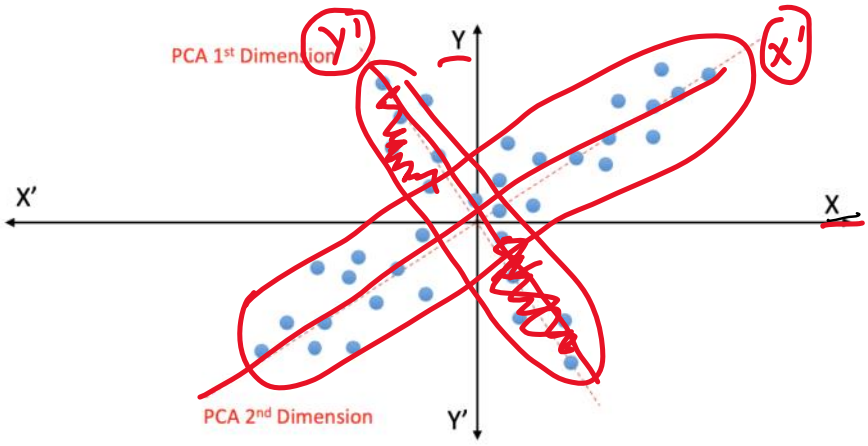
MNIST

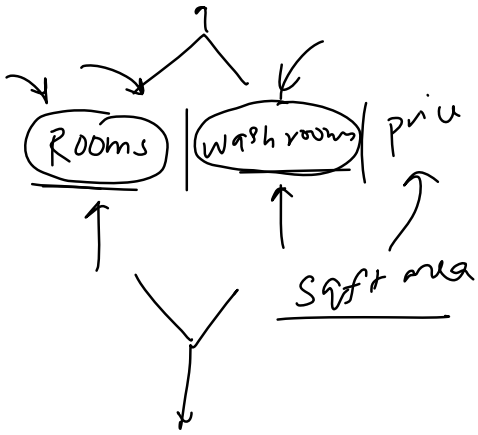
50000 imgs

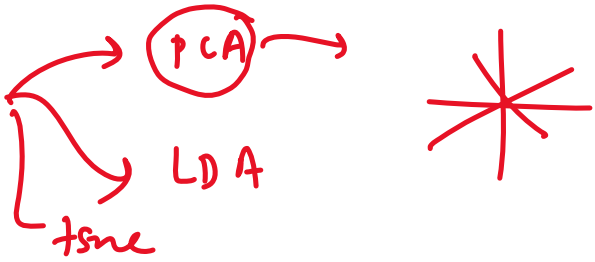
28 × 28 = 784

4. Feature Extraction

Thursday, April 8, 2021 7:13 PM



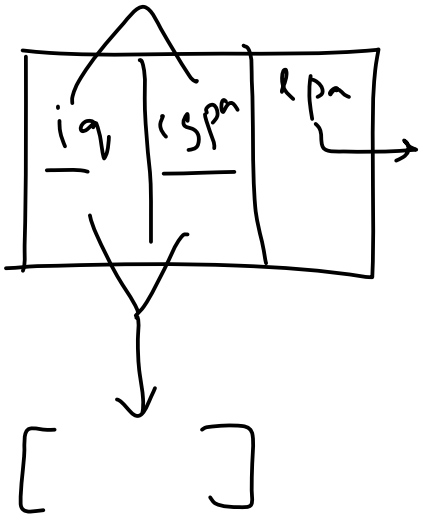


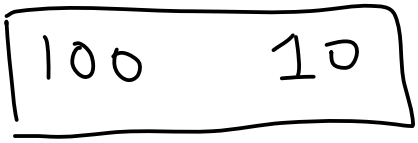


What is Feature Scaling?

Friday, April 9, 2021 4:21 PM

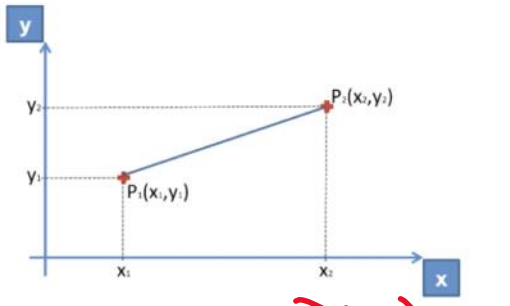
Feature Scaling is a technique to standardize the independent features present in the data in a fixed range





Why do we need Feature Scaling?

Friday, April 9, 2021 4:27 PM



Euclidean Distance between P_1 and $P_2 = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2}$

Let x be the no. of Salary and y be the no. of Age

Example: x1&y1 are in row 2, x2&y2 are in row 9

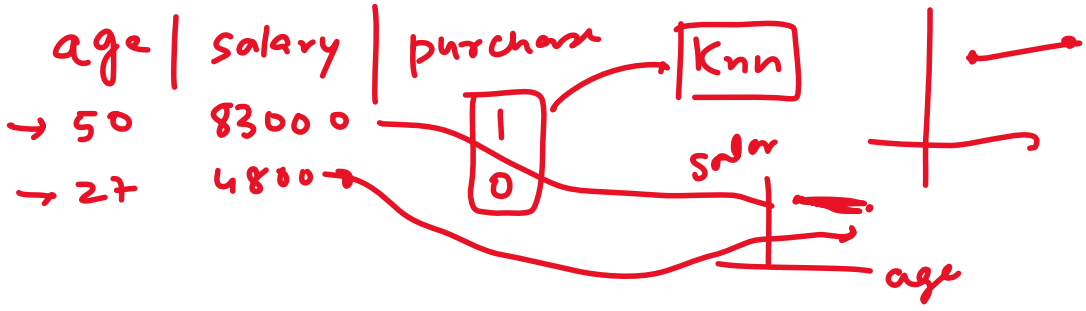
(x2-x1)^2= (83000-48000)^2

=1225000000

(y2-y1)^2=(50-27)^2

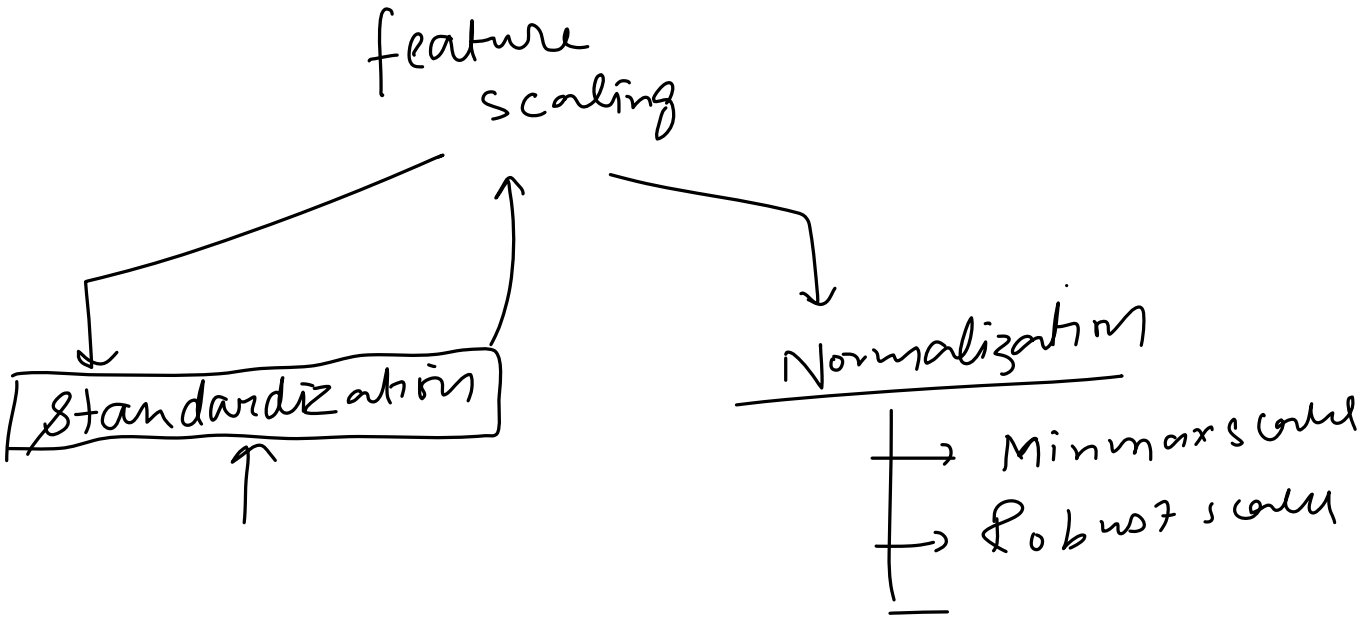
=529

Knn



Types of Feature Scaling

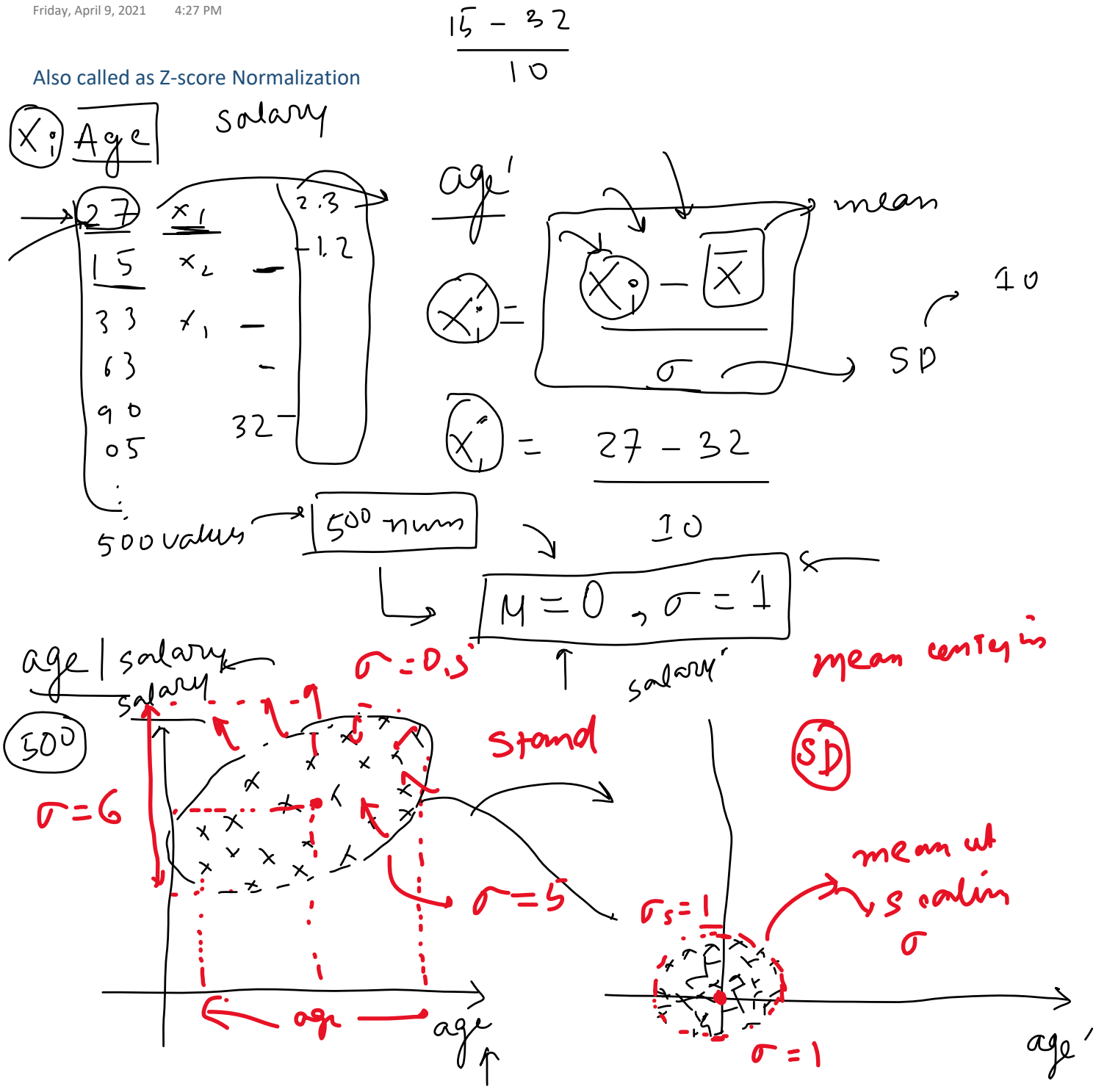
Friday, April 9, 2021 4:27 PM



Standardization - Intuition

Friday, April 9, 2021 4:27 PM

Also called as Z-score Normalization

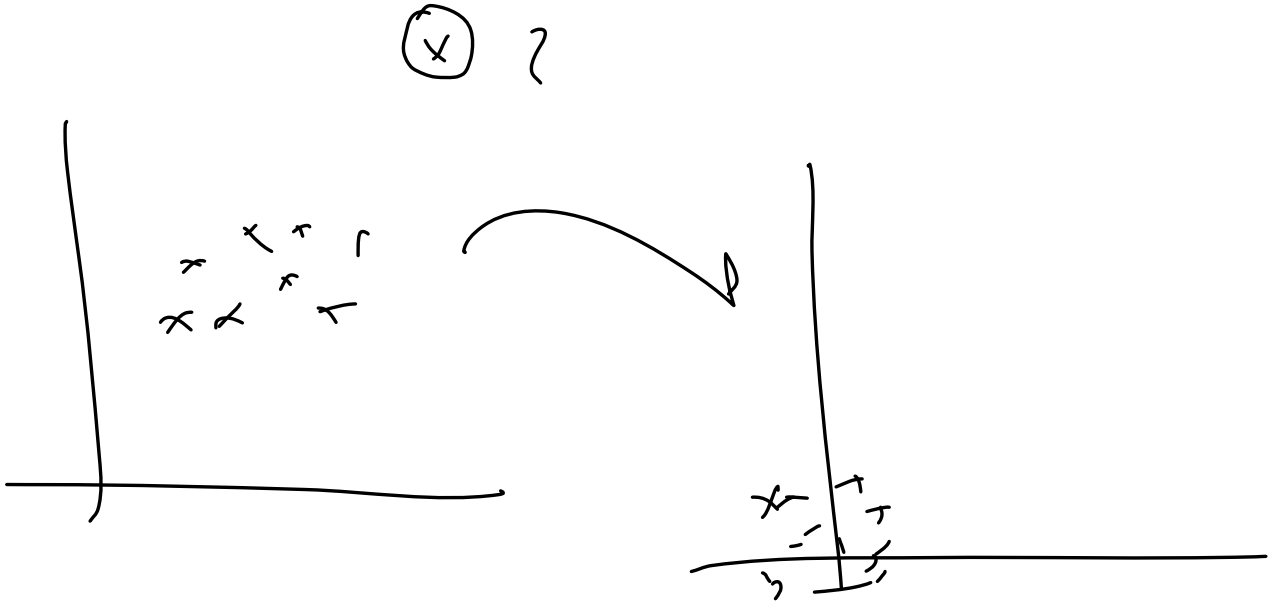


Example

Friday, April 9, 2021 4:28 PM

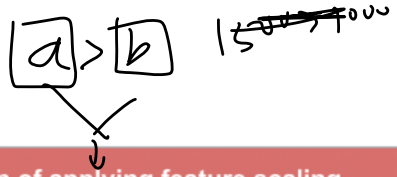
Impact of Outliers

Friday, April 9, 2021 4:28 PM

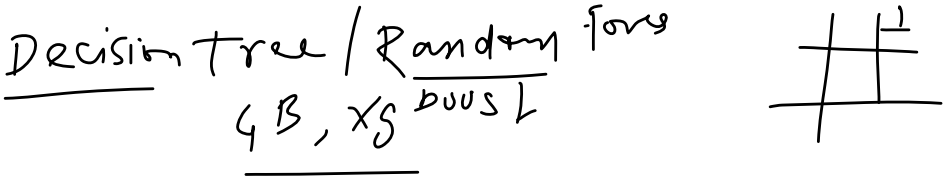


When to use Standardization?

Friday, April 9, 2021 4:28 PM



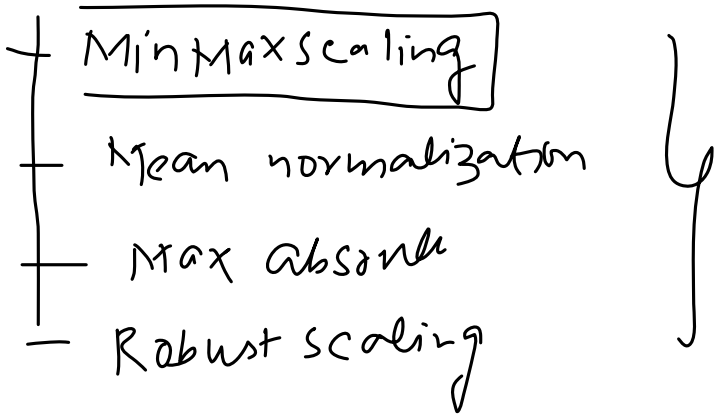
Algorithm(s)	Reason of applying feature scaling
1. K-Means	Use the Euclidean distance measure.
2. K-Nearest-Neighbours	Measure the distances between pairs of samples and these distances are influenced by the measurement units
3. Principal Component Analysis (PCA)	Try to get the feature with maximum variance
4. Artificial Neural Network	Apply Gradient Descent
5. Gradient Descent	Theta calculation becomes faster after feature scaling and the learning rate in the update equation of Stochastic Gradient Descent is the same for every parameter



Normalization

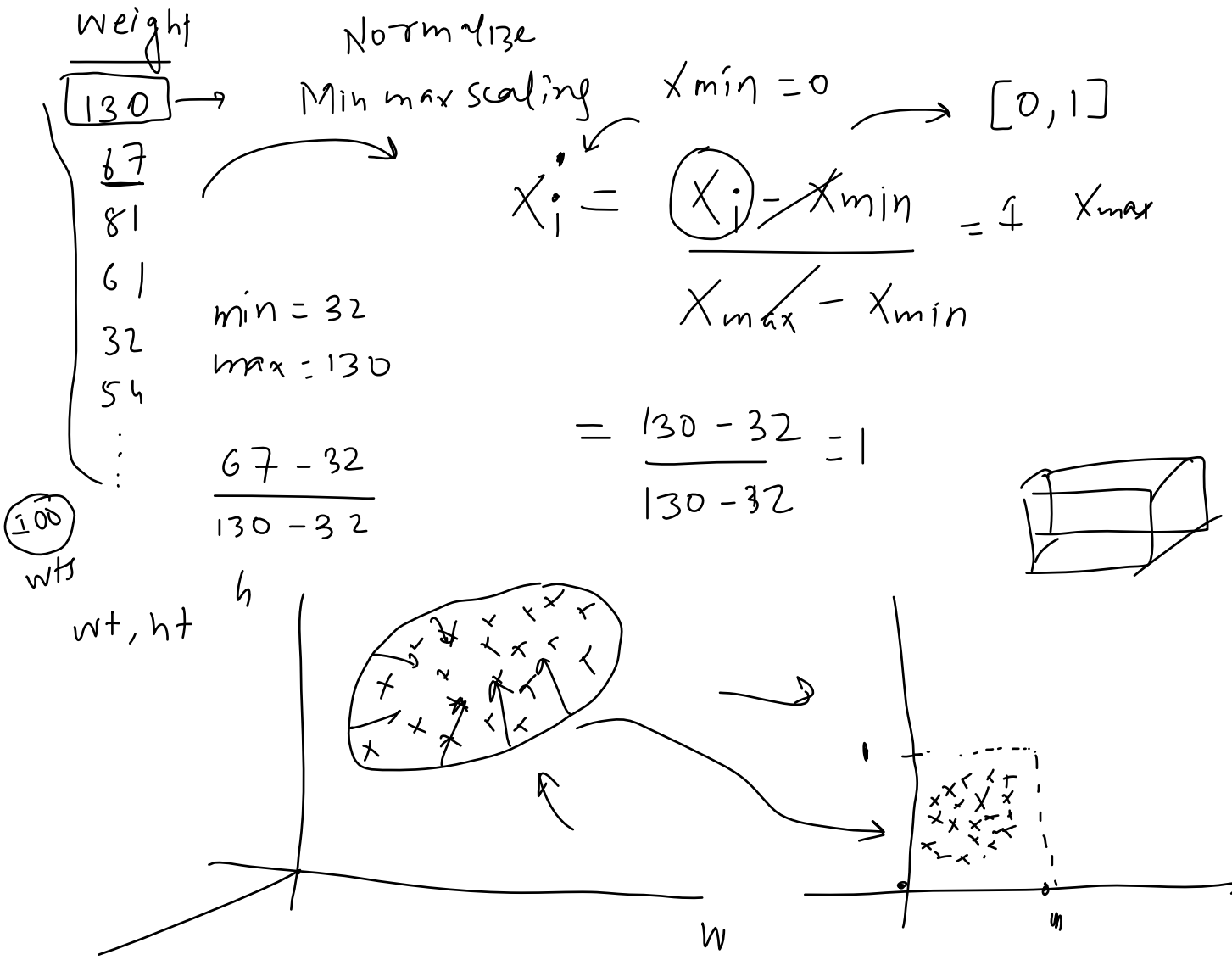
Saturday, April 10, 2021 1:15 PM

Normalization is a technique often applied as part of data preparation for machine learning. The goal of normalization is to change the values of numeric columns in the dataset to use a common scale, without distorting differences in the ranges of values or losing information



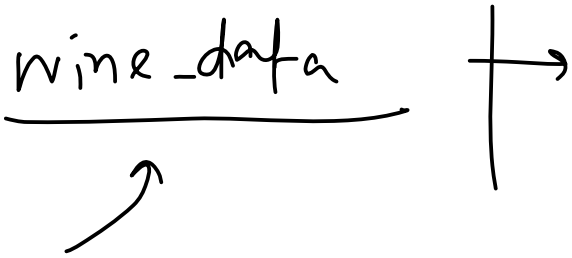
MinMaxScaling - Intuition

Saturday, April 10, 2021 1:16 PM



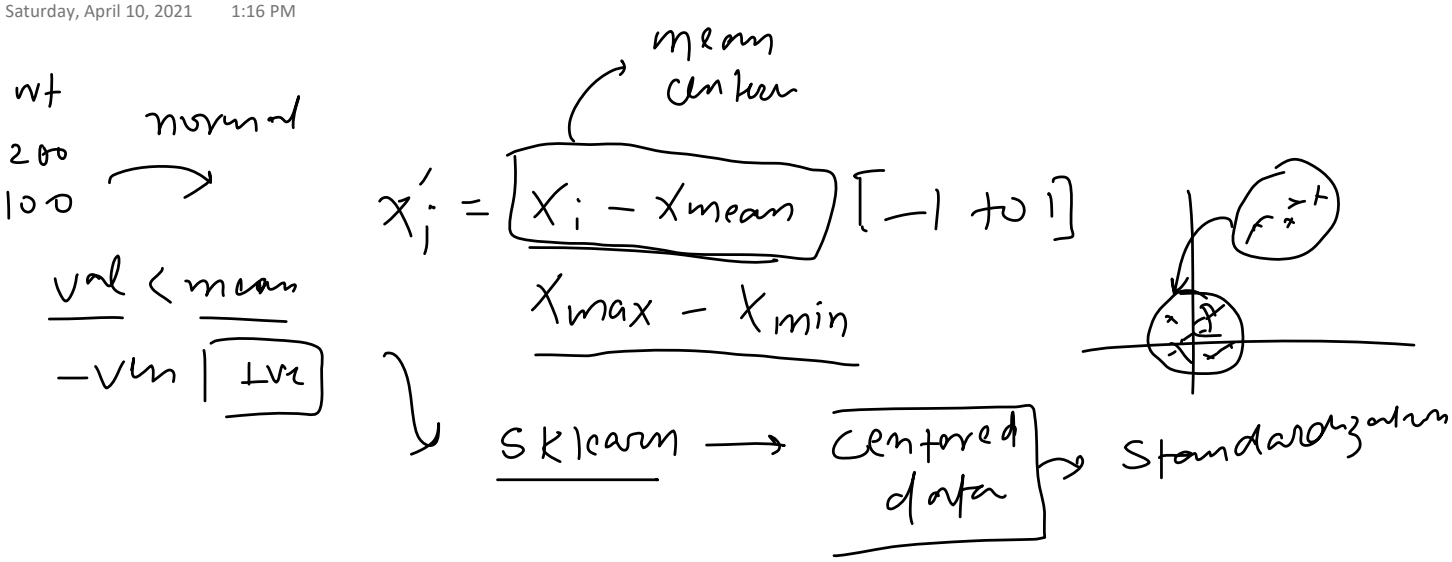
Code Example

Saturday, April 10, 2021 1:17 PM



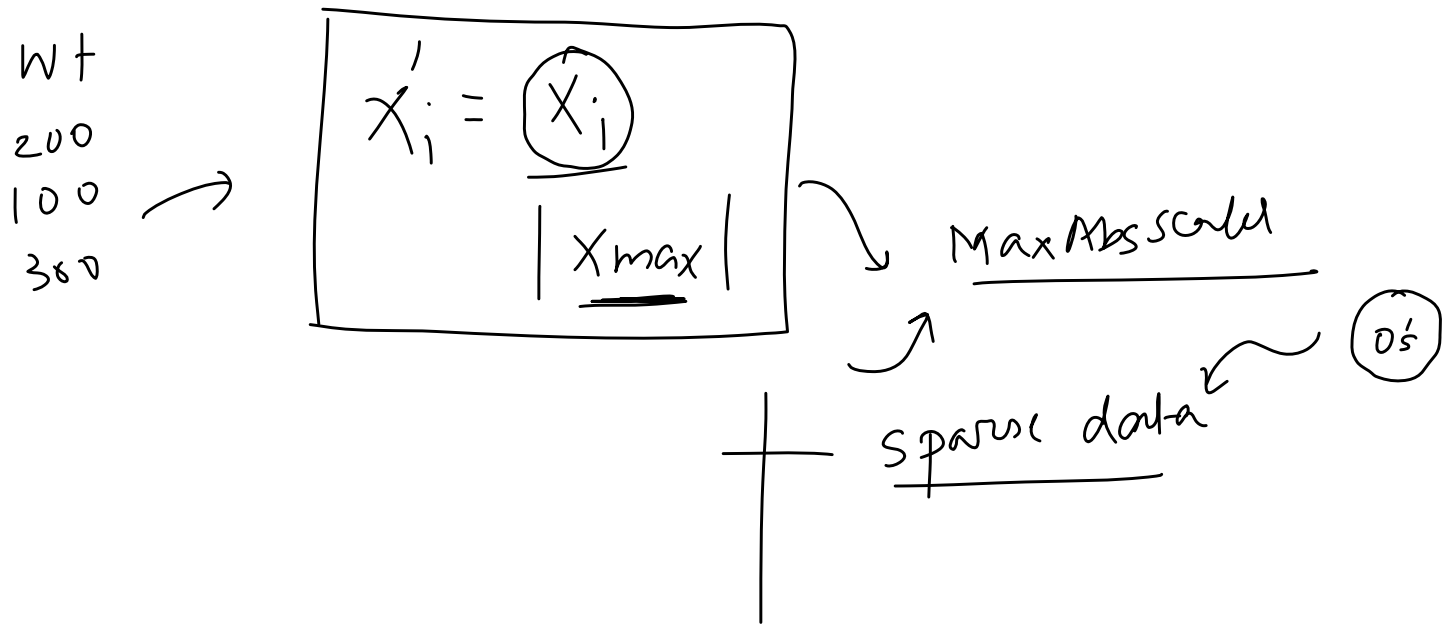
Mean Normalization

Saturday, April 10, 2021 1:16 PM



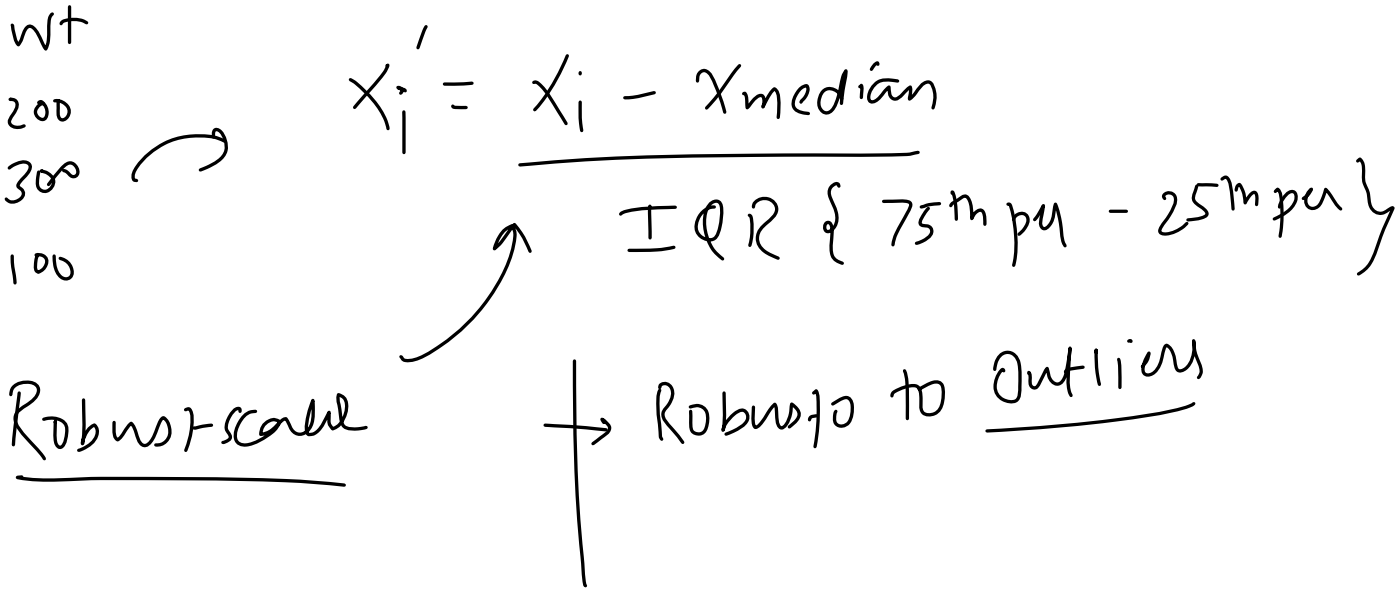
MaxAbsScaling

Saturday, April 10, 2021 1:17 PM



Robust Scaling

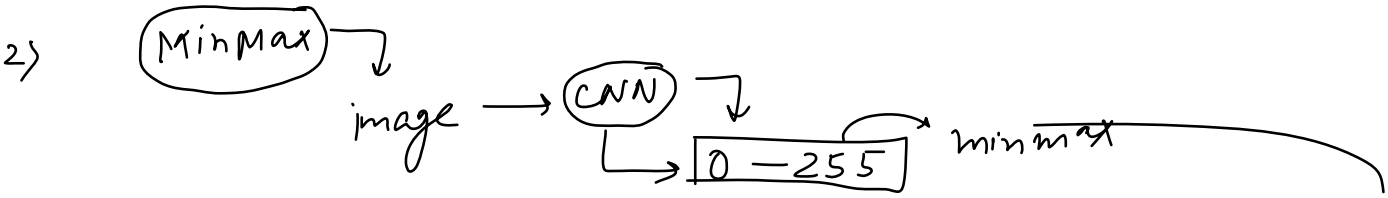
Saturday, April 10, 2021 1:17 PM



Normalization Vs Standardization

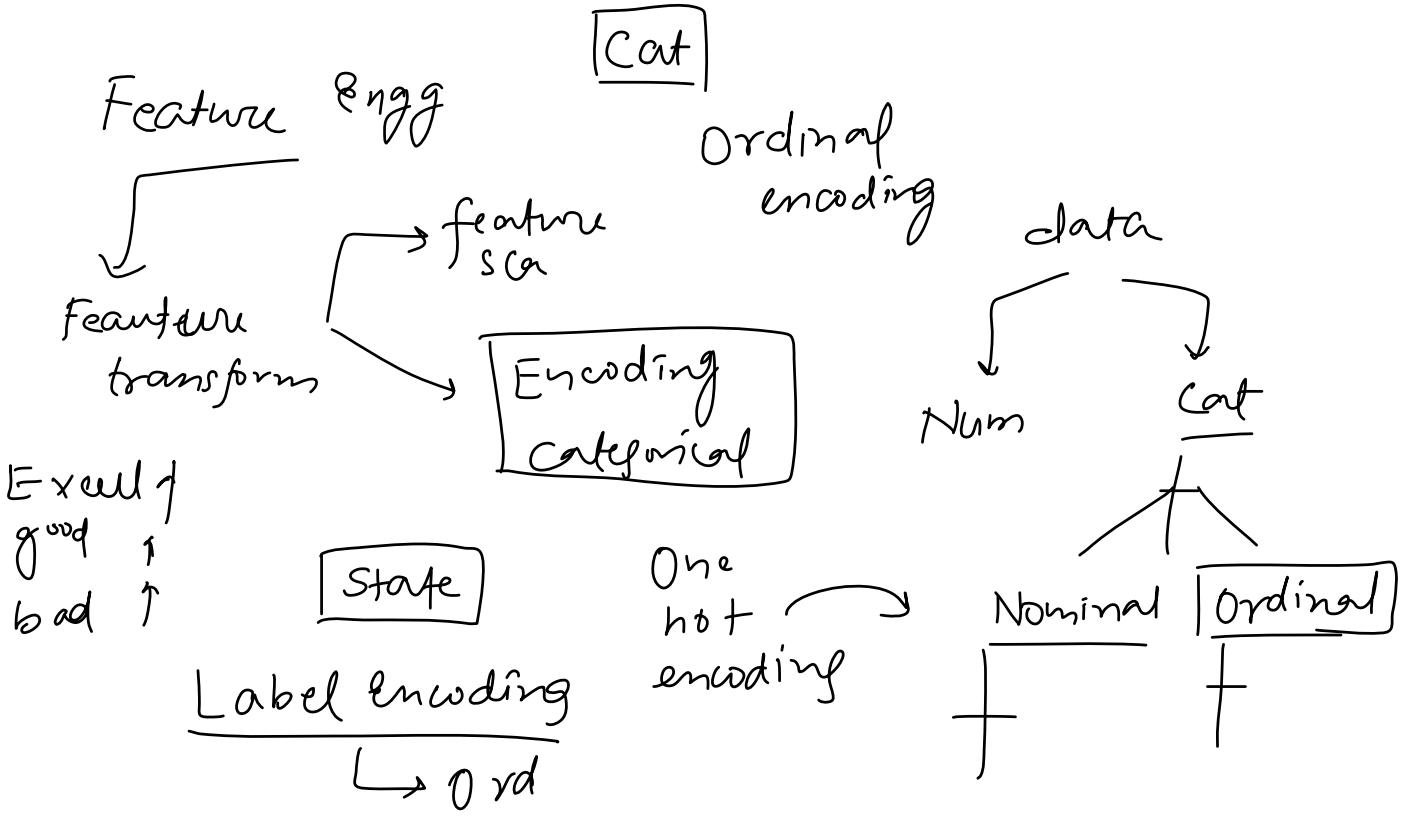
Saturday, April 10, 2021 1:17 PM

1) Is feature scaling required?

2) 

Encoding Categorical Variables

Monday, April 12, 2021 3:53 PM



Ordinal Encoding

Monday, April 12, 2021 3:54 PM

Nominal (OHE)

Ordinal encoding

Education

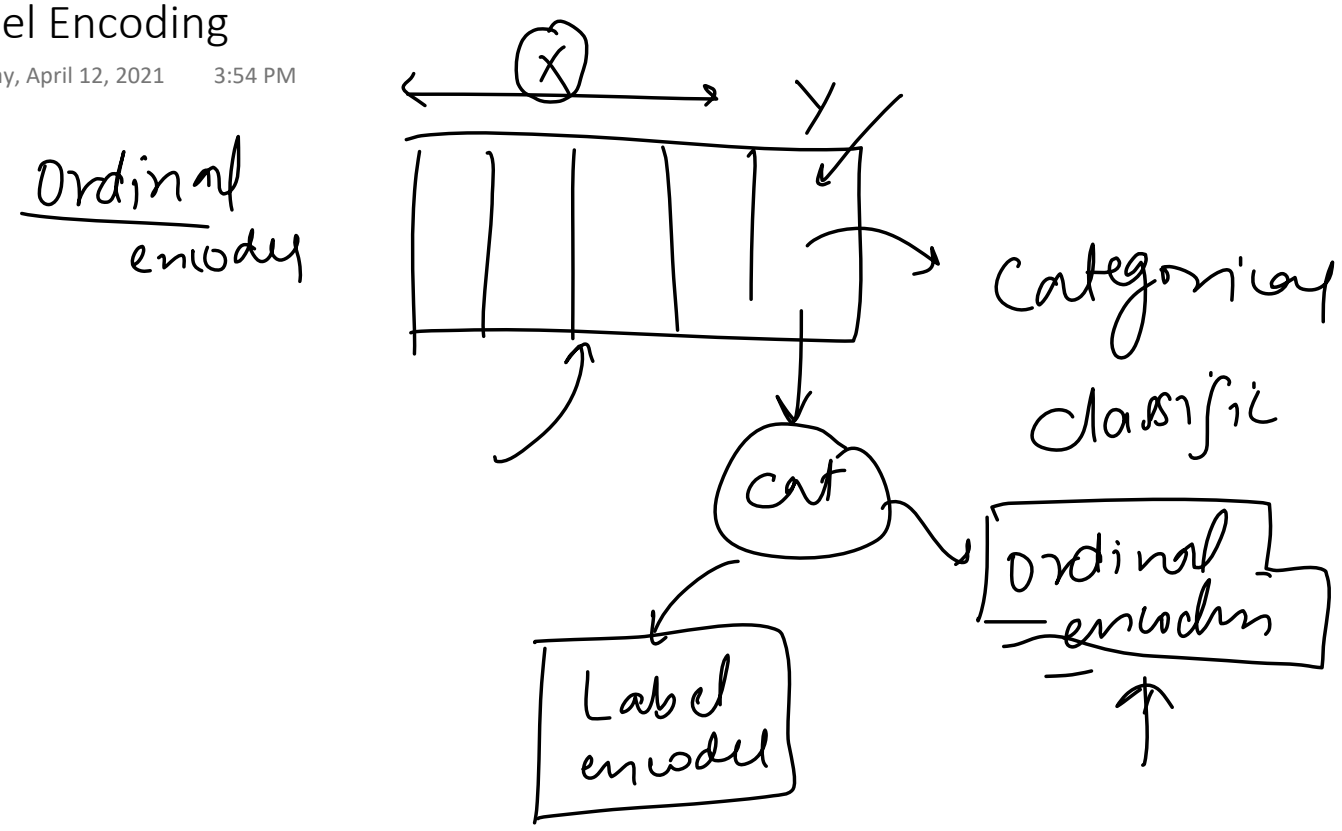
HS	0
UG	1
PG	2
PG	2
UG	1
HS	0
UG	2

PG > UG > HS

PG → 2
UG → 1
HS → 0

Label Encoding

Monday, April 12, 2021 3:54 PM



Example

Monday, April 12, 2021 3:54 PM

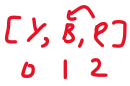
OneHotEncoding

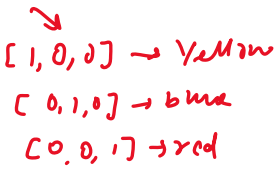
Tuesday, April 13, 2021 12:24 PM

Color	Target
Yellow	0
Yellow	1
Blue	1
Yellow	1
Red	1
Yellow	0
Red	1
Red	0
Yellow	1
Blue	0

Color_Y	Color_B	Color_R	Target
1	0	0	0
1	0	0	1
0	1	0	1
1	0	0	1
0	0	1	1
1	0	0	0
0	0	1	1
0	0	1	0
1	0	0	1
0	1	0	0

One Hot Encoding





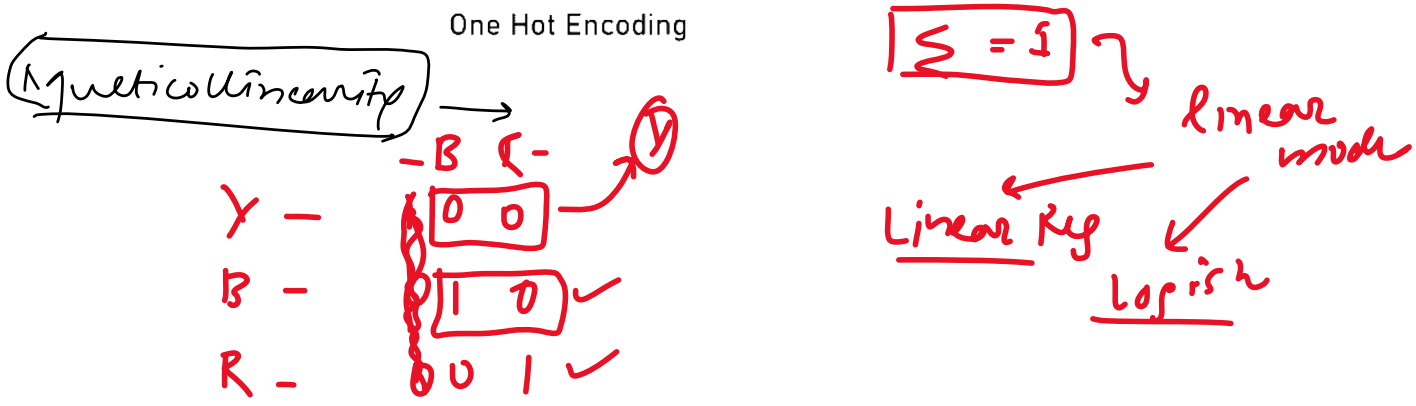
Dummy Variable Trap

Tuesday, April 13, 2021 12:26 PM

Color	Target
Yellow	0
Yellow	1
Blue	1
Yellow	1
Red	1
Yellow	0
Red	1
Red	0
Yellow	1
Blue	0

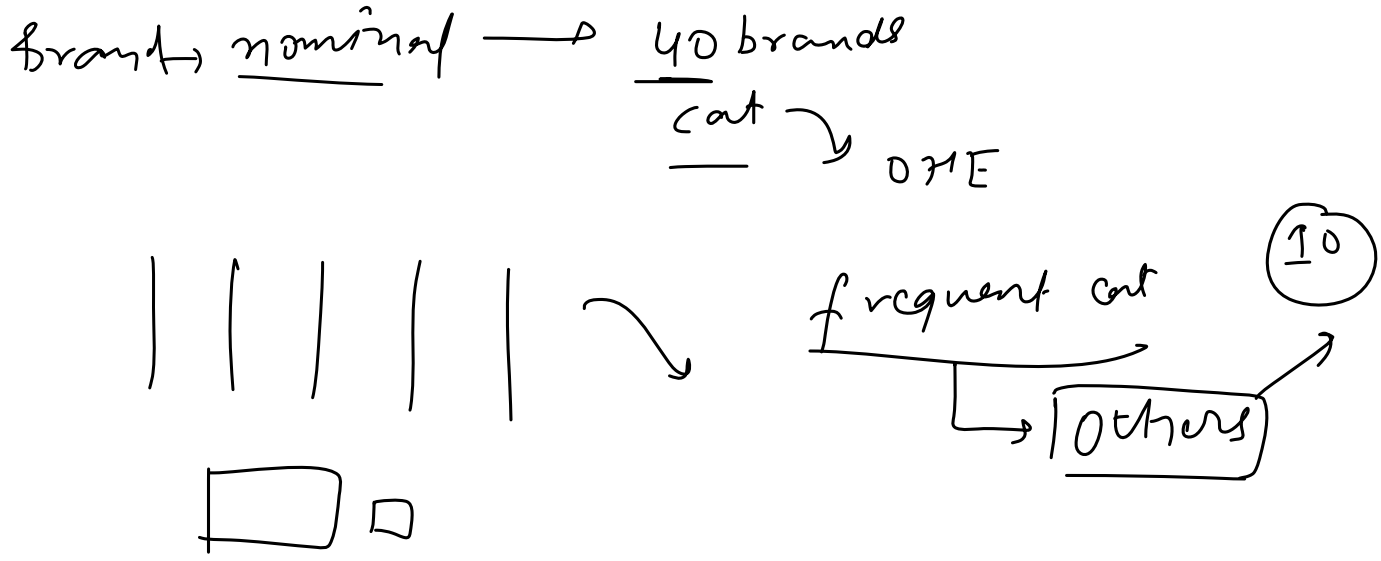
Color_Y	Color_B	Color_R	Target
1	0	0	0
1	0	0	1
0	1	0	1
1	0	0	1
0	0	1	1
1	0	0	0
0	0	1	1
0	0	1	0
1	0	0	1
0	1	0	0

One Hot Encoding



OHE using most frequent variables

Tuesday, April 13, 2021 12:31 PM



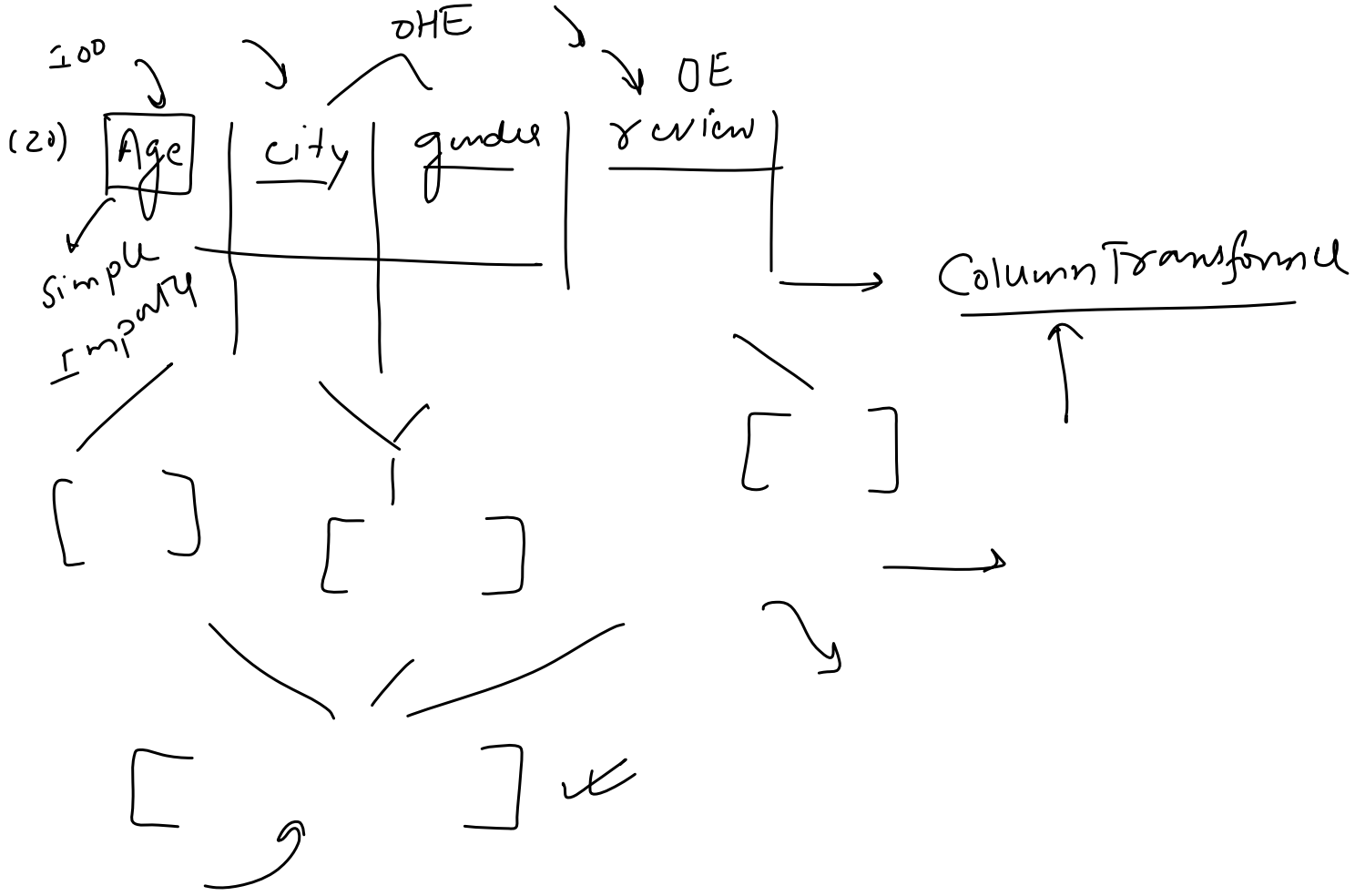
Example

Tuesday, April 13, 2021 12:26 PM

	brand	km_driven	fuel	owner	selling_price
0	Maruti	145500	Diesel	First Owner	450000
1	Skoda	120000	Diesel	Second Owner	370000
2	Honda	140000	Petrol	Third Owner	158000
3	Hyundai	127000	Diesel	First Owner	225000
4	Maruti	120000	Petrol	First Owner	130000

Column Transformer

Wednesday, April 14, 2021 5:16 PM



Internship Project Discussion

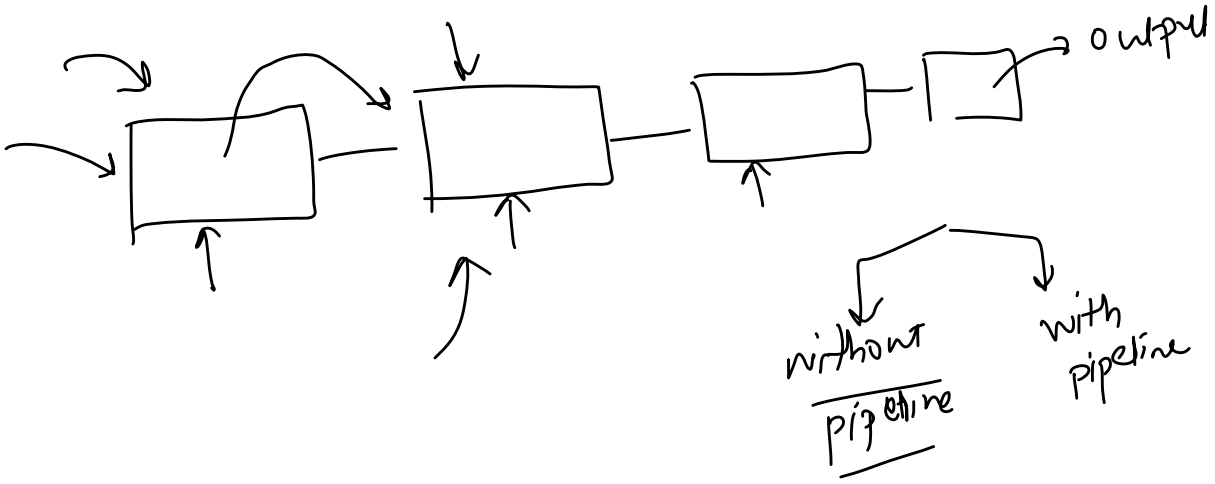
Wednesday, April 14, 2021 6:41 PM

Scikit Learn Pipelines

Thursday, April 15, 2021 5:59 PM

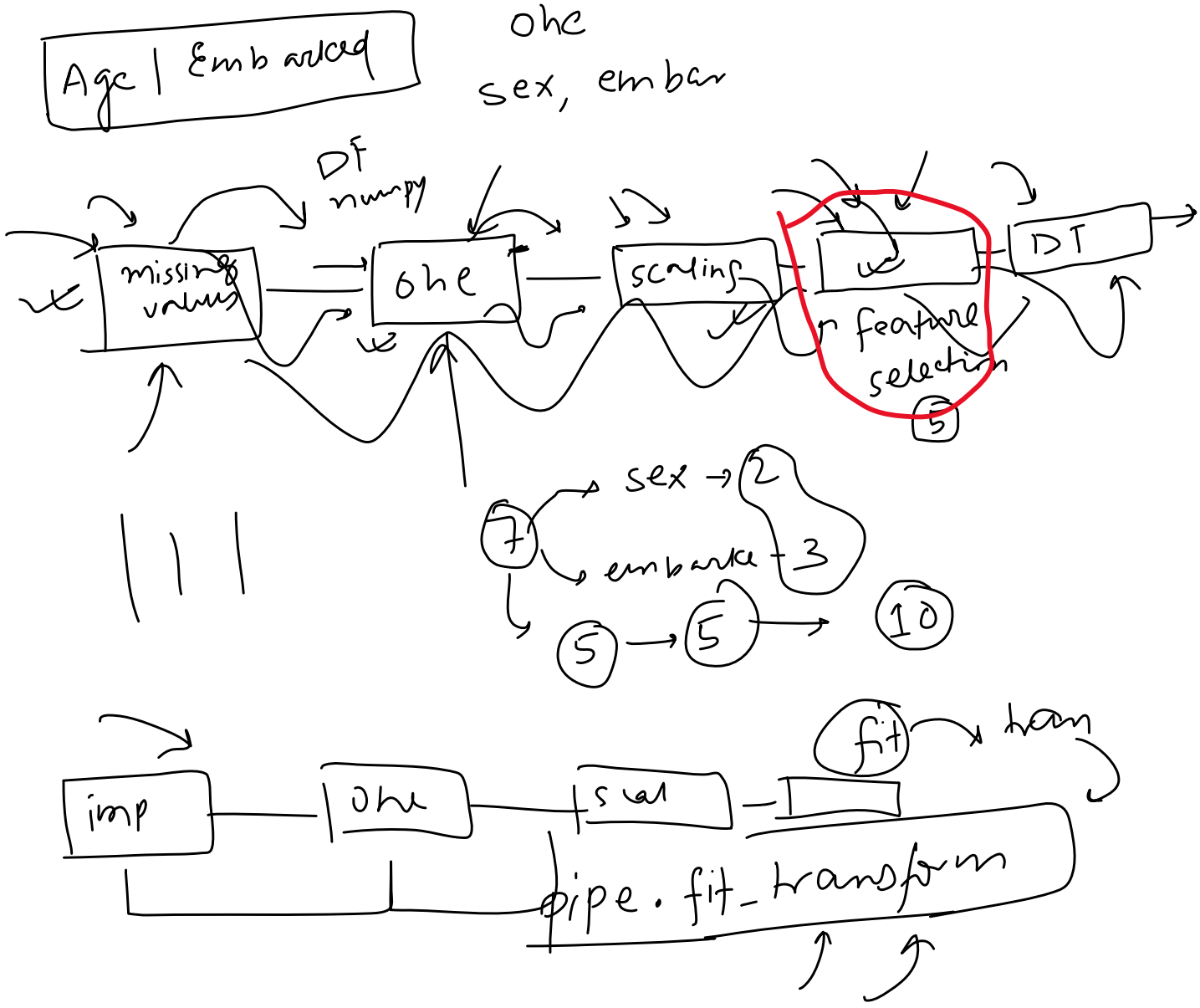
Pipelines chains together multiple steps so that the output of each step is used as input to the next step.

Pipelines makes it easy to apply the same preprocessing to train and test!



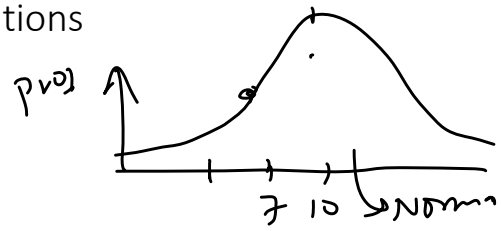
Strategy

Thursday, April 15, 2021 6:58 PM


Age | Embarked
ohe
sex, embar
DF
numpy
missing values
ohe
scaling
DT
feature selection
5
sex → 2
embarke → 3
7
5
5
10

Mathematical Transformations

Friday, April 16, 2021 5:21 PM



FE

↳ feature transformations

Normal distribution

↳ Mathematical trans

1) log trans
2) Reciprocal
3) power (sq / sqrt)

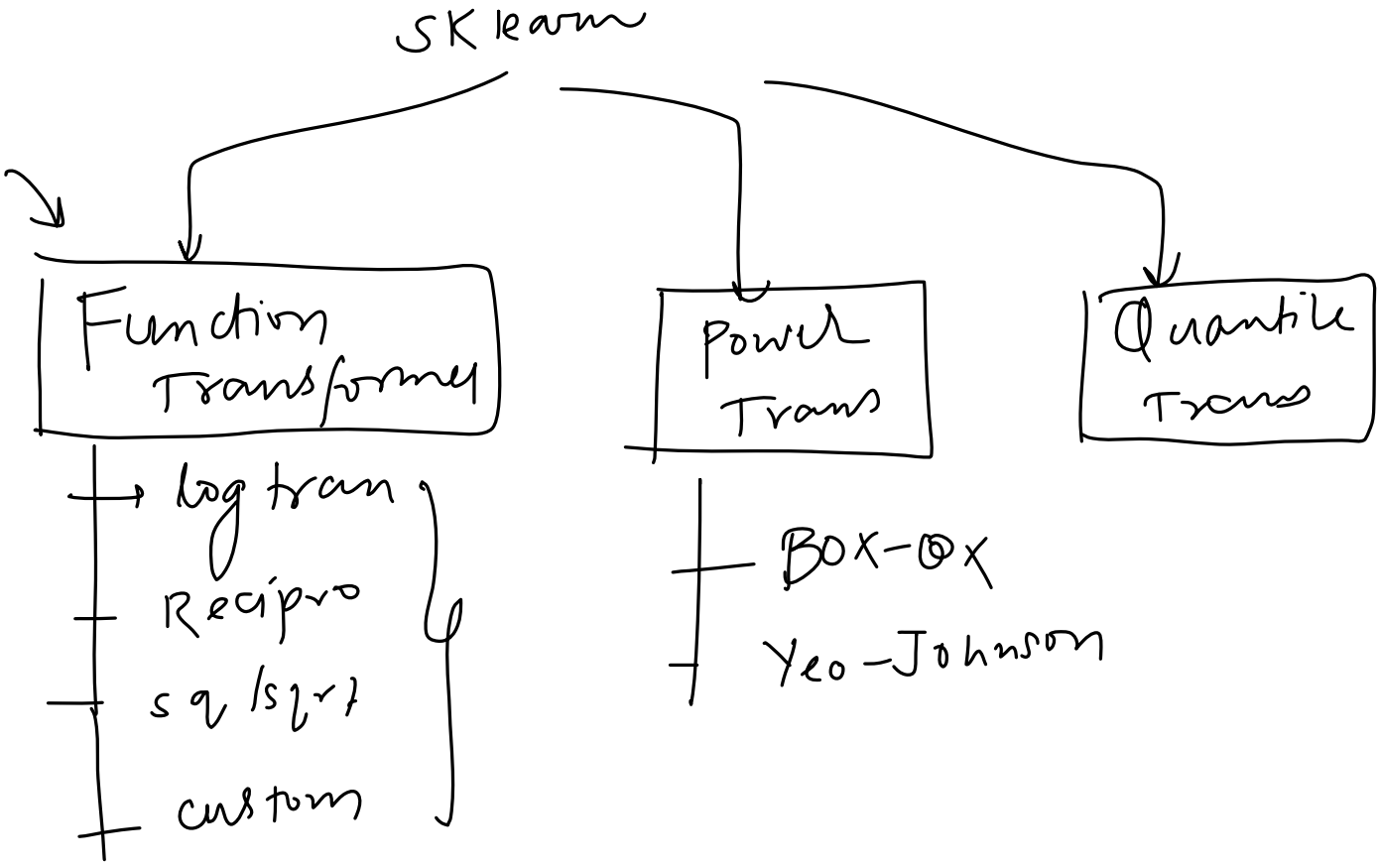
custom

1) Box-Cox
2) Yeo-Johnson

Statistic
↳ (MI) → Linear Reg
Logistic

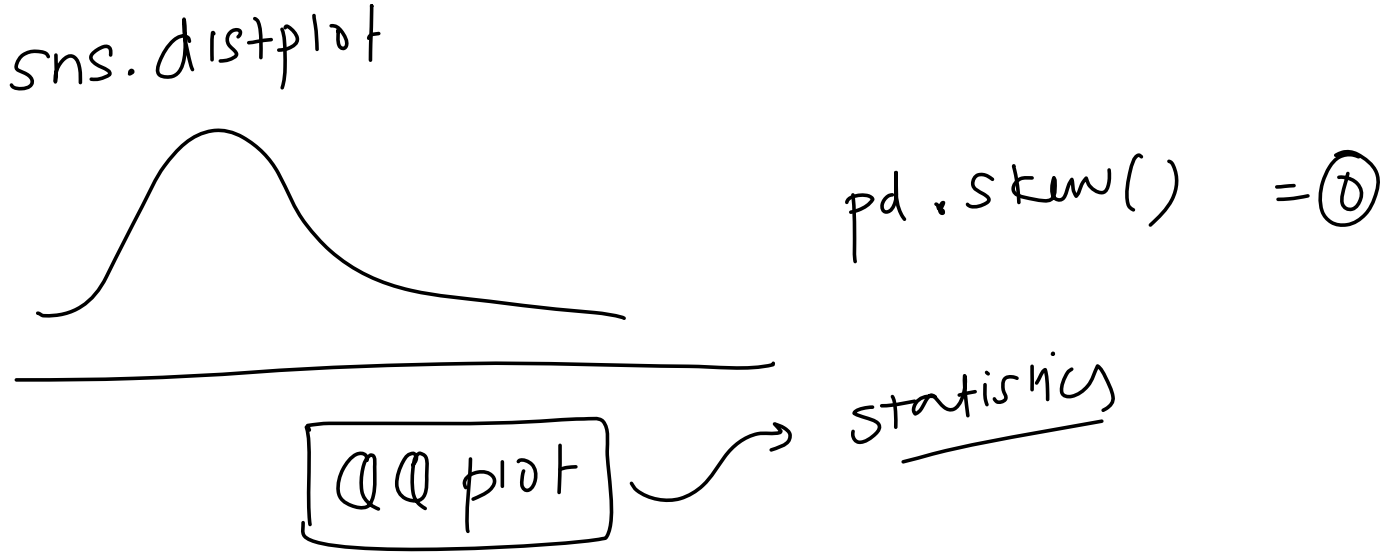
Function Transformer

Friday, April 16, 2021 5:22 PM



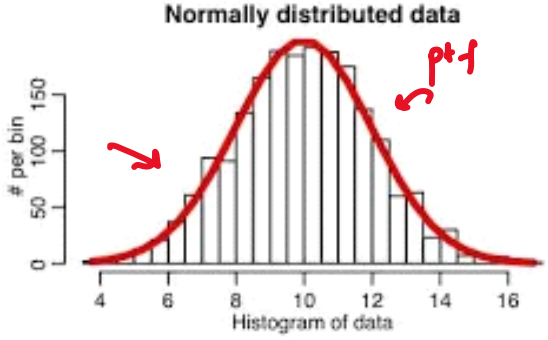
How to find if data is normal?

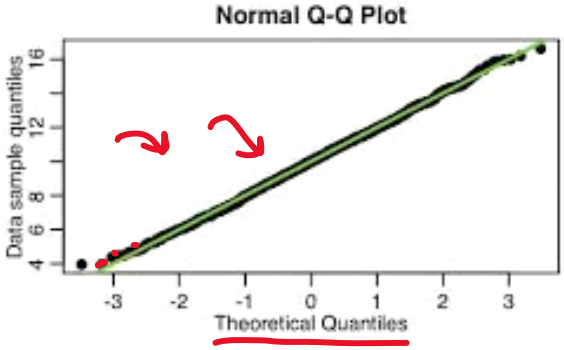
Friday, April 16, 2021 6:30 PM

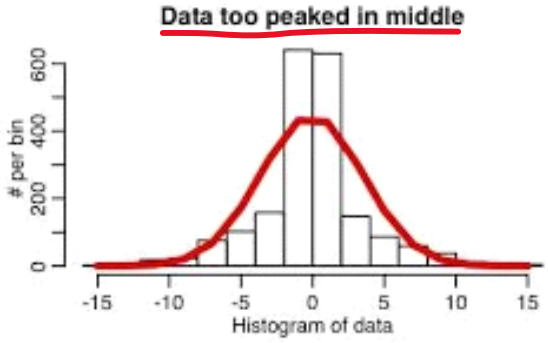


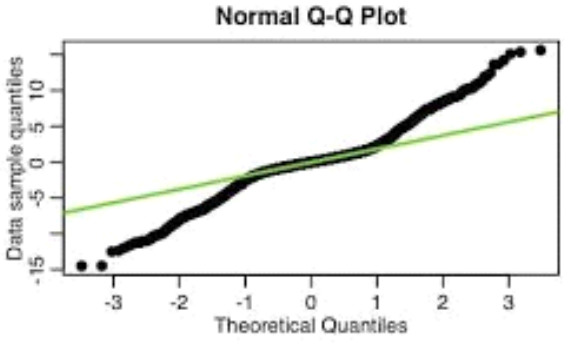
QQ Plots

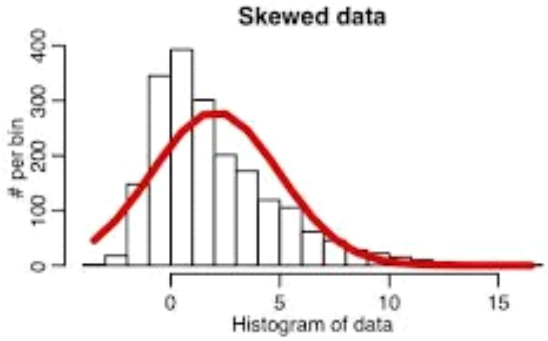
Friday, April 16, 2021 5:23 PM

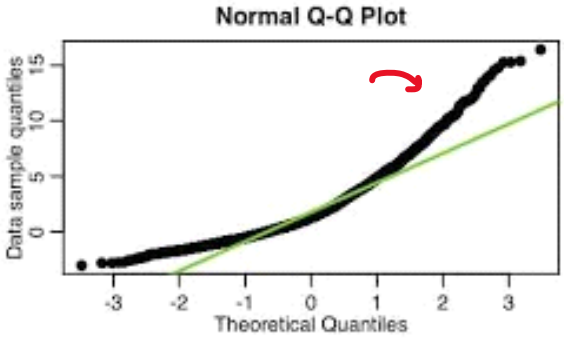


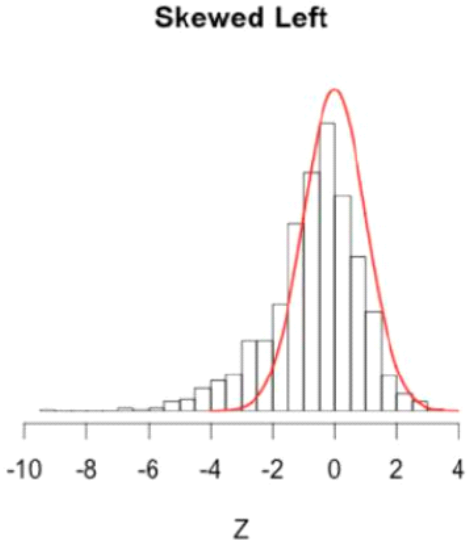


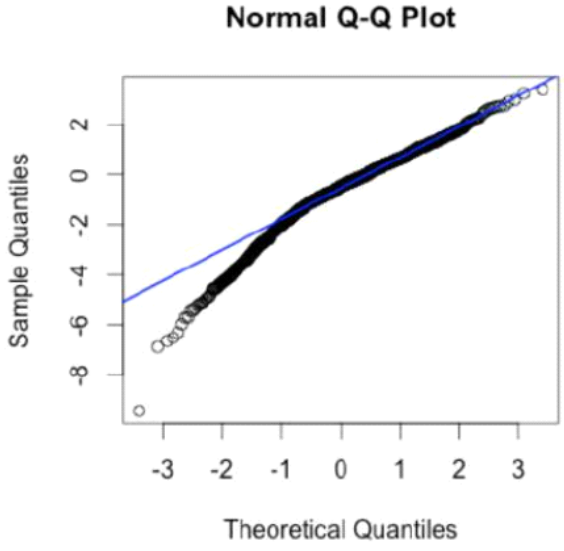


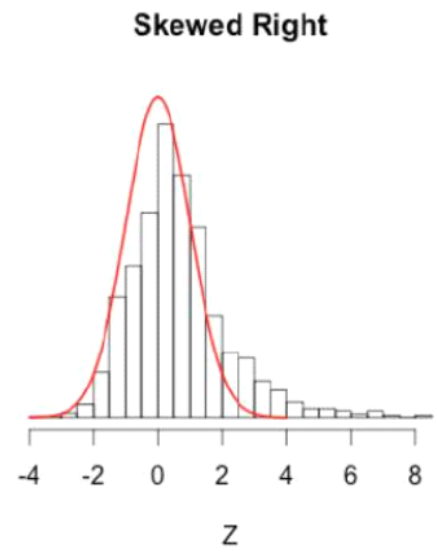


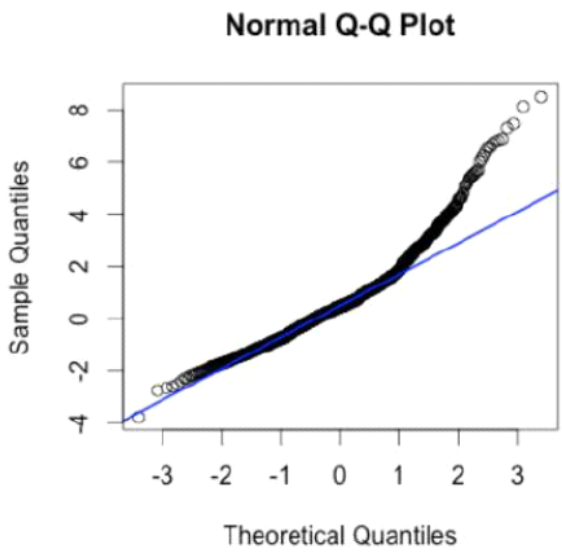


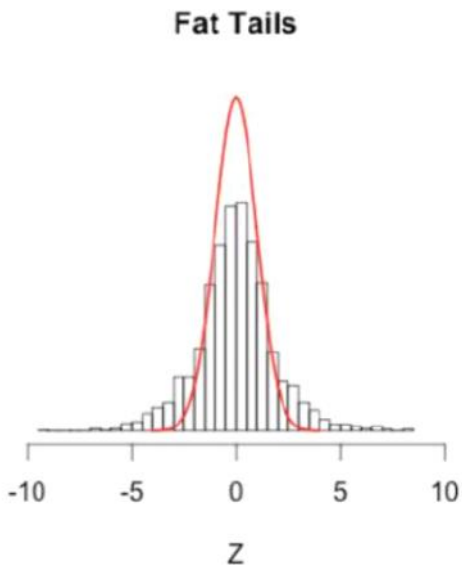


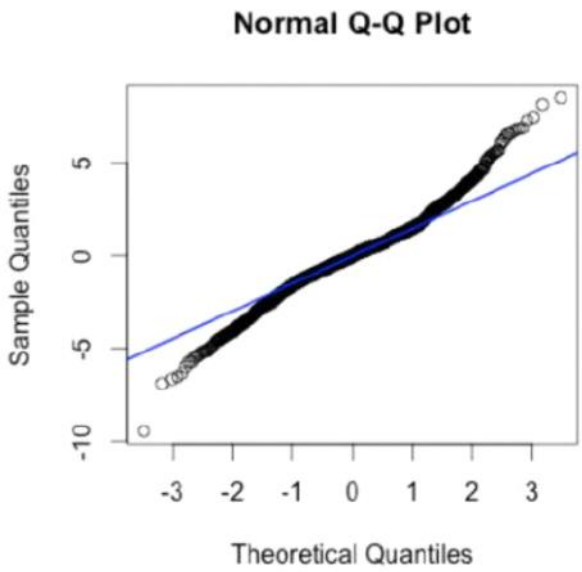


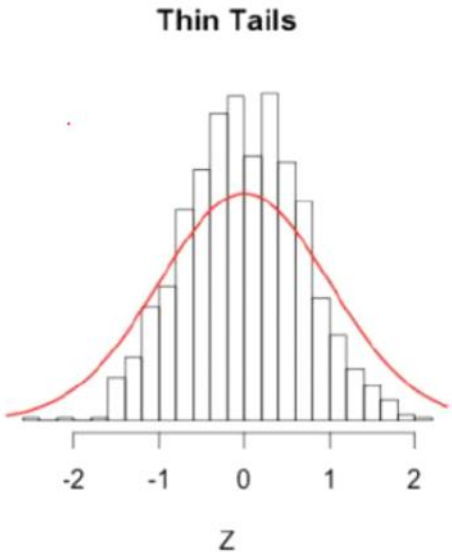


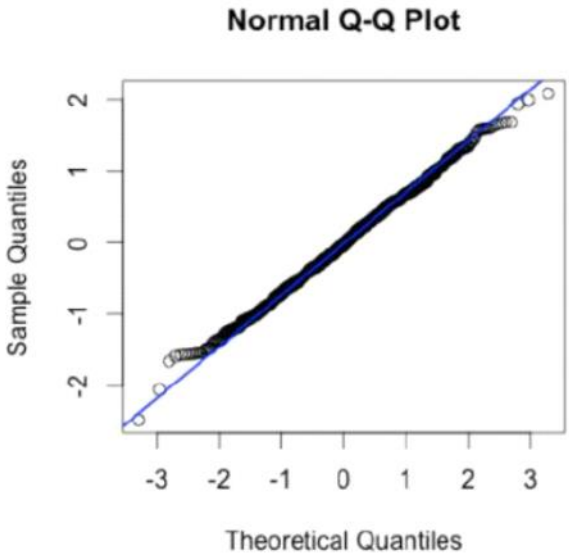

Skewed Right
-4
-2
0
2
4
6
8
Z


Normal Q-Q Plot
Sample Quantiles
-4
-2
0
2
4
6
8
-3
-2
-1
0
1
2
3
Theoretical Quantiles


Fat Tails
-10
-5
0
5
10
Z

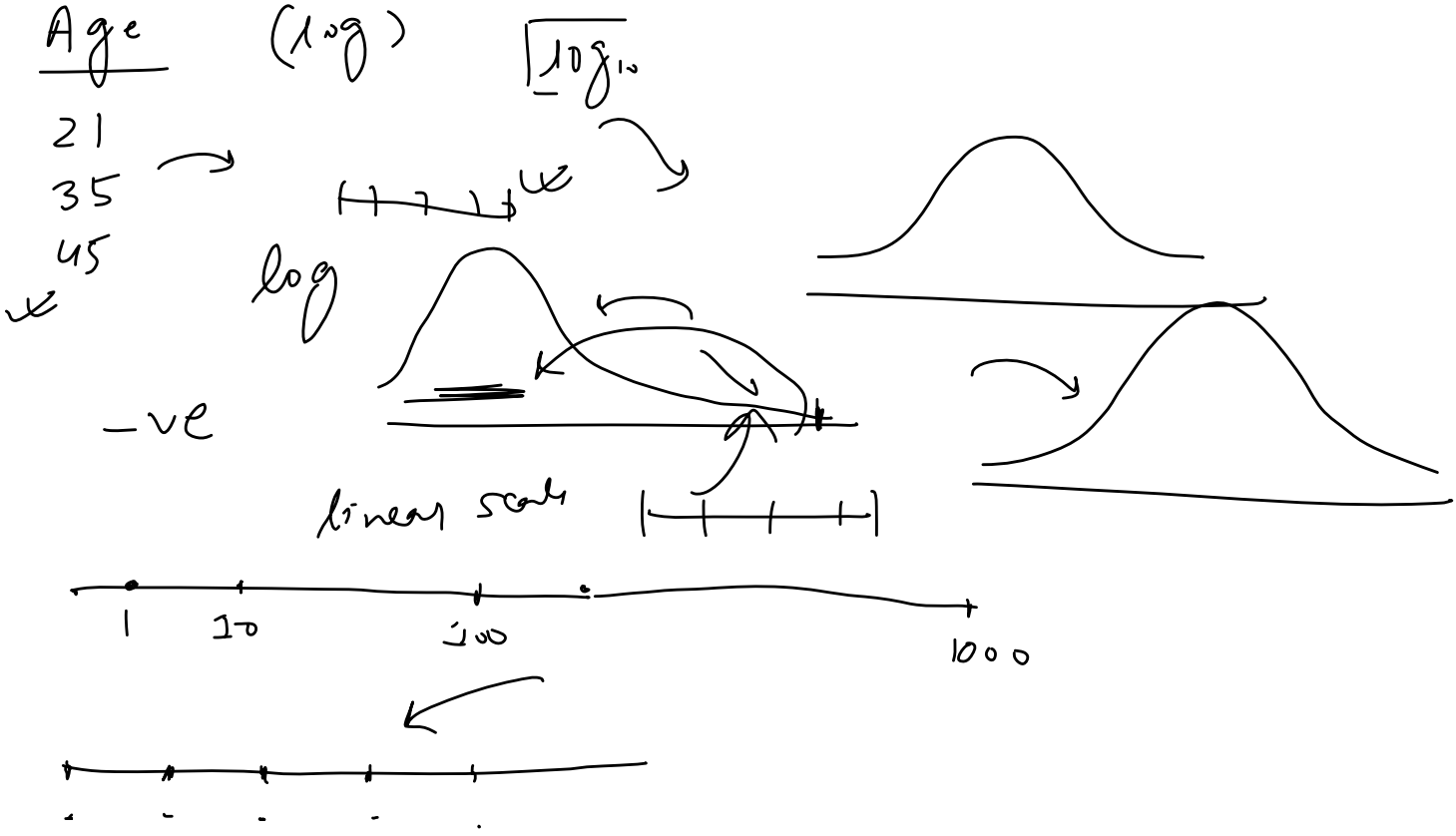

Normal Q-Q Plot
Sample Quantiles
-10
-5
0
5
-3
-2
-1
0
1
2
3
Theoretical Quantiles


Thin Tails
-2
-1
0
1
2
Z


Normal Q-Q Plot
Sample Quantiles
-2
-1
0
1
2
-3
-2
-1
0
1
2
3
Theoretical Quantiles

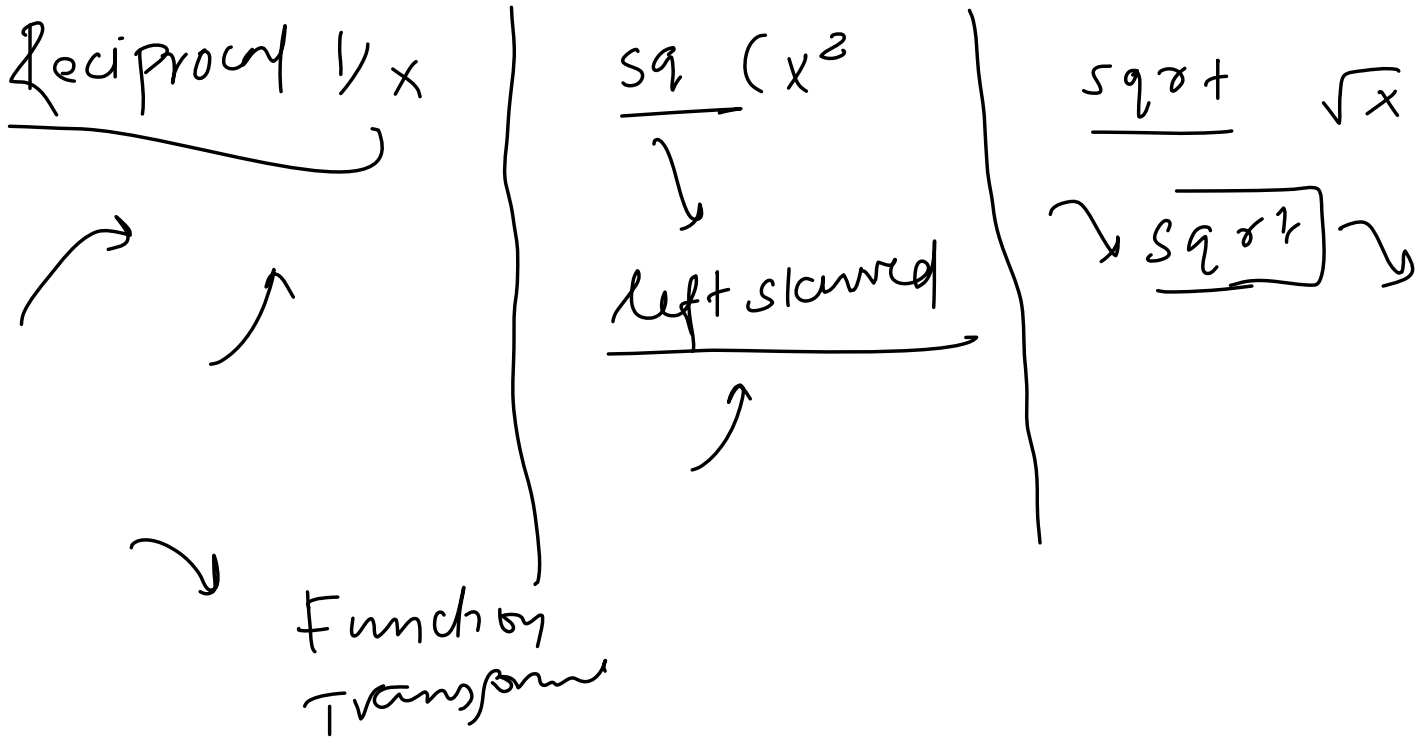
Log Transform

Friday, April 16, 2021 5:22 PM



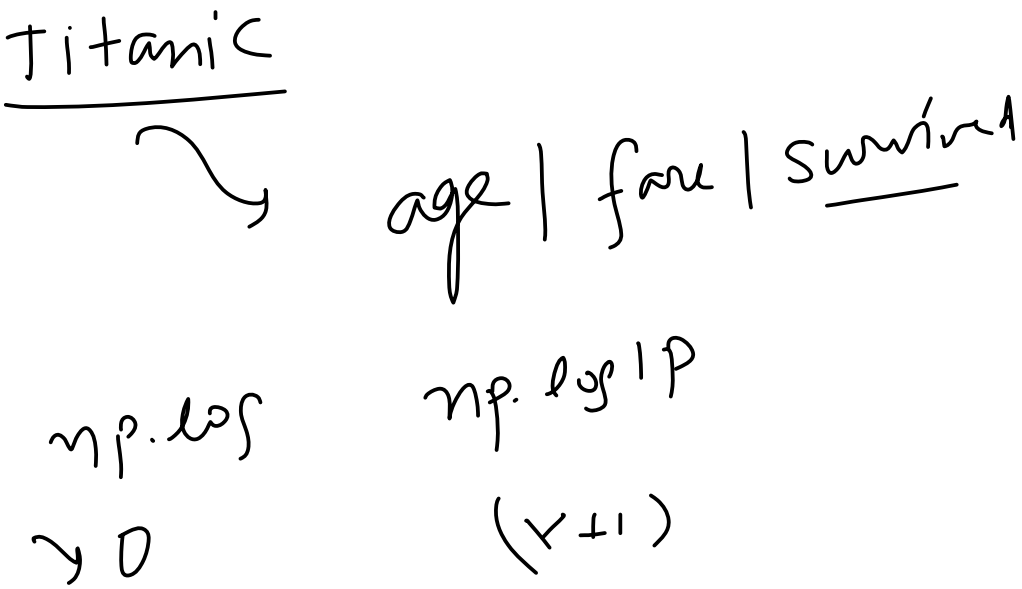
Other Transforms

Friday, April 16, 2021 6:23 PM



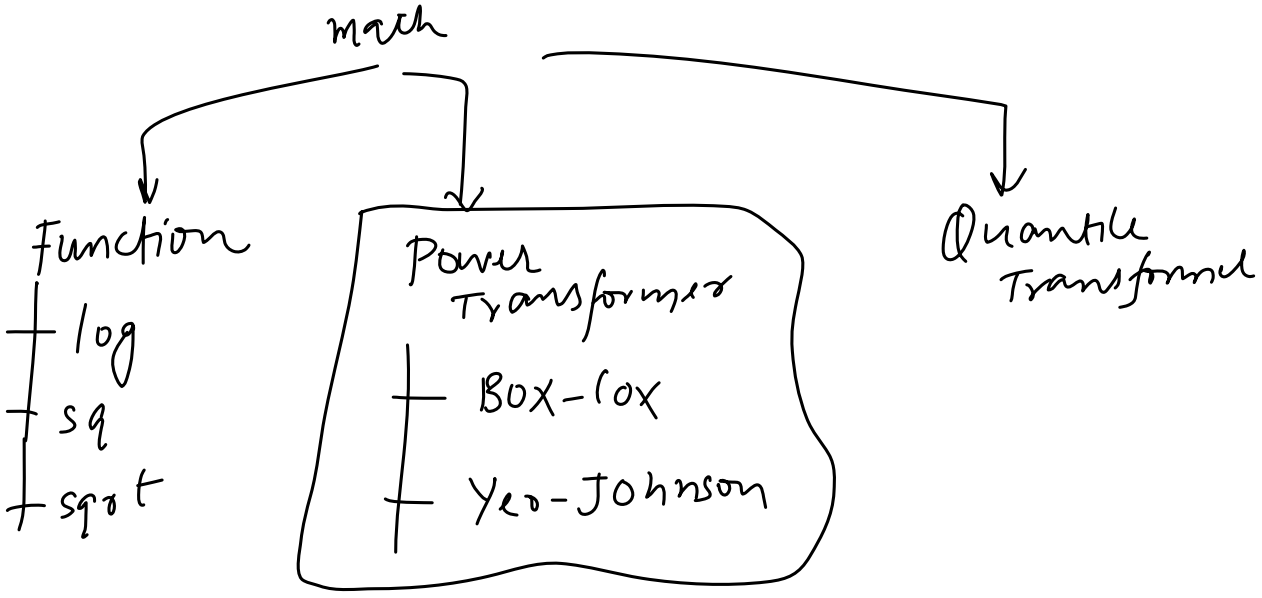
Example

Friday, April 16, 2021 5:22 PM



Power Transformer

Saturday, April 17, 2021 4:55 PM



Box Cox Transform

Saturday, April 17, 2021 4:56 PM

$$x_i^{(\lambda)} = \begin{cases} \dfrac{x_i^{\lambda} - 1}{\lambda} & \text{if } \lambda \neq 0, \\ \ln(x_i) & \text{if } \lambda = 0, \end{cases}$$

given dist.
↳ Normal dist

$x^{\frac{1}{2}}, x^3, x^{1.5}$

1.5, 1.1.

1) Max. likelihood
2) Bayesian

The exponent here is a variable called lambda (λ) that varies over the range of -5 to 5, and in the process of searching, we examine all values of λ. Finally, we choose the optimal value (resulting in the best approximation to a normal distribution) for your variable.

Yeo - Johnson Transform

Saturday, April 17, 2021 4:56 PM

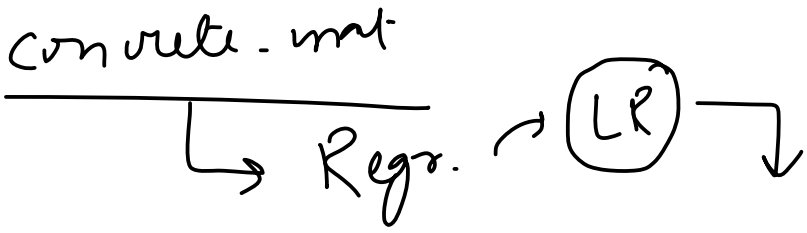
$$x_i^{(\lambda)} = \begin{cases} [(x_i + 1)^{\lambda} - 1]/\lambda & \text{if } \lambda \neq 0, x_i \geq 0, \\ \ln(x_i) + 1 & \text{if } \lambda = 0, x_i \geq 0 \\ -[(-x_i + 1)^{2-\lambda} - 1]/(2 - \lambda) & \text{if } \lambda \neq 2, x_i < 0, \\ -\ln(-x_i + 1) & \text{if } \lambda = 2, x_i < 0 \end{cases}$$

This transformation is somewhat of an adjustment to the Box-Cox transformation, by which we can apply it to negative numbers.

Power Transformer

Example

Saturday, April 17, 2021 4:56 PM



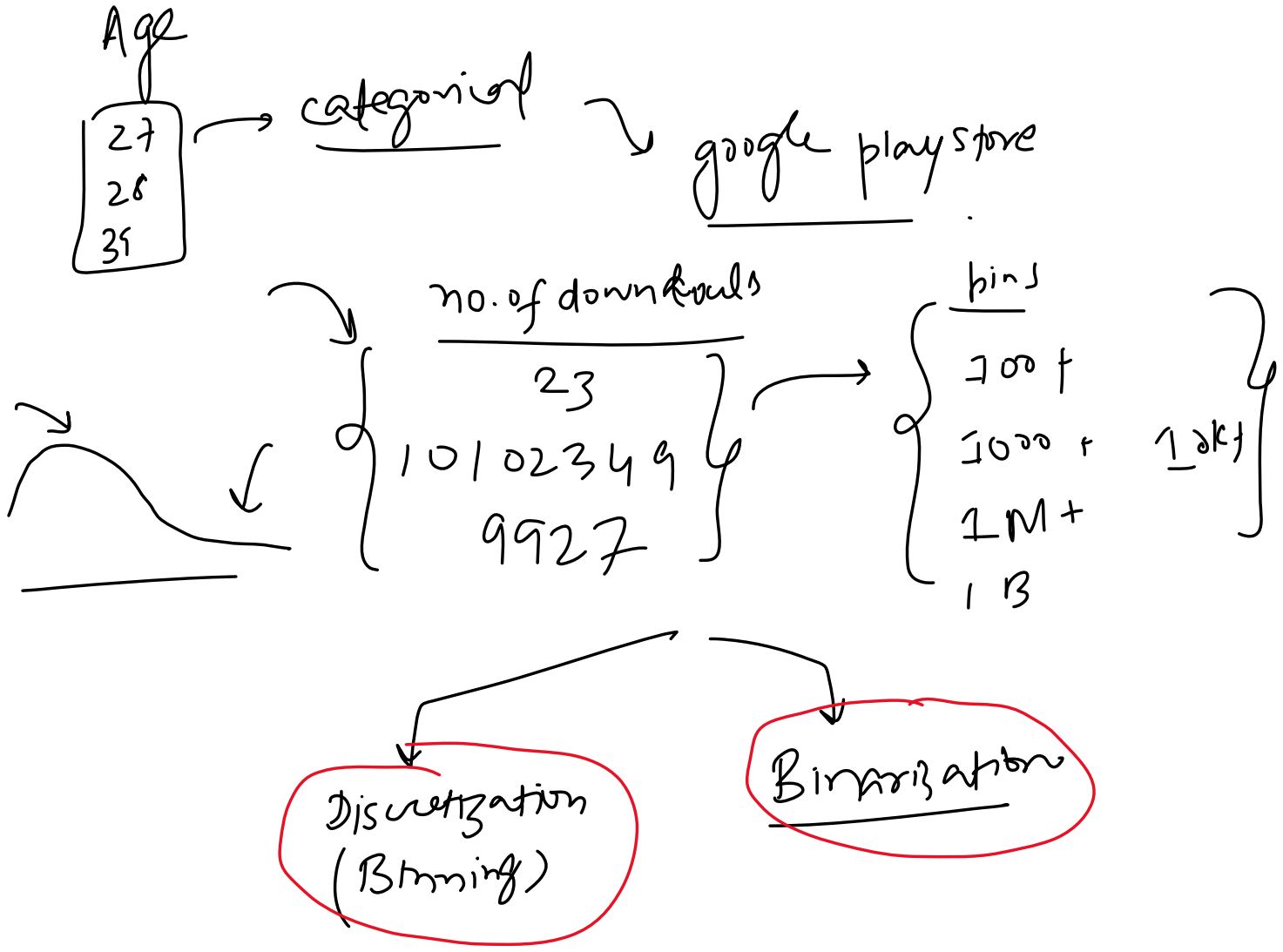
Cement

$(540)^{0.177}$ ⟶

$(30)^{(0.02)}$

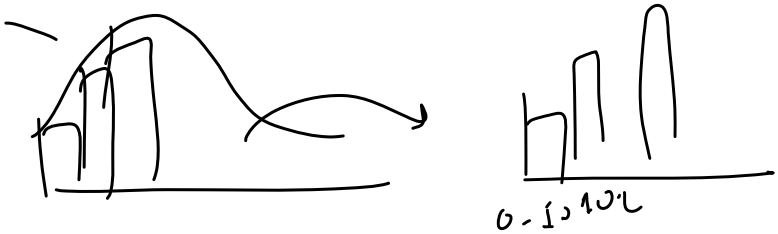
1. Encoding Numerical Features

Monday, April 19, 2021 4:18 PM



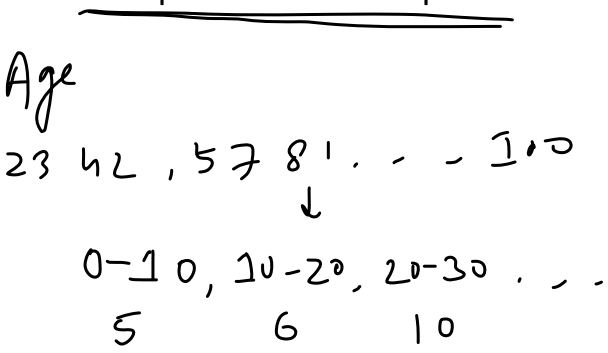
2. Discretization

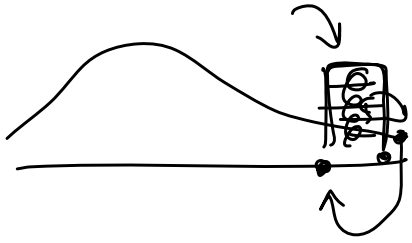
Discretization is the process of transforming continuous variables into discrete variables by creating a set of contiguous intervals that span the range of the variable's values. Discretization is also called binning, where bin is an alternative name for interval.



Why use Discretization:

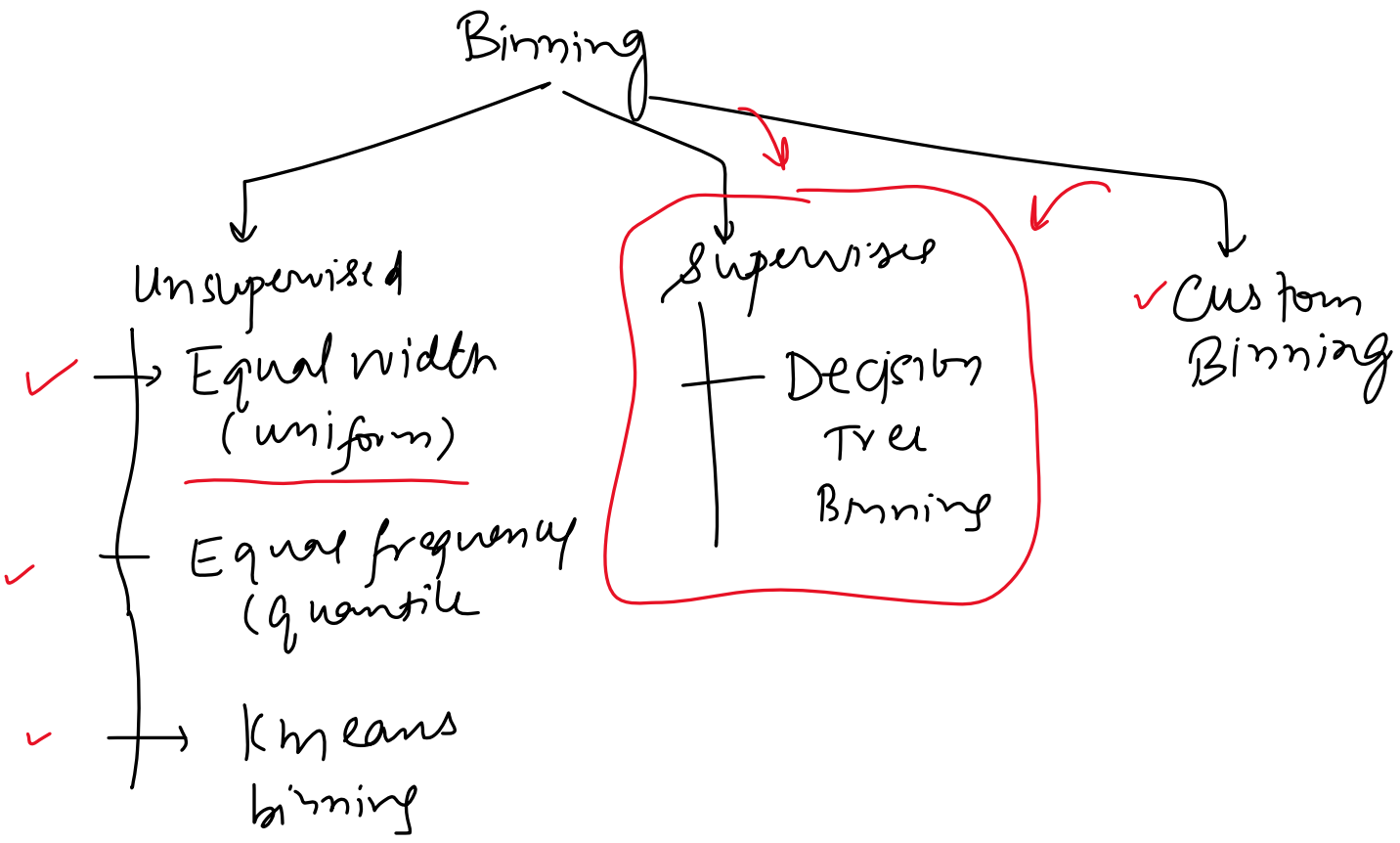
1. To handle Outliers
2. To improve the value spread





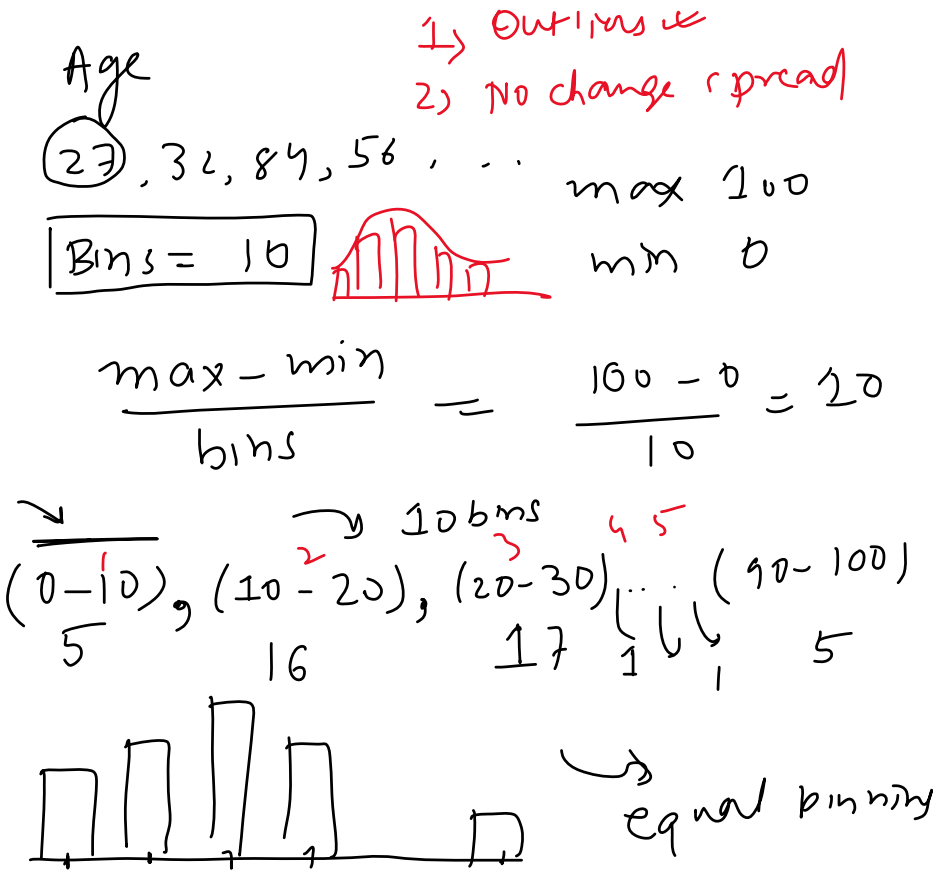
3. Types of Discretization

Monday, April 19, 2021 3:56 PM



4. Equal Width/Uniform Binning

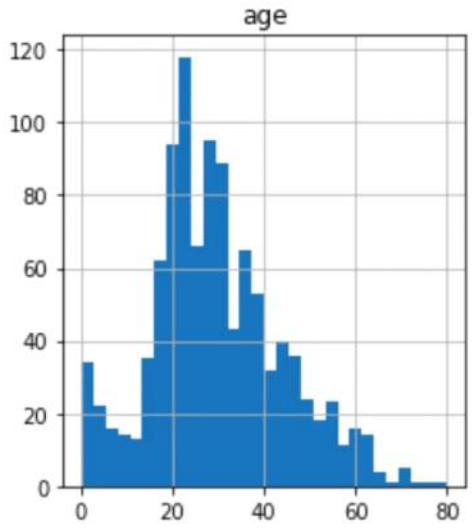
Monday, April 19, 2021 3:56 PM



	age	age_trf	age_labels
314	43.0	5.0	(40.21, 48.168]
523	44.0	5.0	(40.21, 48.168]
352	15.0	1.0	(8.378, 16.336]
534	30.0	3.0	(24.294, 32.252]
211	35.0	4.0	(32.252, 40.21]
530	2.0	0.0	(0.42, 8.378]
786	18.0	2.0	(16.336, 24.294]
827	1.0	0.0	(0.42, 8.378]
372	19.0	2.0	(16.336, 24.294]
518	36.0	4.0	(32.252, 40.21]

5. Equal Frequency/Quantile Binning

Monday, April 19, 2021 3:56 PM



Intervals = 10

Each interval contains 10% of total observations

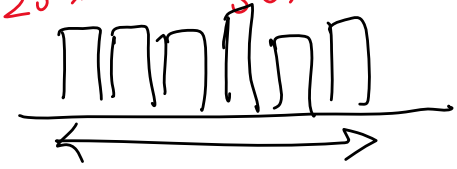
Intervals:
0-16; 16-20; 20-22; 22-25; ...
50-74

1) Outliers
2) Value spread
spread

0 — ? — 10th percentile

0-16 | 16-20 | 20-22

10% 20% 30%

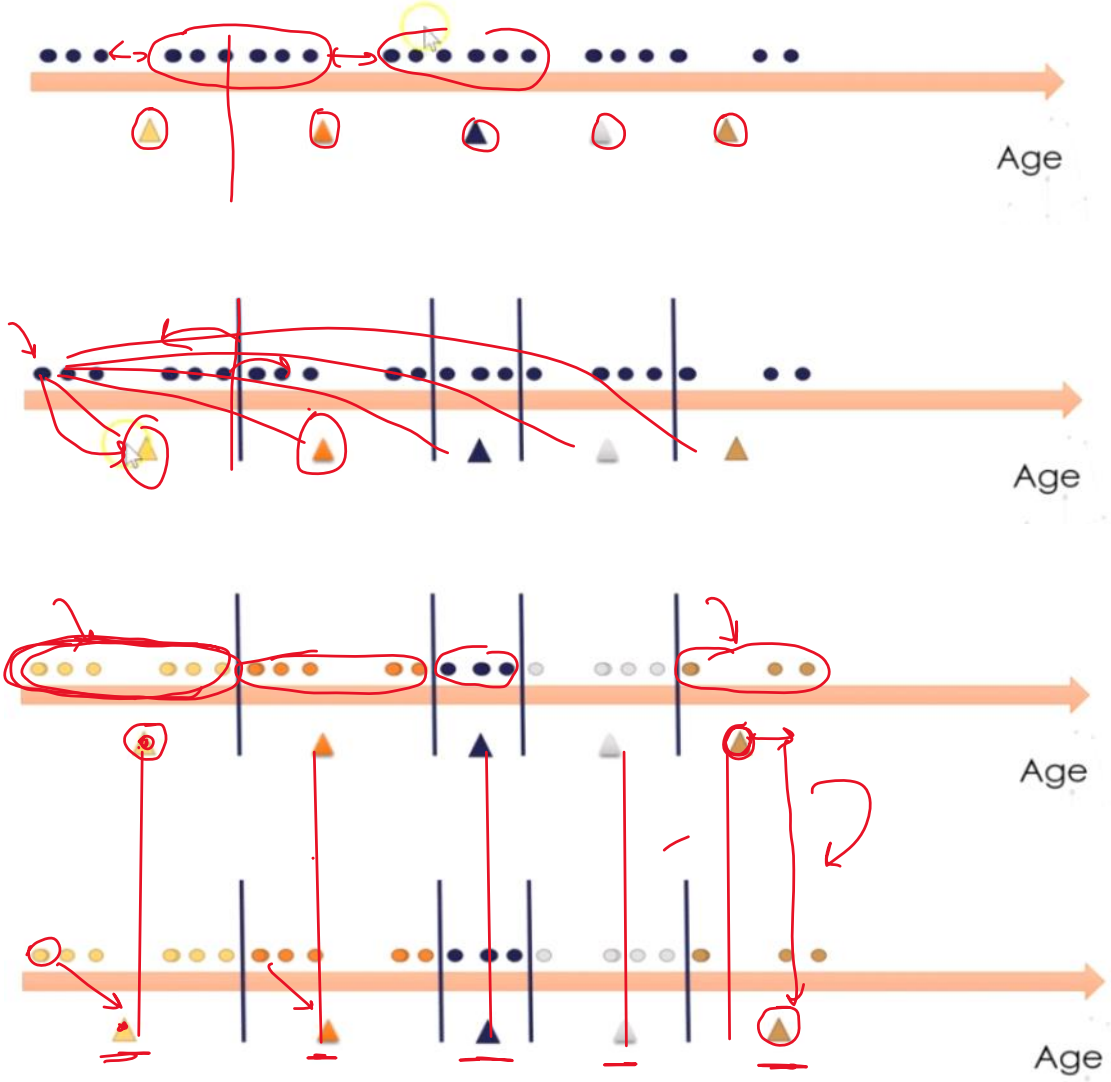


6. KMeans Binning

Monday, April 19, 2021 3:57 PM

centroid

interval-5



KMeans

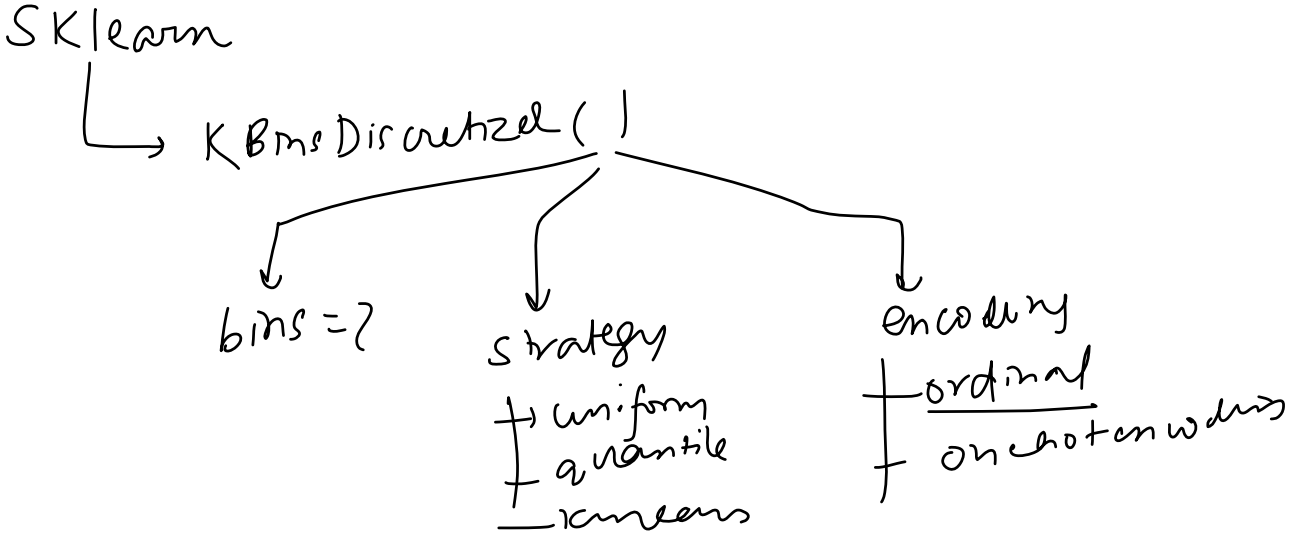
↳ Clustering

2D

4 0D

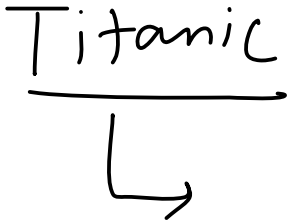
7. Encoding the discretized variable

Monday, April 19, 2021 3:58 PM



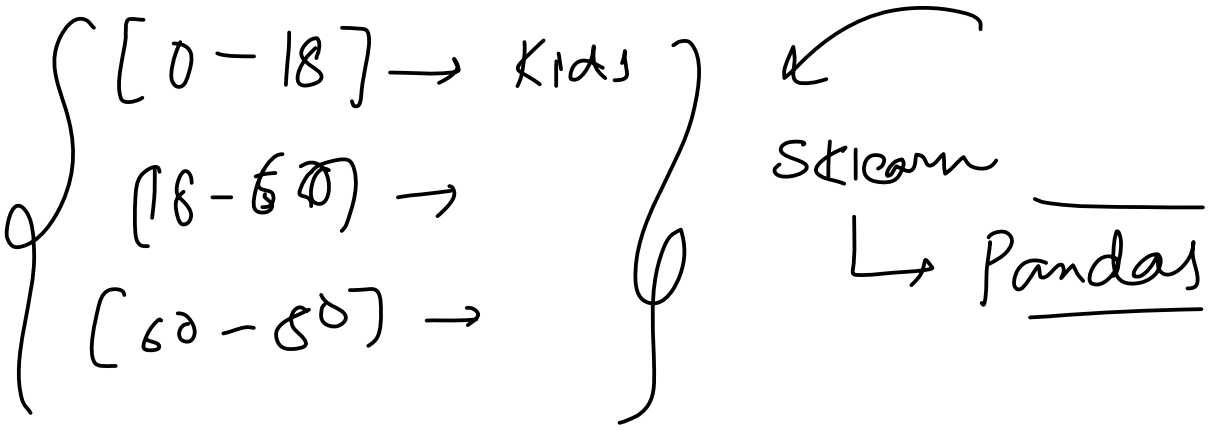
8. Example

Monday, April 19, 2021 3:58 PM



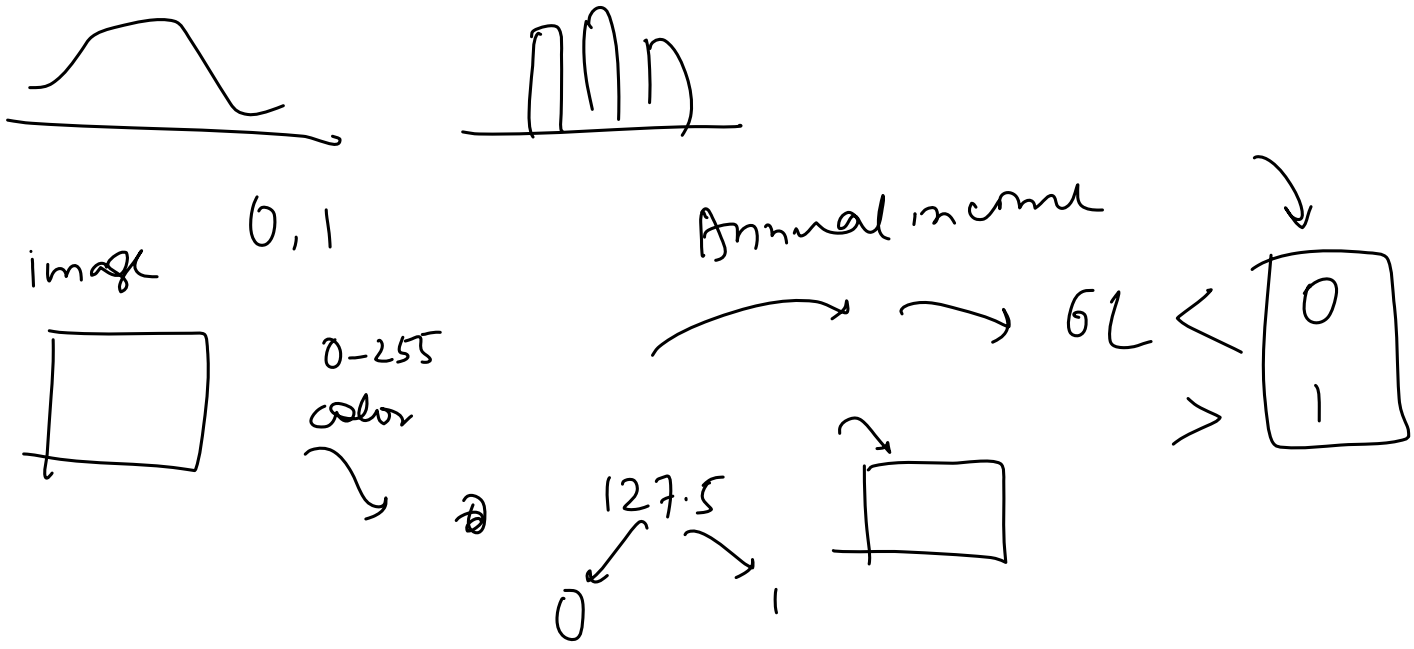
9. Custom/Domain Based Binning

Monday, April 19, 2021 3:58 PM



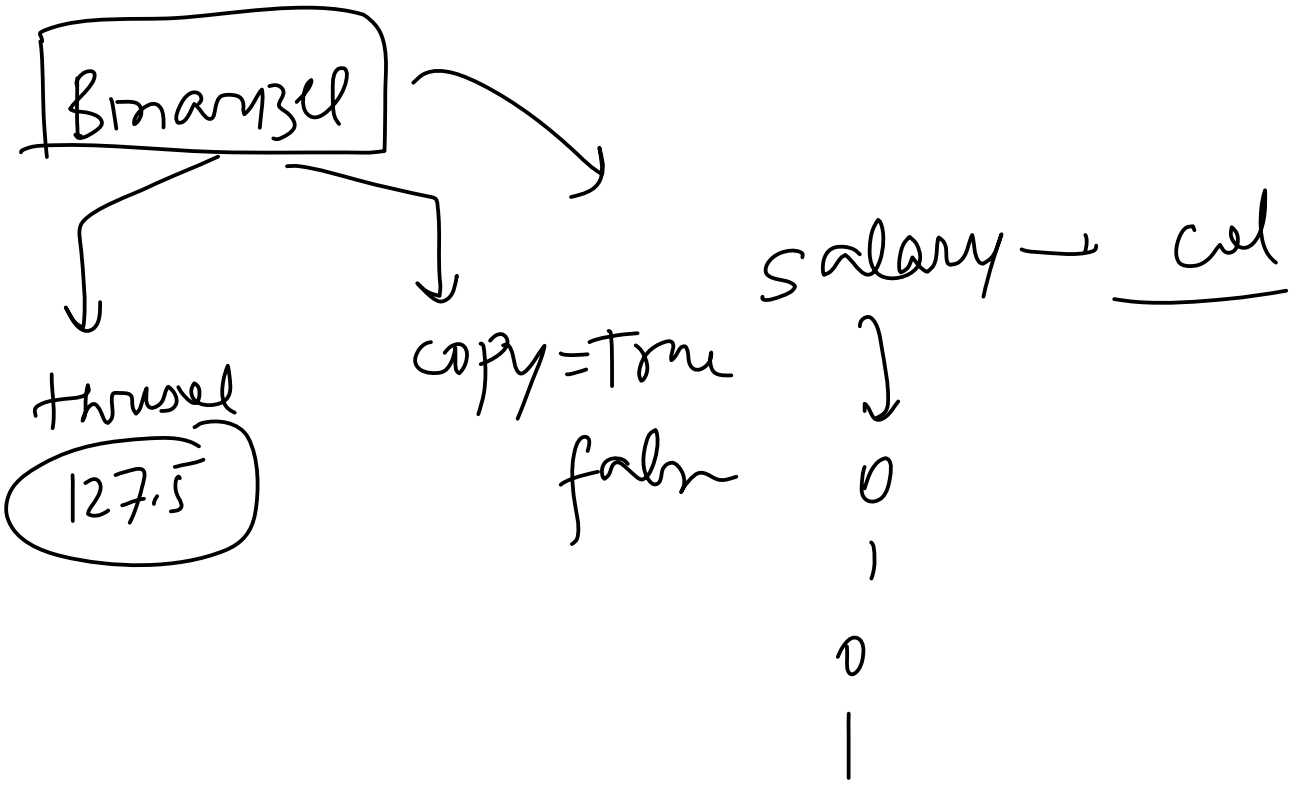
10. Binarization

Monday, April 19, 2021 4:17 PM



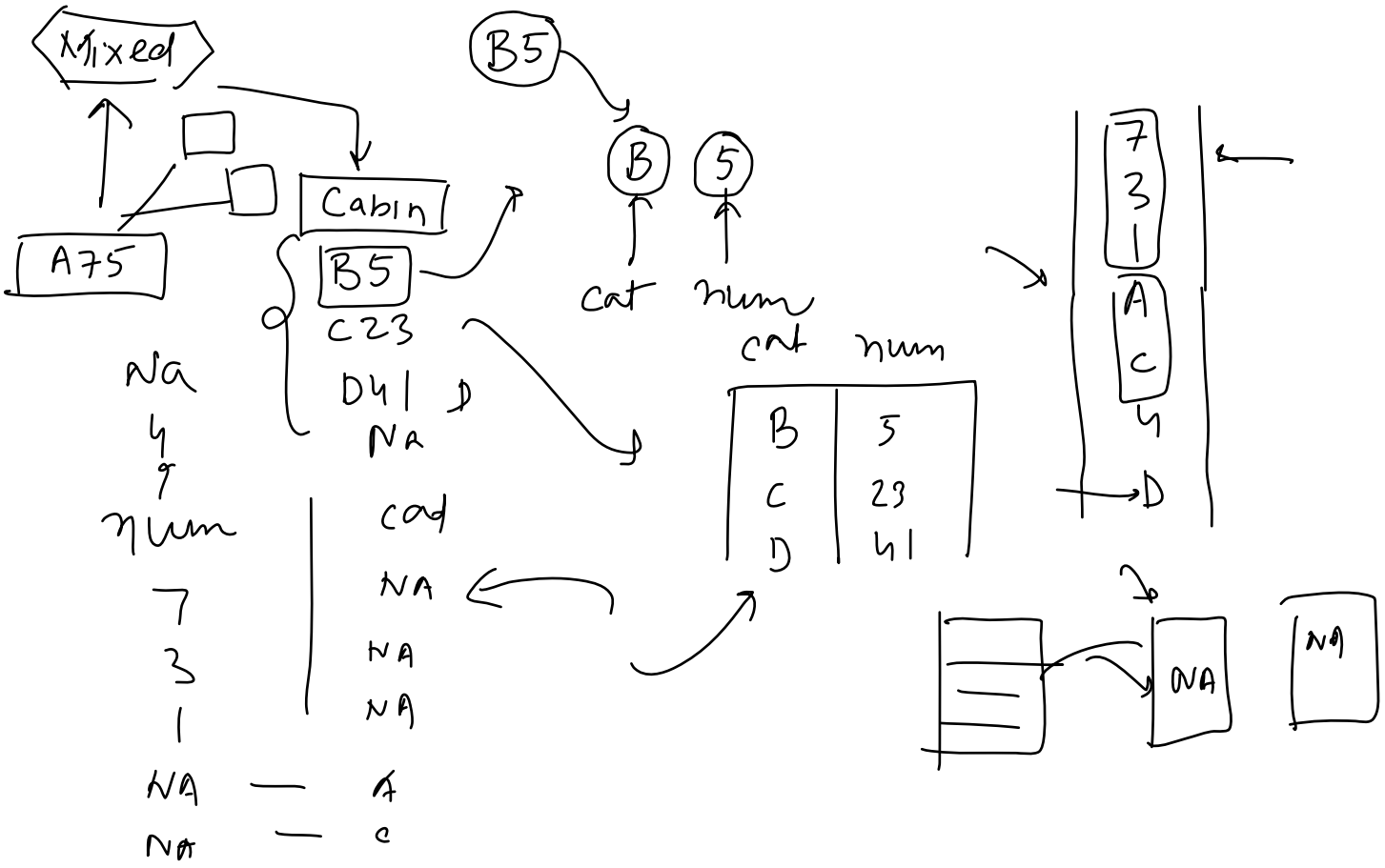
11. Example

Monday, April 19, 2021 4:17 PM



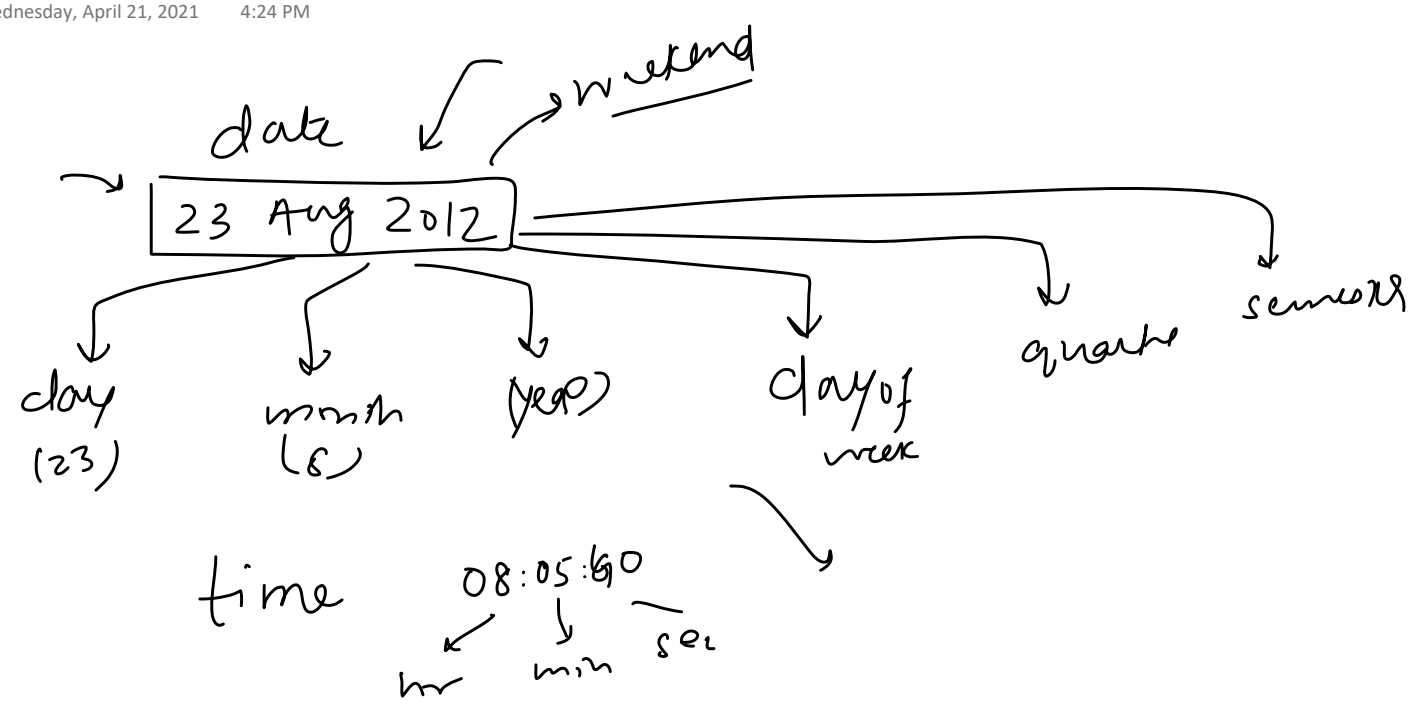
Mixed Data

Tuesday, April 20, 2021 4:24 PM



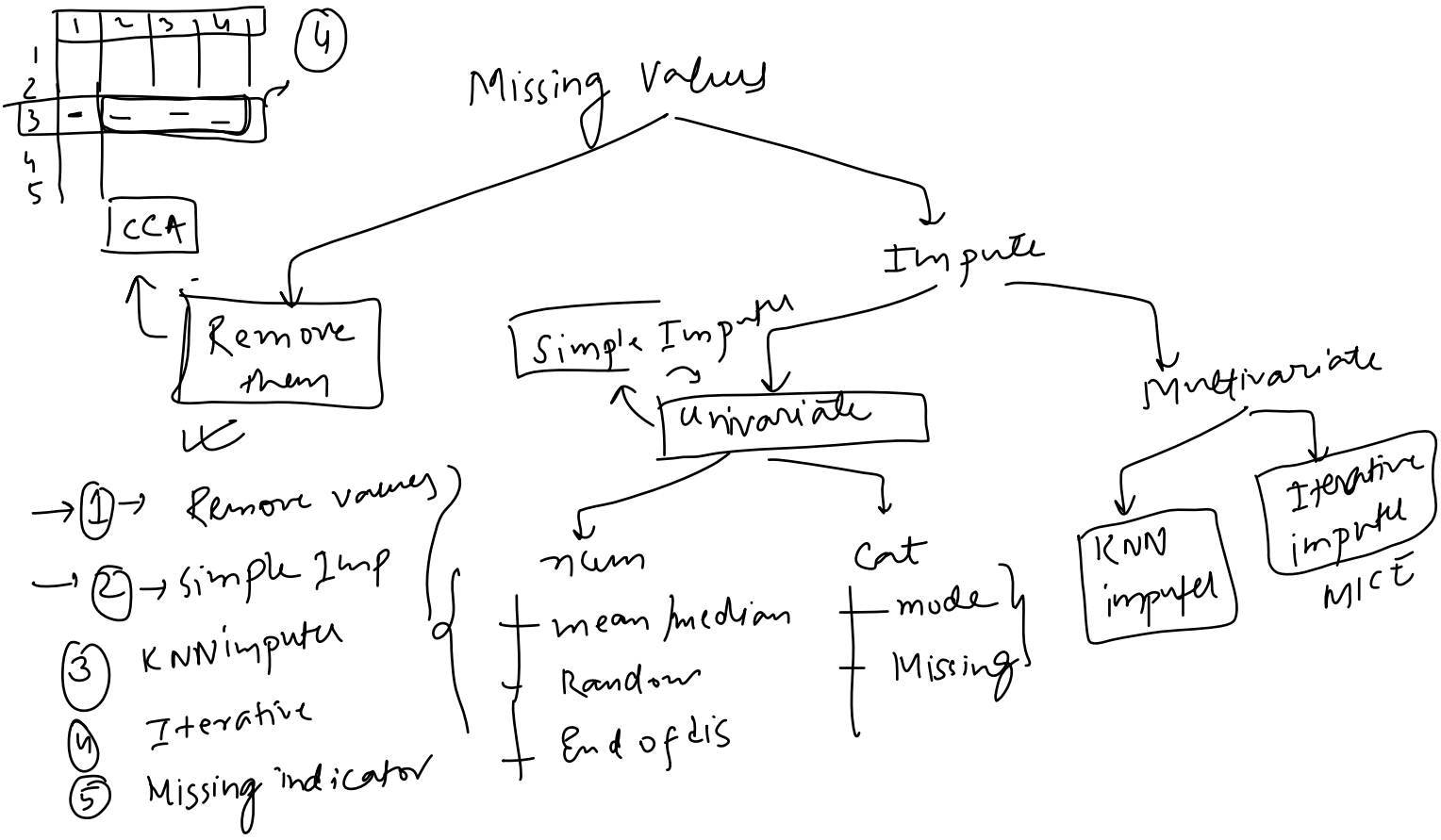
Working with date and time

Wednesday, April 21, 2021 4:24 PM



Handling Missing Data

Thursday, April 22, 2021 1:03 PM



Complete Case Analysis

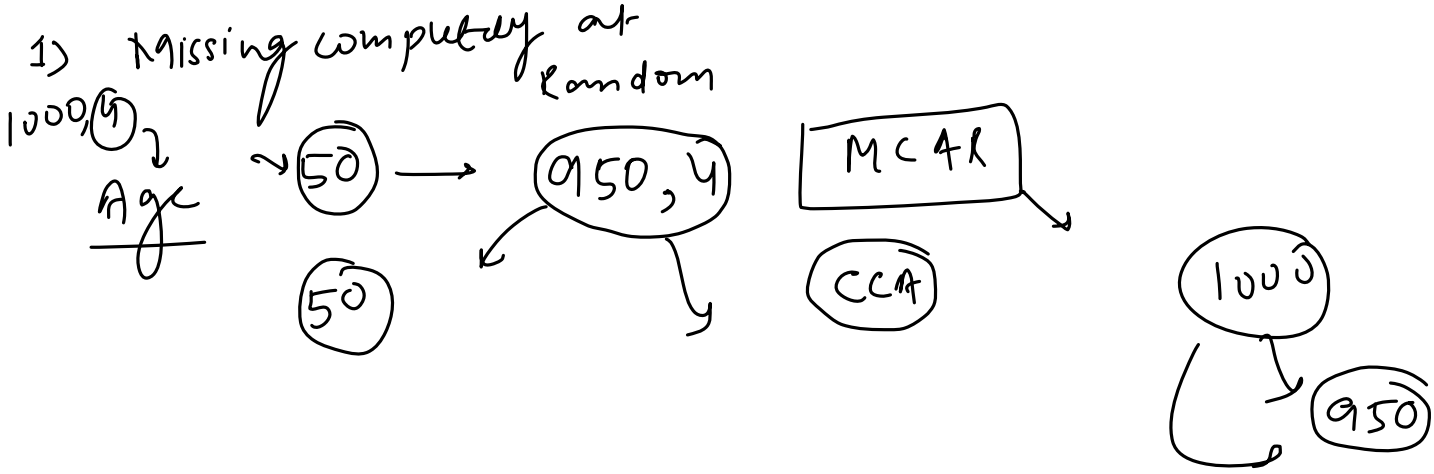
Thursday, April 22, 2021 12:55 PM

Complete-case analysis (CCA), also called "list-wise deletion" of cases, consists in **discarding** observations where values in **any** of the variables are missing.

Complete Case Analysis means literally analyzing only those observations for which there is information in **all** of the variables in the dataset.

Assumption For CCA

Thursday, April 22, 2021 12:58 PM



Advantage/Disadvantage

Thursday, April 22, 2021 12:59 PM

Advantage

1. Easy to implement as no data manipulation required
2. Preserves variable distribution (if data is MCAR, then the distribution of the variables of the reduced dataset should match the distribution in the original dataset

Disadvantage

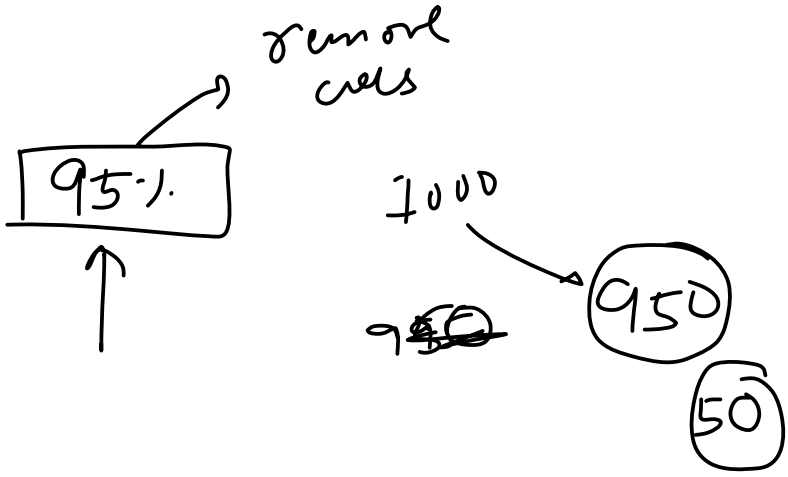
1. It can exclude a large fraction of the original dataset (if missing data is abundant)
2. Excluded observations could be informative for the analysis (if data is not missing at random)
3. When using our models in production, the model will not know how to handle missing data

When to use CCA?

Thursday, April 22, 2021 1:02 PM

1) MCAR ✓

2) 5% <

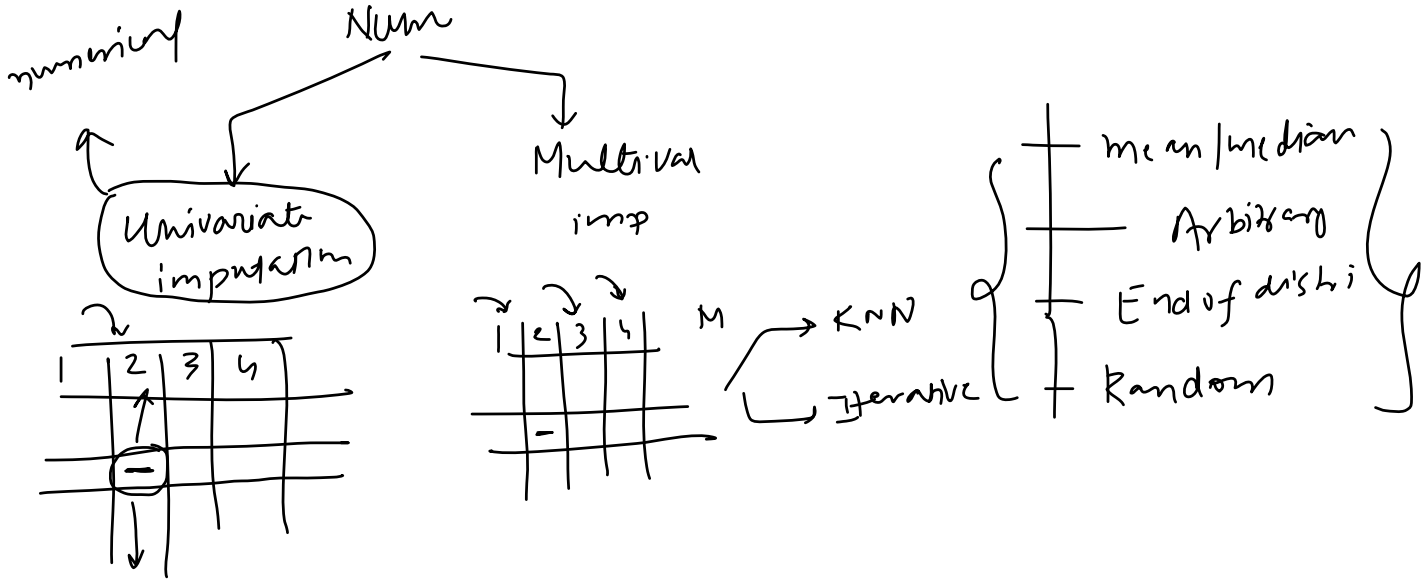


Example

Thursday, April 22, 2021 1:03 PM

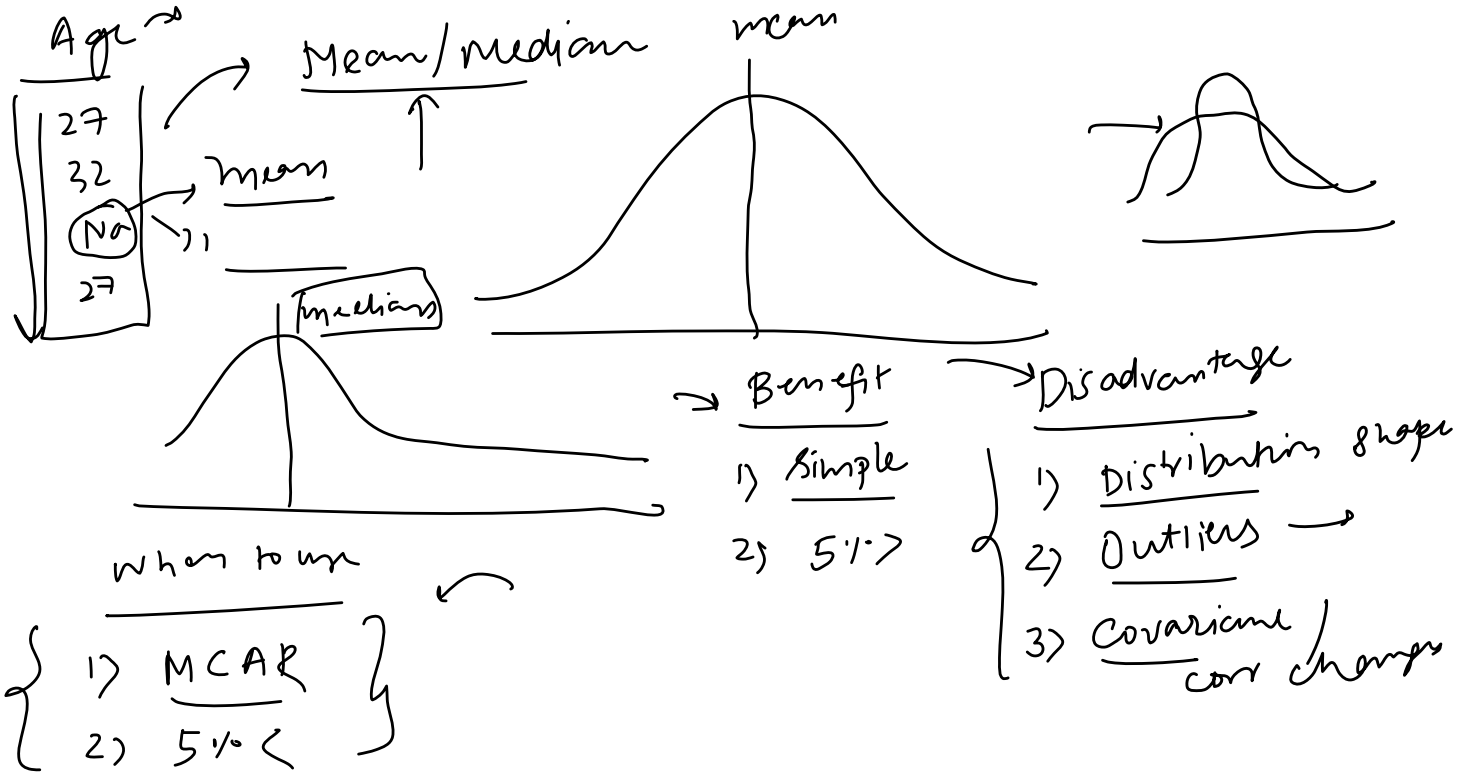
1. Handling Missing Numerical Data

Friday, April 23, 2021 11:28 AM



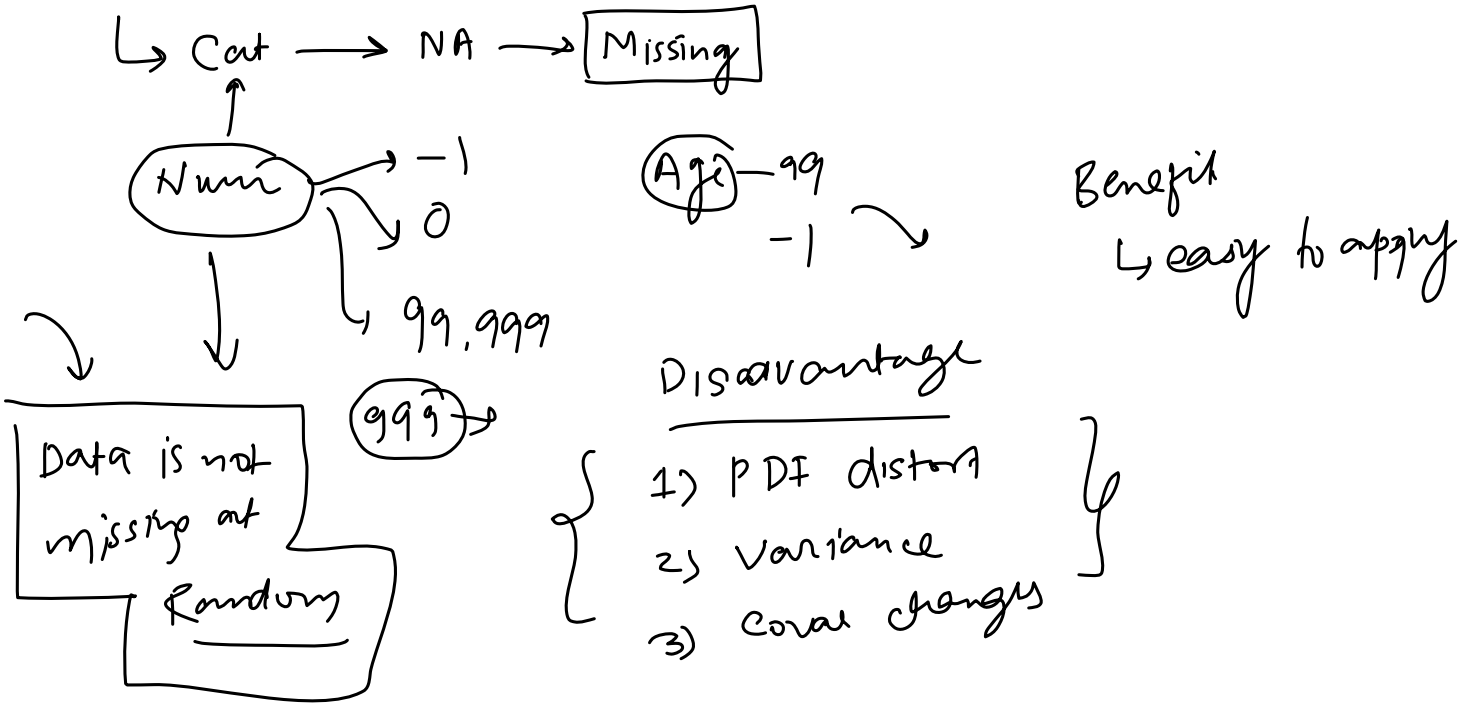
2. Mean/Median Imputation

Friday, April 23, 2021 11:31 AM



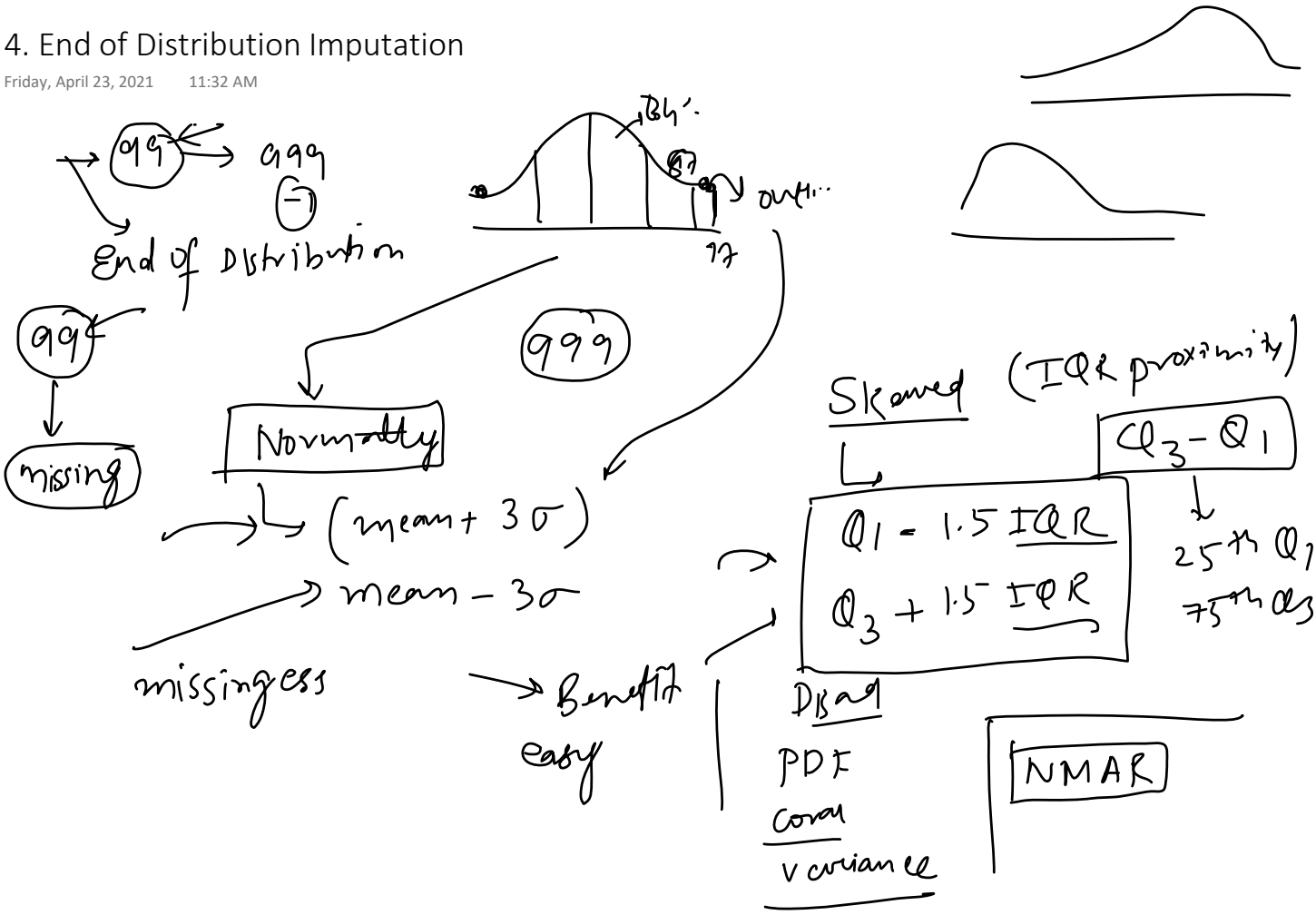
3. Arbitrary Value Imputation

Friday, April 23, 2021 11:32 AM



4. End of Distribution Imputation

Friday, April 23, 2021 11:32 AM



5. Random Sample Imputation

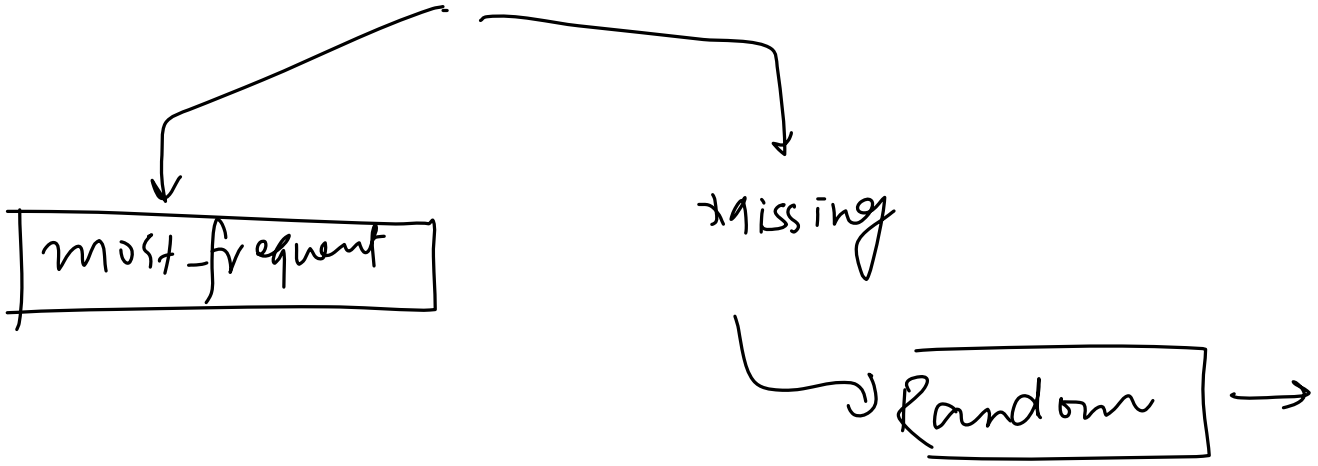
Friday, April 23, 2021 11:33 AM

6. Automatically select best imputation technique

Friday, April 23, 2021 11:32 AM

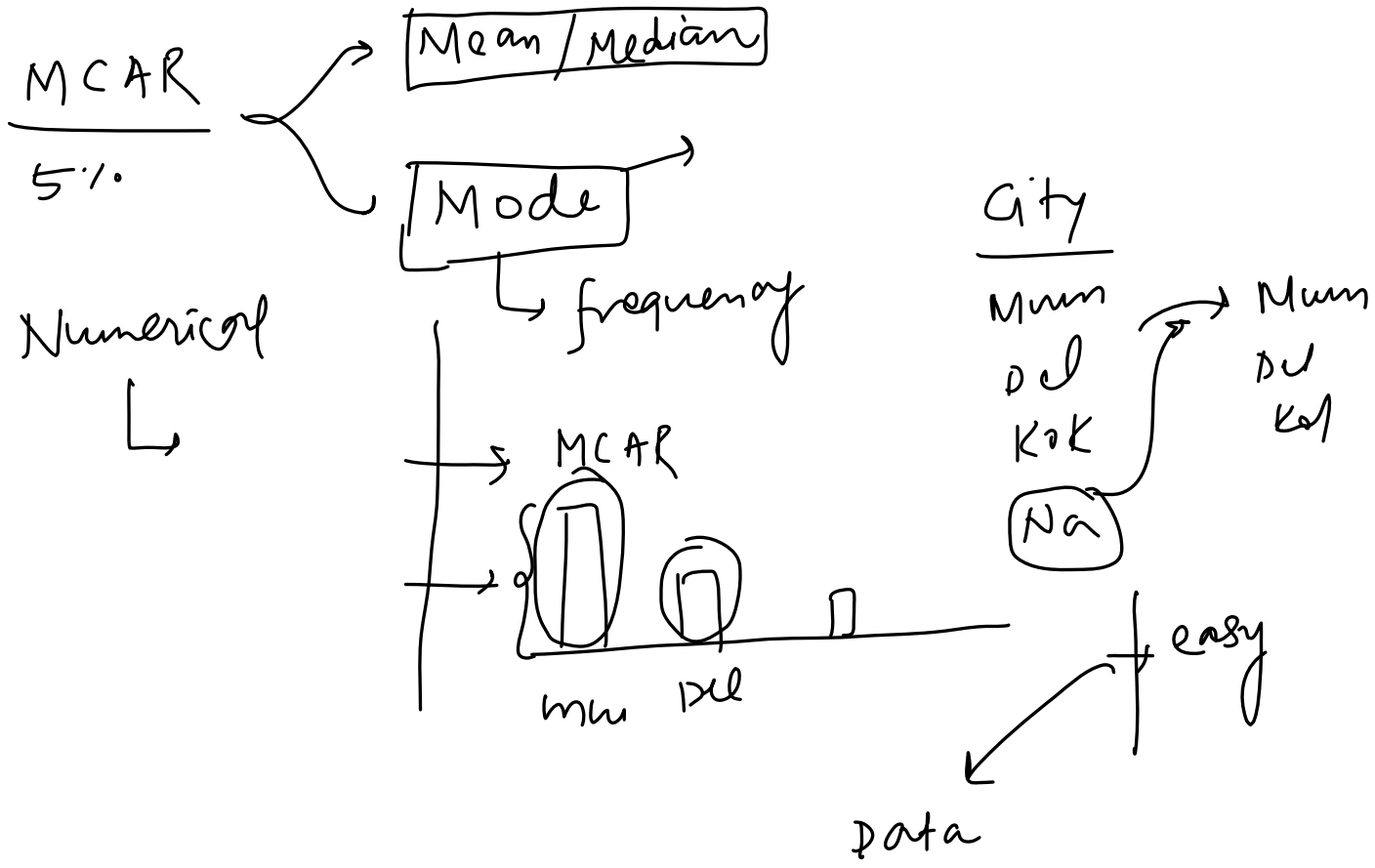
Handling Categorical Missing Data

Saturday, April 24, 2021 5:16 PM



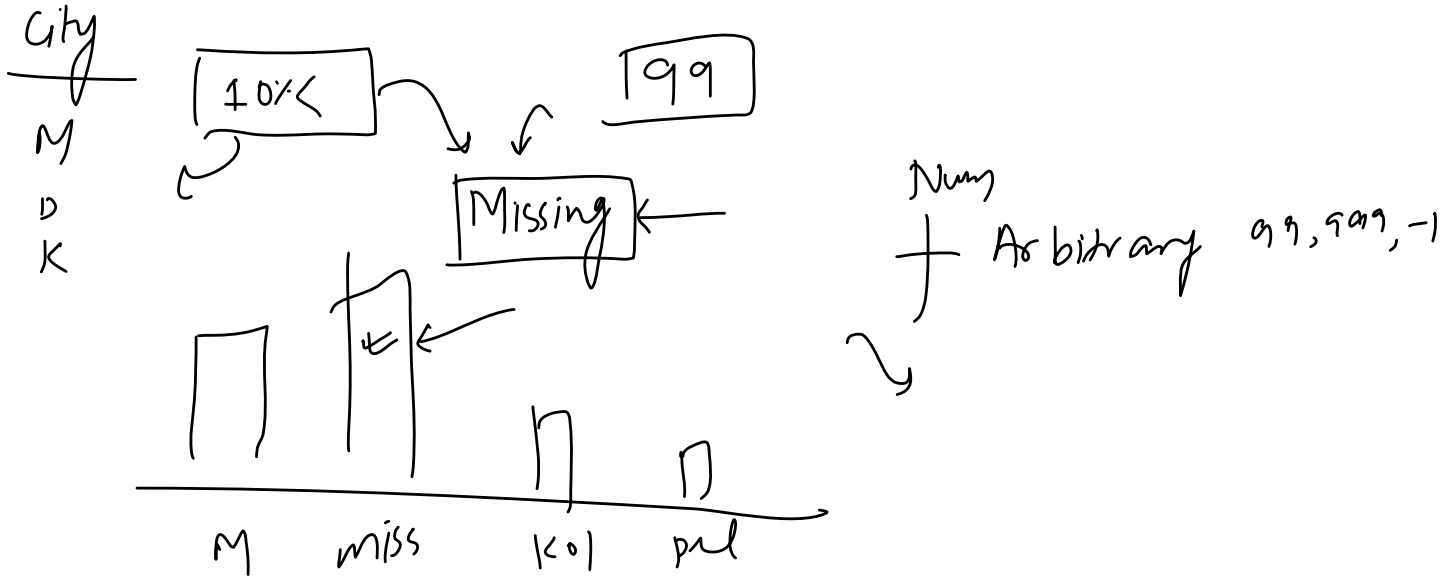
Most Frequent Value Imputation

Saturday, April 24, 2021 5:17 PM



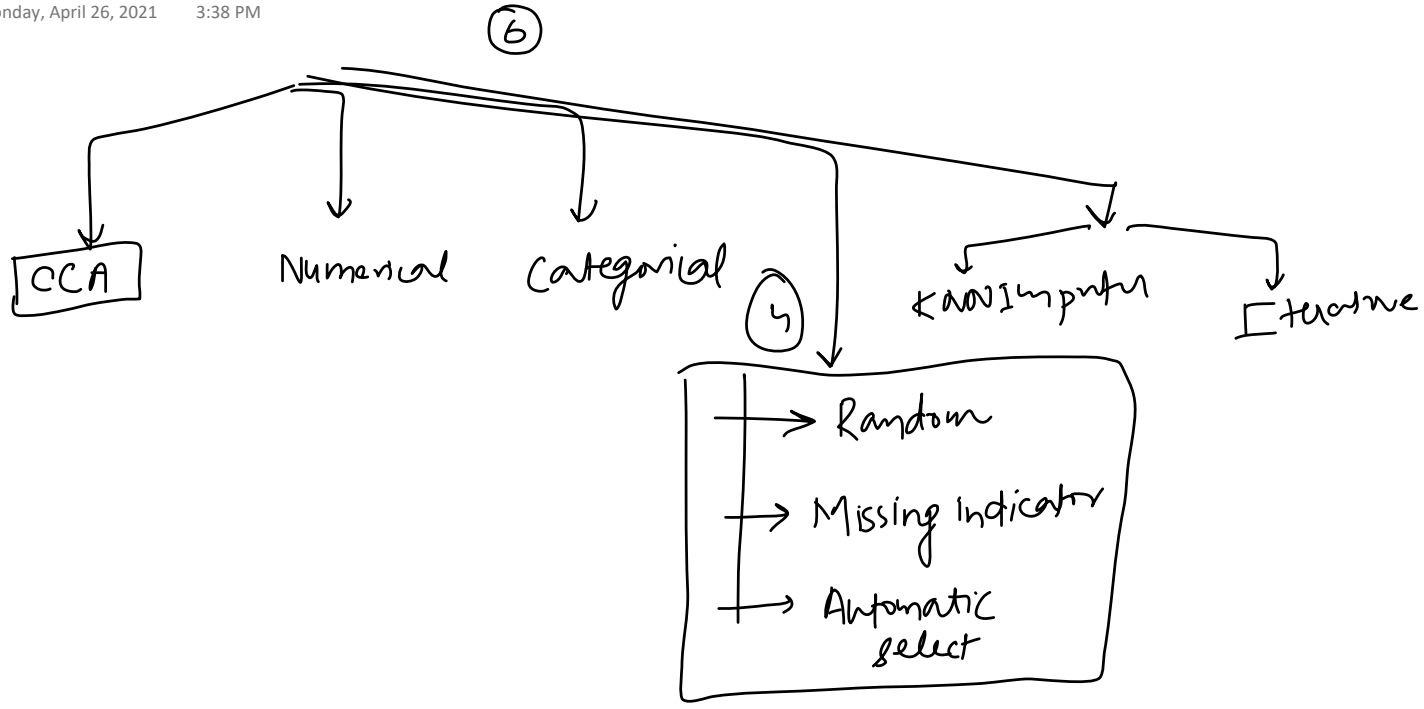
Missing Category Imputation

Saturday, April 24, 2021 5:17 PM



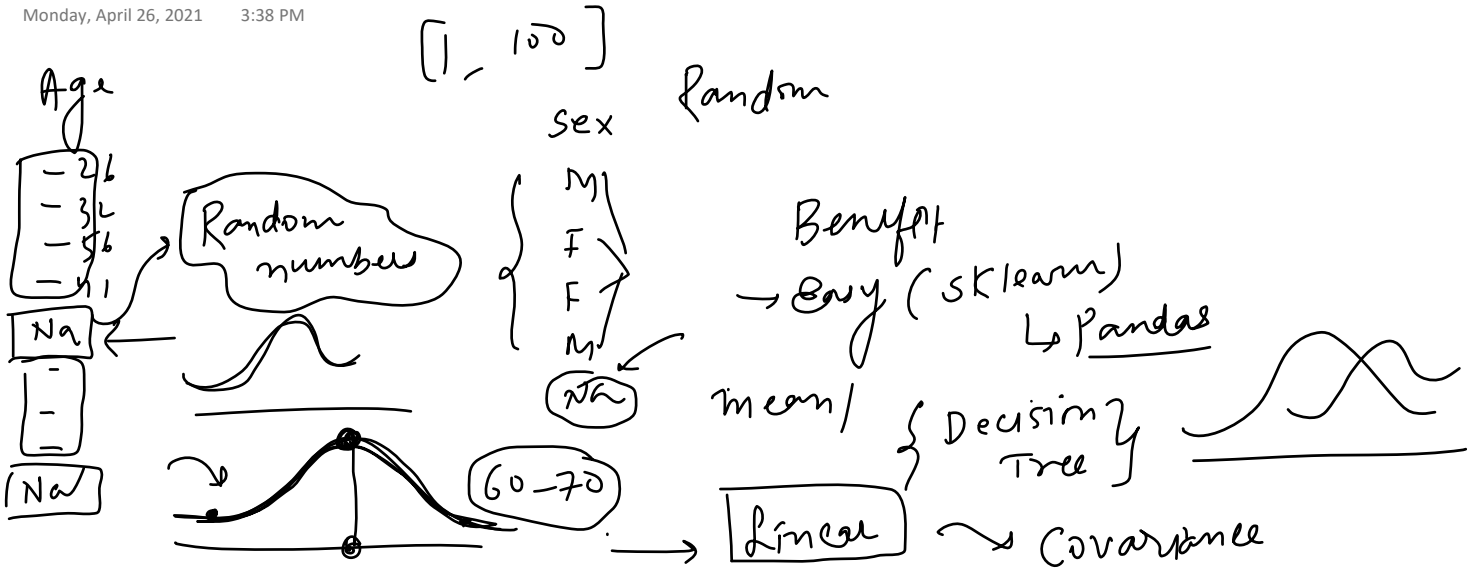
Topics to be covered

Monday, April 26, 2021 3:38 PM

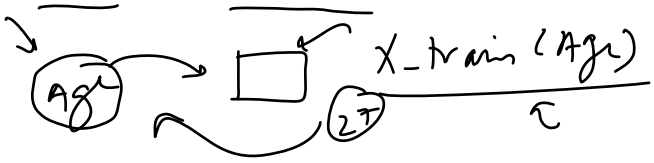


Random Imputation

Monday, April 26, 2021 3:38 PM

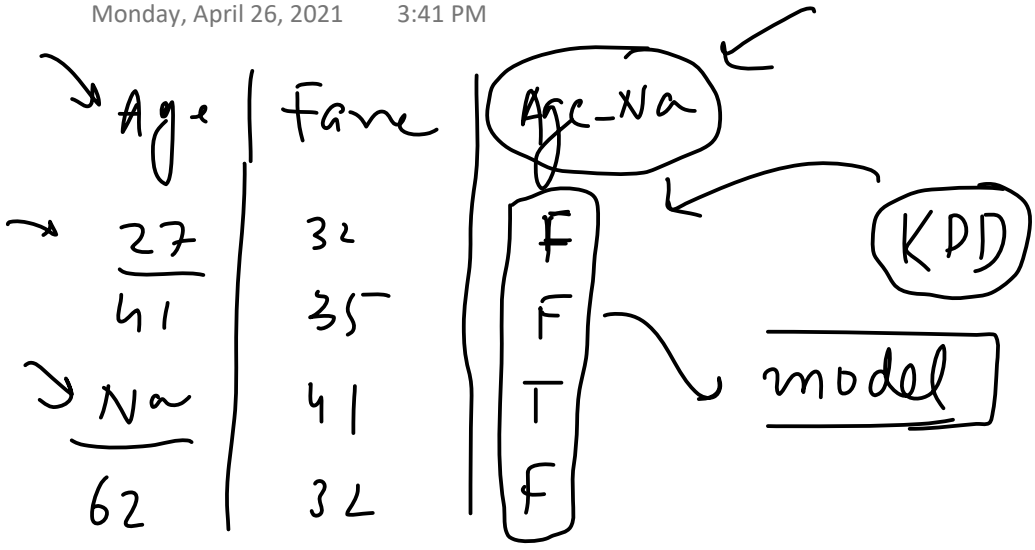


1. Preserves the variance of the variable(But Why?)
2. Memory heavy for deployment, as we need to store the original training set to extract values from and replace the NA in coming observations
3. Well suited for linear models as it does not distort the distribution, regardless of the % of NA



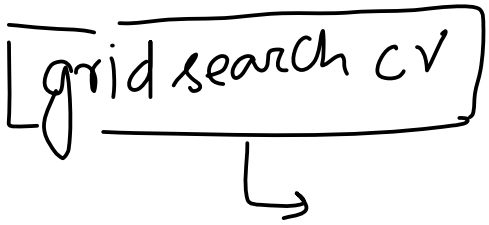
Missing Indicator

Monday, April 26, 2021 3:41 PM



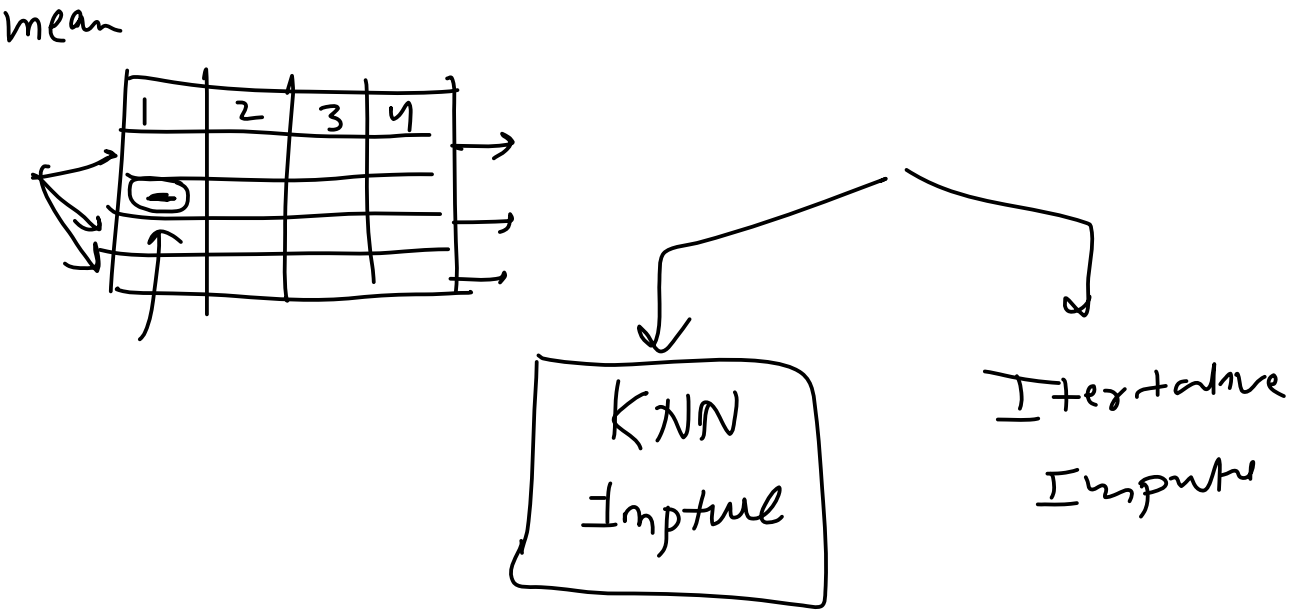
Automatically select value for Imputation

Monday, April 26, 2021 3:41 PM



KNN Imputer

Tuesday, April 27, 2021 12:19 PM



Working

Tuesday, April 27, 2021 7:44 AM

K=2

S NO	Feature 1	Feature 2	Feature 3	Feature 4
1	33	-----	67	21
2	----- 31.5	45	68	12
3 → 2	23	51	71	18
4 → 1	40	-----	81	------
5	35	60	79	------

Distance from Pt 1

$$d = \sqrt{\frac{3}{2}((68-17)^2 + (12-21)^2}$$

$$= \sqrt{1.5(4+81)} = \sqrt{127.5} = 11.29$$

Distance from pt 3

$$d = \sqrt{\frac{3}{3}((51-45)^2 + (71-68)^2 + (18-12)^2}$$

$$= \sqrt{36+9+36} = 9$$

Distance from pt 4

$$d = \sqrt{\frac{3}{1}((81-68))^2}$$

$$d = \sqrt{3(9)} = \sqrt{27} = 5.19$$

Distance from pt 5

$$d = \sqrt{\frac{3}{2}((60-45)^2 + (79-68)^2}$$

$$= \sqrt{1.5(25+121)} = \sqrt{219} = 14.79$$

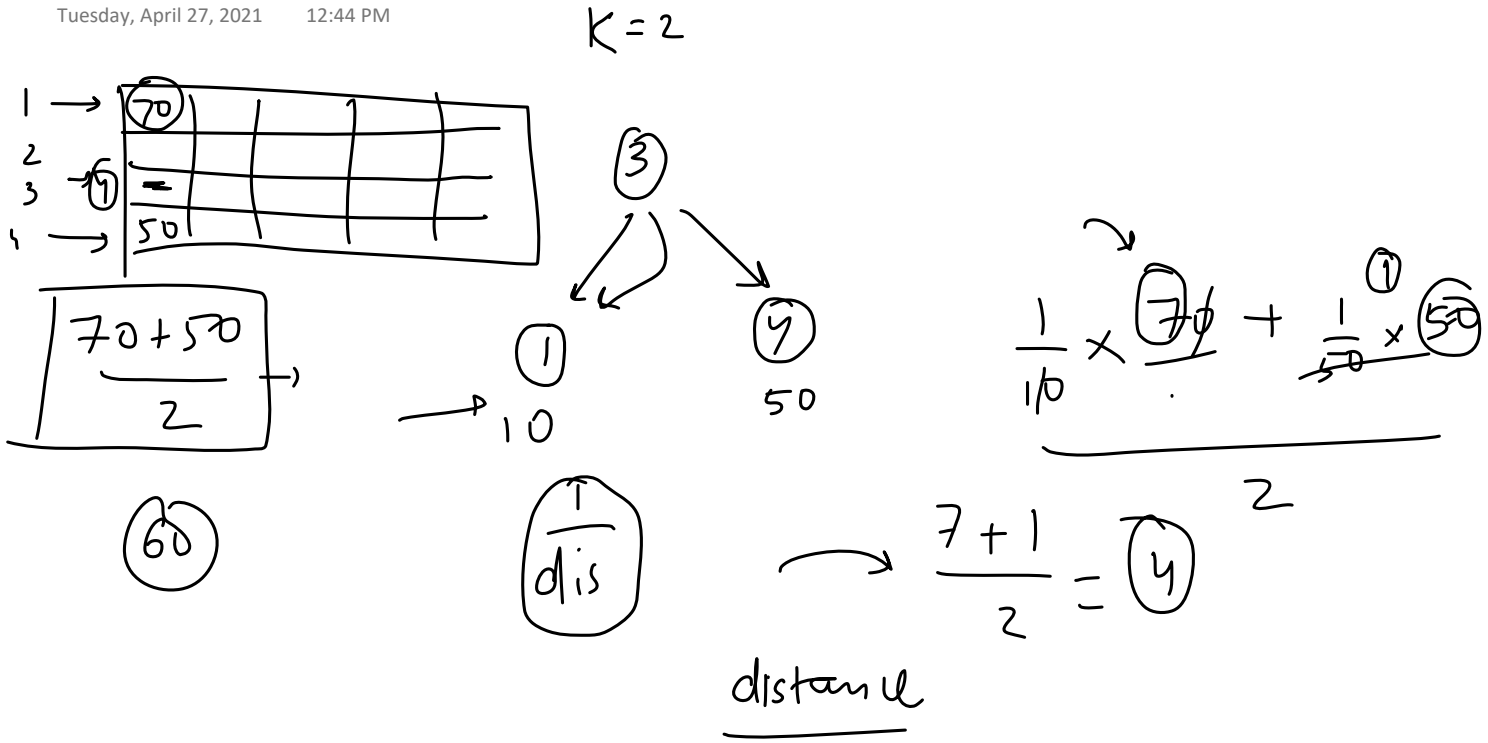
Advantage & Disadvantage

1) More accurate

2) More no. of calculations

3) Prediction → X_train

Tuesday, April 27, 2021 12:44 PM

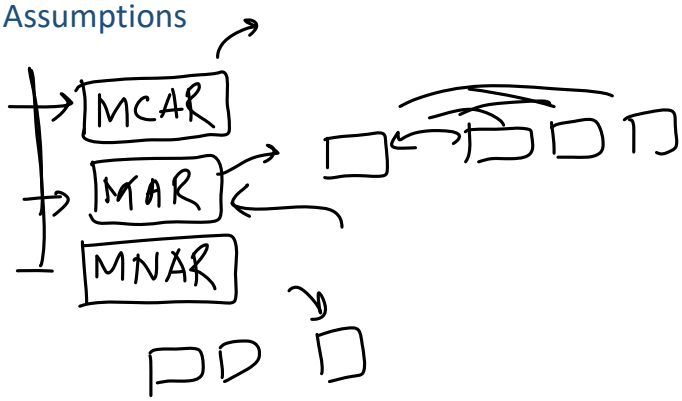


Iterative Imputer/MICE

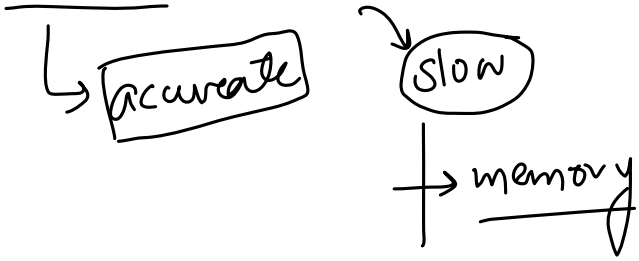
Wednesday, April 28, 2021 11:40 AM

MICE stands for Multivariate Imputation by Chained Equations

Assumptions



Advantages & Disadvantage



How it works?

Wednesday, April 28, 2021 11:41 AM

1. Actual Dataset

	R&D Spend	Administration	Marketing Spend	Profit
21	8.0	15.0	30.0	11.0
37	4.0	5.0	20.0	9.0
2	15.0	10.0	41.0	19.0
14	12.0	16.0	26.0	13.0
44	2.0	15.0	3.0	7.0

2. Removing the target column

	R&D Spend	Administration	Marketing Spend
21	8.0	15.0	30.0
37	4.0	5.0	20.0
2	15.0	10.0	41.0
14	12.0	16.0	26.0
44	2.0	15.0	3.0

3. Introduced some fake nan values

	R&D Spend	Administration	Marketing Spend
21	8.0	15.0	30.0
37	NaN	5.0	20.0
2	15.0	10.0	41.0
14	12.0	NaN	26.0
44	2.0	15.0	NaN

Step 1 - Fill all the NaN values with mean of respective cols

	R&D Spend	Administration	Marketing Spend
21	8.00	15.00	30.00
37	9.25	5.00	20.00
2	15.00	10.00	41.00
14	12.00	11.25	26.00
44	2.00	15.00	29.25

Step 2 - Remove all col1 missing values

Step 3 - Predict the missing values of col1 using other cols

	R&D Spend	Administration	Marketing Spend
21	8.0	15.00	30.00
37	NaN	5.00	20.00
2	15.0	10.00	41.00
14	12.0	11.25	26.00
44	2.0	15.00	29.25

	Administration	Marketing Spend
21	15.00	30.00
2	10.00	41.00
14	11.25	26.00
44	15.00	29.25

	R&D Spend	Administration	Marketing Spend
21	8.00	15.00	30.00
37	23.14	5.00	20.00
2	15.00	10.00	41.00
14	12.00	11.25	26.00
44	2.00	15.00	29.25

Step 4 - Remove all col2 missing values

	R&D Spend	Administration	Marketing Spend
21	8.00	15.0	30.00
37	23.14	5.0	20.00
2	15.00	10.0	41.00
14	12.00	NaN	26.00
44	2.00	15.0	29.25

	R&D Spend	Marketing Spend
21	8.00	30.00
37	23.14	20.00
2	15.00	41.00
44	2.00	29.25

Step 5 - Predict the missing values of col2 using other cols

	R&D Spend	Administration	Marketing Spend
21	8.00	15.00	30.00
37	23.14	5.00	20.00
2	15.00	10.00	41.00
14	12.00	11.06	26.00
44	2.00	15.00	29.25

Step 6 - Remove all col3 values

	R&D Spend	Administration	Marketing Spend
21	8.00	15.00	30.0
37	23.14	5.00	20.0
2	15.00	10.00	41.0
14	12.00	11.06	26.0
44	2.00	15.00	NaN

	R&D Spend	Administration
21	8.00	15.00
37	23.14	5.00
2	15.00	10.00
14	12.00	11.06

	R&D Spend	Administration	Marketing Spend
21	8.00	15.00	30.00
37	23.14	5.00	20.00
2	15.00	10.00	41.00
14	12.00	11.06	26.00
44	2.00	15.00	31.56

Iteration 0

	R&D Spend	Administration	Marketing Spend
21	8.00	15.00	30.00
37	9.25	5.00	20.00
2	15.00	10.00	41.00
14	12.00	11.25	26.00
44	2.00	15.00	29.25

Iteration 1

	R&D Spend	Administration	Marketing Spend
21	8.00	15.00	30.00
37	[illegible]	5.00	20.00
2	15.00	10.00	41.00
14	12.00	[illegible]06	26.00
44	2.00	15.00	[illegible]56

Difference

	R&D Spend	Administration	Marketing Spend
21	0.00	0.00	0.00
37	13.89	0.00	0.00
2	0.00	0.00	0.00
14	0.00	-0.19	0.00
44	0.00	0.00	2.31

Iteration 1

	R&D Spend	Administration	Marketing Spend
21	8.00	15.00	30.00
37	23.14	5.00	20.00
2	15.00	10.00	41.00
14	12.00	11.06	26.00
44	2.00	15.00	31.56

Iteration 2

	R&D Spend	Administration	Marketing Spend
21	8.00	15.00	30.00
37	23.78	5.00	20.00
2	15.00	10.00	41.00
14	12.00	11.22	26.00
44	2.00	15.00	31.56

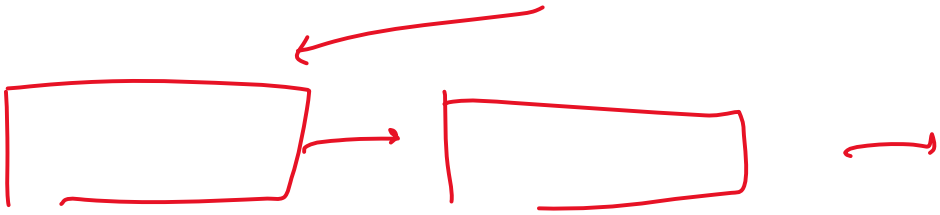
Difference

	R&D Spend	Administration	Marketing Spend
21	0.00	0.00	0.0
37	0.64	0.00	0.0
2	0.00	0.00	0.0
14	0.00	0.16	0.0
44	0.00	0.00	0.0

	R&D Spend	Administration	Marketing Spend
21	8.00	15.00	30.00
37	23.78	5.00	20.00
2	15.00	10.00	41.00
14	12.00	11.22	26.00
44	2.00	15.00	31.56

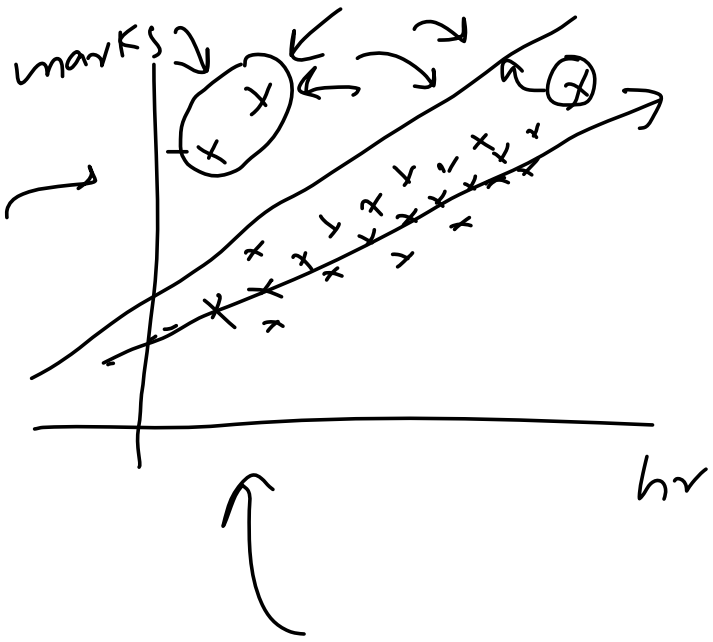
	R&D Spend	Administration	Marketing Spend
21	8.00	15.00	30.00
37	24.57	5.00	20.00
2	15.00	10.00	41.00
14	12.00	11.37	26.00
44	2.00	15.00	45.53

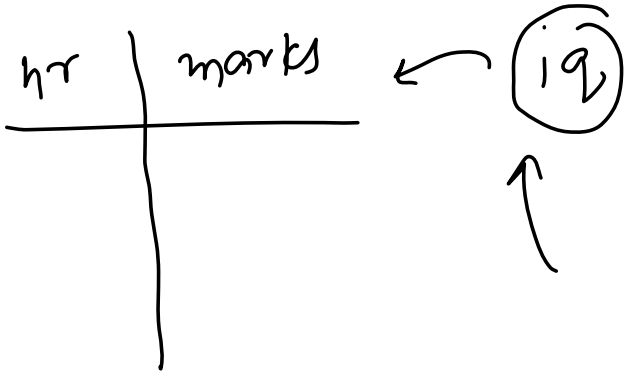
	R&D Spend	Administration	Marketing Spend
21	0.00	0.00	0.00
37	0.79	0.00	0.00
2	0.00	0.00	0.00
14	0.00	0.15	0.00
44	0.00	0.00	13.97



What are Outliers?

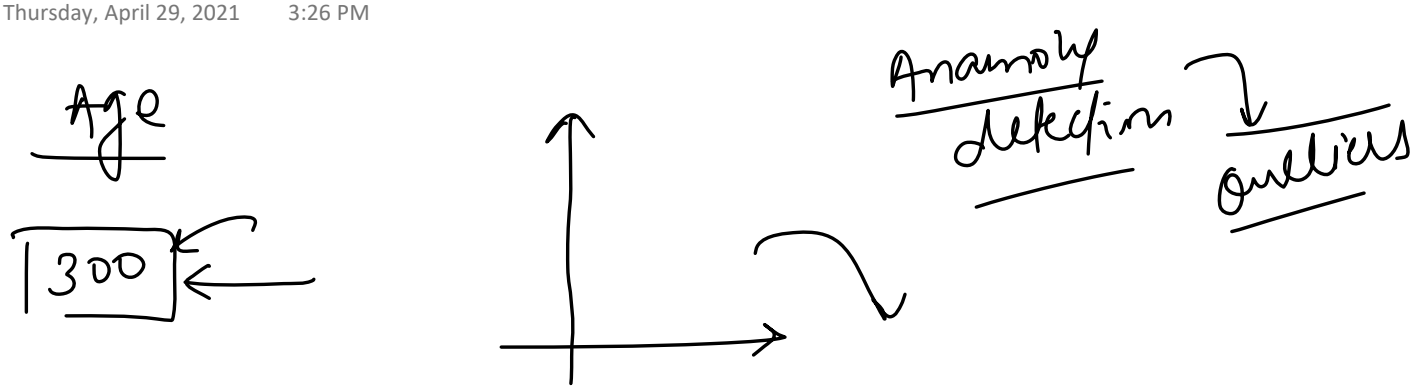
Thursday, April 29, 2021 3:25 PM





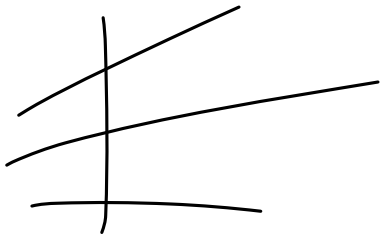
When is outlier dangerous?

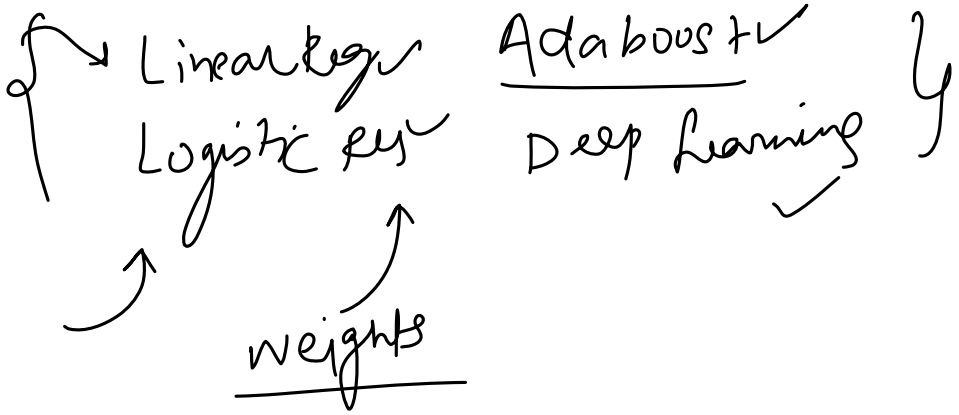
Thursday, April 29, 2021 3:26 PM



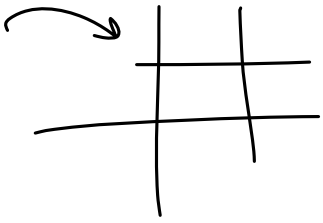
Effect of Outliers on ML algorithms

Thursday, April 29, 2021 3:27 PM



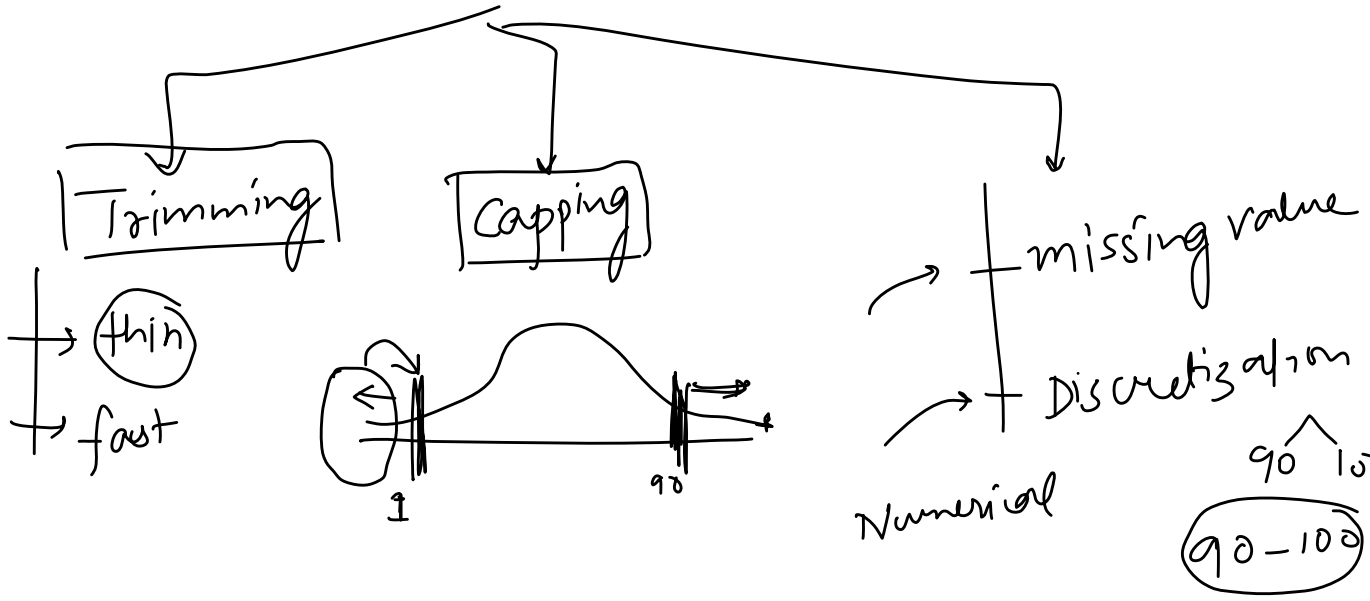


Tree based



How to treat Outliers?

Thursday, April 29, 2021 3:27 PM



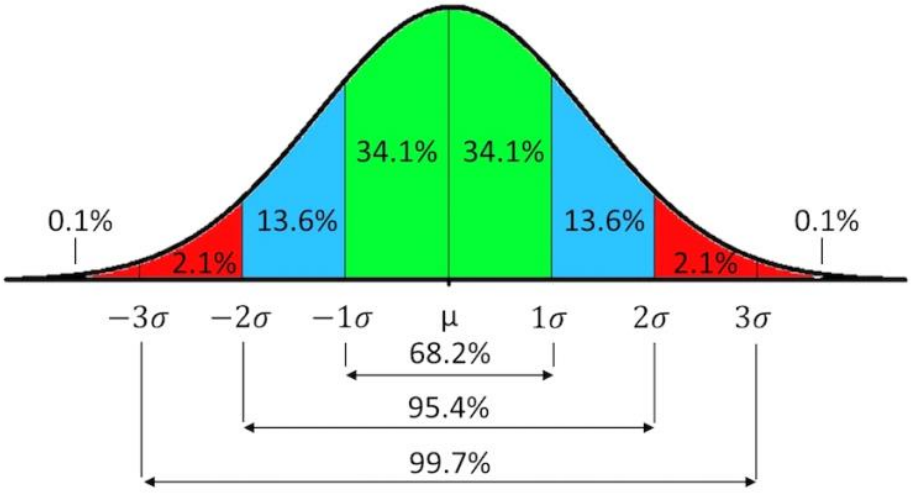
How to detect Outliers?

Thursday, April 29, 2021 3:27 PM

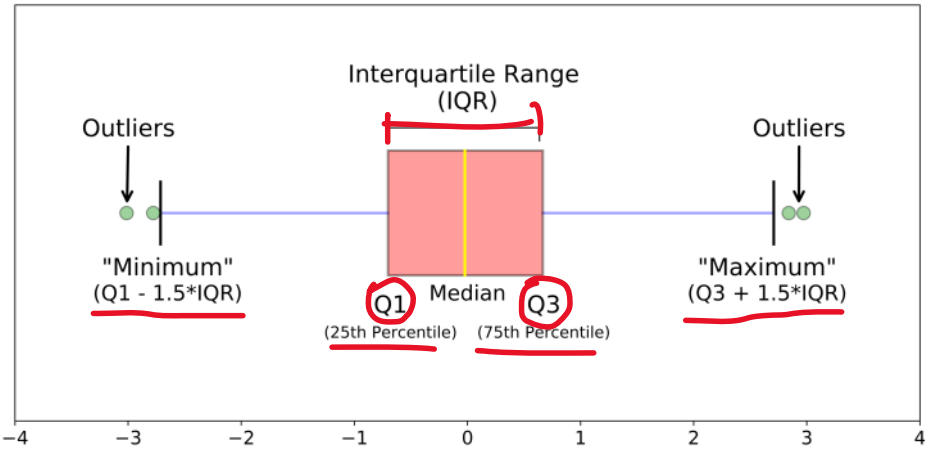
$(\mu + 3\sigma) >$

$(\mu - 3\sigma) <$

1. Normal Distribution



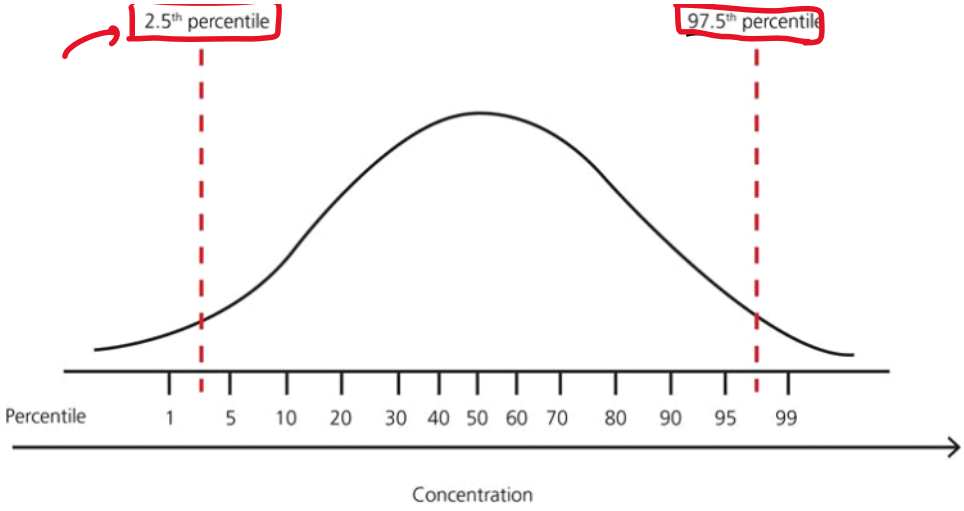
2. Skewed Distribution



$min = Q1 - 1.5IQR$

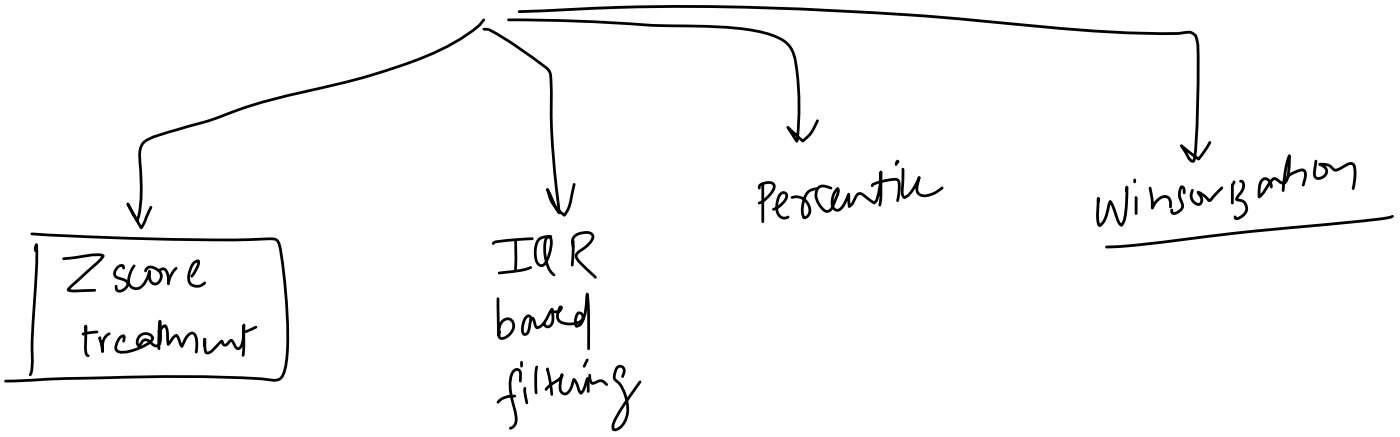
$Q_3 + 1.5IQR$

3. Other Distributions


2.5th percentile
97.5th percentile
Percentile
1
5
10
20
30
40
50
60
70
80
90
95
99
Concentration

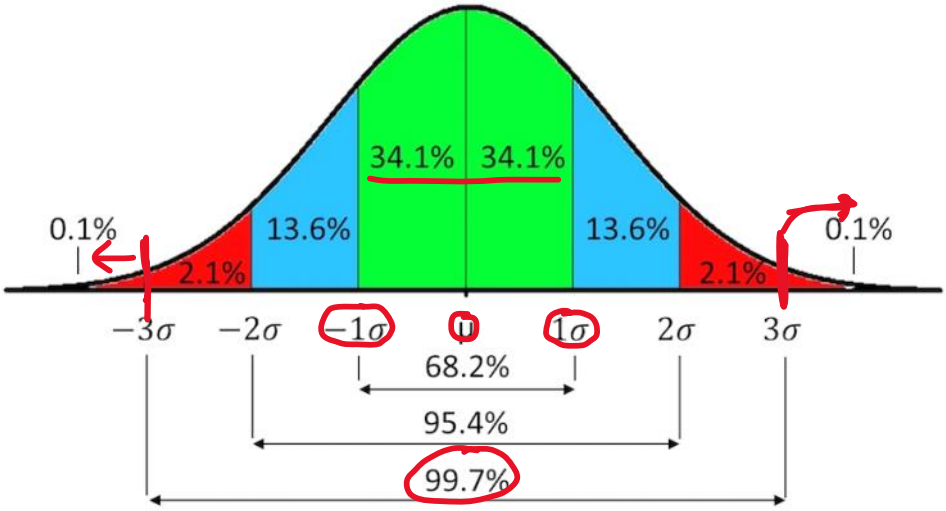
Techniques for Outlier Detection and Removal

Thursday, April 29, 2021 3:35 PM



Outlier removal using Z score

Friday, April 30, 2021 1:33 PM



$\{\mu+\sigma, \mu-\sigma\}$ 68%

$\{\mu+2\sigma, \mu-2\sigma\}$ 95%

Z score

$x_i' = \frac{x_i - \mu}{\sigma}$

Age

27

32

−3 to 3

3

1000

5

995

Outlier Treatment

Trimming

$\mu+3\sigma$

80

$\mu-3\sigma$

5

Capping

5

3 −3

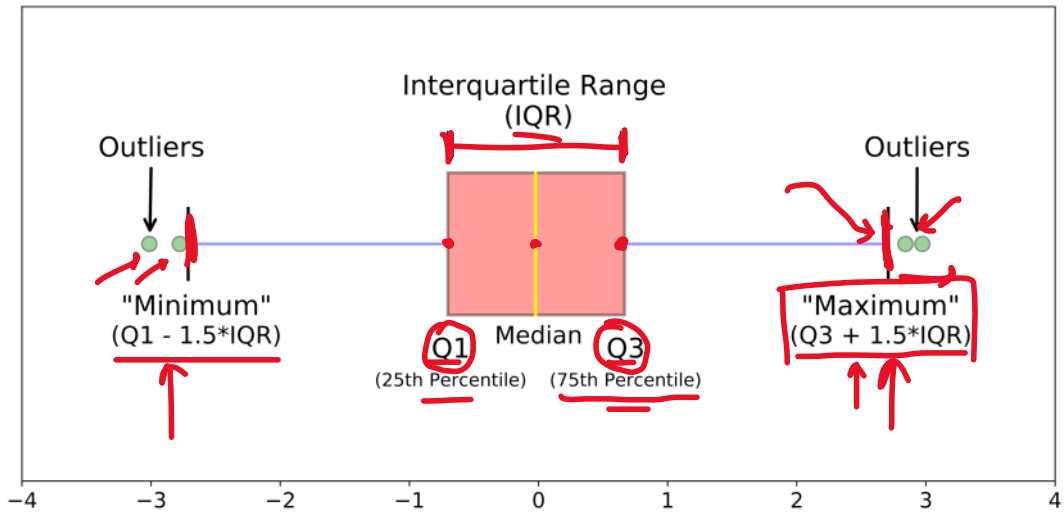
85, 3, 90

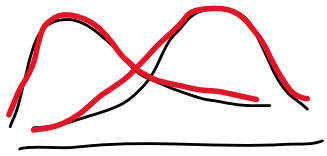
80 5 80

Outlier Detection using IQR

Friday, April 30, 2021 3:48 PM

Boxplot
IQR





percentiles

25
50 Median
75
100

Age
72
89
15
16

100% → 89 ← max

0% → min (16)

25th { ±IQR Proximity Rule }

Intro

Monday, May 3, 2021 11:49 AM

max - 95 → 100%.

min - 10 → 0%.

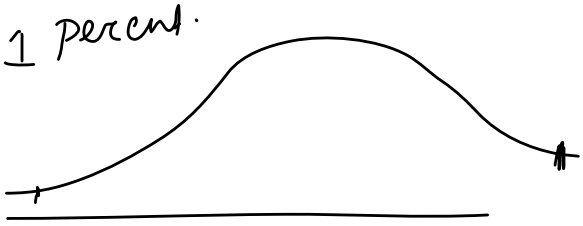
50% → median

Age

27

35

72



99 10 → 1 per

5

95 per

5 per

95 per

68

70

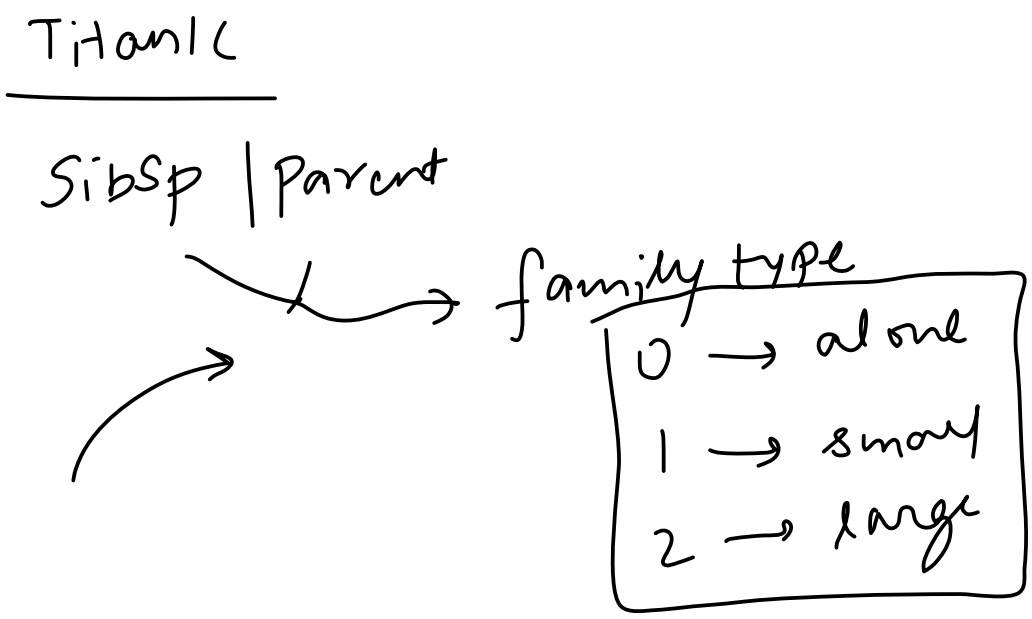
remove

capping

winsorization

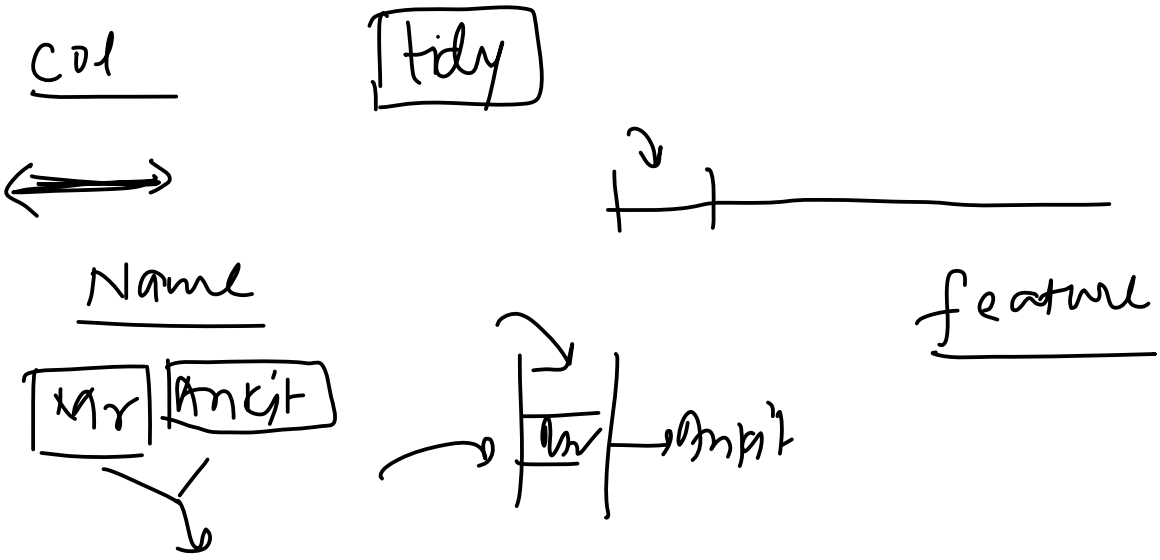
Feature Construction

Tuesday, May 4, 2021 11:13 AM



Feature Splitting

Tuesday, May 4, 2021 11:13 AM

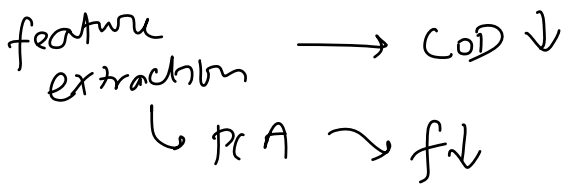


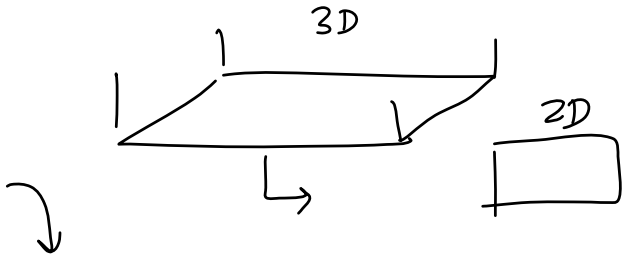
Curse of Dimensionality

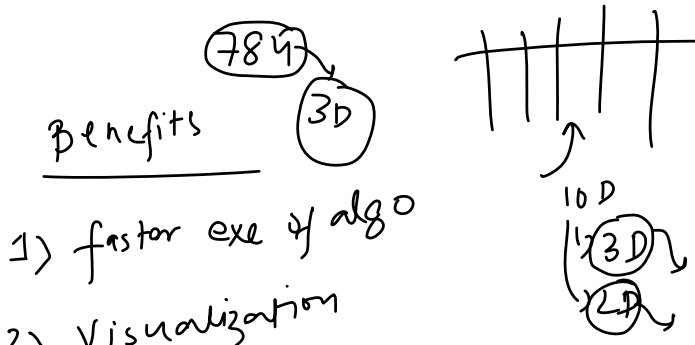
Wednesday, May 5, 2021 1:02 PM

1. Introduction

Wednesday, May 5, 2021 1:02 PM







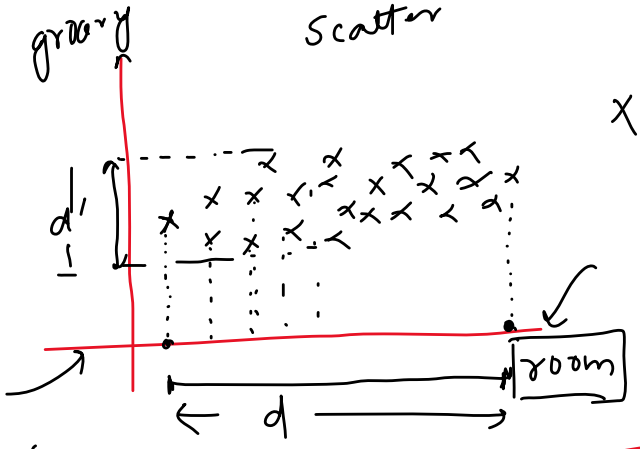
2. Geometric Intuition

Wednesday, May 5, 2021 1:02 PM

PCA

feature selection

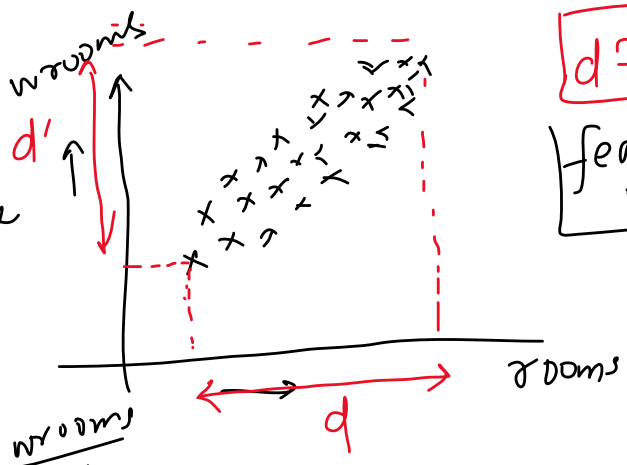
No. of rooms	No. of grocery shops	price (L)
3	2	60
4	0	130
5	6	170
2	10	90



X → rooms

variance

$d > d'$



$d = d'$

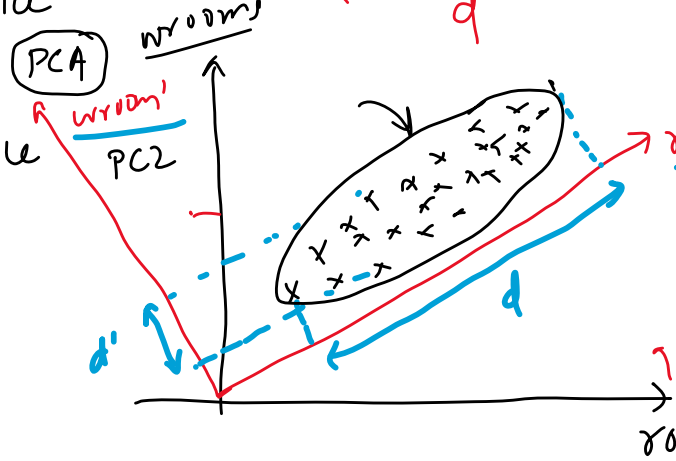
feature extraction

No. of rooms	No. of washrooms	Price

ID	Size	price

2D

Room	WRooms	price



PCA

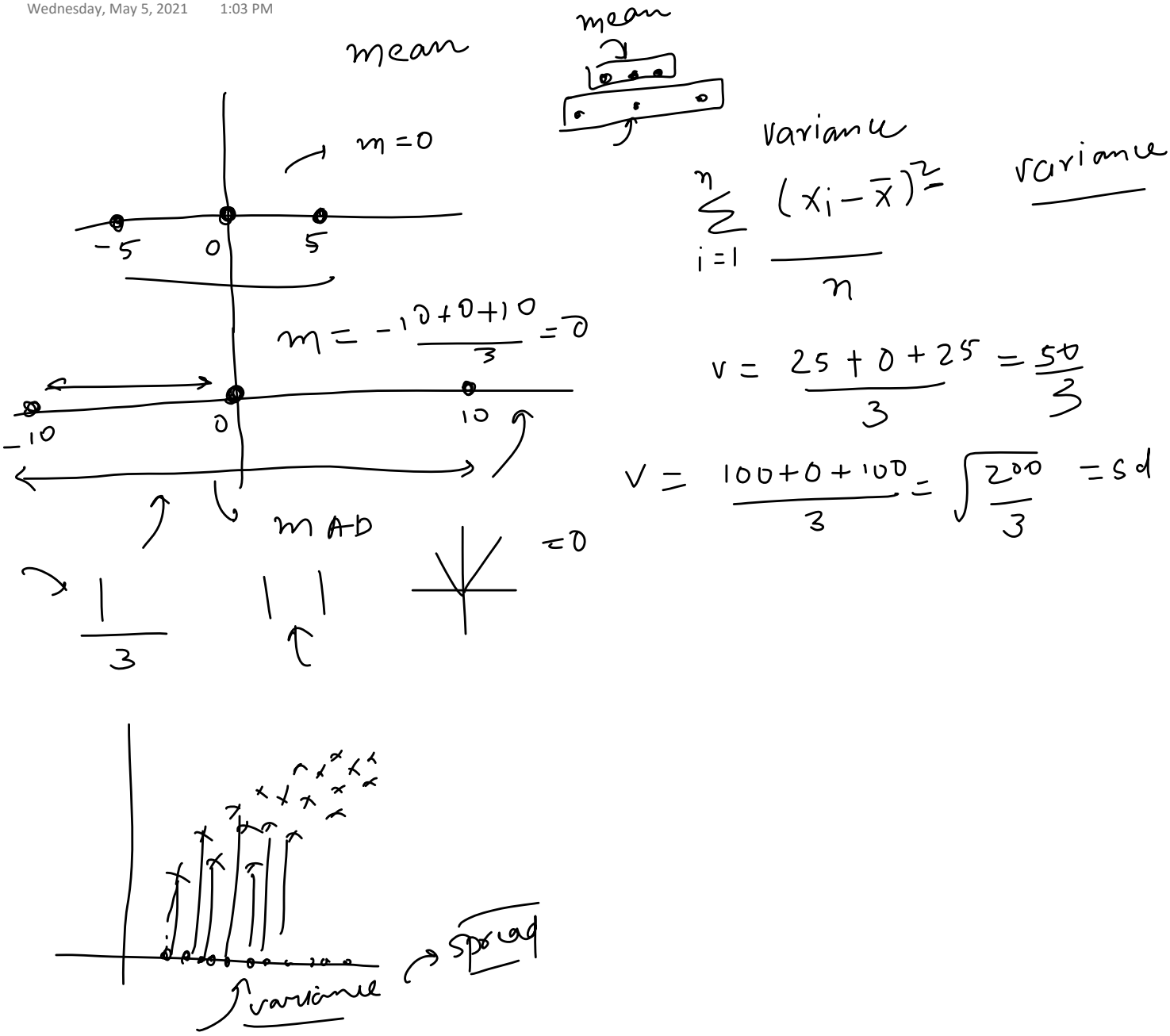
no. of PC $\leq$ (n)

$d > d'$

→ Variance / spread → imp?

3. Why Variance is Important?

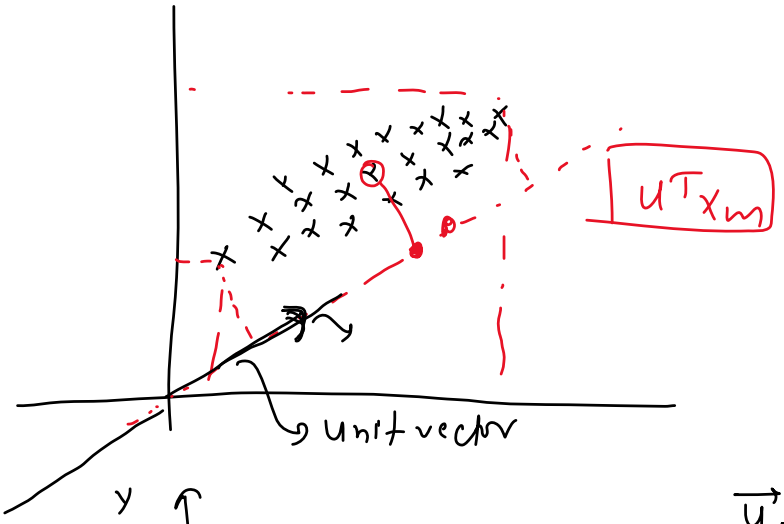
Wednesday, May 5, 2021 1:03 PM



4. Problem Formulation

Wednesday, May 5, 2021 1:03 PM

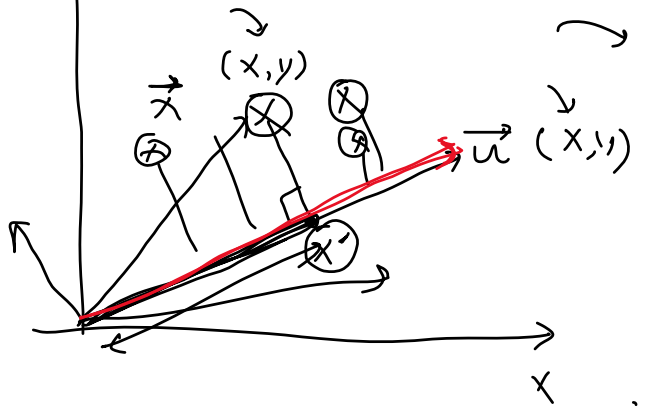
2D → 1D



u

$\vec{u} \cdot \vec{x} = \vec{u} \cdot \vec{x} = u^T x$

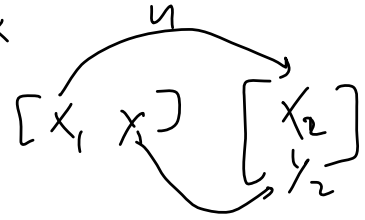
$\rightarrow |u| = 1$



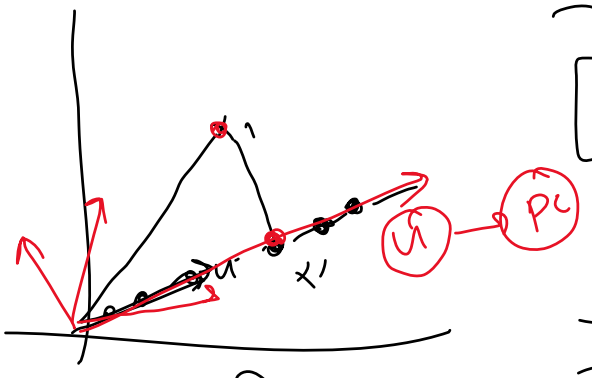
variance maxi

$[x\ y]\ [x\ y]$

$\times$



$= x_1 x_2 + y_1 y_2$ – scalar



$[u^T x_1]\ [u^T x_2] \ldots [u^T x_n]$

$\frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})^2$

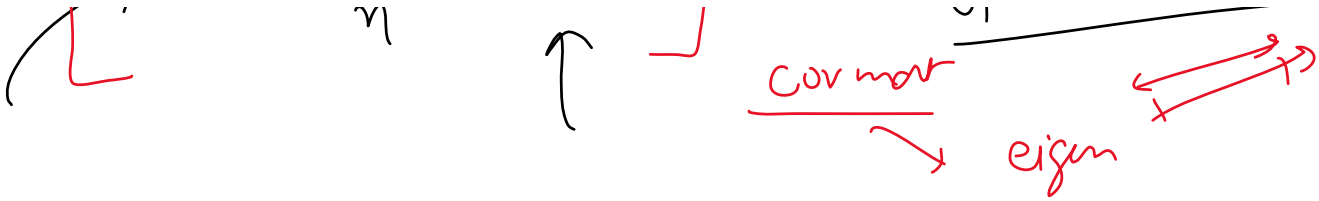
$\rightarrow \max$

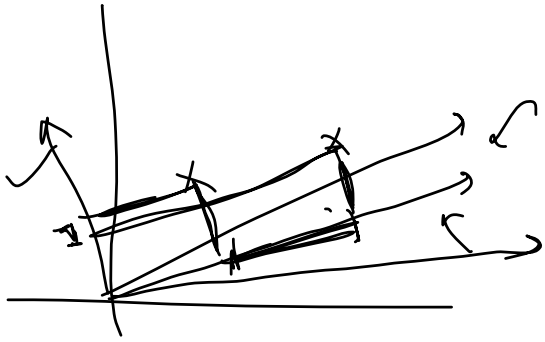
$\vec{u}$

$$\left[\frac{\sum_{i=1}^{n} (u^T x_i - u^T \bar{x})^2}{n} \right] = \text{variance}$$

Optimization

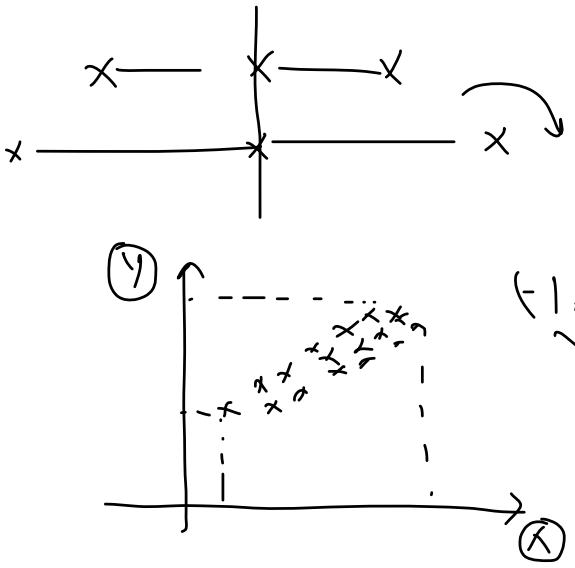
Covariance


cov mat
eigen

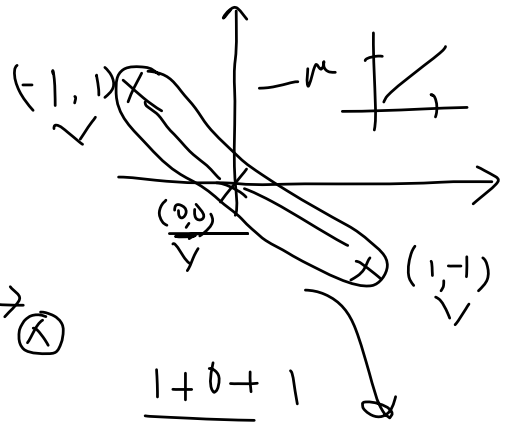


Covariance and Covariance Matrix

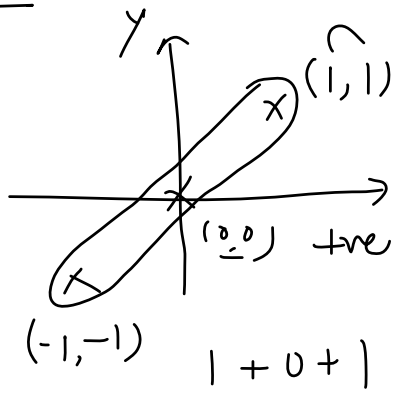
Thursday, May 6, 2021 7:02 AM



variance



Covariance



Correlation

−1 to 1

$\frac{1+0+1}{3}$

$-\frac{2}{3}$

$\frac{-1+0-1}{3}$

$\frac{1+0+1}{3}$

$-\frac{2}{3}$

x_1	x_2

cov mat

$[2\times 2]$

-ve

	x_1	x_2
x_1	$\text{cov}(x_1,x_1)$	$\text{cov}(x_2,x_1)$
x_2	$\text{cov}(x_1,x_2)$	$\text{cov}(x_2,x_2)$

cov mat

$\text{cov}(x_1,x_1)$

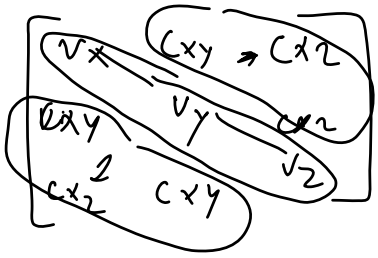
$\downarrow$

$\text{var}(x_1)$

$$\begin{bmatrix} \text{Var } x_1 & \text{cov}(x_2,x_1) \\ \text{cov}(x_1,x_2) & \text{var } x_2 \end{bmatrix}$$

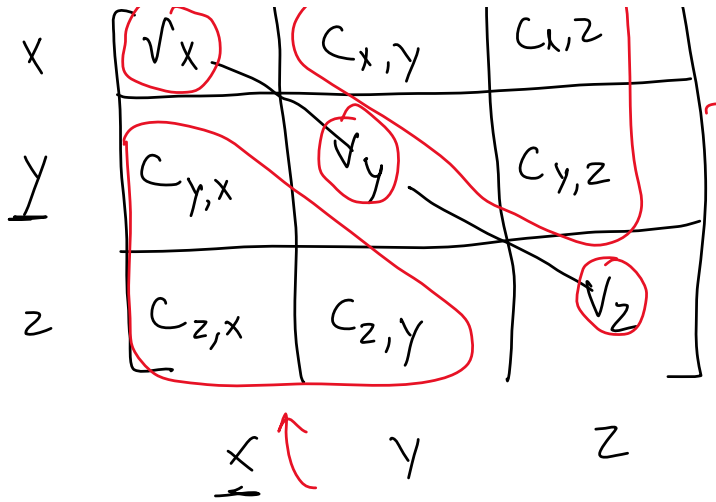
square symmetric

$\text{Cov}(a,b) = \text{cov}(b,a)$



x	y	z

cov matrix

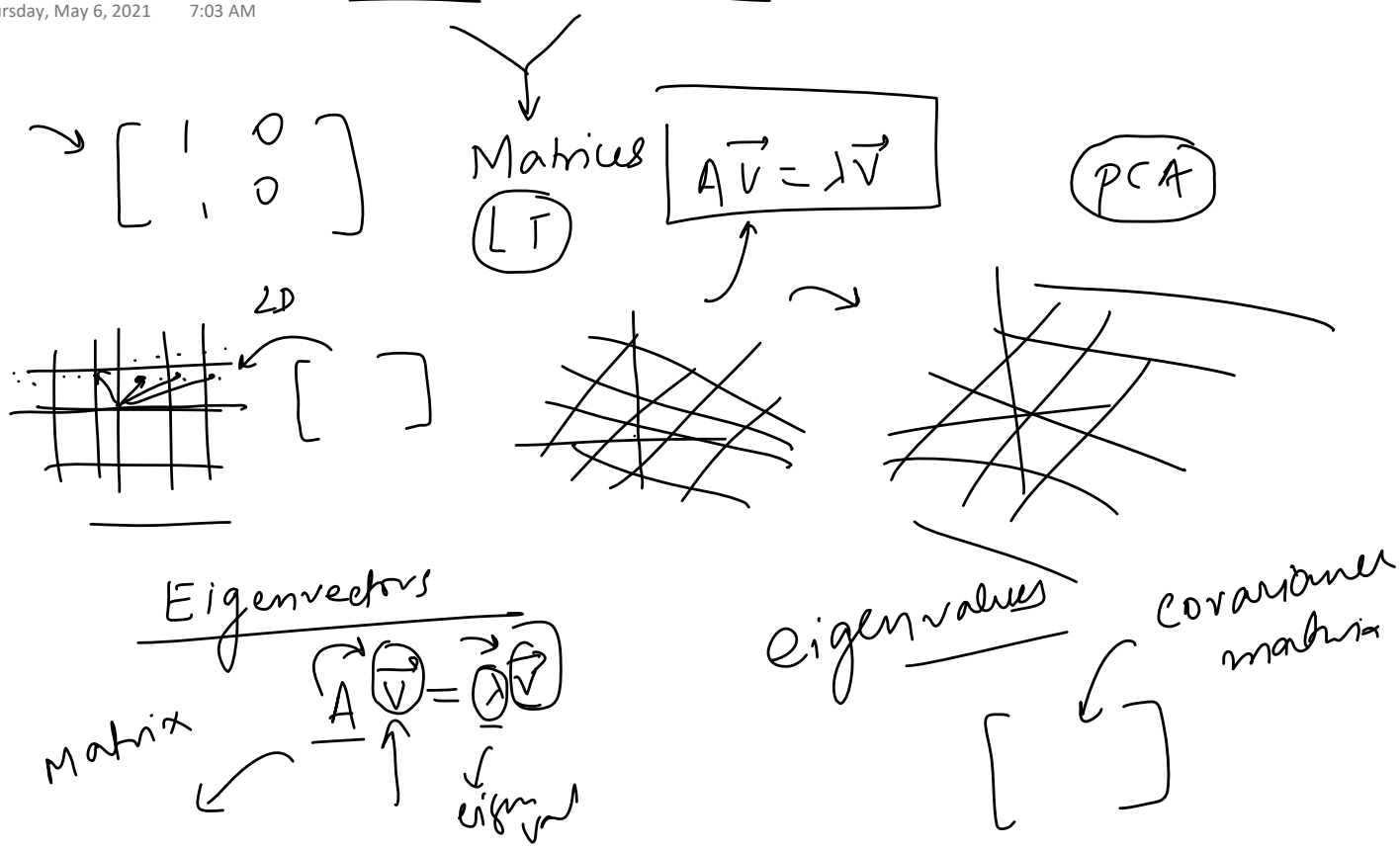


→ cov matrix

eigen decomposition of covariance matrix

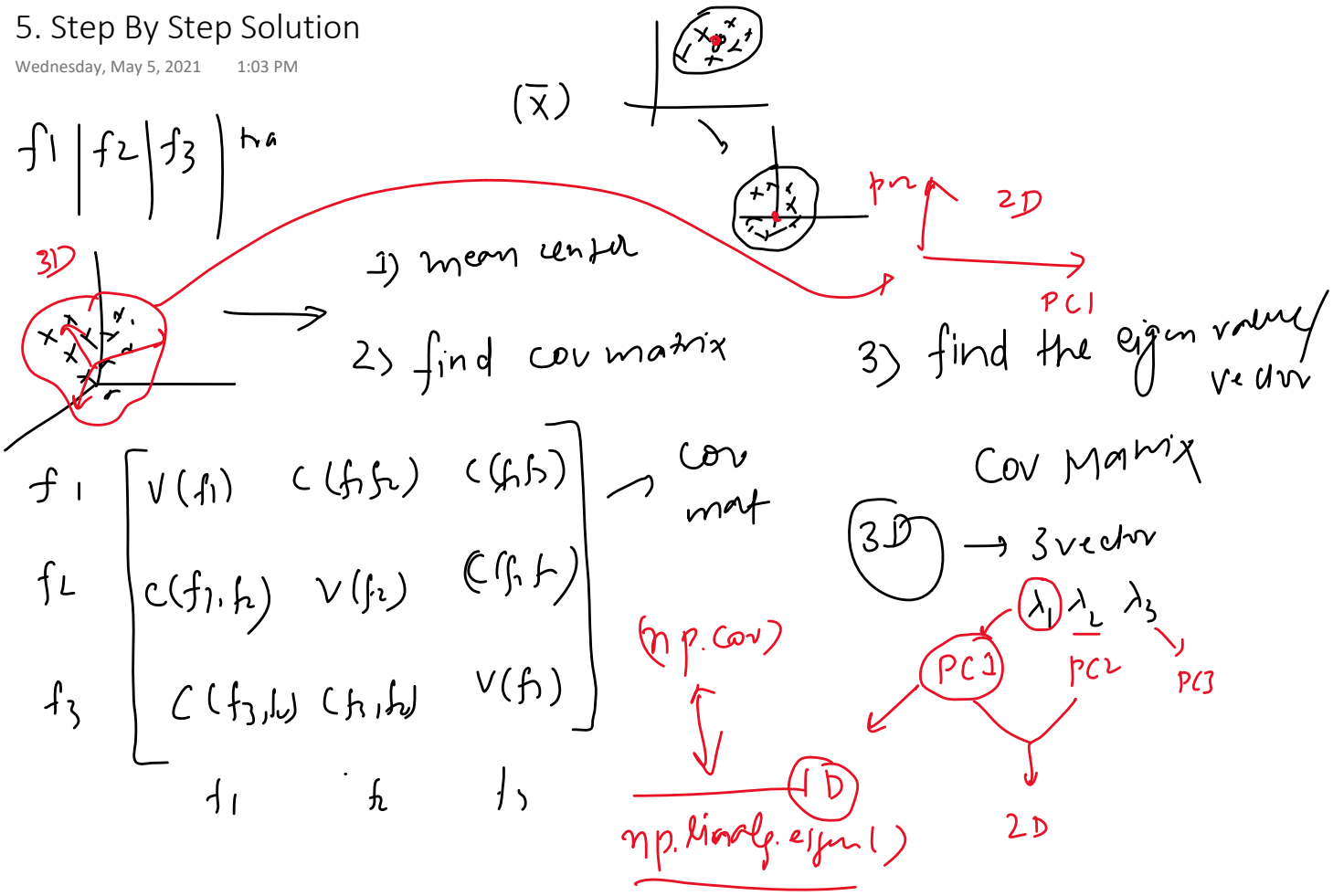
Linear Transformation, Eigen Vectors and Eigen Values

Thursday, May 6, 2021 7:03 AM



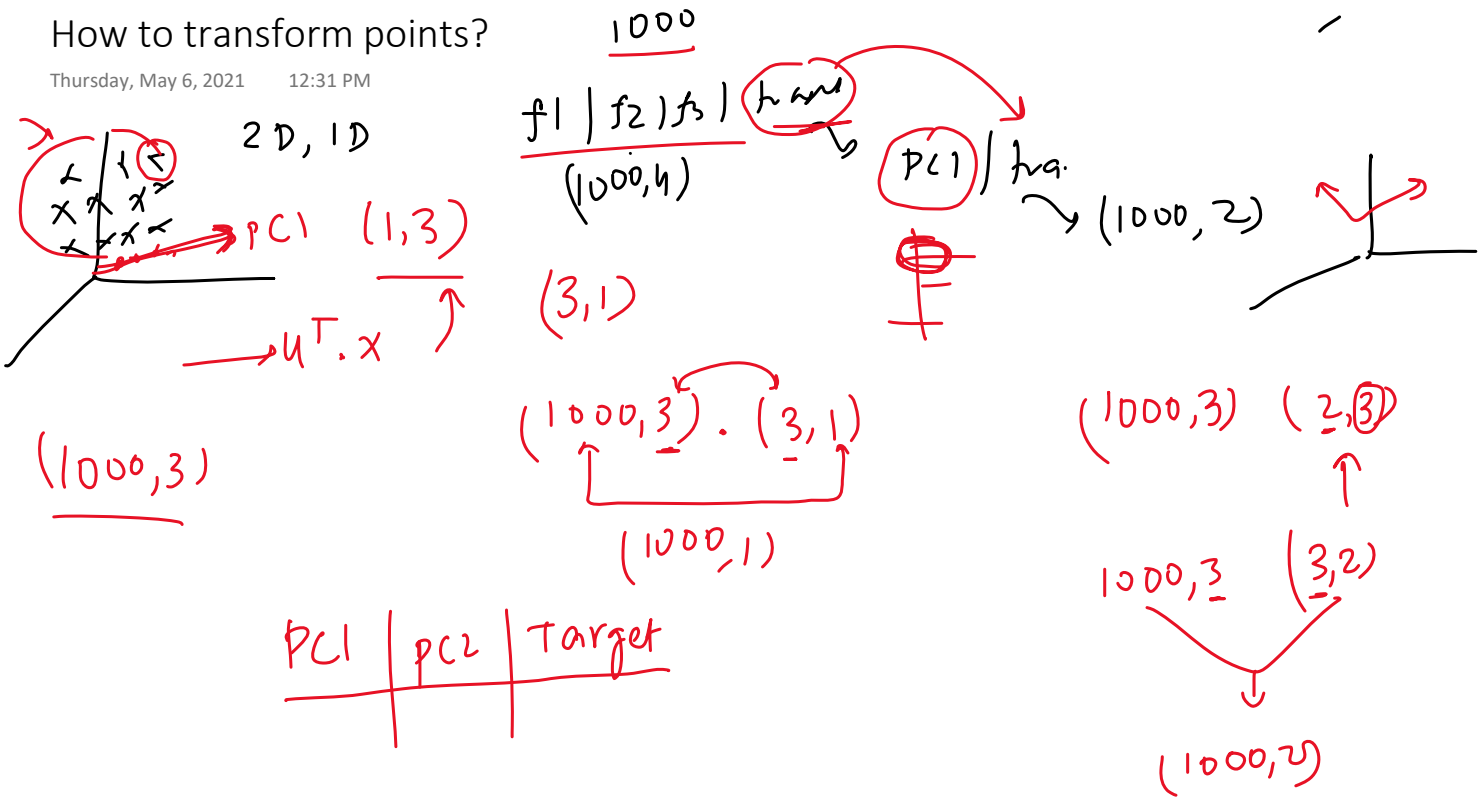
5. Step By Step Solution

Wednesday, May 5, 2021 1:03 PM



How to transform points?

Thursday, May 6, 2021 12:31 PM



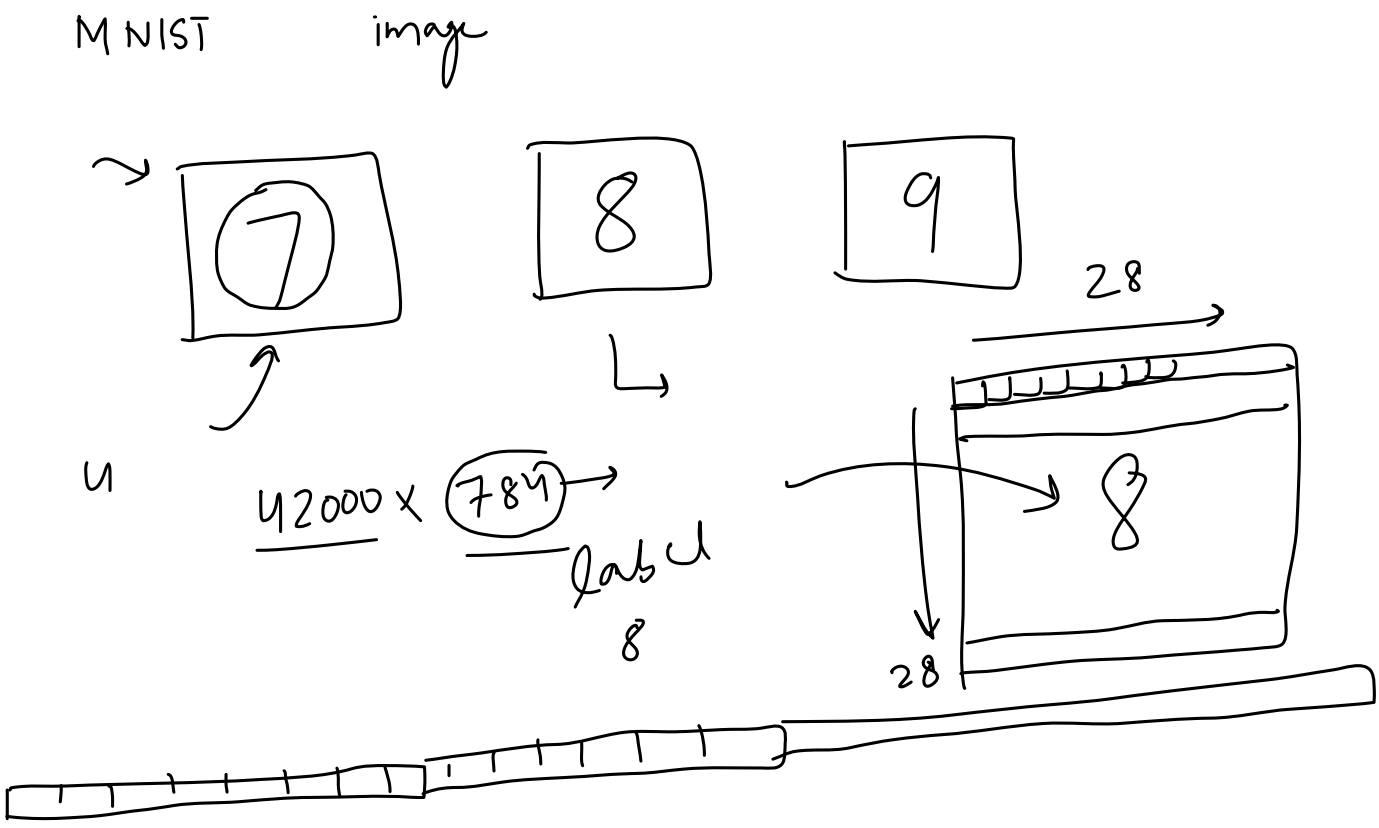
PC1	PC2	Target

6. Coding the Steps

Wednesday, May 5, 2021 1:03 PM

7. Practical Example on MNIST Dataset

Wednesday, May 5, 2021 1:04 PM

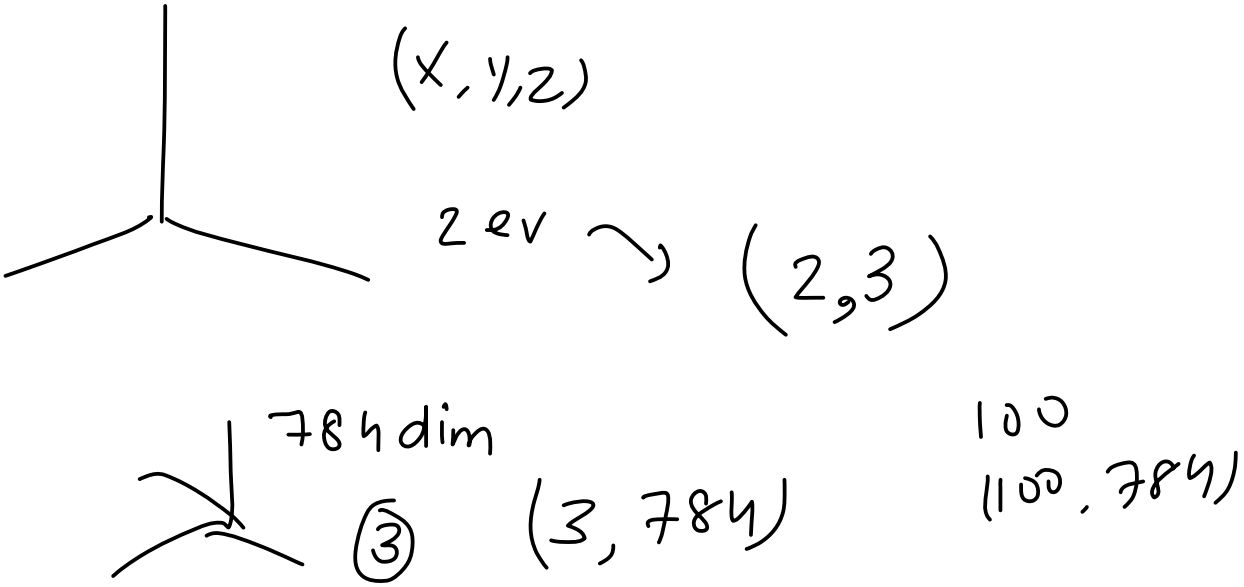


8. Visualization of MNIST Dataset

Wednesday, May 5, 2021 1:04 PM

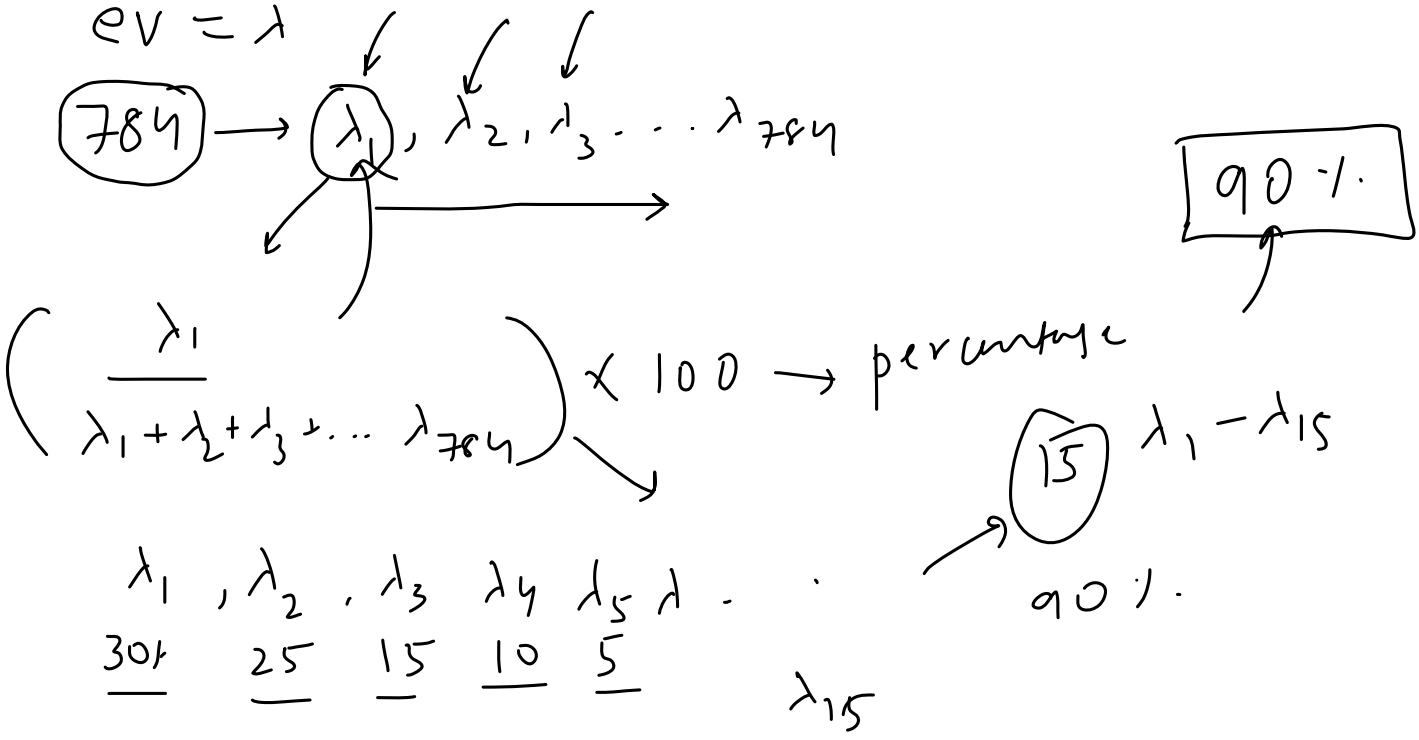
9. Explained Variance

Wednesday, May 5, 2021 1:04 PM



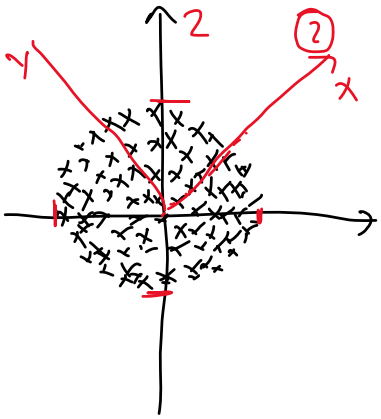
10. Finding optimum number of Principle Components

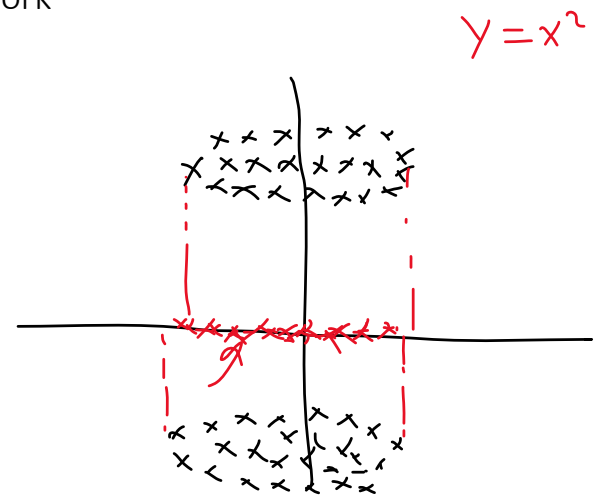
Wednesday, May 5, 2021 1:05 PM



11. When PCA does not work

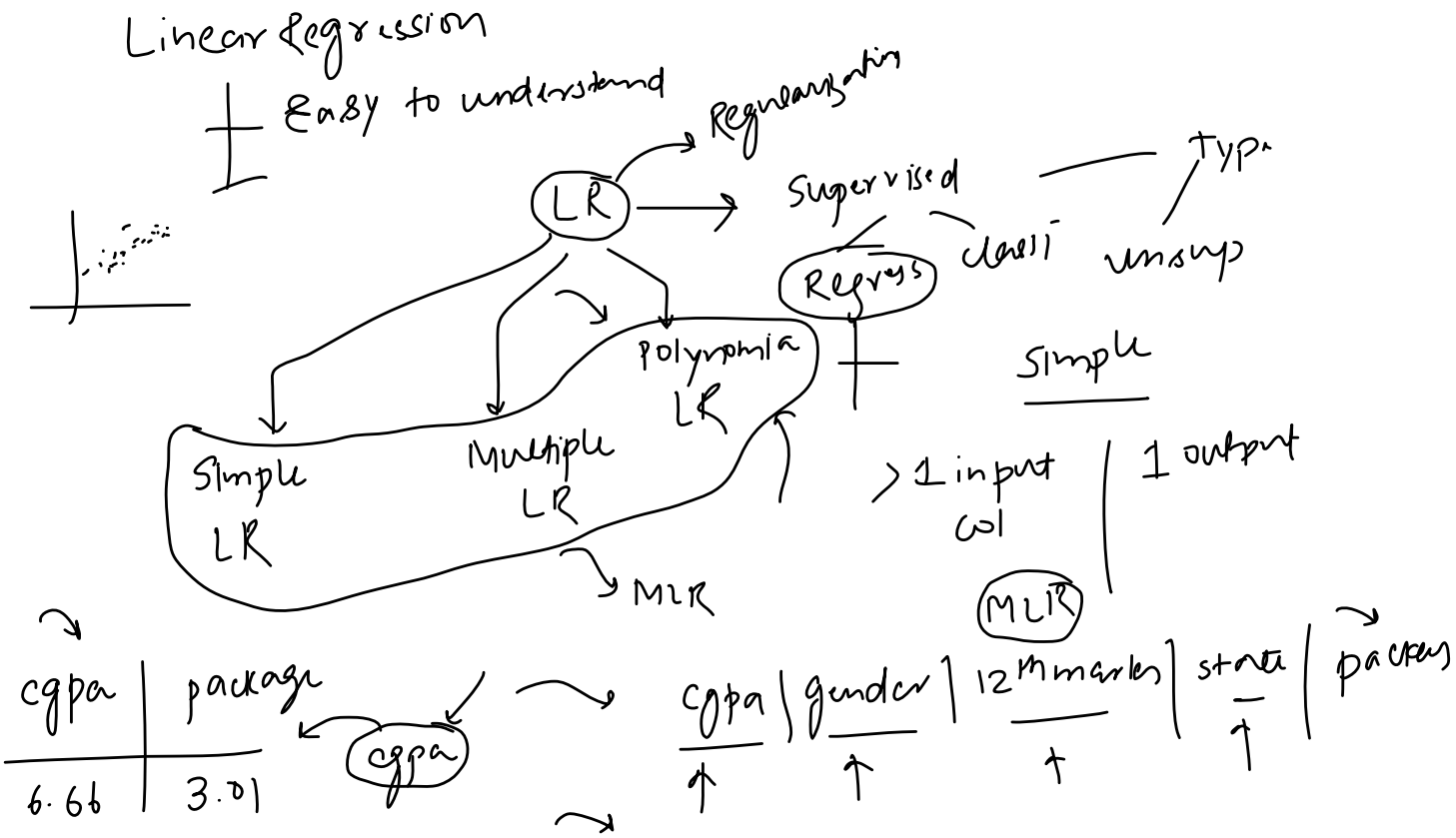
Thursday, May 6, 2021 7:03 AM





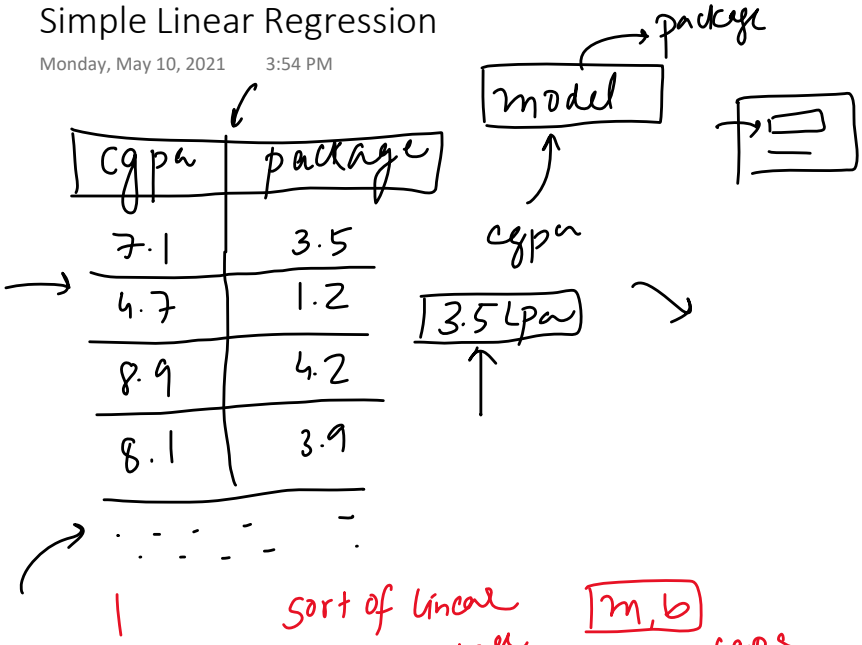
Introduction

Monday, May 10, 2021 3:53 PM



Simple Linear Regression

Monday, May 10, 2021 3:54 PM

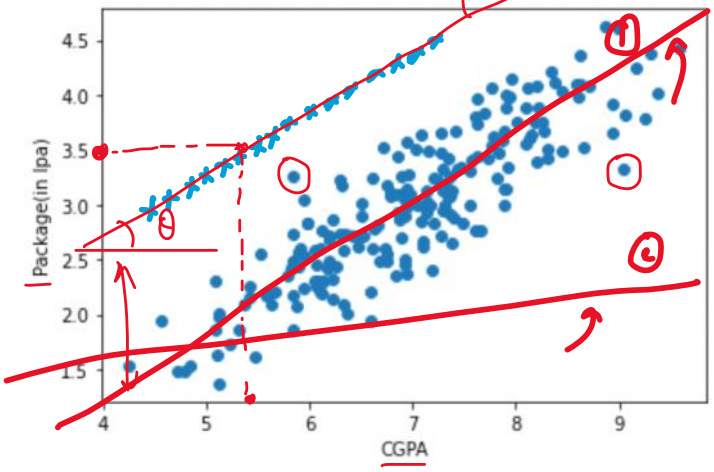


1 sort of linear m, b

package cgpa

$y = m(x) + b$

200 students



Why?

↳ Real word dataset

Best fit line

m → slope

b – y intercept

LR

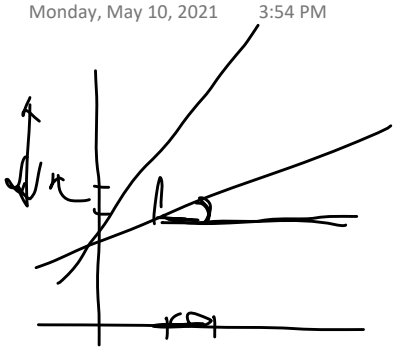
→ m, b

Code Example

Monday, May 10, 2021 3:54 PM

Intuition

Monday, May 10, 2021 3:54 PM



m / b

$y = mx + b$

package = m x cgpa + b → b

b = 0

package = m x cgpa exp → 0

m → weightage

m ↓ ↑

b

package

package = (m x exp) + b = 0

exp | package

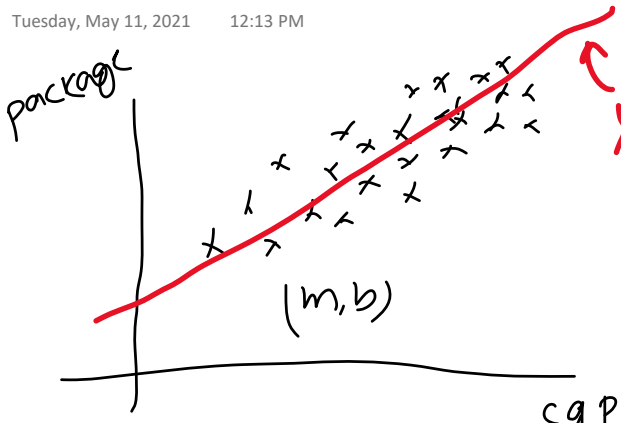
offset

p = 0

p = m x exp

How to find m and b?

Tuesday, May 11, 2021 12:13 PM



$y = mx + b$

(m, b) →

Higher 1-D → Linear Reg → SGD Regressor

direct formula → Closed form solution → OLS

Non-closed form → Gradient descent

$y \rightarrow$ package

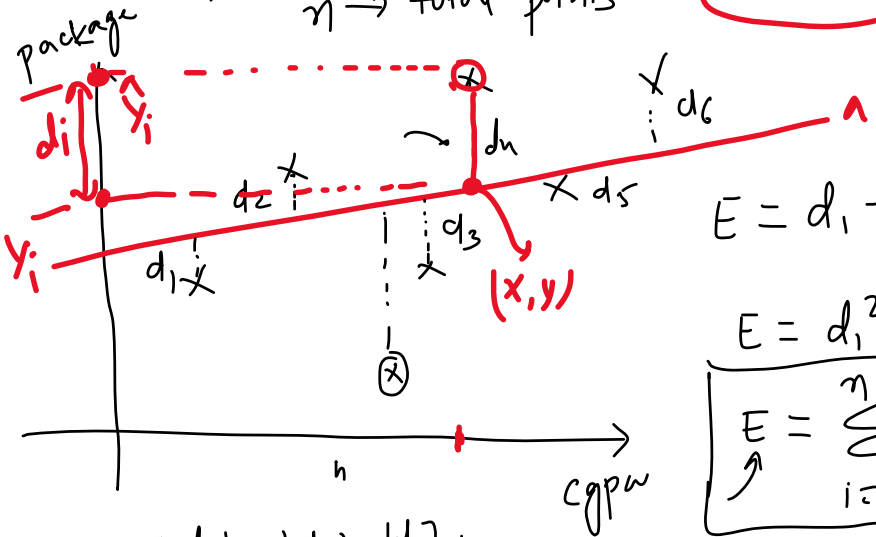
$x \rightarrow$ cgpa

$$b = \bar{y} - m\bar{x}$$

$$m = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2}$$

$\bar{x}, \bar{y} \rightarrow$ mean (m, b)

$n \rightarrow$ total points



$$E = d_1 + d_2 + d_3 + \ldots + d_n$$

$$E = d_1^2 + d_2^2 + d_3^2 + \ldots + d_n^2$$

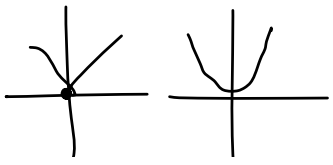
$$E = \sum_{i=1}^{n} d_i^2$$ ← Error function (J)

(m, b)

$$E = |d_1| + |d_2| + |d_3| + \ldots$$

R1

R2 →



$$d_i = (y_i - \hat{y}_i)$$

$R2 \rightarrow$

$d_i = (y_i - \hat{y}_i)$

(m, b)

$$E = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

(m, b)

$\hat{y}_i$

$mx_i + b$

$$\hat{y}_i = m x_i + b$$

(m, b)

$$E(m, b) = \sum_{i=1}^{n} (y_i - m x_i - b)^2$$

$E(m, b)$

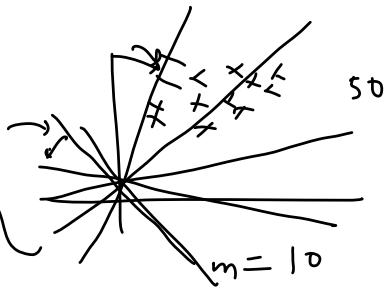
$b \uparrow \quad m = 0 \rightarrow$

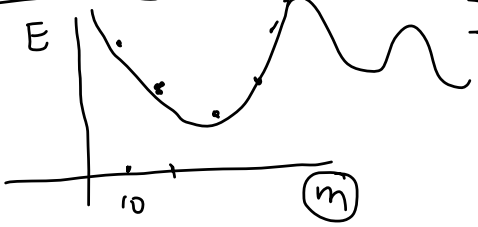
$y = f(x)$

$$E(m, b) = \sum_{i=1}^{n} (y_i - m x_i - b)^2$$

$b = 0$

minimum





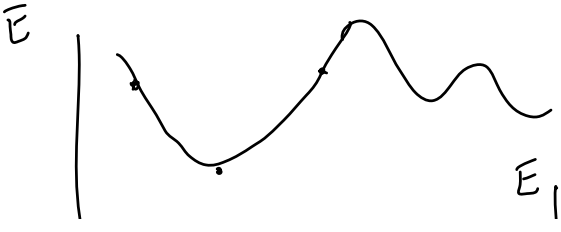
$m = 1$

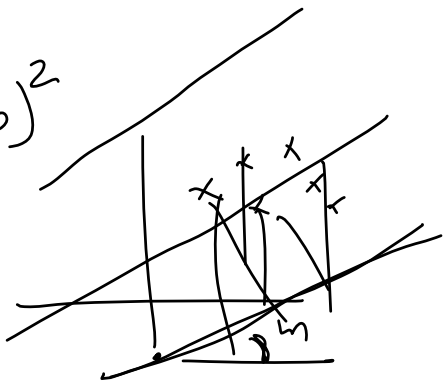
$y = f(x) \quad (m, b)$

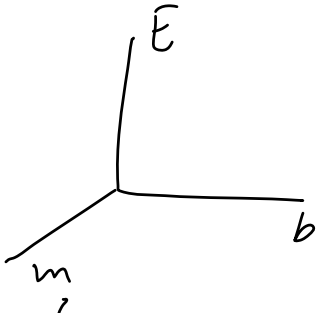
$b = 0$

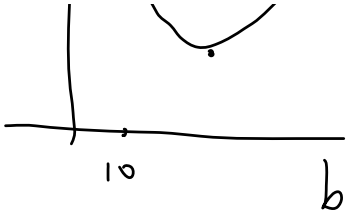
$$E(m) = \sum_{i=1}^{n} (y_i - m x_i)^2$$

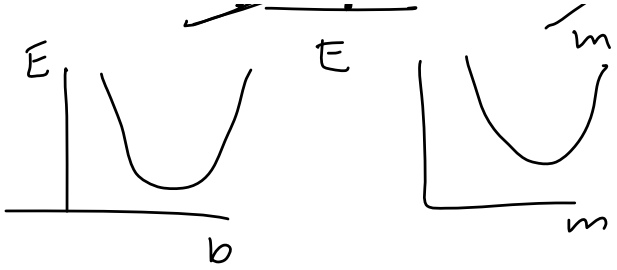
$$E(b) = \sum_{i=1}^{n} (y_i - x_i - b)^2$$

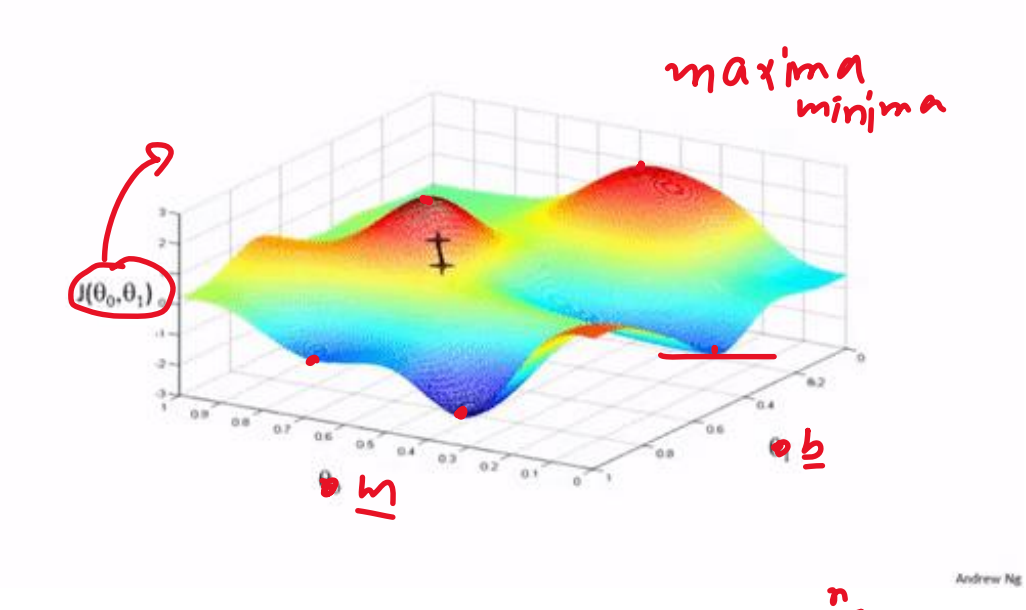












$\rightarrow E(x)$

$\frac{dE}{dx} = 0$

$f(x,y)$

$\frac{\partial E}{\partial m} = 0, \frac{\partial E}{\partial b} = 0$

(m, b) n

$$\frac{\partial E}{\partial b} = \frac{\partial}{\partial b} \sum_{i=1}^{n} (y_i - mx_i - b)^2 = 0$$

$$= \sum \frac{\partial}{\partial b} (y_i - mx_i - b)^2 = 0$$

$$\Rightarrow \sum -2 (y_i - mx_i - b) = 0$$

$$\Rightarrow \sum (y_i - mx_i - b) = 0$$

$$b + b + b + b + \ldots + b = nb$$

n times

$\sum_{i=1}^{n} y_i$

$\frac{\sum y_i}{n} - \frac{\sum mx_i}{n} - \frac{\sum b}{n} = 0$

$\bar{y} - m\bar{x} - \frac{nb}{n} = 0$

$\bar{y} - m\bar{x} = b$

$$\boxed{b = \bar{y} - m\bar{x}}$$

$$E = \sum (y_i - mx_i - \bar{y} + m\bar{x})^2$$

$$\frac{\partial E}{\partial m} = \sum \frac{\partial}{\partial m} (y_i - mx_i - \bar{y} + m\bar{x})^2 = 0$$

$$\Rightarrow \sum 2 (y_i - mx_i - \bar{y} + m\bar{x}) (-x_i + \bar{x}) = 0$$

$$= \sum -2\left(y_i - mx_i - \bar{y} + m\bar{x}\right)\left(x_i - \bar{x}\right) = 0$$

$$= \sum \left(y_i - mx_i - \bar{y} + m\bar{x}\right)\left(x_i - \bar{x}\right) = 0$$

$$= \sum \left[\left(y_i - \bar{y}\right) - m\left(x_i - \bar{x}\right)\right]\left(x_i - \bar{x}\right) = 0$$

$$= \sum \left[\left(y_i - \bar{y}\right)\left(x_i - \bar{x}\right) - m\left(x_i - \bar{x}\right)^2\right] = 0$$

$$= \sum \left(y_i - \bar{y}\right)\left(x_i - \bar{x}\right) = m \sum \left(x_i - \bar{x}\right)^2$$

$$m = \frac{\sum_{i=1}^{n} \left(x_i - \bar{x}\right)\left(y_i - \bar{y}\right)}{\sum_{i=1}^{n} \left(x_i - \bar{x}\right)^2}$$

Code from scratch

Tuesday, May 11, 2021 12:14 PM

Regression Metrics

Thursday, May 13, 2021 11:56 AM

1) MAE

2) MSE

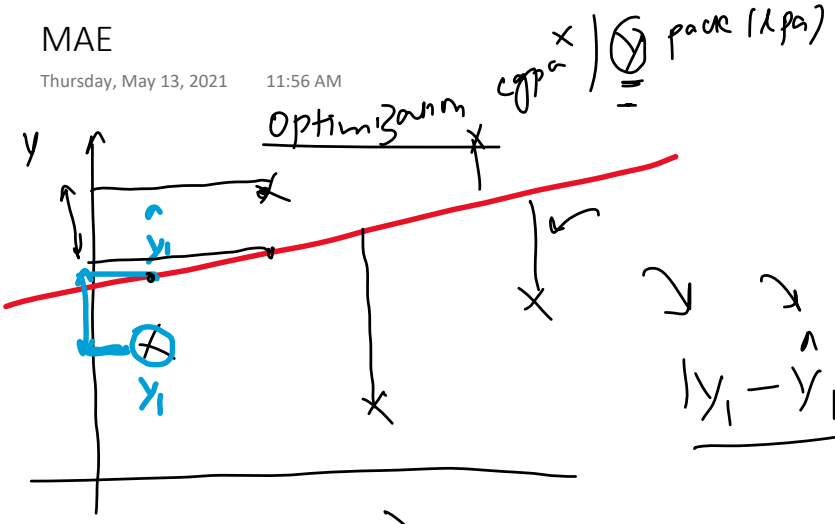
3) RMSE

4) R2 score

5) Adjusted R2 score

MAE

Thursday, May 13, 2021 11:56 AM



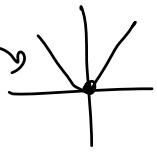
$$|y_1 - \hat{y}_1| + |y_2 - \hat{y}_2| + \dots + |y_n - \hat{y}_n|$$

$$mae = \frac{\sum_{i=1}^{n} |y_i - \hat{y}_i|}{n}$$

Advantage

1) same unit
2) Robust outliers

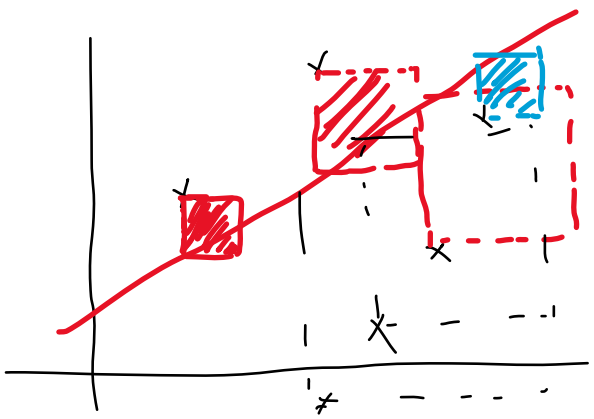
Disadvantage



MSE

Thursday, May 13, 2021 11:56 AM

mean squared error



11.25

$$mae = \frac{\sum_{i=1}^{n} |y_i - \hat{y}_i|}{n}$$

mse

L function

$(y_1 - \hat{y}_1)^2$

Advantage

→ L function

Disadvantage

≠ Robust to outliers

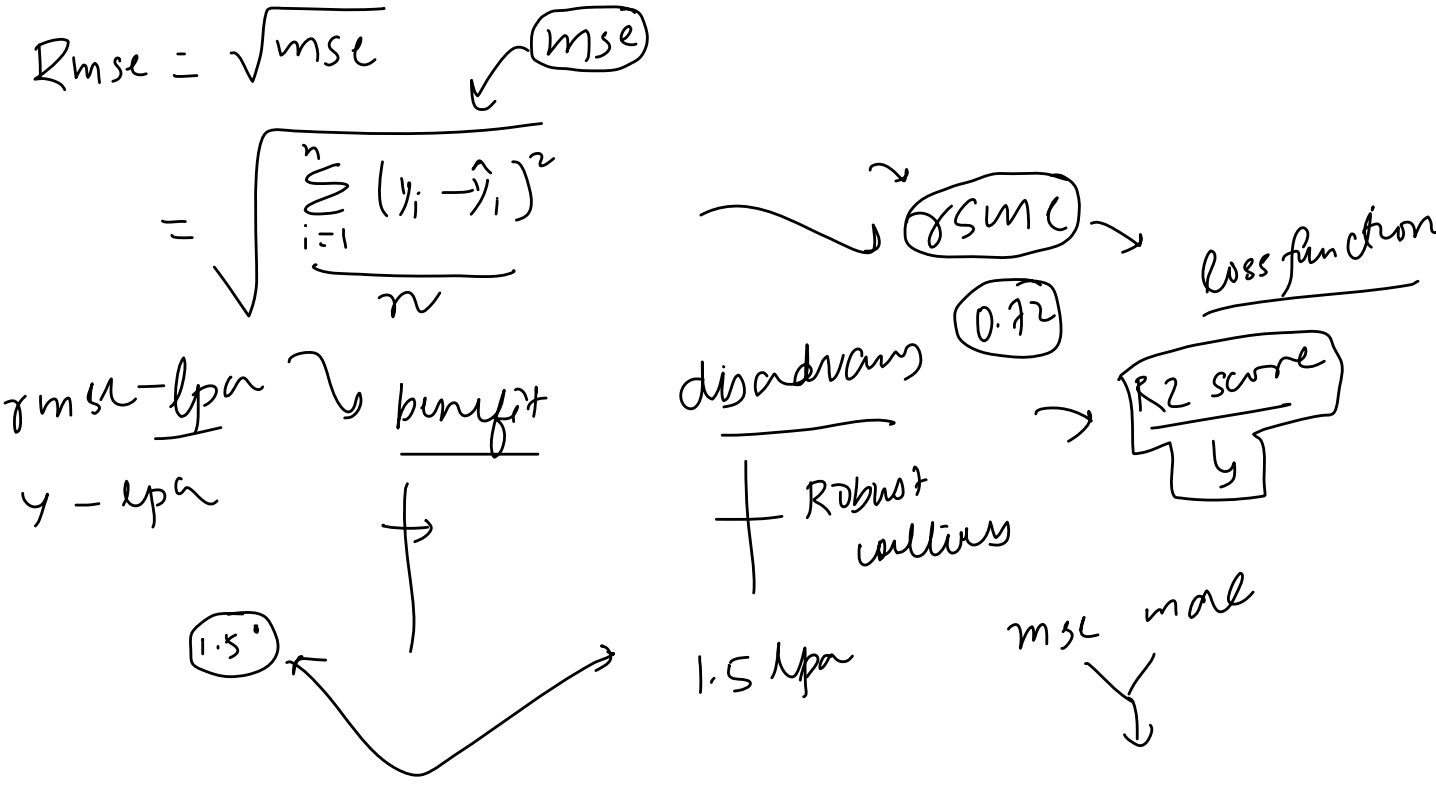
$$mse = \frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{n}$$

y_lpa

mse - (lpa)²

RMSE

Thursday, May 13, 2021 11:56 AM



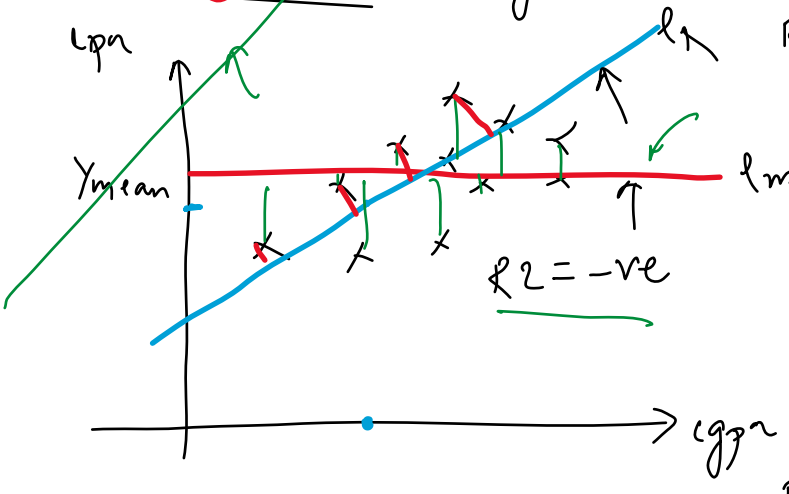
R2 Score

Thursday, May 13, 2021 11:56 AM

100 → mean → 3.4

cgpa | package (lpa)

coeff of det
↳ goodness of fit



$R2 = 1 - \frac{SS_R}{SS_m}$ ≥ 1 $= 1$ $= 0$

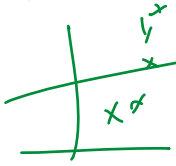
$R2 = 1 - \frac{\left[\sum_{i=1}^{n} (y_i - \hat{y}_i)^2\right]_{Reg}}{\left[\sum_{i=1}^{n} (y_i - \hat{y}_i)^2\right]_{m}}$

R2 = -ve

cgpa → 0

$R_2 = 0$

Sum of squared error Reg



$R_2 = 1$

20%

$0 \rightarrow R2 \rightarrow 1$

$SSR > SSm$

80%

R2 → 0.80 → 20%

cgpa | lpa

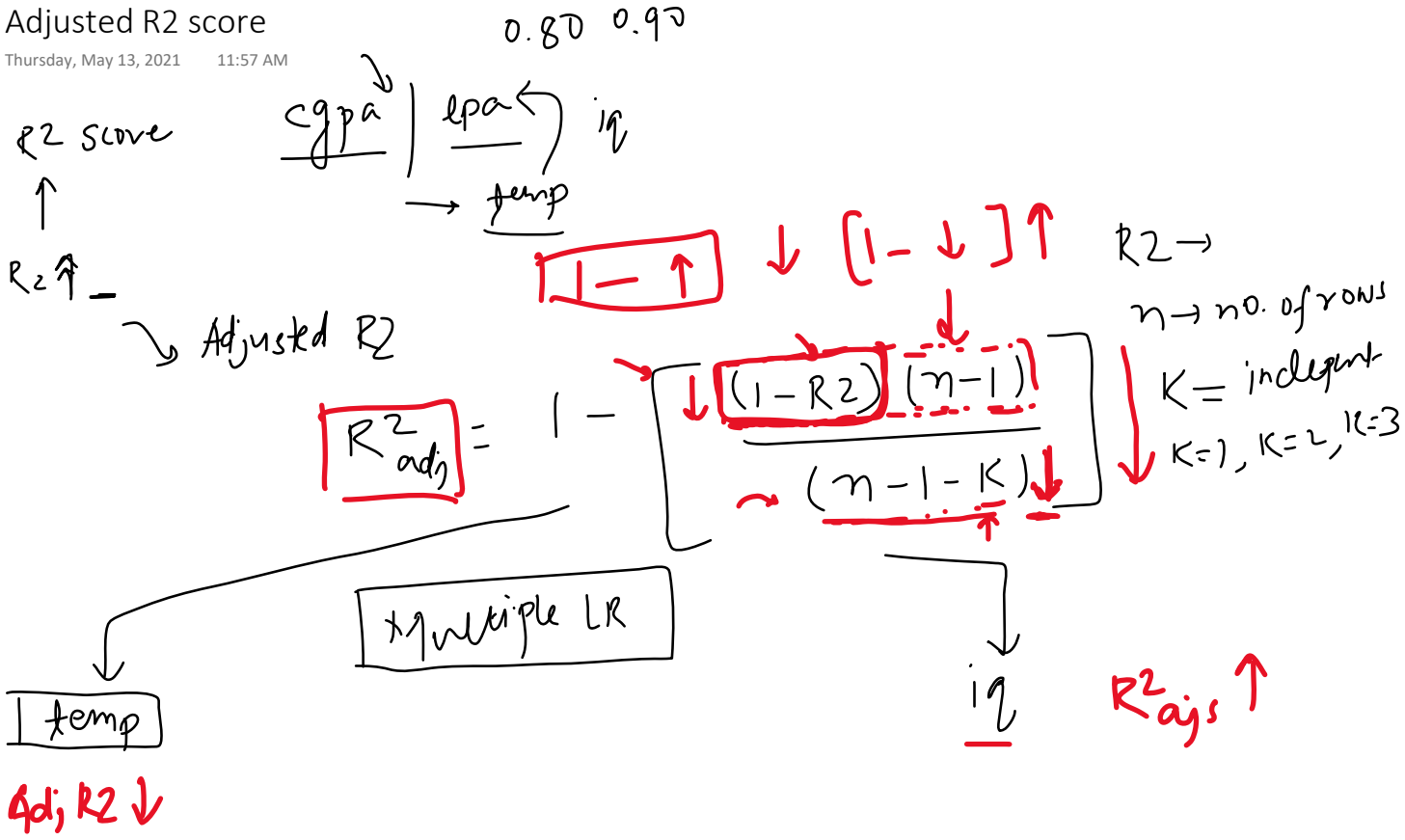
cgpa → lpa explains 80% of variance in lpa

cgpa | ia | lpa

lpa

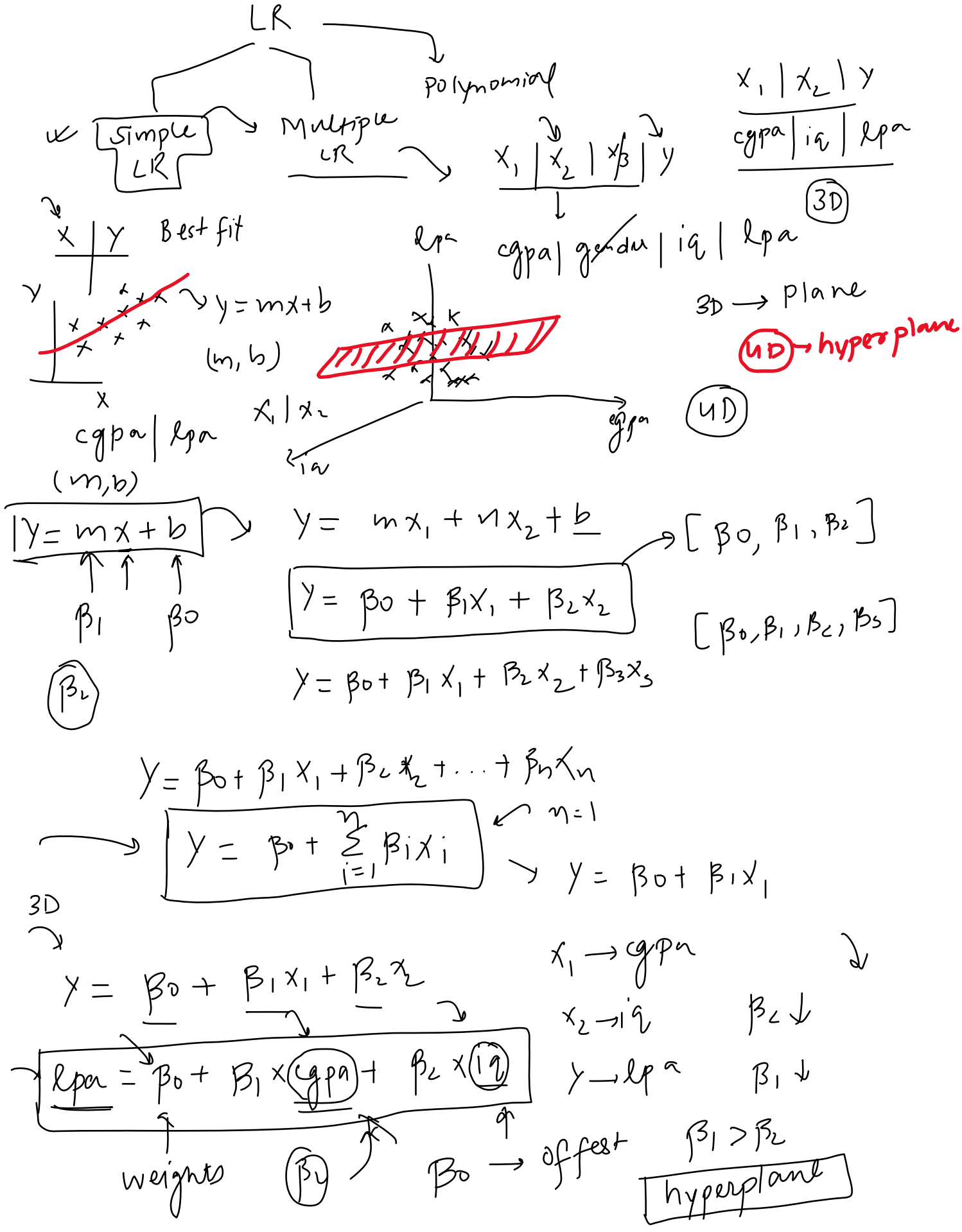
Adjusted R2 score

Thursday, May 13, 2021 11:57 AM



Multiple Linear Regression

Friday, May 14, 2021 4:31 PM

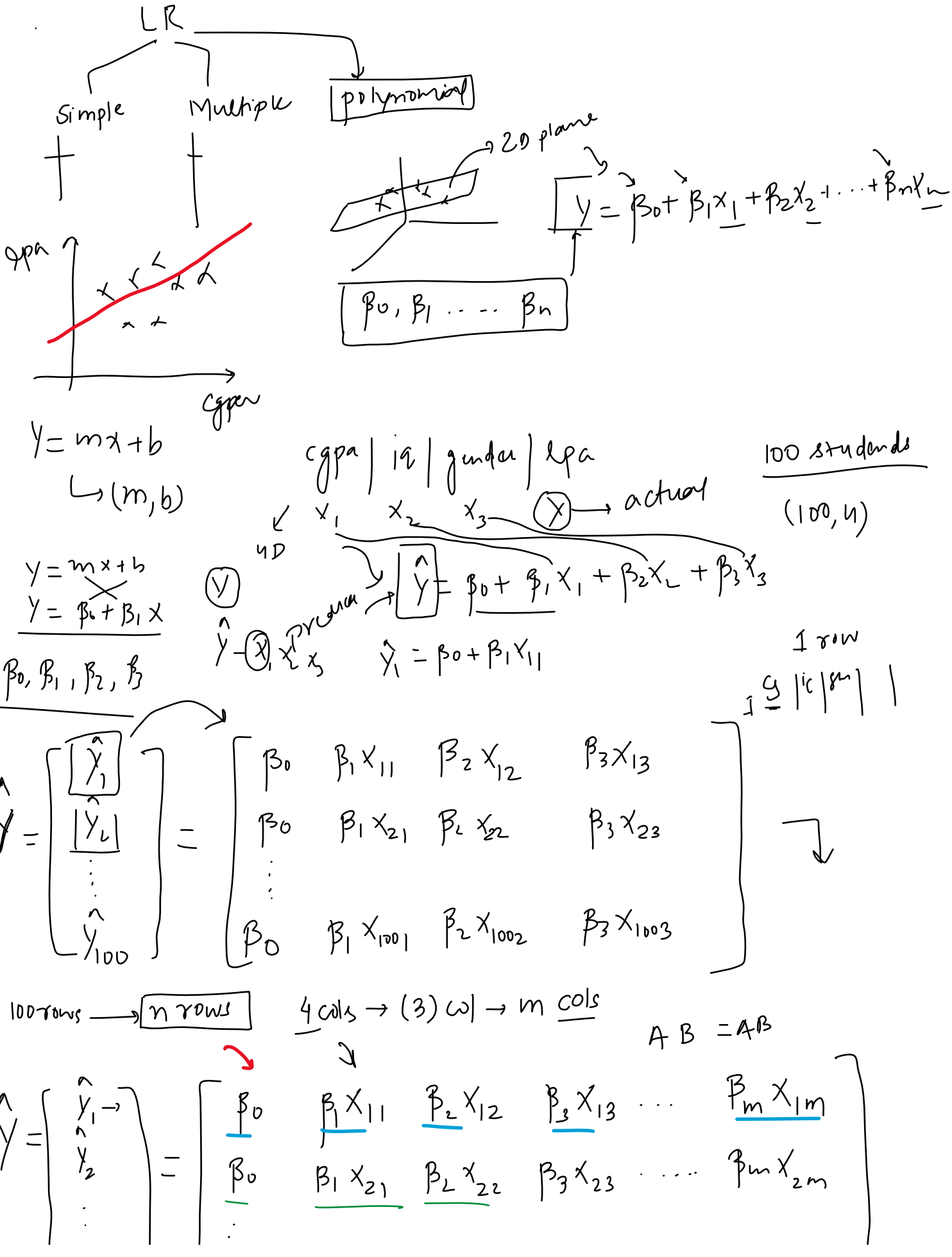


Code Example

Friday, May 14, 2021 4:31 PM

Mathematical Formulation

Saturday, May 15, 2021 4:02 PM



$$\begin{bmatrix} \hat{y}_1 \\ \vdots \\ \hat{y}_n \end{bmatrix} \begin{bmatrix} 1 & x_{11} & x_{12} & x_{13} & \cdots & x_{1m} \\ \vdots & & & & & \\ \beta_0 & \beta_1 x_{n1} & \beta_2 x_{n2} & \beta_3 x_{n3} & \cdots & \beta_m x_{nm} \end{bmatrix}$$

$$= \underbrace{\begin{bmatrix} 1 & x_{11} & x_{12} & x_{13} & \cdots & x_{1m} \\ 1 & x_{21} & x_{22} & x_{23} & \cdots & x_{2m} \\ \vdots & & & & & \\ 1 & x_{n1} & x_{n2} & x_{n3} & \cdots & x_{nm} \end{bmatrix}}_{A} \underbrace{\begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_m \end{bmatrix}}_{B}$$

$$\boxed{\hat{Y} = X\beta} \quad -(1)$$

β $\begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix}$ $\begin{bmatrix} b \\ m \end{bmatrix}$ $\begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{bmatrix}$

X cgpa | gmat | iq

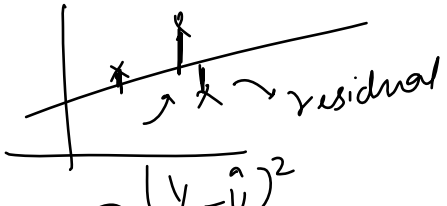
$X \rightarrow$ x_1 x_2 x_3

x_{11}	x_{12}	x_{13}
x_{21}	x_{22}	x_{23}
x_{31}	x_{32}	x_{33}

df → X, y

$$Y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} \qquad e = Y - \hat{Y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} - \begin{bmatrix} \hat{y}_1 \\ \hat{y}_2 \\ \vdots \\ \hat{y}_n \end{bmatrix}$$

$$\boxed{e} = \begin{bmatrix} y_1 - \hat{y}_1 \\ y_2 - \hat{y}_2 \\ \vdots \\ y_n - \hat{y}_n \end{bmatrix}$$



Single LR

Single LR

$$E = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \quad \rightarrow \quad E = e^T e$$

$$\begin{bmatrix} y_1 - \hat{y}_1 \\ y_2 - \hat{y}_2 \\ \vdots \\ y_n - \hat{y}_n \end{bmatrix} = e$$

$$\underset{(1 \times n)}{\begin{bmatrix} (y_1 - \hat{y}_1) & (y_2 - \hat{y}_2) & (y_3 - \hat{y}_3) \dots (y_n - \hat{y}_n) \end{bmatrix}} \begin{bmatrix} (y_1 - \hat{y}_1) \\ (y_2 - \hat{y}_2) \\ (y_3 - \hat{y}_3) \\ \vdots \\ (y_n - \hat{y}_n) \end{bmatrix}_{(n \times 1)} \rightarrow (1) \times 1$$

$$(y_1 - \hat{y}_1)^2 + (y_2 - \hat{y}_2)^2 + (y_3 - \hat{y}_3)^2 + \dots + (y_n - \hat{y}_n)^2$$

$$\sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

$$(A + B)^T = A^T + B^T$$

$$(A - B)^T = A^T - B^T$$

$$E = e^T e = (y - \hat{y})^T (y - \hat{y})$$

$$= (y^T - \hat{y}^T)(y - \hat{y})$$

$$= \left[y^T - (X\beta)^T \right] (y - X\beta)$$

$$E = y^T y - y^T X\beta - (X\beta)^T y + (X\beta)^T X\beta$$

$$E = y^T y - 2y^T X\beta + \beta^T X^T X\beta$$

matrix diff

$$\frac{dE}{d\beta} = \frac{d}{d\beta} \left[y^T y - 2y^T X\beta + \beta^T X^T X\beta \right] = 0$$

$$\frac{du}{d\beta} = \frac{d}{d\beta}\left[\underbrace{y^T y}_{\Downarrow 0} - 2y^T X\beta + \beta^T X^T X \beta \right] - u$$

$$= 0 - 2y^T X + \frac{d}{d\beta}\left[\beta^T X^T X \beta \right] = 0$$

$\frac{dy}{dA} = 2XA^T$ if $y = A^T X A$

$$= -2y^T X + 2X^T X \beta^T = 0$$

$$= 2X^T X \beta^T = 2y^T X \quad \frac{1}{(X^T X)}$$

$$\beta^T = y^T X (X^T X)^{-1}$$

$$(\beta^T)^T = \left[y^T X (X^T X)^{-1} \right]^T$$

$$\beta = \left[(X^T X)^{-1} \right]^T (y^T X)^T$$

$X = n \times (m+1)$

$(m+1) \times n \quad n \times (m+1)$

$X^T X$

$(m+1) \times (m+1)$

$$\beta = \left[(X^T X)^{-1} \right]^T X^T y$$

$$\beta = (X^T X)^{-1} X^T y$$

$$\boxed{\beta = (X^T X)^{-1} X^T y}$$

$X \rightarrow X\_train$

$Y = y\_train$

$(X^T X)^{-1}$

$\begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_m \end{bmatrix}$

$(m+1) \times 1$

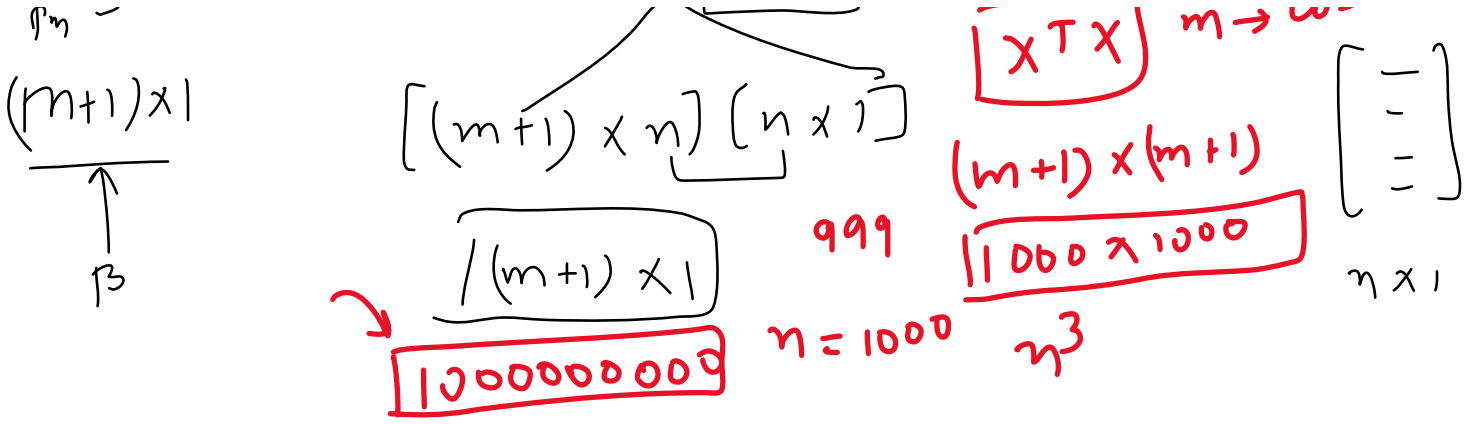
$[(m+1) \times (m+1)][(m+1) \times n][n \times 1]$

$n \times (m+1)^T$

$(m+1) \times n$

$X^T X$ $m \rightarrow$ cols

$\begin{bmatrix} - \end{bmatrix}$


(m+1)x1
β
[(m+1) x n] [n x 1]
/(m+1) x 1
1000000000
999
n = 1000
XTX
m →
(m+1) x (m+1)
1000 x 1000
n3
n x 1

Detour

Saturday, May 15, 2021 4:56 PM

$(AB)^T = B^T A^T$ $\quad Y = A \quad X\beta = B \quad A^{T^T} = A$

$\boxed{Y^T X\beta} = (X\beta)^T Y$

$(A^T B)^T = B^T A$ — (1)

RHS

$Y?$

$A^T B = B^T A$ — (1)

RHS

$A^T B = (A+B)^T \longrightarrow A^T B = C$

$\boxed{C = C^T}$ $\quad A+B \quad C = \boxed{Y^T X\beta}$ — (X)

n students

$(Y^T X\beta)^T = Y^T X\beta$

$Y = \begin{bmatrix} - \\ - \\ - \\ - \\ - \end{bmatrix}$

$(n \times 1) \rightarrow (1 \times n)$

$(1 \times n)\ (n \times (m+1))\ [(m+1) \times 1]$

$1 \times (m+1) \quad (m+1) \times 1$

$1 \times 1 = [7]^T$

$[7]$

: n rows in cols

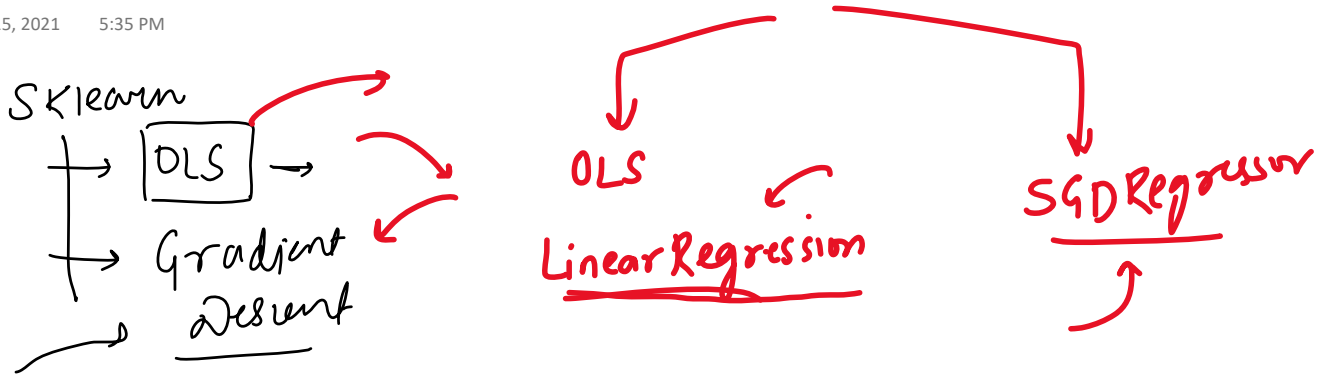
$n \times (m+1)$

$\begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_m \end{bmatrix}$ $m+1$

$(m+1) \times 1$

Why Gradient Descent

Saturday, May 15, 2021 5:35 PM



Code From Scratch

Monday, May 17, 2021 12:15 PM

$$\beta = (X^T X)^{-1} X^T y$$

(100,3) (100)

100 (100,4)

X → matrix

X → X_train

y → y train

	cgpa	iq	gender	lpa
1	—	—	—	→
1	—	-	-	—
				—

diabetes → sklearn → R2- → same

→ my class → R2 →

X test

3

β0 β1 → β10

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3$$

$$= \beta_0 + np.dot(\beta, x\_test)$$

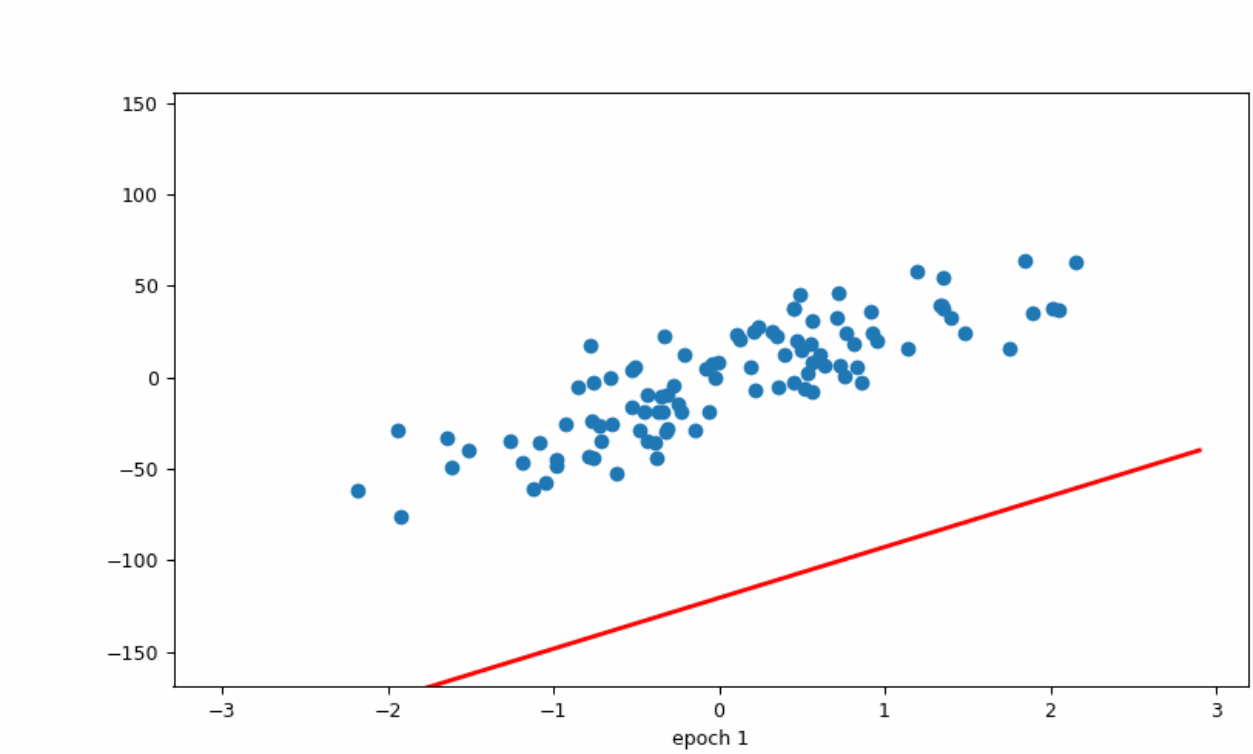
X_test (coeff

(89,10) (10,1)

(89,1) + β0

89 → 1 (89,1)

y pred



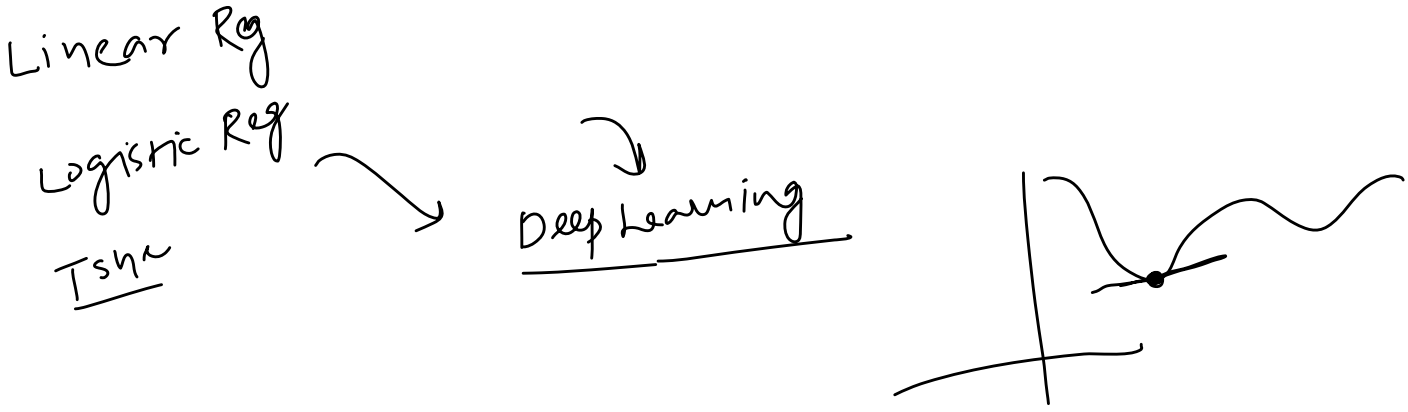
$$\sum (y_i - mx_i - b)^2$$

$$\sum -2 (y_i - mx_i - b)\, x_i$$

$$-2$$

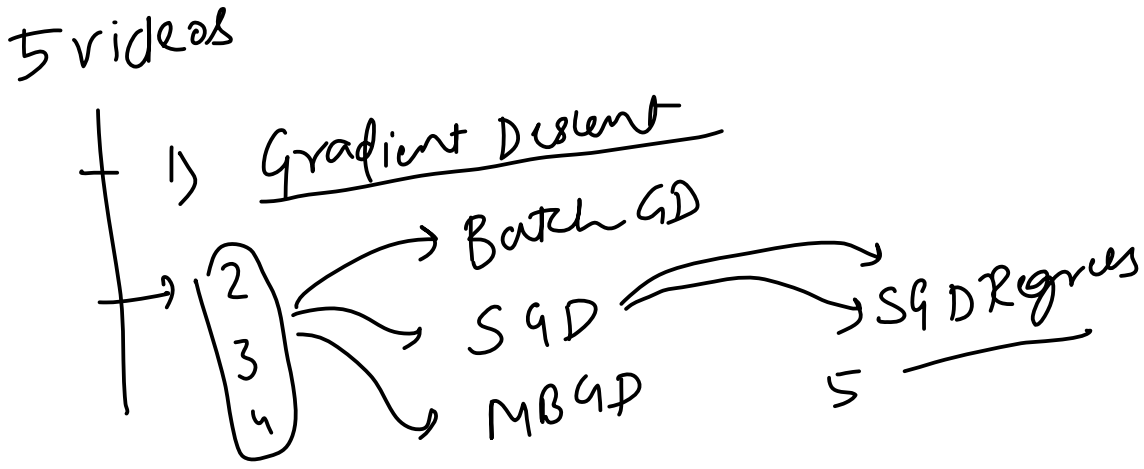
What is Gradient Descent?

Thursday, May 20, 2021 1:46 PM



The Plan

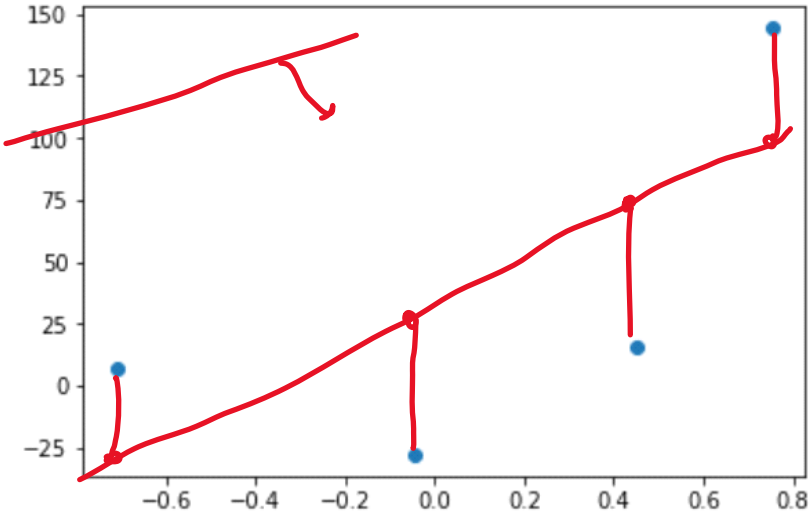
Thursday, May 20, 2021 1:46 PM



Intuition

Thursday, May 20, 2021 1:46 PM

2 cols
4 rows



cgpa | lpa

OLS →

$\hat{y}_i = mx_i + b$

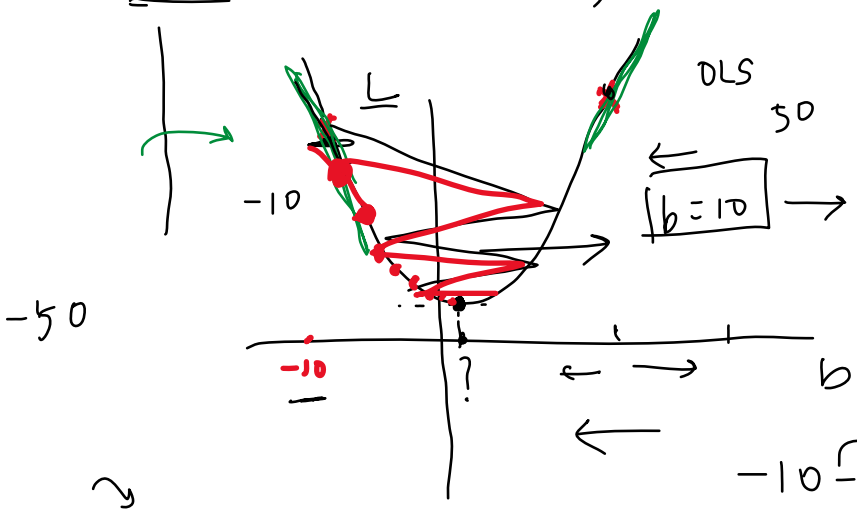
$m = 78.35$

$L = \sum_{i=1}^{4} (y_i - \hat{y}_i)^2$

$L = \sum_{i=1}^{4} (y_i - mx_i - b)^2$

$L \to b^2$

L(b)



$\hat{b} = \min \sum_{i=1}^{4} (y_i - 78.35 * x_i - b)^2$

Step 1 - select a random b_{min}

b increment

b decrement

L_{min}

b = −10

x = 5

b = −10 → Slope = −ve

$b_{new} = b_{old} - slope$

$b_{new} = b_{old} - \eta \, slope$

−10 − (−50)

40

−40

10 − (50)

0.01

0.1, 0.001

× 50

−10

$b_{new} = -10 + (0.01 \times 50)$

$= -10 + 0.5 = -9.5$

$y = x^2 + 2$

$\frac{dy}{dx} = 2x$

= 10

−40

−9.5

$b_{new} = -9.5 - (0.01 * -40) =$

$-9.5 + 0.4 = -9.1$

When to stop

1) diff b/w b_{old} b_{new}

=> 0.0001

$b_{new} - b_{old} = 0.0001$

=> 0.0001

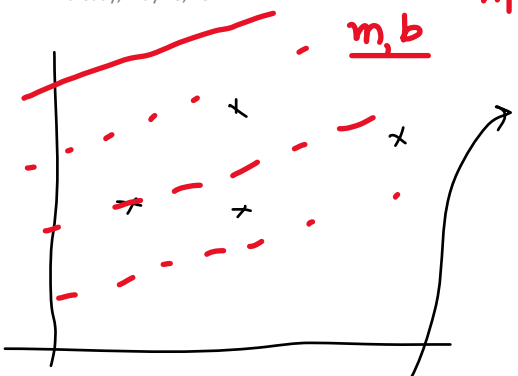
2) Iteration => 1000, 100,

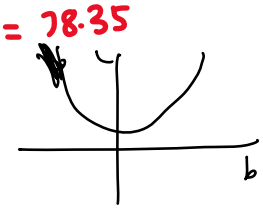
epochs

$b_{new} - b_{old} = 0$

Mathematical Formulation

Thursday, May 20, 2021 1:47 PM





Step → Start with a random $b = b'$

for i in epochs; 1000, 100

$i = 0$ 0,1,

$\eta = 0.01$

$b = 0$

$$L = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

$$b_{new} = b_{old} - \eta * slope \quad (b = 0)$$

$$\frac{dL}{db} = \frac{d}{db}\left(\sum_{i=1}^{n} (y_i - \hat{y}_i)^2\right) = 2\sum_{i=1}^{n} (y_i - mx_i - b)(-1)$$

$$\frac{d}{db}\sum_{i=1}^{n} (y_i - mx_i - b)^2 \quad \text{slope} = -2\sum_{i=1}^{n} (y_i - mx_i - b)$$

$$= -2\sum_{i=1}^{n} (y_i - 78.35 * x_i - 0)$$

$i = 1$

epoch

slope $(b=0)$

$$b_{new} = b_{old} - \eta \, slope_{b=b_{old}}$$

b_{new}

step size

Example

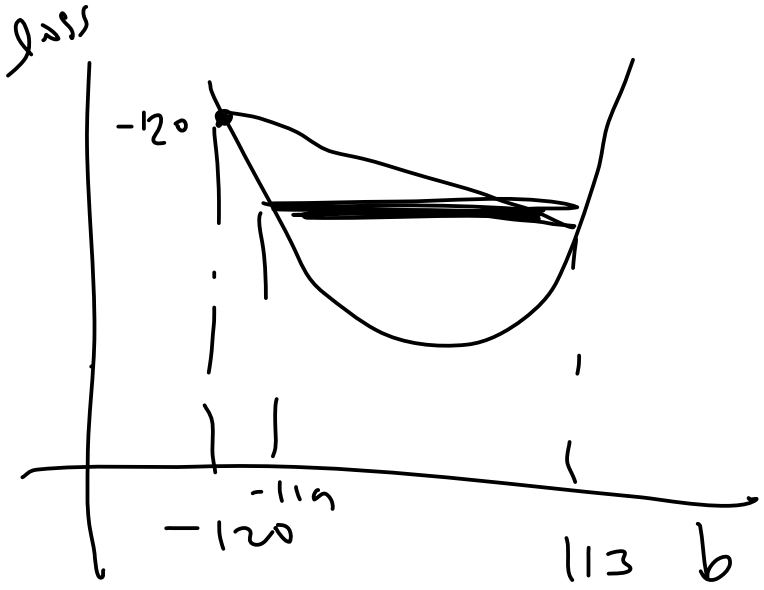
Thursday, May 20, 2021 1:47 PM

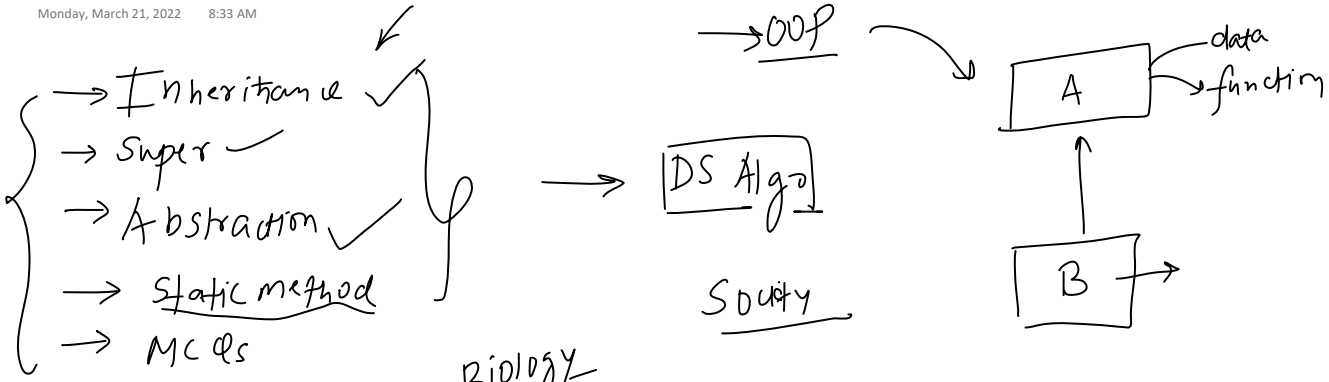
Code from Scratch

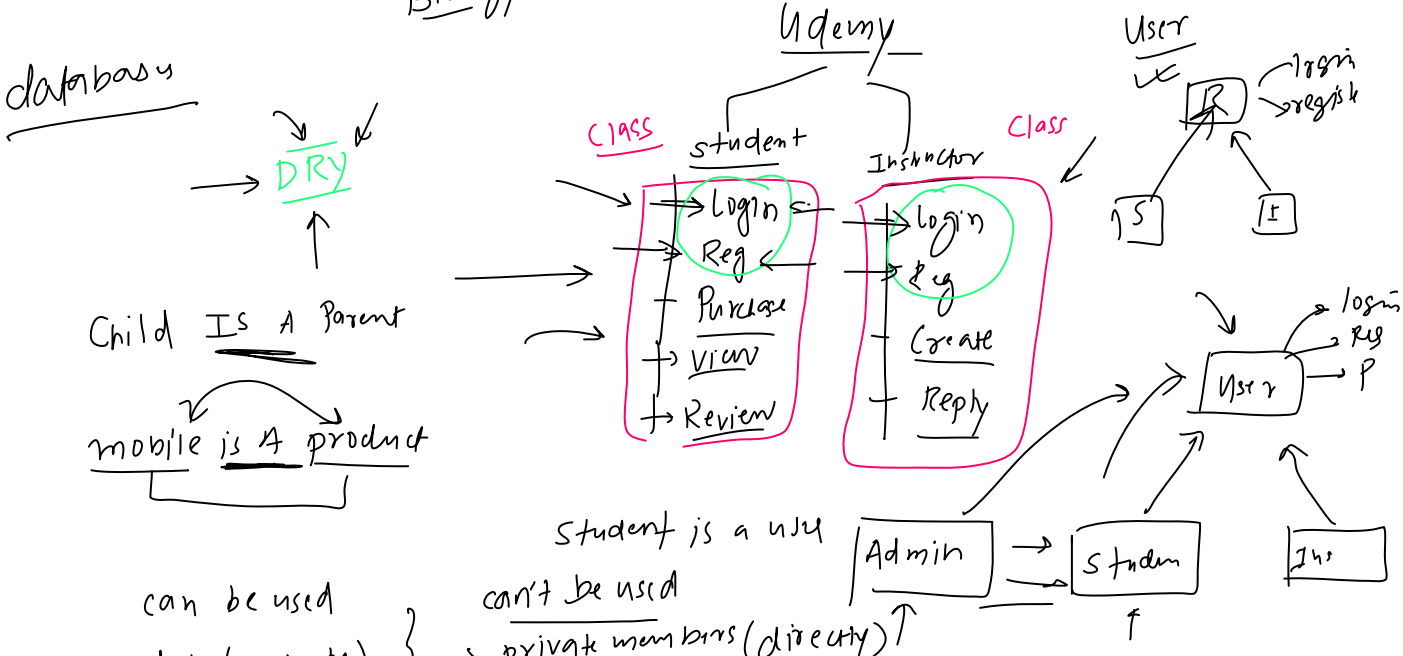
Thursday, May 20, 2021 1:48 PM

Visualization 1

Thursday, May 20, 2021 1:52 PM

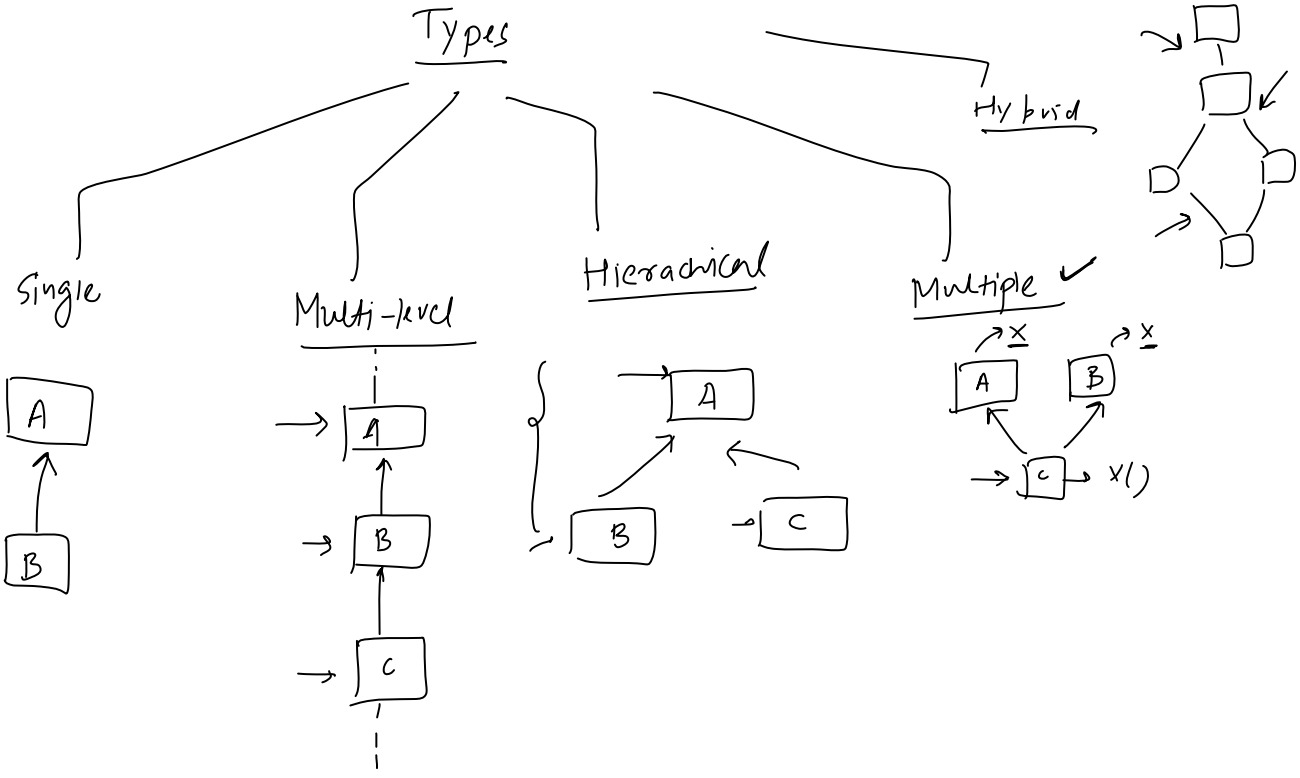


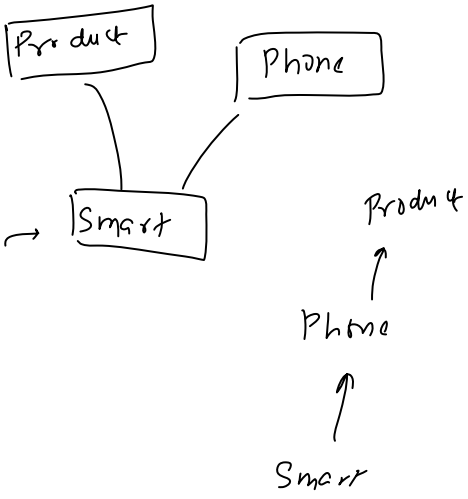


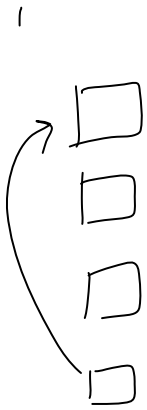


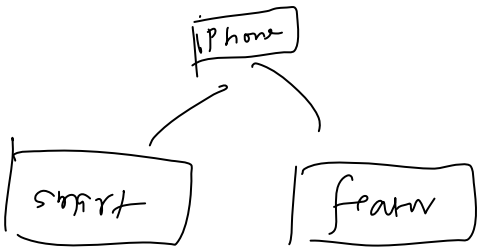
can be used
- data (property)
- function
- constructor

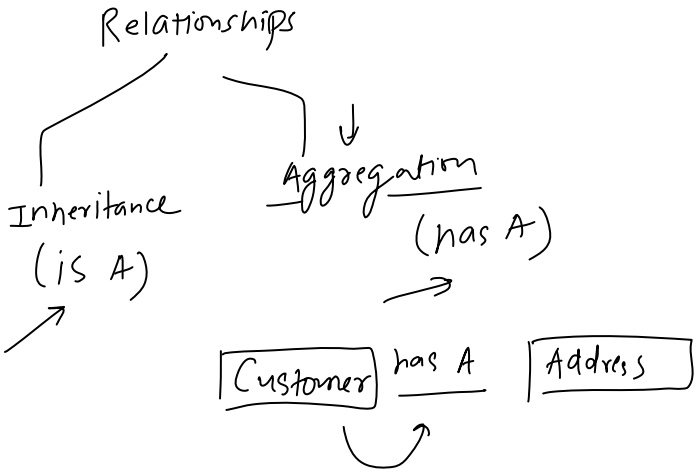
can't be used
- private members (directly)



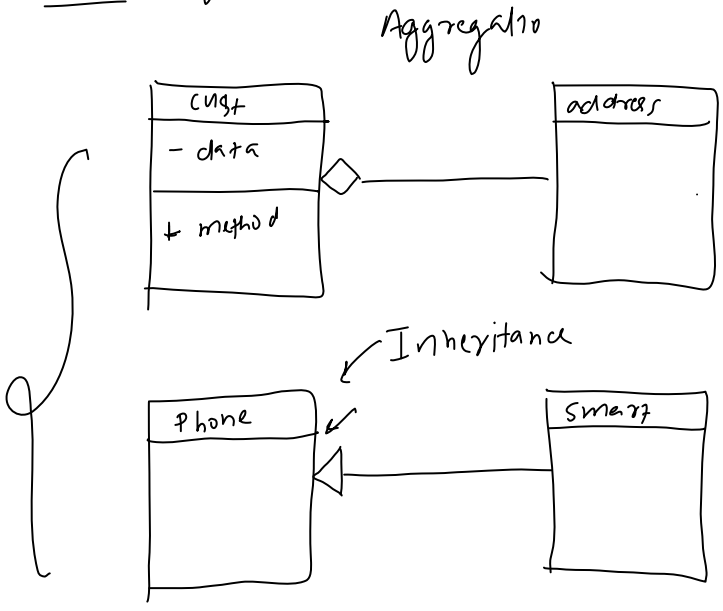




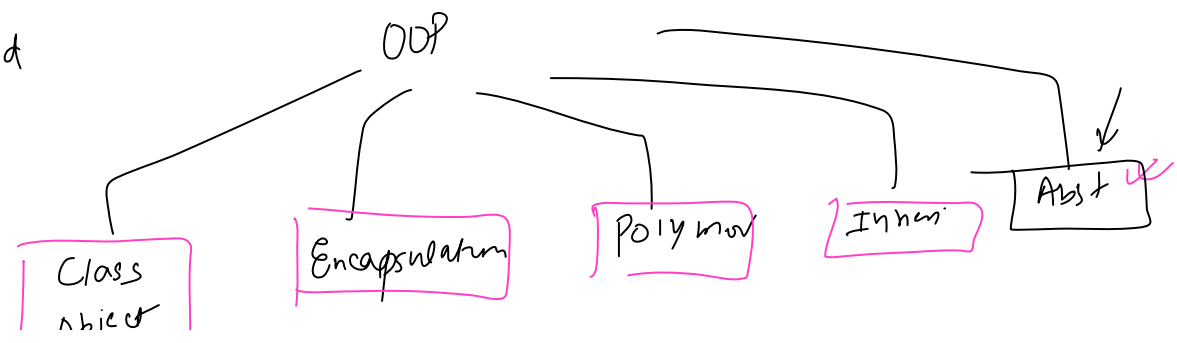




Class diagram



Construct
static method
inst vs stat
super



Class
Object

Encapsulation

20 → 9 → DS Algo → 8 classes

Class 9 → Time Complexity
Class 10 → Dynamic Arrays (List)
Class 11 → LL
Class 12 → Stacks + Queues
Class 13 → Hashing → Searching and sorting
Class 14 → Tree + Graph
Class 15 → Dynamic + greedy
Class 16 → Exam + MCQ + qns

Python → Except
→ file
→ Learn

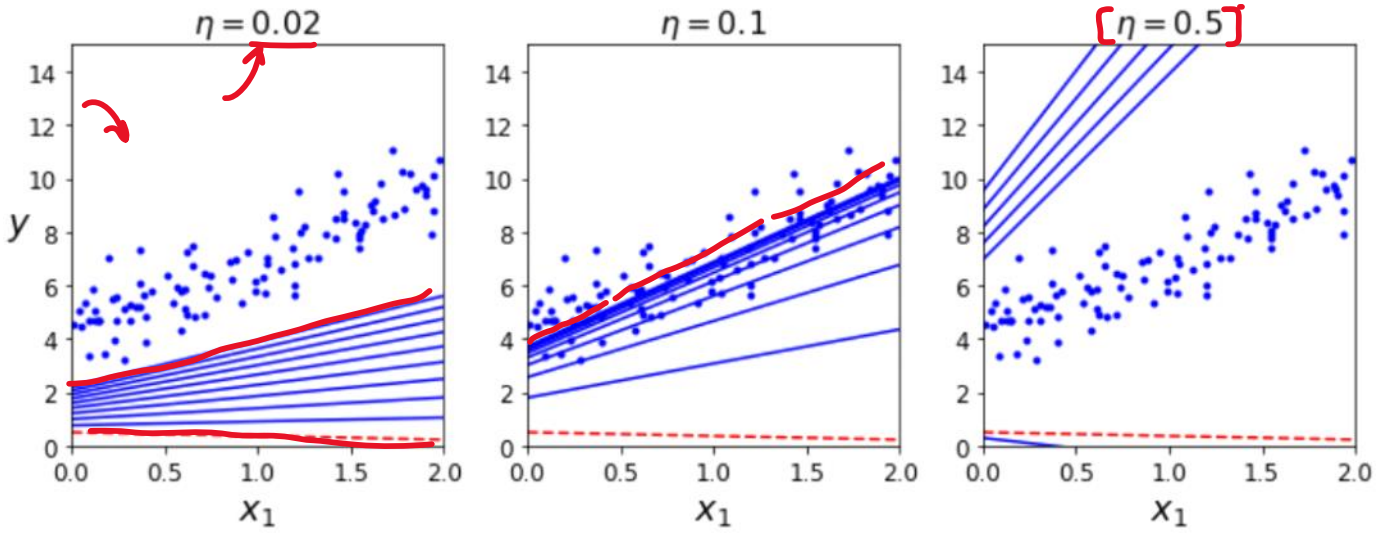
{ 17 18 19 20 }
SQL

Few Discussions

Thursday, May 20, 2021 4:47 PM

1. Effect of Learning rate
2. The universality of Gradient Descent

epoch = 10



$b = 0$

$b = b_{old} - \eta \cdot \text{slope}$

$\frac{dL}{db} = [\quad]$

$\sum (y_i - \hat{y}_i)^2$ (LR)

Deep →

LOR → [function

Adding m into the mix

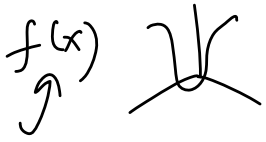
Thursday, May 20, 2021 1:48 PM

Steps

1) init random vals for m and b
m = 1 and b = 0

2) epochs = 100, lr = 0.01
for i in epochs:
b = b − η slope
m = m − η slope

z = f(x, y)



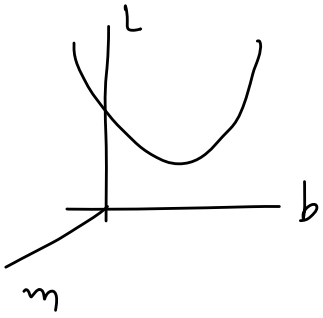
b = 0

$L = \sum (y_i - \hat{y}_i)^2 = \sum (y_i - mx_i - b)^2$

L(m, b)

$\frac{dL}{db}$

b_slope = $\frac{\partial L}{\partial b}$



$m\_slope = \frac{\partial L}{\partial m}$

$\sum (y_i - mx_i - b)^2$

2D graph (m, b)

$\frac{\partial L}{\partial b} = -2 \sum (y_i - mx_i - b)$

$(y_i - mx_i - b)$: 1, 0, 0, −1

$= -2 \sum (y_i - mx_i - b)$

= slope_b at b = 0

$\frac{\partial L}{\partial m} = 2 \sum (y_i - mx_i - b)$

$(y_i - mx_i - b)$: 0, $-x_i$, 0

$= -2 \sum (y_i - mx_i - b) x_i$

slope_m at m = 1

Code

Thursday, May 20, 2021 1:48 PM

Visualization 2

Thursday, May 20, 2021 1:52 PM

Effect of Learning Data

Thursday, May 20, 2021 1:53 PM

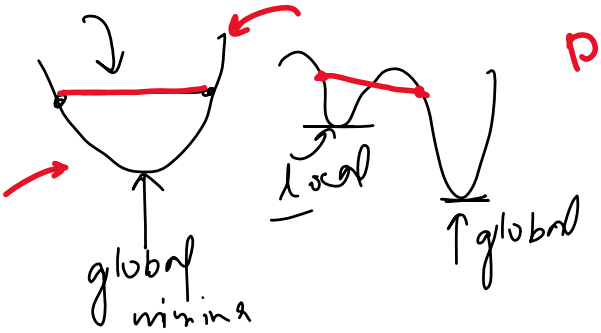
Effect of Loss Function

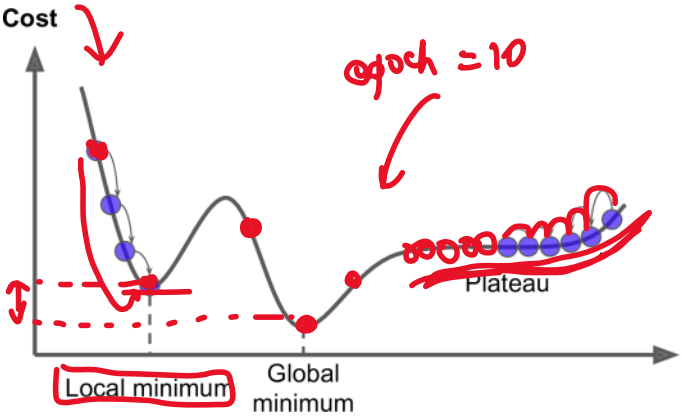
Thursday, May 20, 2021 1:53 PM

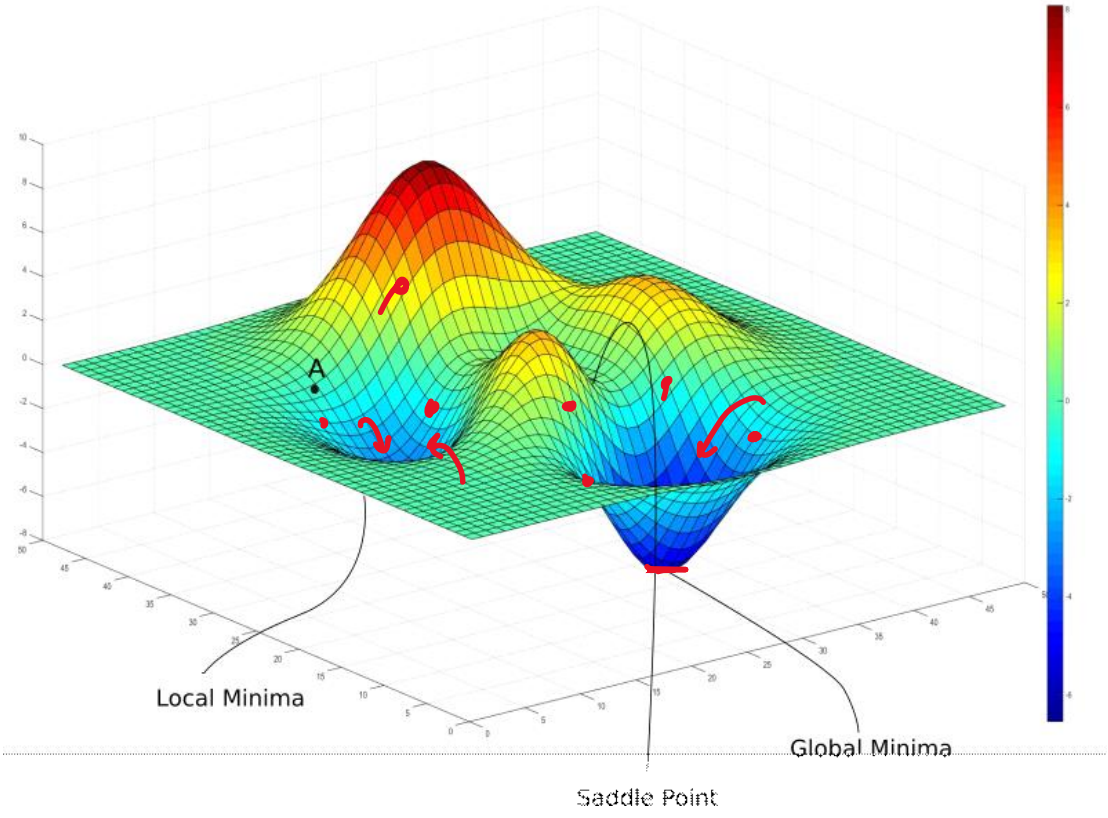
$L = \sum (y_i - \hat{y}_i)^2$ → convex function

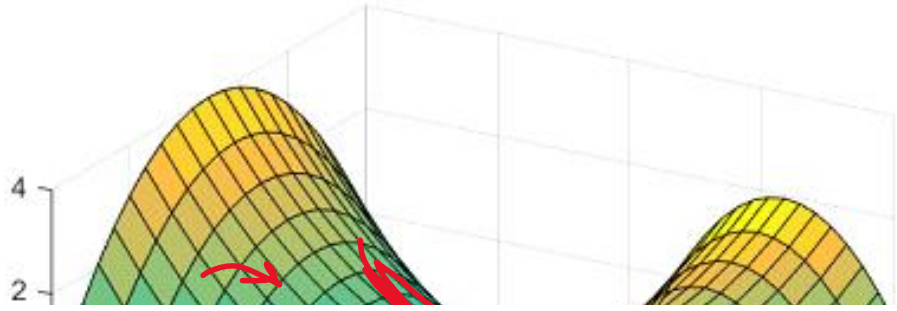
10 cols

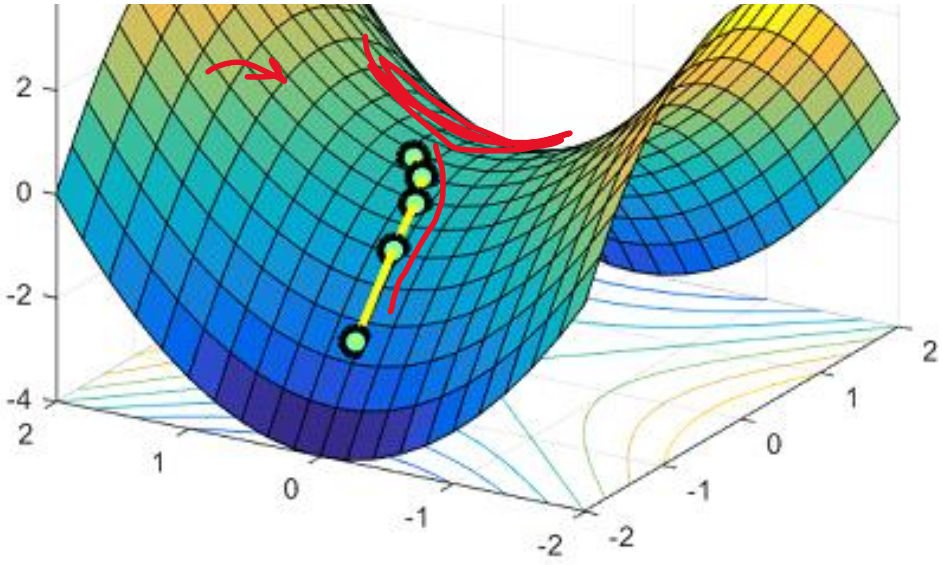
deep learn





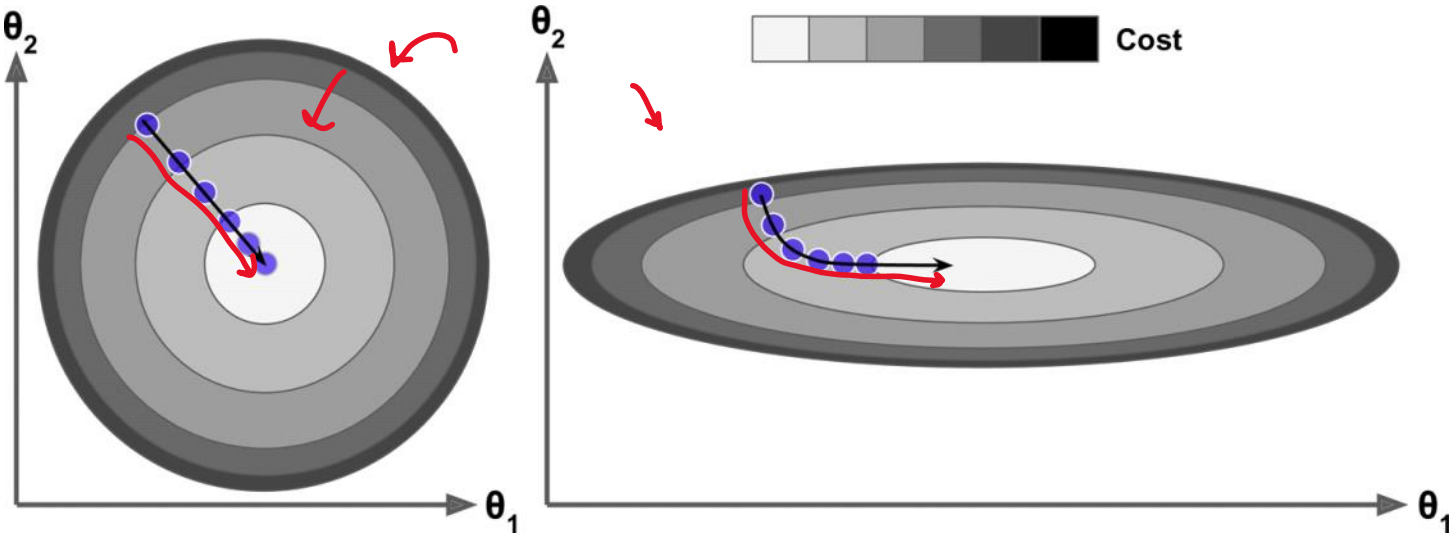





2
0
-2
-4
2
1
0
-1
-2
2
1
0
-1
-2

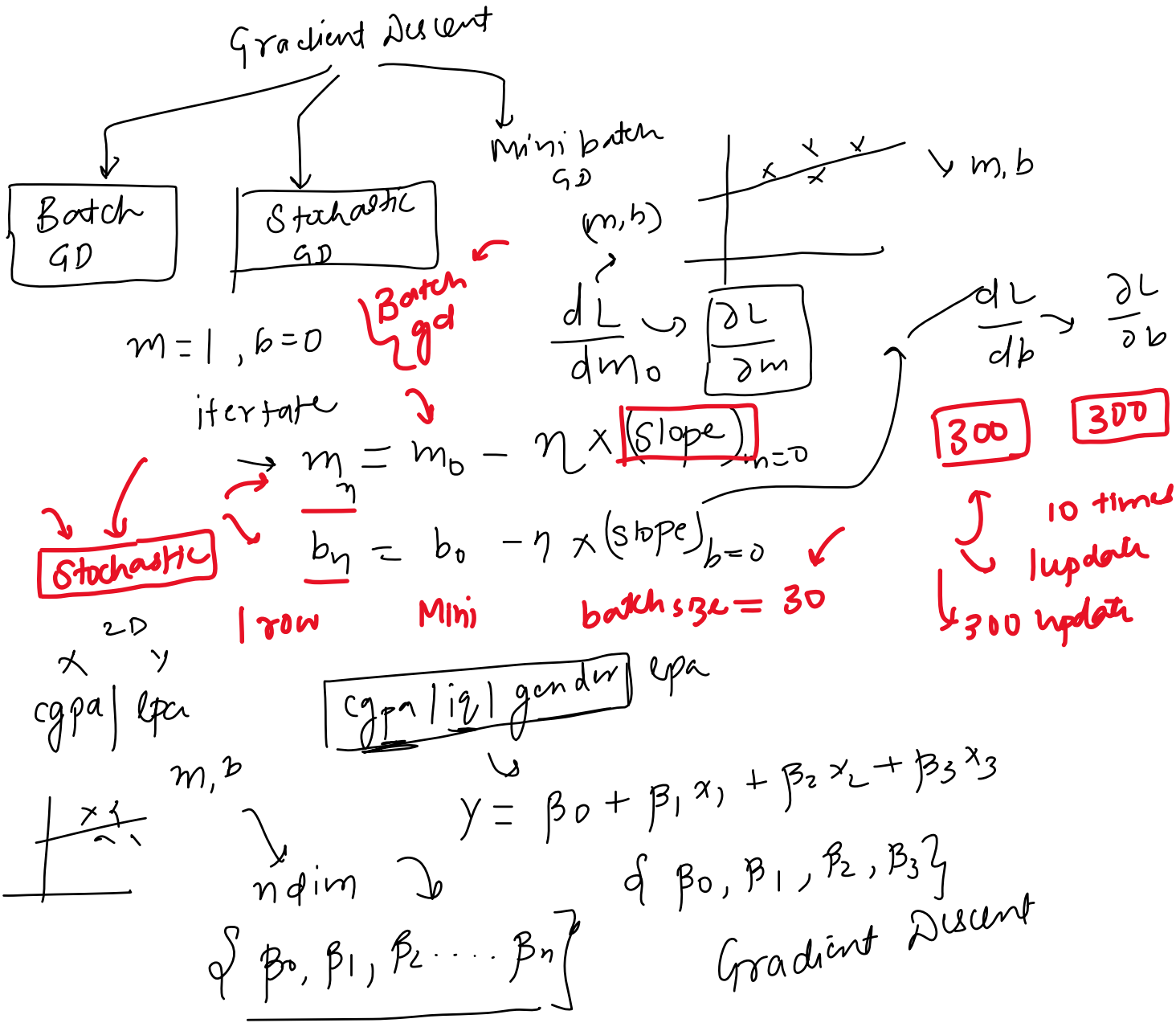
Effect of Data

Thursday, May 20, 2021 1:53 PM



Types of Gradient Descent

Saturday, May 22, 2021 1:30 PM



Batch GD

Saturday, May 22, 2021 3:38 PM

n-dim-dataset 3-cols

$\hat{y}_1$ x_{11} | x_n

$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2$ x_{21} | x

(lpa) (cgpa) (iq)

x cgpa | iq | lpa

(2,3) .

{m, b}

{ $\beta_0, \beta_1, \beta_2$ }

L ($\beta_0, \beta_1, \beta_2$)

1) Random values

$\beta_0 = 0$, $\beta_1, \beta_2 = 1$

2) epoch = 100, α = 0.1

$\beta_0 = \beta_0 - \eta$ slope

$\beta_1 = \beta_1 - \eta$ slope

$\beta_2 = \beta_2 - \eta$ slope

$\frac{\partial L}{\partial \beta_0}$ $\frac{\partial L}{\partial \beta_1}$ $\frac{\partial L}{\partial \beta_2}$ $\frac{\partial L}{\partial \beta_3}$ β_n β_n

n-dim

(n+1) $\beta_0 - \beta_n$

MSE

$L = \frac{1}{2n} \sum_{i=1}^{n=2} (y_i - \hat{y}_i)^2$

{ row = 2, cols = 2+1 }

$\hat{y}_1 = \beta_0 +$

$= \frac{1}{2} \left[(y_1 - \hat{y}_1)^2 + (y_2 - \hat{y}_2)^2 \right]$

x_1	x_2	y
8.1	93	3.2
7.5	95	3.5

$L = \frac{1}{2} \left[(y_1 - \beta_0 - \beta_1 x_{11} - \beta_2 x_{12})^2 + (y_2 - \beta_0 - \beta_1 x_{21} - \beta_2 x_{22})^2 \right]$

$\hat{y} = \beta_0 + \beta_1 x_1 + \beta_2 x_2$

$\hat{y}_1 = \beta_0 + \beta_1 x_{11} + \beta_2 x_{12}$

$\hat{y}_2 = \beta_0 + \beta_1 x_{21} + \beta_2 x_{22}$

$\frac{\partial L}{\partial \beta_0} = \frac{1}{2} \left[2 (y_1 - \hat{y}_1)(-1) + 2 (y_2 - \hat{y}_2)(-1) \right]$

$\frac{\partial L}{\partial \beta_0} = -\frac{2}{2} \left[(y_1 - \hat{y}_1) + (y_2 - \hat{y}_2) \right]$

$= \frac{-2}{n} \left[(y_1 - \hat{y}_1) + (y_2 - \hat{y}_2) + (y_3 - \hat{y}_3) + \dots + (y_n - \hat{y}_n) \right]$

y → array

$= \frac{-2}{n} [(y_1 - \hat{y}_1) + (y_2 - \hat{y}_2) + \dots$

$= \frac{-2}{n} \sum_{i=1}^{n} (y_i - \hat{y}_i) = \frac{\partial L}{\partial \beta_0}$

$L = \frac{1}{2} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$ $\quad \frac{1}{n}\sum_{i=1}^{n} x_i$ $\quad \frac{\partial}{\partial \beta} - \beta_1 x_{11} = -x_{11}$

$L = \frac{1}{2} [(y_1 - \hat{y}_1)^2 + (y_2 - \hat{y}_2)^2]$

$L = \frac{1}{2} [(y_1 - \beta_0 - \beta_1 x_{11} - \beta_2 x_{12})^2 +$

$(y_2 - \beta_0 - \beta_1 x_{21} - \beta_2 x_{22})^2]$

$\frac{\partial L}{\partial \beta_1} = \frac{1}{2} [2(y_1 - \hat{y}_1)(-x_{11}) + 2(y_2 - \hat{y}_2)(-x_{21})]$

$11, 21, 31 \dots n1$

$\frac{\partial L}{\partial \beta_1} = \frac{-2}{n} [(y_1 - \hat{y}_1)(x_{11}) + (y_2 - \hat{y}_2)(x_{21}) + (y_3 - \hat{y}_3)(x_{31}) + \dots + (y_n - \hat{y}_n)(x_{n1})]$

$\frac{\partial L}{\partial \beta_1} = \frac{-2}{n} \sum_{i=1}^{n} (y_i - \hat{y}_i) x_{i1}$

x_{i1} → 1 col data

β_1 → values of 1 col.

m cols

$\beta_0 - \beta_m$

$\frac{\partial L}{\partial \beta_2} = \frac{-2}{n} \sum_{i=1}^{n} (y_i - \hat{y}_i) x_{i2}$

$\frac{\partial L}{\partial \beta_m} = \frac{-2}{n} \sum_{i=1}^{n} (y_i - \hat{y}_i) x_{im}$ code

$\hat{y}_1 = \beta_0 + \beta_1 x_{11} + \beta_2 x_{12} + \beta_3 x_{13}$

$\hat{y}_2 = \beta_0 + \beta_1 x_{21} + \beta_2 x_{22} + \beta_3 x_{23}$

$\beta_0 + np.dot(X\_train, coef)$ $\hat{y} = \beta_0 + [x_{11}\ x_{12}\ x_{13}] \begin{bmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \end{bmatrix}$

(353,10) (10,1)

$\hat{y} = np.dot(coef\_, Xtrain) + \beta_0$

$\beta_0 + (353,1)$

$\hat{y}$ (353,1)

y_pred

x_1	x_2	y	$\hat{y}$
1	2	5	6
3	4	7	8

y $\hat{y}$

$[5\ \ 7]\ [6\ \ 8]$

$y - \hat{y} = [-1\ \ -1]$

$\frac{\partial L}{\partial \beta_1} = -\frac{2}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)x_{i1} = -\frac{2}{n}\left[[-1\ -1]\begin{bmatrix}1\\3\end{bmatrix}\right]$

$[-4 + -4]$

$-8 = 8$

(1,2)

$\frac{\partial L}{\partial \beta_2} = [y - \hat{y}]\begin{bmatrix}2\\4\end{bmatrix}$

$= [-1\ -1]\begin{bmatrix}2\\4\end{bmatrix} \times \frac{-2}{n}$

(1,2) (2,2)

$\left[[y - \hat{y}]\begin{bmatrix}1 & 2\\ 3 & 4\end{bmatrix}\right] \times \frac{-2}{n}$

[x y]

$\frac{\partial L}{\partial \beta_1} \cdots \frac{\partial L}{\partial \beta_{10}} = \left[(y_i - \hat{y}_i)\, X\_train\right] \times \frac{-2}{n}$

y_train

$y_i = 353$

353

(353,1) (353,10) (1,353)

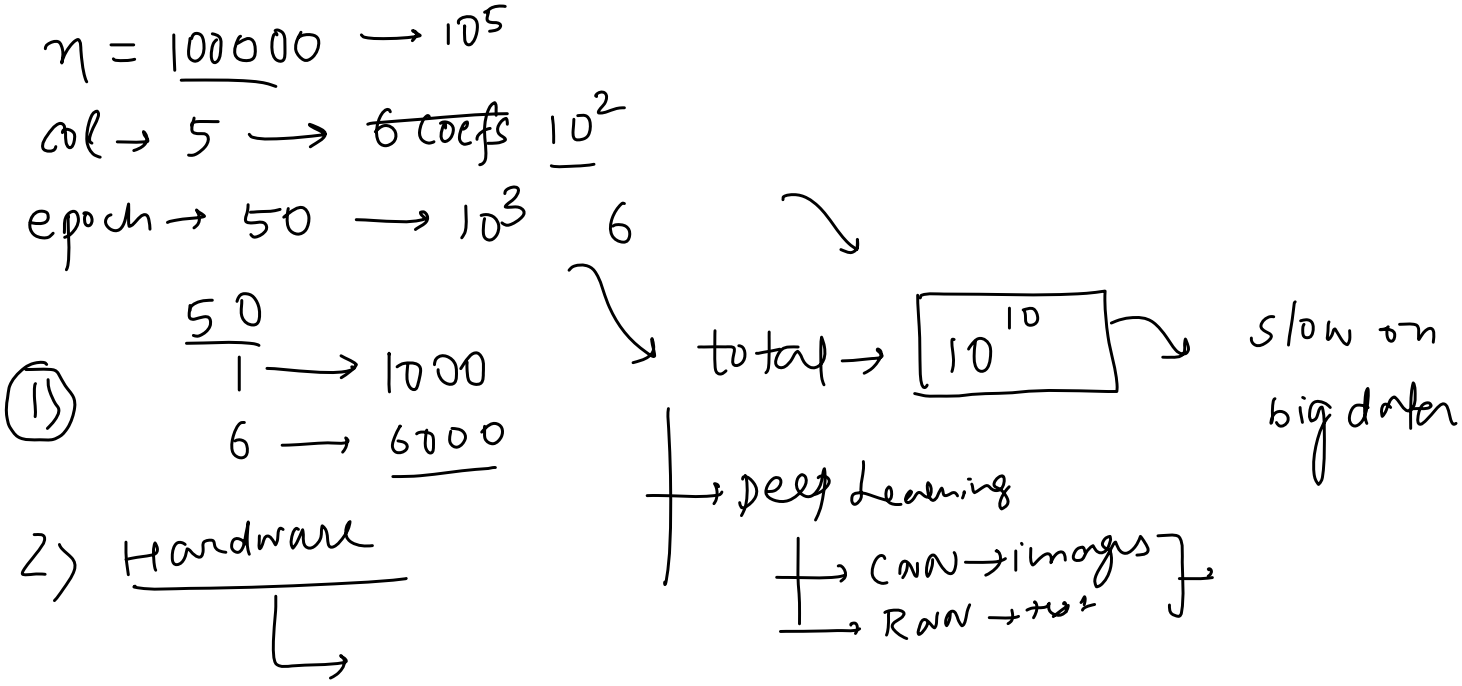
(1,353) (353,10)

coef_del

(1,10) $\times \frac{-2}{n}$ = (1,10)

The Problem with Batch GD

Tuesday, May 25, 2021 6:43 AM



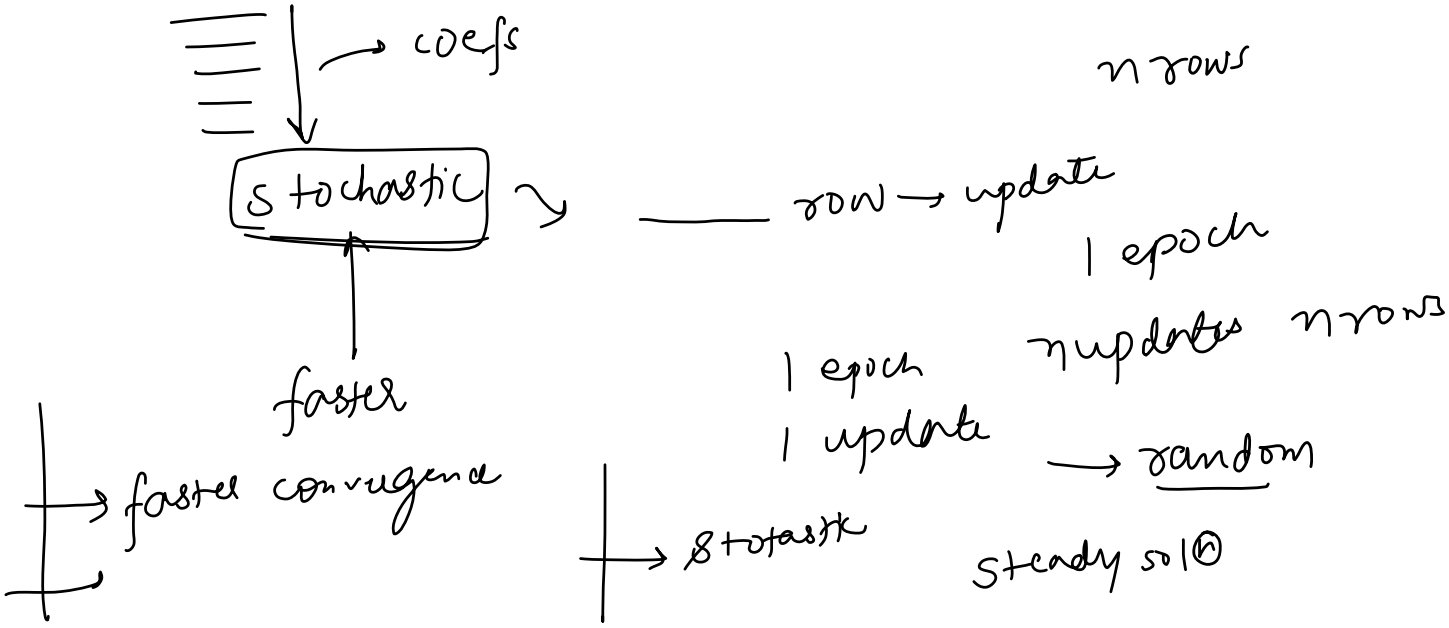
Stochastic GD

Tuesday, May 25, 2021 6:44 AM

5, 10

100 epochs

Batch



Code

Tuesday, May 25, 2021 6:44 AM

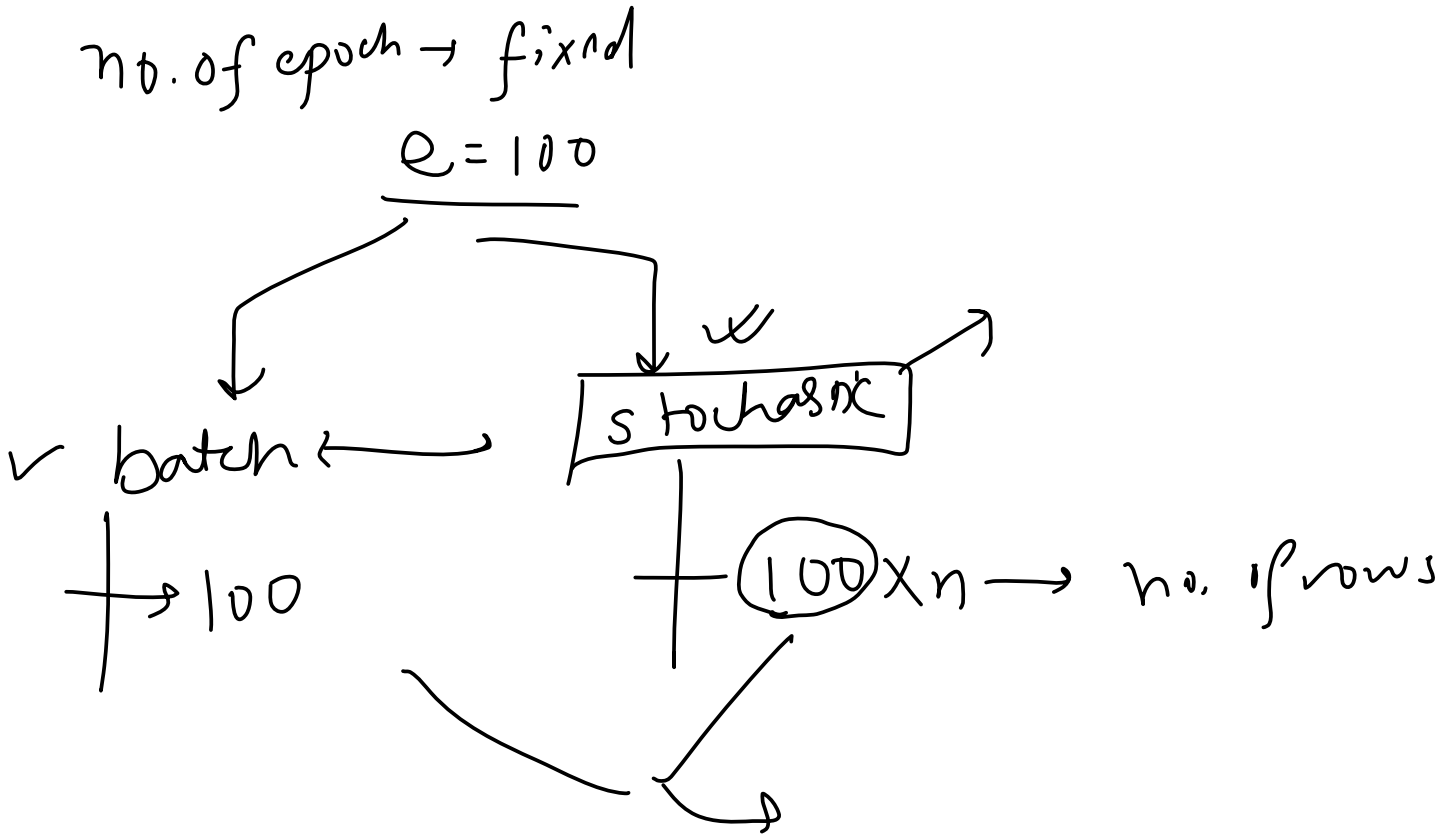
$$\frac{\partial L}{\partial \beta_0} = \frac{-2}{m} \sum_{i=1}^{m} (y_i - \hat{y}_i)$$

$i = idx$

$$= -2 \left(y_i - \hat{y}_i \right)$$

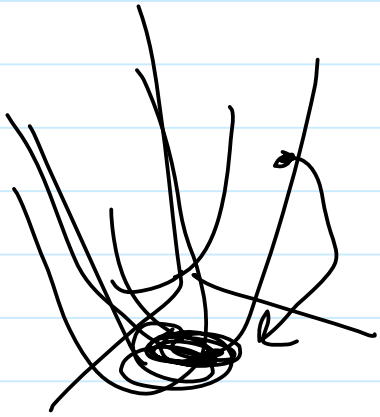
Time Comparison

Tuesday, May 25, 2021 12:39 PM



Visualizations

Tuesday, May 25, 2021 6:44 AM

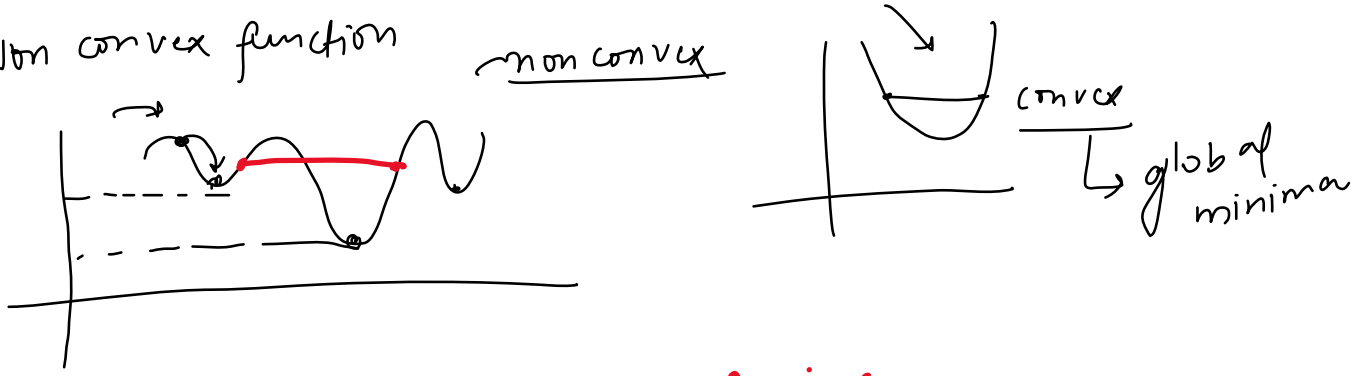


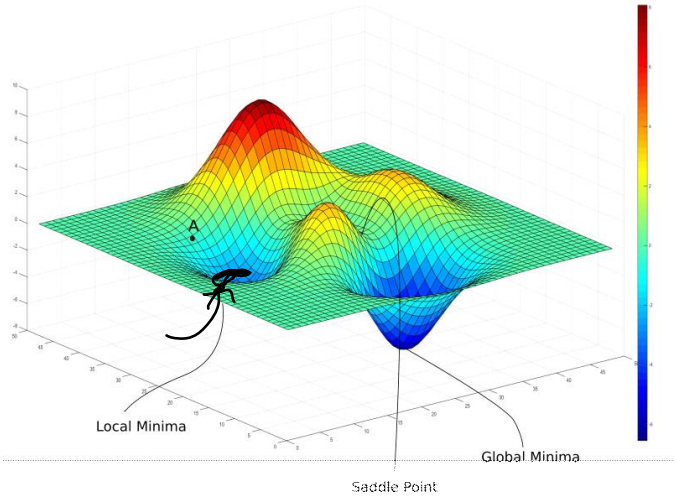
When to use Stochastic GD

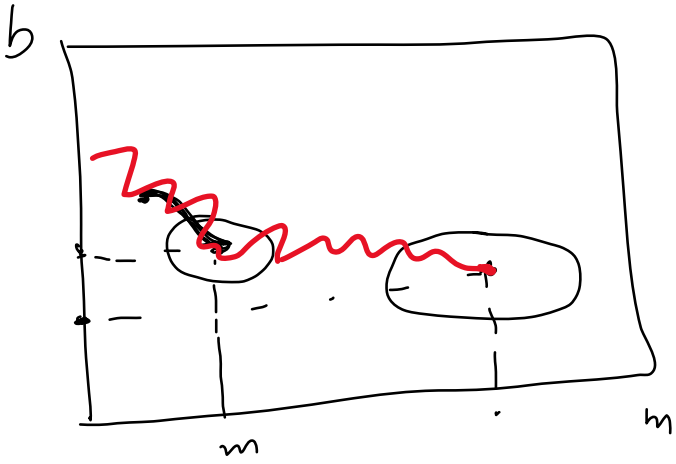
Tuesday, May 25, 2021 6:44 AM

1) Big data → sgd

2) Non convex function

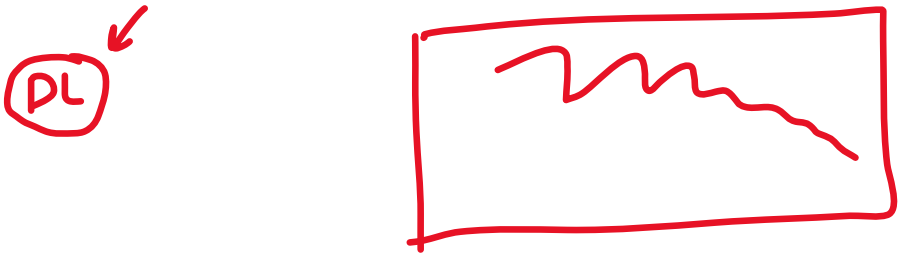


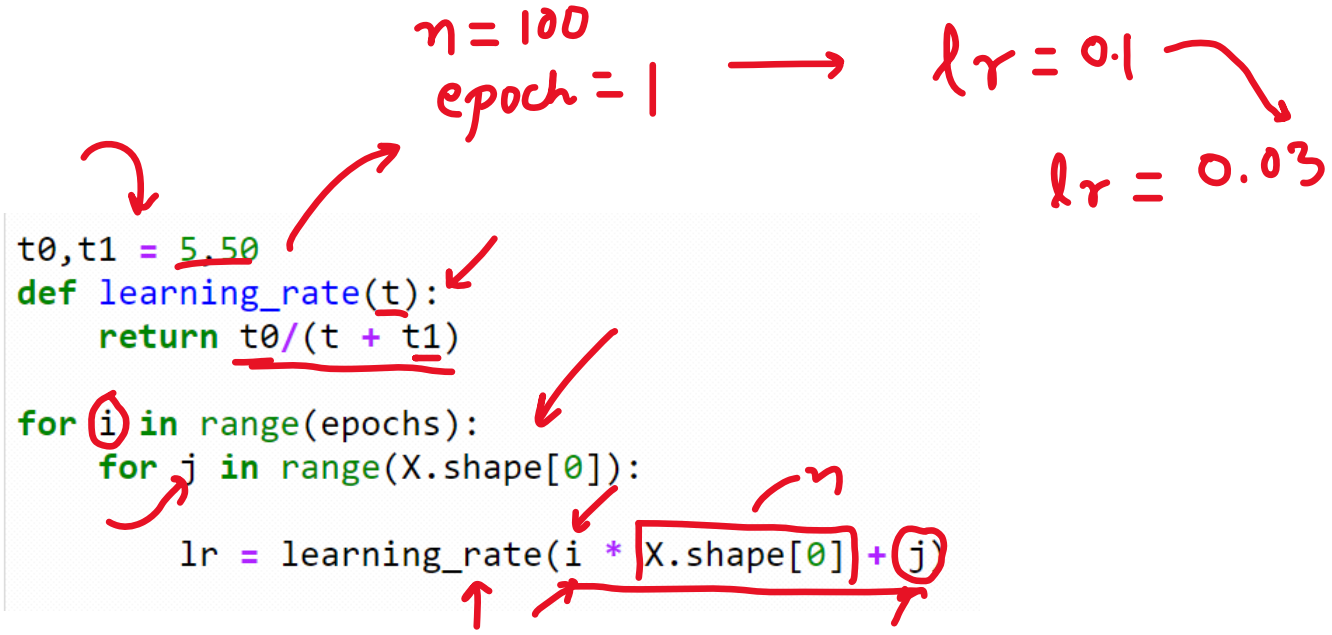




Learning Schedules

Tuesday, May 25, 2021 7:00 AM



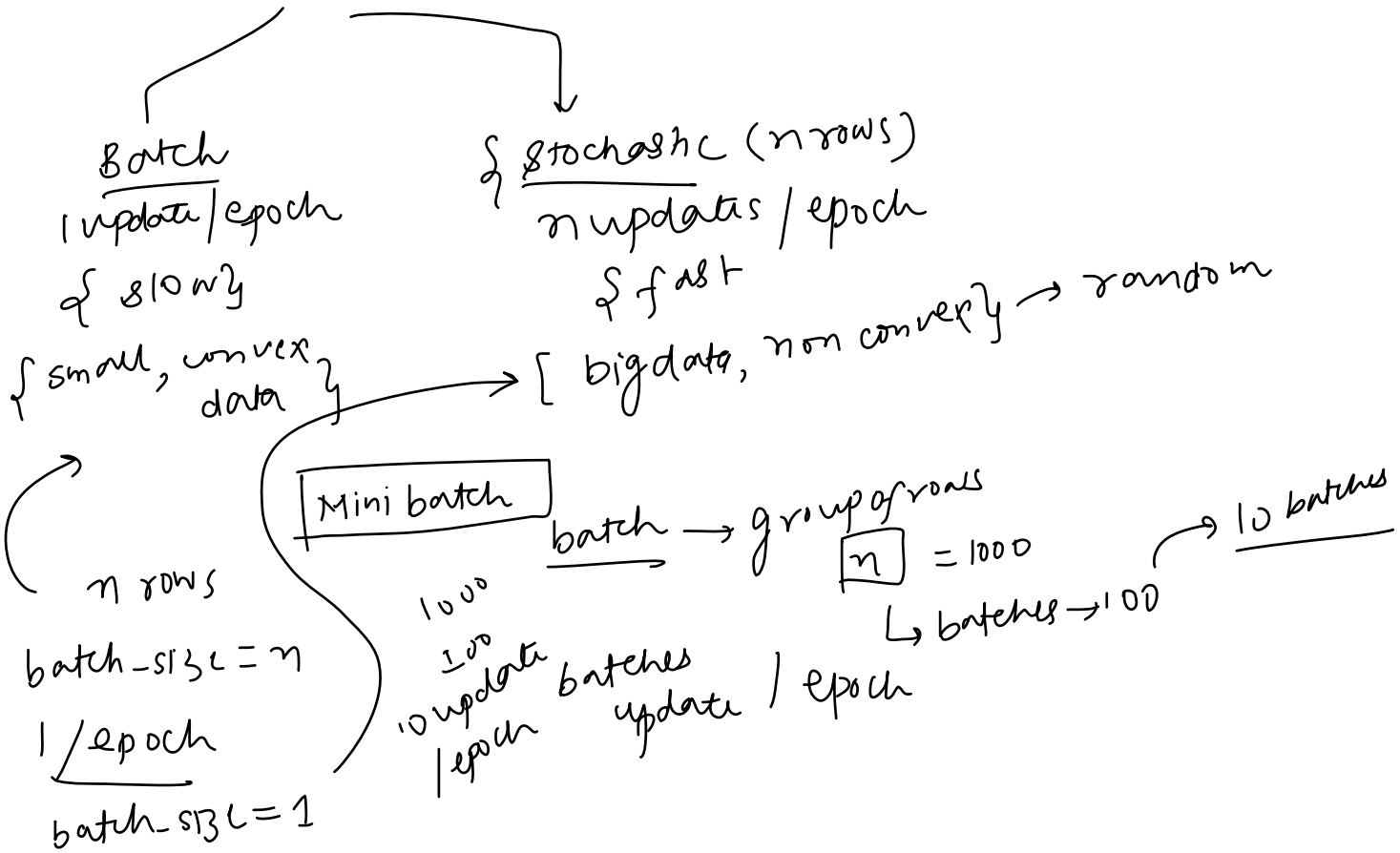


Sklearn Implementation

Tuesday, May 25, 2021 2:16 PM

Mini-Batch Gradient Descent

Wednesday, May 26, 2021 4:47 PM



Code

Wednesday, May 26, 2021 4:47 PM

Visualization

Wednesday, May 26, 2021 4:47 PM

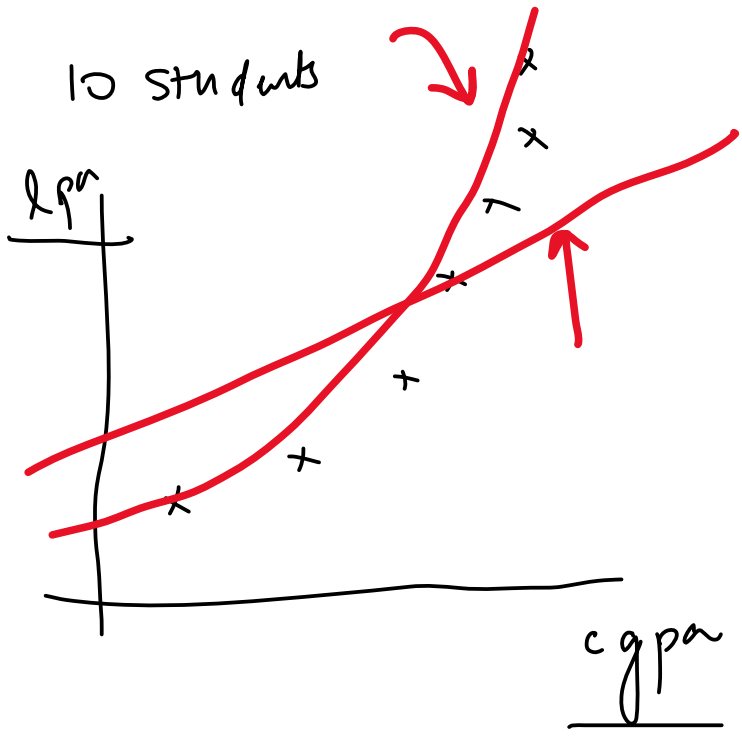
Sklearn Implmentation

Wednesday, May 26, 2021 4:47 PM

Introduction

Friday, May 28, 2021 1:47 PM

cgpa | lpa

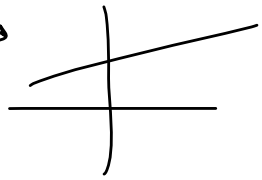


Intuition

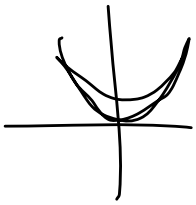
Friday, May 28, 2021 1:47 PM

polynomials

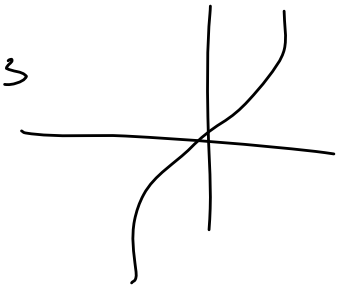
$y = mx + b$



$y = x^2$ →



$y = x^3$



$x^3 + x^2$

$y = ax^4 + bx^3 + cx^2 + dx + 2$

degree

1

x	y
2	

3

transform → degree = 2 polynomial

$y = mx + b$

3

x^0	x^1	x^2
1	2	4
1	3	9

$y = \beta_0 + \beta_1 x^0 + \beta_2 x^2$

Code

Friday, May 28, 2021 1:48 PM

Multiple Polynomial Regression

Friday, May 28, 2021 1:48 PM

Why is it linear

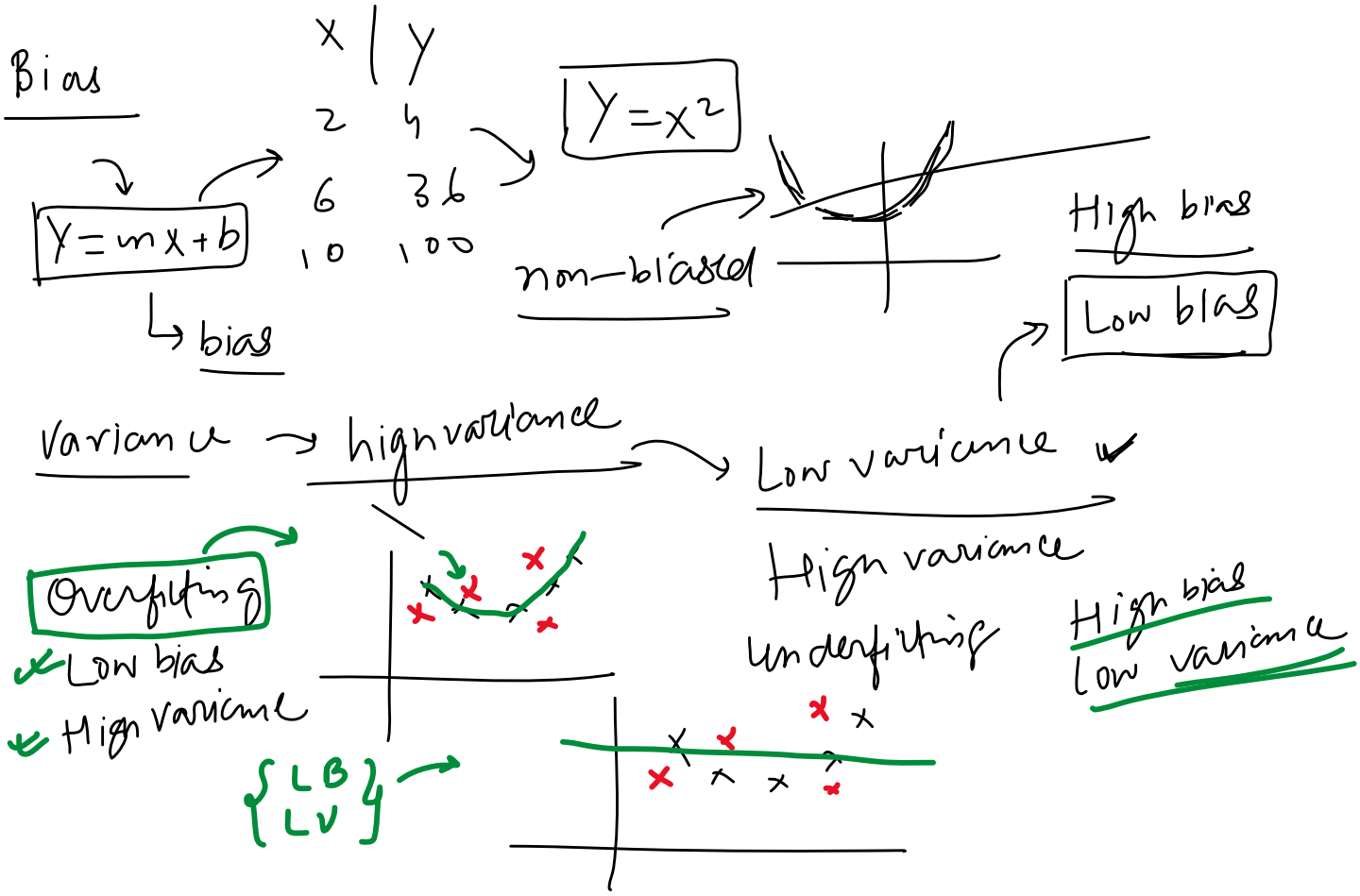
Friday, May 28, 2021 1:48 PM

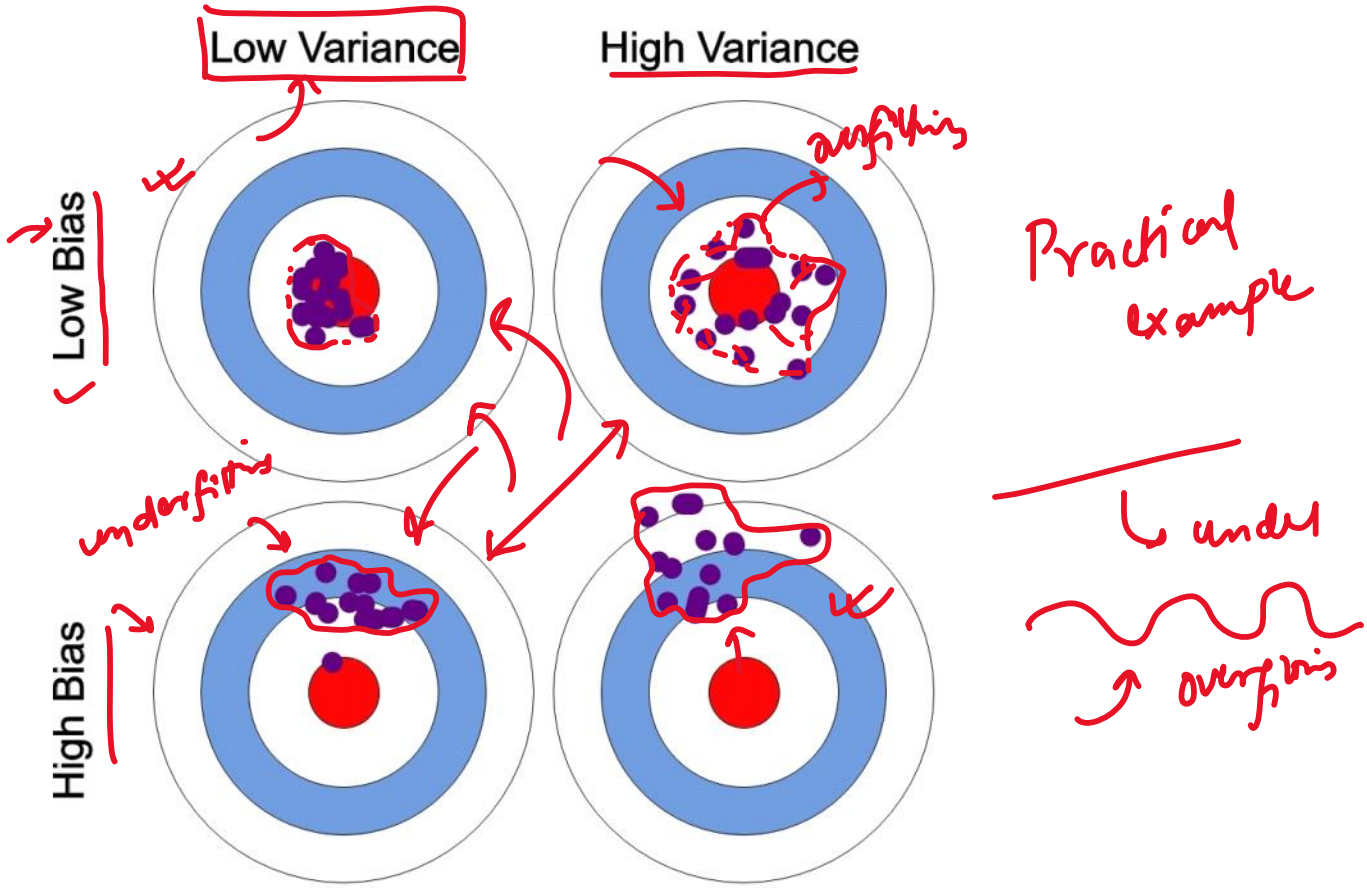
When to use Polynomial Regression

Friday, May 28, 2021 1:48 PM

Bias And Variance

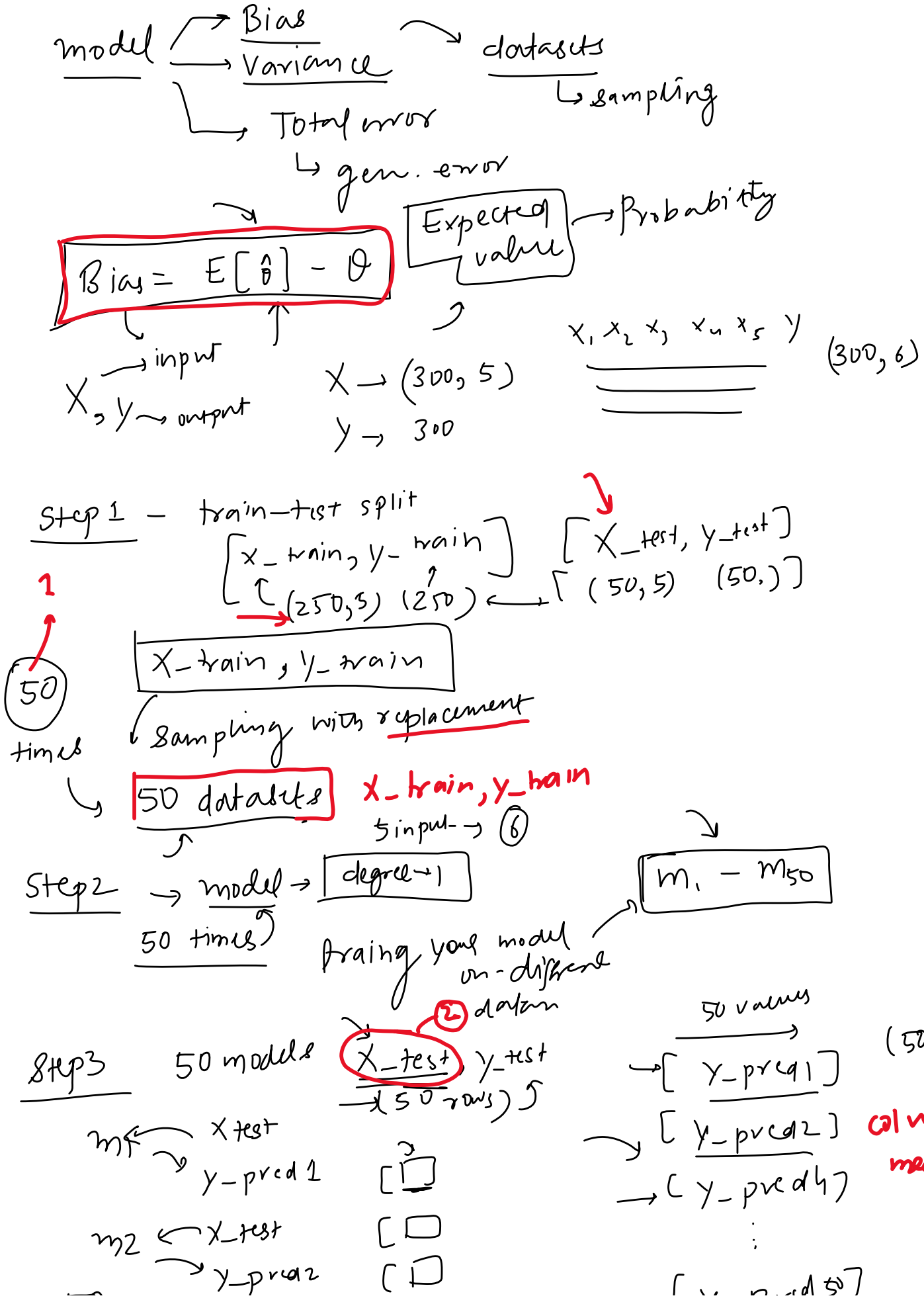
Tuesday, June 1, 2021 12:20 PM

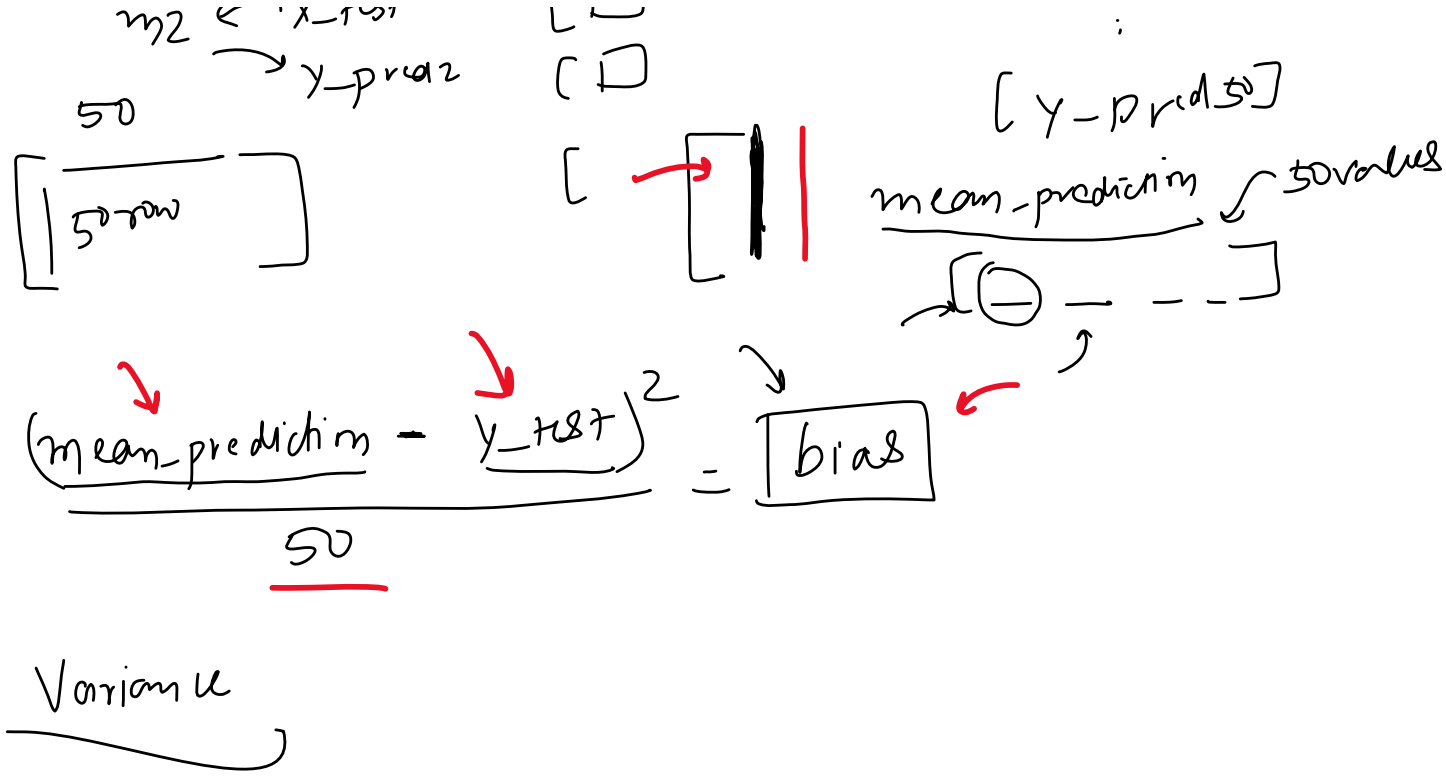




How to calculate Bias and Variance

Tuesday, June 1, 2021 1:44 PM




y_pred2
50
50 row
[y_pred50]
mean_prediction
50 values
(mean_prediction - y_test)^2
50
= bias

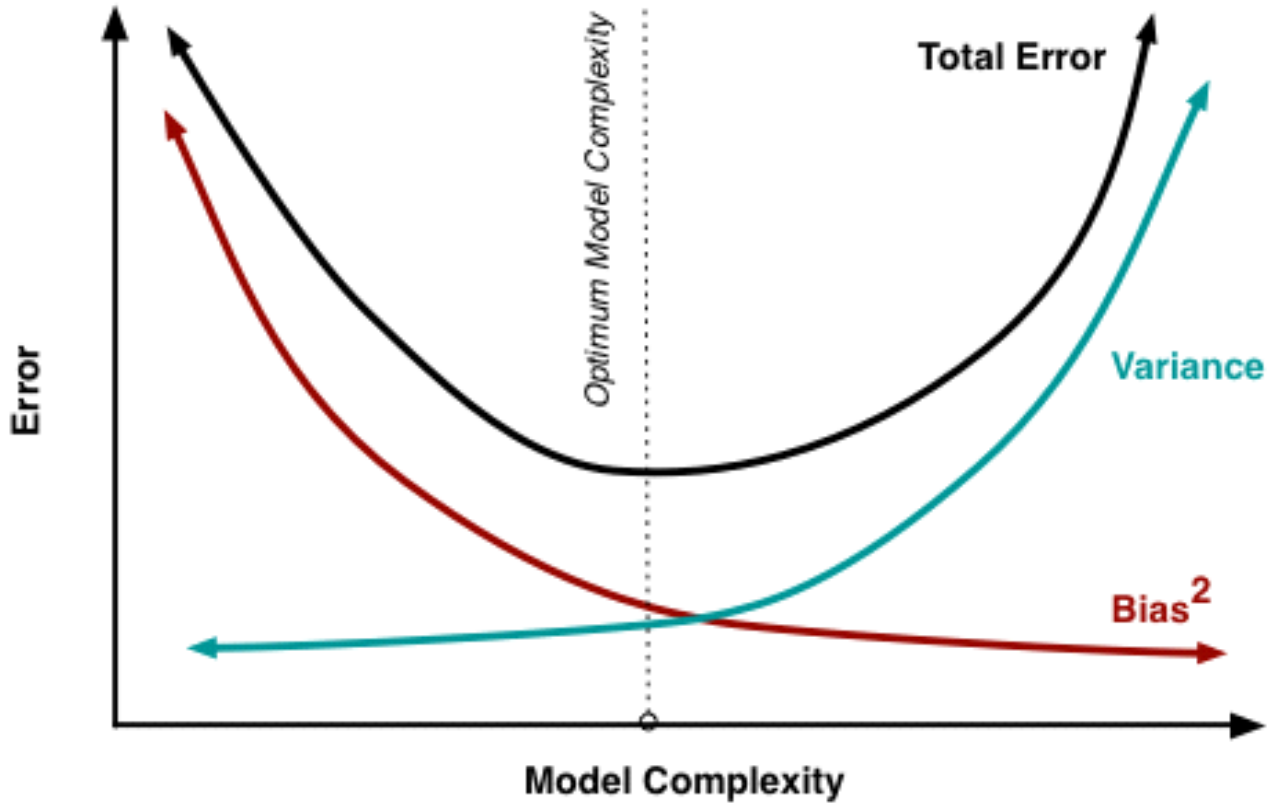
Variance

Code

Monday, May 31, 2021 9:26 AM

Bias Variance Decomposition for Squared Error

Monday, May 31, 2021 9:26 AM

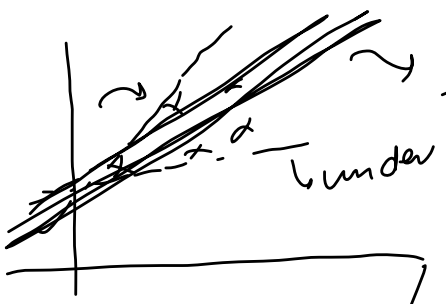


Relationship between Bias and Variance

Monday, May 31, 2021 9:26 AM

Ridge Regression

Wednesday, June 2, 2021 11:13 AM



bias ↑ variance ↓

under

$y = mx + b$ 2D line

$$L = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 + \lambda m^2$$

m slope

$$L = \sum_{i=1}^{n} (y_i - m x_i - b)^2 + \lambda m^2$$

$\frac{\partial L}{\partial b} = 0 \rightarrow b$

$\frac{\partial L}{\partial m} = 0 \rightarrow m$

$$b = \bar{y} - m\bar{x}$$

$\bar{y} \rightarrow$ y-mean

$\bar{x} \rightarrow$ x-mean

m → slope

$$L = \sum_{i=1}^{n} (y_i - m x_i - \bar{y} + m\bar{x})^2 + \lambda m^2$$

$$\frac{\partial L}{\partial m} = 2 \sum_{i=1}^{n} (y_i - m x_i - \bar{y} + m\bar{x})(-x_i + \bar{x}) + 2\lambda m = 0$$

$$= -2 \sum_{i=1}^{n} (y_i - \bar{y} - m x_i + m\bar{x})(x_i - \bar{x}) + 2\lambda m = 0$$

$$= \lambda m - \sum_{i=1}^{n} [(y_i - \bar{y}) - m(x_i - \bar{x})](x_i - \bar{x}) = 0$$

n $(x_i - \bar{x})^2 - n$

$$= \lambda m - \sum_{i=1}^{n} (y_i - \bar{y})(x_i - \bar{x}) - m(x_i - \bar{x})^2 = 0$$

$$= \lambda m - \sum_{i=1}^{n} (y_i - \bar{y})(x_i - \bar{x}) + m \sum_{i=1}^{n} (x_i - \bar{x})^2 = 0$$

$$= \lambda m + m \sum_{i=1}^{n} (x_i - \bar{x})^2 = \sum_{i=1}^{n} (y_i - \bar{y})(x_i - \bar{x})$$

$$= m = \frac{\sum_{i=1}^{n} (y_i - \bar{y})(x_i - \bar{x})}{\sum_{i=1}^{n} (x_i - \bar{x})^2 + \lambda}$$

λ → hyperparam. alpha

$\lambda = 0 \quad \lambda = 10,000$

$$m = \frac{\sum_{i=1}^{n} (y_i - \bar{y})(x_i - \bar{x})}{\sum_{i=1}^{n} (x_i - \bar{x})^2}$$

$m \rightarrow b = \bar{y} - m\bar{x}$

Ridge Regression for 2D data

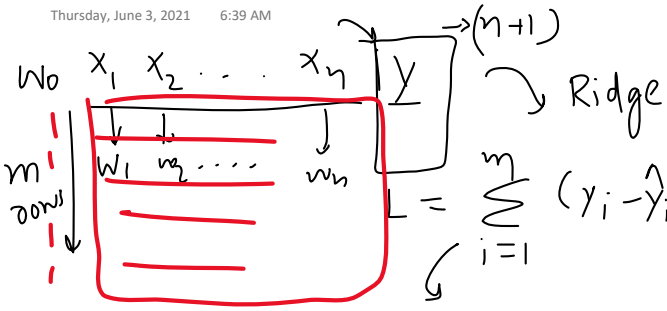
Thursday, June 3, 2021 6:38 AM

Code

Thursday, June 3, 2021 6:39 AM

Ridge Regression for nD data

Thursday, June 3, 2021 6:39 AM



$$L = \sum_{i=1}^{m} (y_i - \hat{y}_i)^2$$

$$= (Xw - y)^T (Xw - y)$$

m rows

$$y = \begin{bmatrix} - \\ - \\ - \\ - \end{bmatrix} \quad w = \begin{bmatrix} w_0 \\ w_1 \\ \vdots \\ w_n \end{bmatrix} \quad (n \times 1)$$

$$X = \begin{bmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1n} \\ 1 & x_{21} & x_{22} & \cdots & x_{2n} \\ 1 & \cdot & \cdot & & \cdot \\ \vdots & \cdot & & \ddots & \\ 1 & x_{m1} & x_{m2} & \cdots & x_{mn} \end{bmatrix}$$

Normal LR → Ridge

$$L = (Xw - y)^T (Xw - y) + \lambda \|w\|^2$$

$\lambda w_0^2 + \lambda w_1^2 + \lambda w_2^2 + \ldots + \lambda w_n^2$

$$[w_0\ w_1\ w_2 \ldots w_n] \begin{bmatrix} w_0 \\ w_1 \\ \vdots \\ w_n \end{bmatrix} \quad w^T w \rightarrow \lambda(w_0^2 + w_1^2 + w_2^2 \ldots w_n^2)$$

1 $(a-b)^T = a^T - b^T$

$\frac{dL}{dw}$

$(ab)^T = b^T a^T$

2 $L = (Xw - y)^T (Xw - y) + \lambda w^T w$

$y^T w x \rightarrow w x^T y^T$

3 $L = [(Xw)^T - (y)^T](Xw - y) + \lambda w^T w$

4 $= (w^T X^T - y^T)(Xw - y) + \lambda w^T w$

5 $= w^T X^T X w - w^T X^T y - y^T X w + y^T y + \lambda w^T w$

$$L = w^T X^T X w - 2 w^T X^T y + y^T y + \lambda w^T w$$

$$\frac{dL}{dw} = 2 X^T X w - 2 X^T y + 0 + 2\lambda w = 0$$

$$X^T X w + \lambda w = X^T y$$

3 (4×4)

$$X^T X w + \lambda w = X^T y$$

$$(X^T X + \lambda I) w = X^T y$$

3 (4x4)

(nx1, nx1)

$$w = (X^T X + \lambda I)^{-1} X^T y$$

$$w = (X^T X)^{-1} X^T y$$

Code

Thursday, June 3, 2021 6:39 AM

Ridge Regression using Gradient Descent

Friday, June 4, 2021 2:17 PM

Vector form loss

$$L = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \rightarrow L = (Xw - y)^T (Xw - y) + \lambda \|w\|^2$$

$$L = (Xw - y)^T (Xw - y) + \lambda w^T w$$

$X_1 X_2 \cdots X_n$ (y) — m rows

$$X = \begin{bmatrix} 1 & x_{11} & x_{12} & x_{13} & \cdots & x_{1n} \\ 1 & x_{21} & x_{22} & x_{23} & \cdots & x_{2n} \\ 1 & x_{m1} & x_{m2} & x_{m3} & \cdots & x_{mn} \end{bmatrix} \quad \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_m \end{bmatrix} \quad \begin{bmatrix} w_0 \\ w_1 \\ w_2 \\ \vdots \\ w_n \end{bmatrix} \rightarrow (n+1)$$

$w_0, w_1, \ldots w_n$ (parameters)

$$w_0 = w_0 - \eta \frac{\partial L}{\partial w_0}; \quad w_1 = w_1 - \eta \frac{\partial L}{\partial w_1} \quad \ldots \quad w_n = w_n - \eta \frac{\partial L}{\partial w_n}$$

$$w_{new} = w_{old} - \eta \left[\frac{\Delta L}{\Delta w}\right] \rightarrow \text{gradient} \begin{bmatrix} \partial L/\partial w_0 \\ \partial L/\partial w_1 \\ \vdots \\ \partial L/\partial w_n \end{bmatrix}$$

(1) w

$$L = \frac{1}{2}(Xw - y)^T (Xw - y) + \frac{1}{2}\lambda w^T w$$

$$= \frac{1}{2}(w^T X^T - y^T)(Xw - y) + \frac{1}{2}\lambda w^T w$$

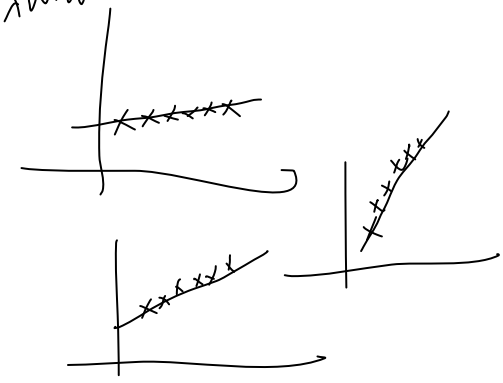
$$= \frac{1}{2}\left[w^T X^T X w - w^T X^T y - y^T w X + y^T y \right] + \frac{1}{2}\lambda w^T w$$

$$= \frac{1}{2}\left[w^T X^T X w - 2 w^T X^T y + y^T y \right] + \frac{1}{2}\lambda w^T w$$

$$\frac{dL}{dw} = \frac{1}{2}\left[2 X^T X w - 2 X^T y \right] + \frac{1}{2} 2\lambda w$$

$$= X^T X w - X^T y + \lambda w = \frac{dL}{dw} \quad \left(\frac{\Delta L}{\Delta w}\right)$$

$w = [w_0, w_1, \ldots w_n]$ starting



$W = [w_0, w_1, \ldots, w_m]$ starting

$W = [0, 1, \ldots, 1]$

epochs

$W = W - \eta \frac{dL}{dW}$

$W \rightarrow$ final answer

epoch times

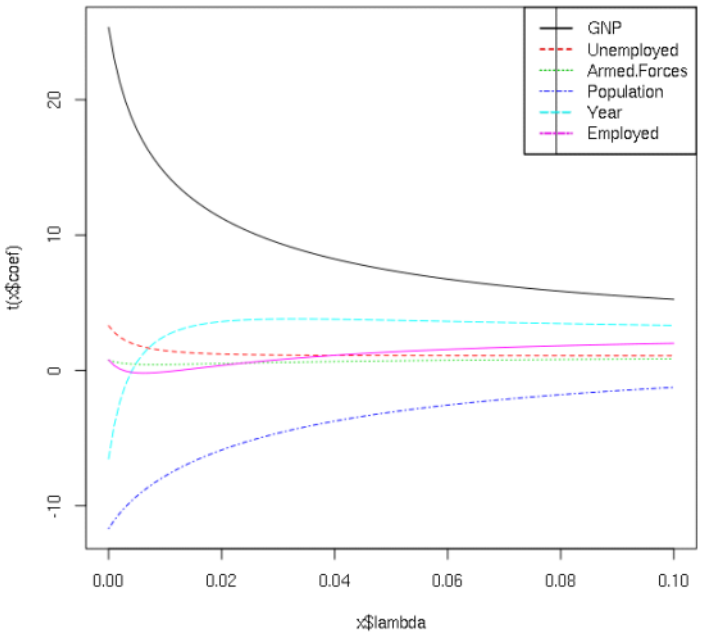
↳ $W = W - \eta \frac{dL}{dW}$

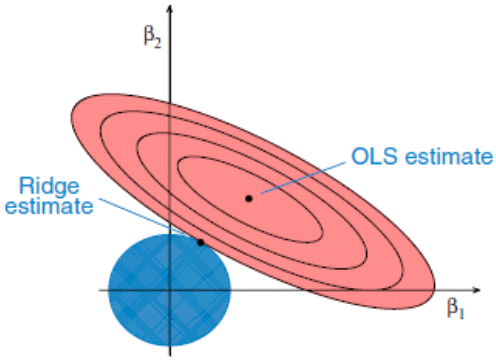
$$\frac{dL}{dW} = X^T X W - X^T y + \lambda W$$

Notes

Friday, June 4, 2021 4:47 PM

Why is it called ridge





5 Key Understandings

Saturday, June 5, 2021 4:20 PM

$$\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 + \boxed{\lambda \|w\|^2}$$

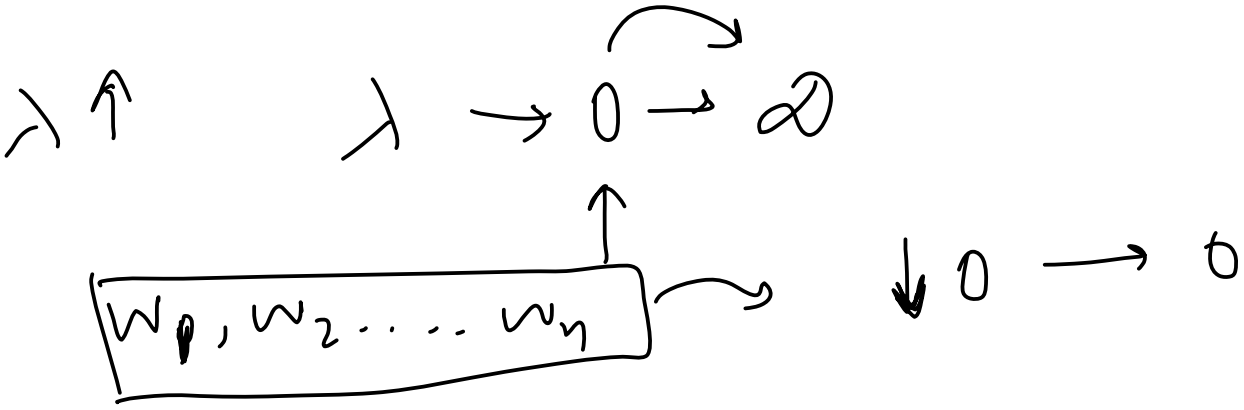
$\lambda(w_1^2 + w_2^2 + \ldots + w_n^2)$

Shrinkage coef

Overfitting ↓

1. How the coefficients get affected?

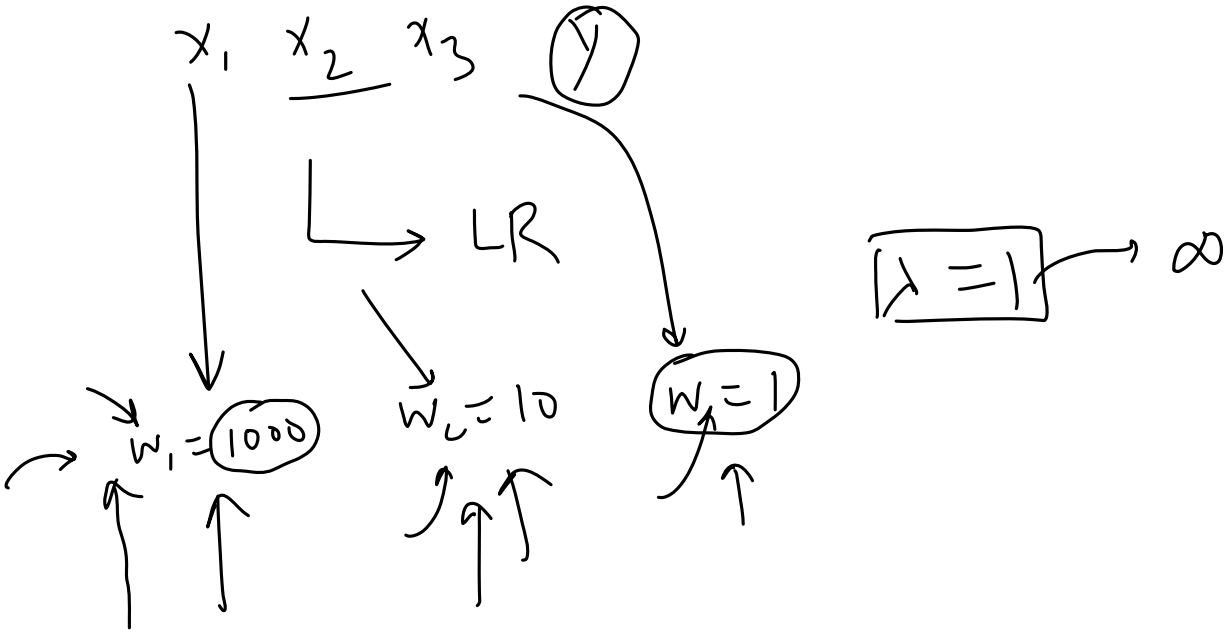
Saturday, June 5, 2021 4:20 PM



2. Higher Values are impacted more

Saturday, June 5, 2021 4:21 PM

Never reaches 0



3. Bias Variance Tradeoff

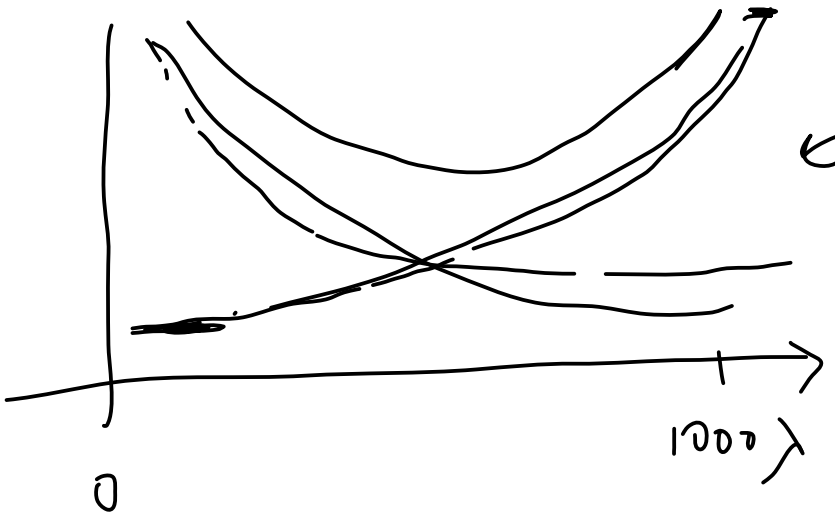
Saturday, June 5, 2021 4:21 PM

Bias Variance

λ ↓ ↑

Bias ↓ overfit Variance ↑

Bias ↑ underfitting Variance ↓



← graph

4. Impact on the Loss Function

Saturday, June 5, 2021 4:21 PM

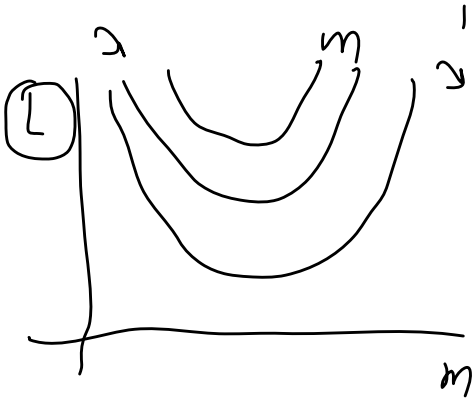
$\lambda \rightarrow$ $L = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 + \lambda \|w\|^2$

X_1 y → [m, b] → b is constant $b = -2.29$

m

$b = 0$

$L = \sum_{i=1}^{n} (y_i - m x_i)^2 + \lambda m^2$

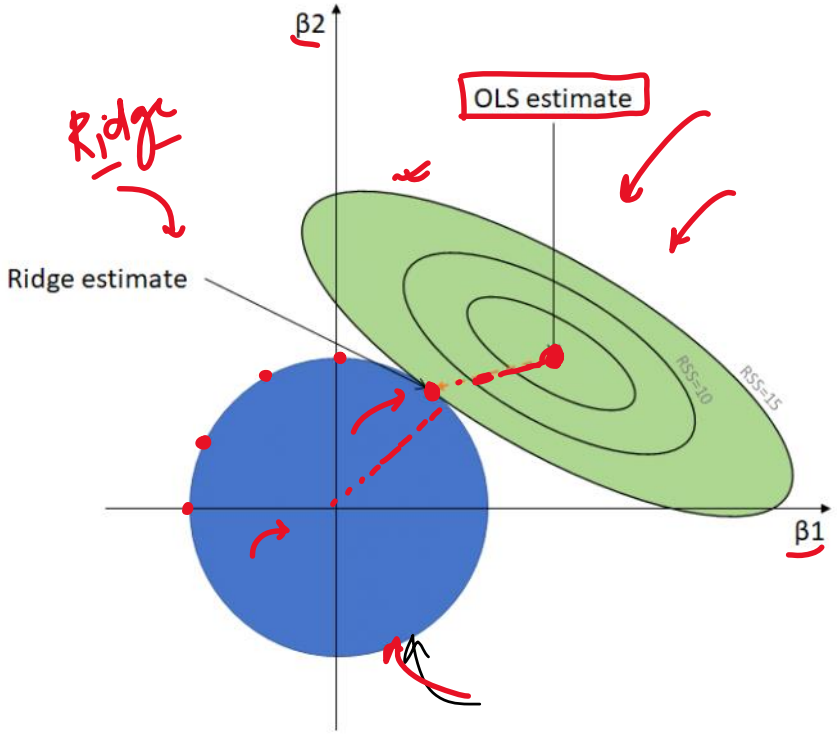


$\lambda = 1$

$\lambda =$

5. Why called Ridge

Saturday, June 5, 2021 4:22 PM



Hard constraint
Ridge constrain

2 coefi β_1 β_2 β_0

$L = MSE + \lambda \|w\|^2$

Contour

$(y_i - \hat{y}_i)^2$

$\sum_{i=1}^{n} (y_i - (\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2}))^2$

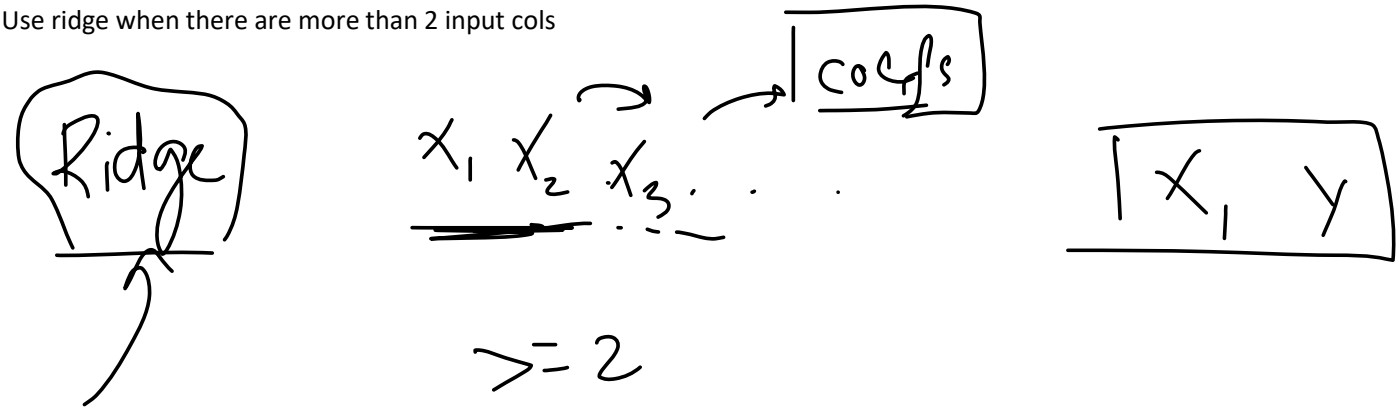
MSE

$\lambda(\beta_1^2 + \beta_2^2)$

Practical Tip

Monday, June 7, 2021 1:20 PM

Use ridge when there are more than 2 input cols



Lasso Regression

Thursday, June 10, 2021 6:42 AM

$y = mx + b$ → $\hat{y} = b$ (underfit)

L1 Regularization | overfitting → L2 reg

$$L = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 + \lambda \|w\|^2$$

$\lambda \geq 0$

$\lambda (w_1^2 + w_2^2 + \dots + w_n^2)$

→ 0 $w_1 \to w_n$ → coeff

overfitting | underfit

alpha

↳ Lasso

L1 norm → $[|w_1| + |w_2| + |w_3| + \dots + |w_n|]$

$$L = MSE + \lambda \|w\|$$

→ code implement (sklearn)

→ key point →

→ difference Ridge Vs Lasso

underfitting

$\lambda = 0 \to$ Linear

overfitting

degree

X_1 X_2 X_3 0 0 X_n (Y)

0.5 (0.2)

ridge

lasso → $\lambda \to \uparrow$

dim ↓

feature selection

0 decrease

lasso

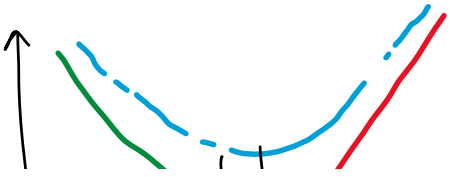
$\lambda \uparrow$

overfitting ↓ Bias ↑

Variance ↓

Variance ↑ Bias ↓

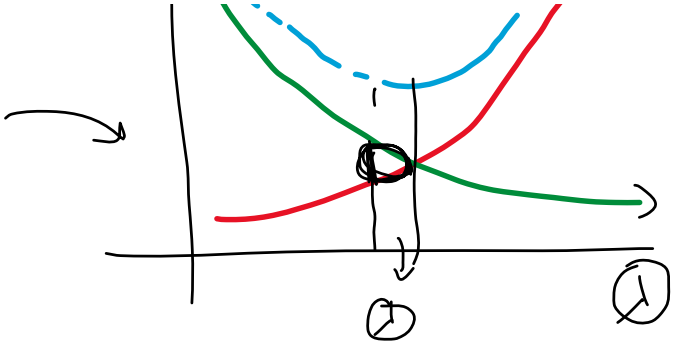
value



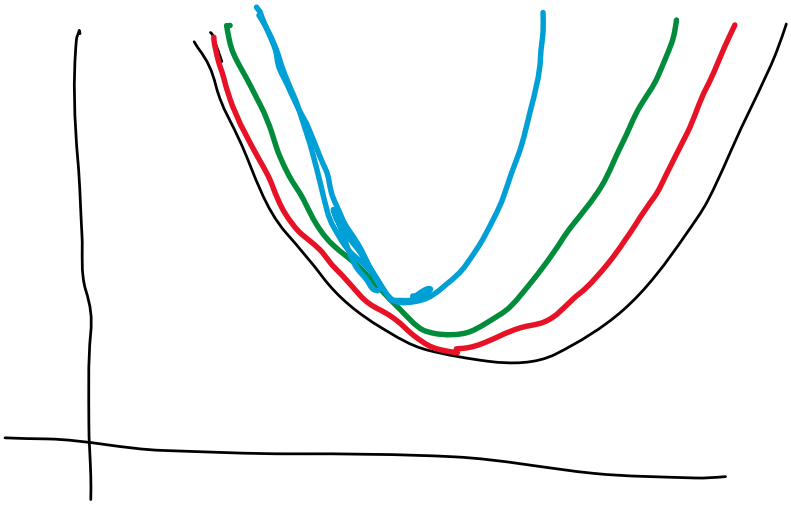
Red → bias

green – variance

blue – total error



blue – total error



Understanding Sparsity

Friday, June 11, 2021 11:21 AM

Lasso
sparsity
$\lambda \uparrow \quad w \to 0$

Ridge

alpha	age	sex	bmi	bp	s1	s2	s3	s4	s5	s6
0.0000	-9.160885	-205.462260	516.684624	340.627341	-895.543609	561.214533	153.884786	126.734316	861.121400	52.419828
0.0001	-9.118336	-205.337133	516.880570	340.556792	-883.415291	551.553259	148.578680	125.355917	856.480254	52.467627
0.0010	-8.763583	-204.321125	518.371729	339.975385	-787.690766	475.274718	106.786540	114.632063	819.739542	52.872100
0.0100	-6.401088	-198.669767	522.048548	336.348363	-383.709187	152.663678	-66.060583	75.611090	659.869402	55.828128
0.1000	6.642753	-172.242166	485.523872	314.682122	-72.939323	-80.590053	-174.466515	83.616653	484.363285	73.584154
1.0000	42.242217	-57.305508	282.170831	198.061386	14.363544	-22.551274	-136.930053	102.023193	260.104308	98.552274
10.0000	21.174004	1.659796	63.659772	48.493240	18.421492	12.875448	-38.915435	38.842464	61.612405	35.505355
100.0000	2.858979	0.629452	7.540604	5.849997	2.710879	2.142134	-4.834047	5.108223	7.448466	4.576129
1000.0000	0.295726	0.069290	0.769004	0.597829	0.282900	0.225936	-0.495607	0.527031	0.761497	0.471029
10000.0000	0.029674	0.006995	0.077054	0.059915	0.028412	0.022715	-0.049686	0.052870	0.076321	0.047241

single $x|y \to$

alpha	age	sex	bmi	bp	s1	s2	s3	s4	s5	s6
0.0000	-9.160885	-205.462260	516.684624	340.627341	-895.543596	561.214523	153.884780	126.734314	861.121395	52.419828
0.0001	-9.071288	-205.337332	516.780313	340.539730	-888.652320	555.952271	150.585260	125.453044	858.639860	52.379002
0.0010	-8.264924	-204.213177	517.641106	339.751339	-826.653342	508.609613	120.899583	113.924518	836.314382	52.011583
0.0100	-1.361404	-192.944226	526.348511	332.649058	-430.205495	191.277876	-44.048113	68.990747	688.384976	47.939528
0.1000	0.000000	-113.976046	526.737112	292.635423	-82.691928	-0.000000	-152.691332	0.000000	551.077200	7.169852
1.0000	0.000000	0.000000	363.882636	27.278420	0.000000	0.000000	-0.000000	0.000000	336.135971	0.000000
10.0000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	-0.000000	0.000000	0.000000	0.000000
100.0000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	-0.000000	0.000000	0.000000	0.000000
1000.0000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	-0.000000	0.000000	0.000000	0.000000
10000.0000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	-0.000000	0.000000	0.000000	0.000000

→ feature selection

$$m = \frac{\sum_{i=1}^{n} (y_i - \bar{y})(x_i - \bar{x})}{\sum_{i=1}^{n} (x_i - \bar{x})^2 + \lambda}$$

simple
$x|y \to y = mx + b$

$b = \bar{y} - m\bar{x}$

$\bar{y} \to \text{mean}(y)$

$\bar{x} \to \text{mean}(x)$

$$m = \frac{\sum_{i=1}^{n} (y_i - \bar{y})(x_i - \bar{x})}{\sum_{i=1}^{n} (x_i - \bar{x})^2}$$

$|m| = m$

$m > 0$

$|m| = -m$

$b = \bar{y} - m\bar{x}$ $\quad m = ?$

$$L = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \pm \lambda |m|$$

$$\frac{d}{dm} \sum_{i=1}^{n} (y_i - mx_i - \bar{y} + m\bar{x})^2 + 2\lambda m$$

$$\frac{dL}{dm} = \sum_{i=1}^{n} (y_i - mx_i - \bar{y} + m\bar{x})^2 + 2\lambda|m| = 2\sum (y_i - mx_i - \bar{y} + m\bar{x})(-x_i + \bar{x}) + 2\lambda = 0$$

$$-2\sum \left[(y_i - \bar{y}) - m(x_i - \bar{x}) \right](x_i - \bar{x}) + 2\lambda = 0$$

$$m\sum (x_i - \bar{x})^2 = \sum (y_i - \bar{y})(x_i - \bar{x}) - \lambda$$

$$m = \sum (y_i - \bar{y})(x_i - \bar{x}) - \lambda$$

$$-\sum [(y_i-\bar{y})(x_i-\bar{x}) - m(x_i-\bar{x})^2] + \lambda = 0$$

$$-\sum (y_i-\bar{y})(x_i-\bar{x}) + m\sum (x_i-\bar{x})^2 + \lambda = 0$$

$$\boxed{m = \frac{\sum (y_i-\bar{y})(x_i-\bar{x}) - \lambda}{\sum (x_i-\bar{x})^2}}$$

Lasso — coeff sparsity

for $m>0$: $\lambda > 0$

$$m = \frac{\sum (y_i-\bar{y})(x_i-\bar{x}) - \lambda}{\sum (x_i-\bar{x})^2}$$

for $m=0$

$$m = \frac{\sum (y_i-\bar{y})(x_i-\bar{x})}{\sum (x_i-\bar{x})^2}$$

for $m<0$

$$m = \frac{\sum (y_i-\bar{y})(x_i-\bar{x}) + \lambda}{\sum (x_i-\bar{x})^2}$$

$m \rightarrow$ (2), $\frac{9}{5}$, 1, (0), -1, .5, 0

$\lambda \uparrow$

$$m = \frac{(YX) - \lambda}{X^2}$$

$YX = 100$, $X^2 = 50$

$\lambda = 100$, $m = 0$

$$m = \frac{YX + \lambda}{X_2} = \frac{100+\lambda}{50}$$

$$= \frac{100+150}{50} =$$

$m = -5$

$$m = \frac{100-\lambda}{50}$$

$\lambda = 0$, $m = 2$ | $\lambda = 10$, $m = \frac{9}{5}$

$\lambda = 50$, $m = 1$

$\lambda > 100$, $m = -1$

150 → 150

$m<0$, $\lambda > 0$

$$m = \frac{\sum (y_i-\bar{y})(x_i-\bar{x}) + \lambda}{\sum (x_i-\bar{x})^2}$$

(+ve, −ve)

$$m = \frac{-100-\lambda}{50}$$

$$= \frac{-100-150}{50} = -5$$

$$m = \frac{-100+\lambda}{50}$$

$\lambda = 0$, $m = -2$

$\lambda = 50$, $m = -1$

$\lambda = 100$, $m = 0 \rightarrow 1$

$\lambda = 150$

$\lambda = 150$, $m = -5$

1) 0 →
2) stop

$m = 1$

$$m = \frac{\sum (y_i-\bar{y})(x_i-\bar{x})}{\sum (x_i-\bar{x})^2 + \lambda}$$

0.1 → 0 X

$\lambda \rightarrow$ Ridge $<<0$ $==0$

Lasso → $\lambda \rightarrow$ numerator

$\lambda = 10000000$ ← Denominator

ElasticNet Regression

Saturday, June 12, 2021 1:04 PM

Ridge $\lambda(w_1^2+w_2^2+..+w_n^2)$ Lasso $\lambda(|w_1|+|w_2|+..+|w_n|)$

$L = \sum_{i=1}^{n}(y_i-\hat{y}_i)^2 + \lambda\|w\|^2$ (L_2)

mse → overfitting

$L = \sum_{i=1}^{n}(y_i-\hat{y}_i)^2 + \lambda\|w\|$ (L_1)

mse → feature selection

$\lambda\uparrow$ $w\downarrow 0$ → 100 cols

$\lambda\uparrow$ $w \to 0$

ENReg

$L = \sum(y_i-\hat{y}_i)^2 + a\|w\|^2 + b\|w\|$

$\{\lambda, l1\_ratio\}$ $[l_1, \lambda]$

$\lambda = a+b$

$l1\_ratio = \frac{a}{a+b}$

$l1 = \frac{a}{\lambda}$

$a = l_1 \times \lambda$

$b = \lambda - a$

$\lambda=1$ $l1\_ratio = 0.5$

$a = 0.5$ $b = 0.5$

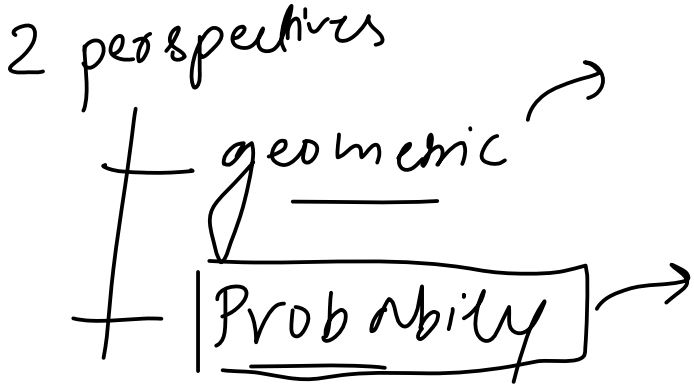
$l1\_ratio > 0.9$ 90% ridge 10% lasso

Input cols → multicollinearity (ElasticNet)

X_1 | X_2

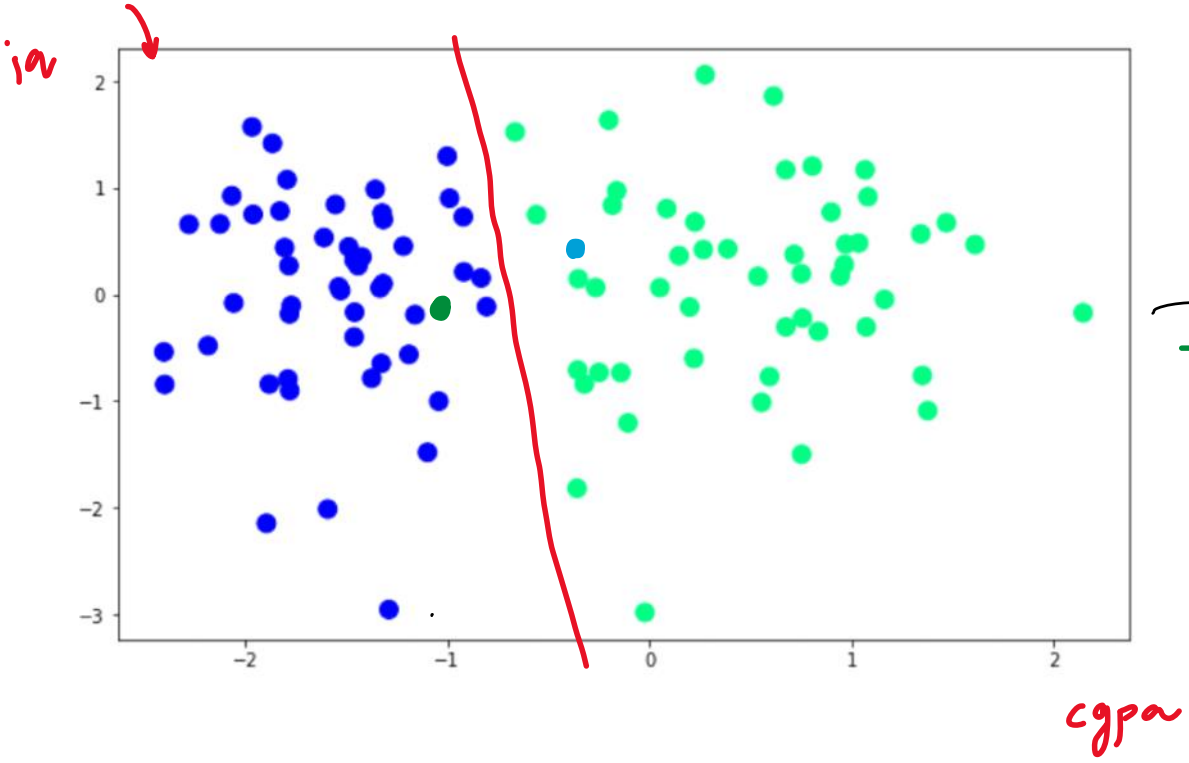
Introduction

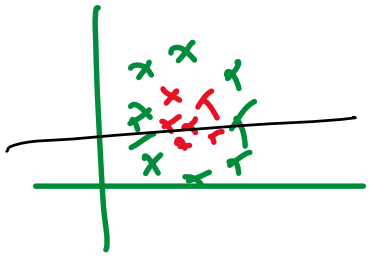
Tuesday, June 15, 2021 12:07 PM



Requirement

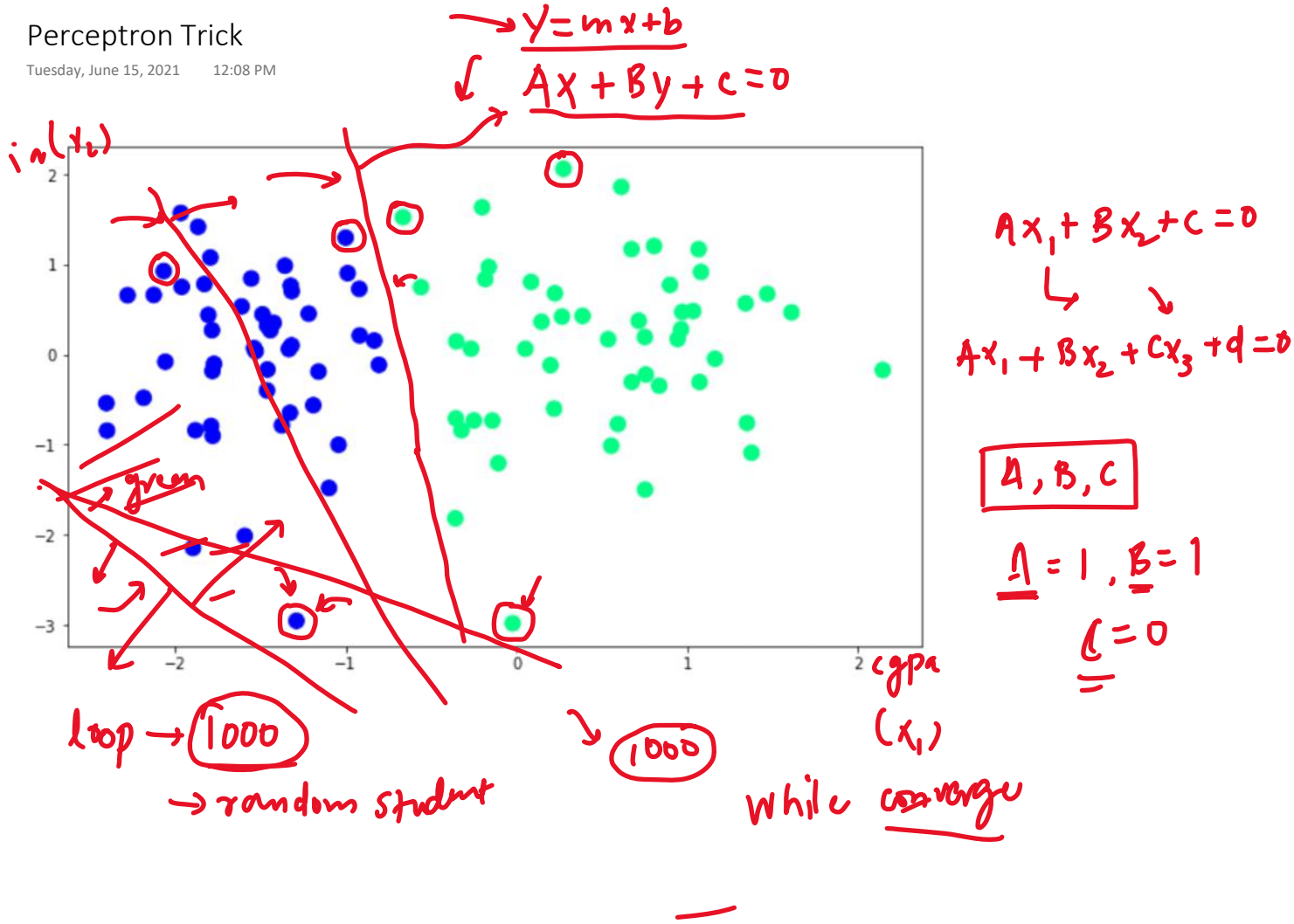
Tuesday, June 15, 2021 12:07 PM





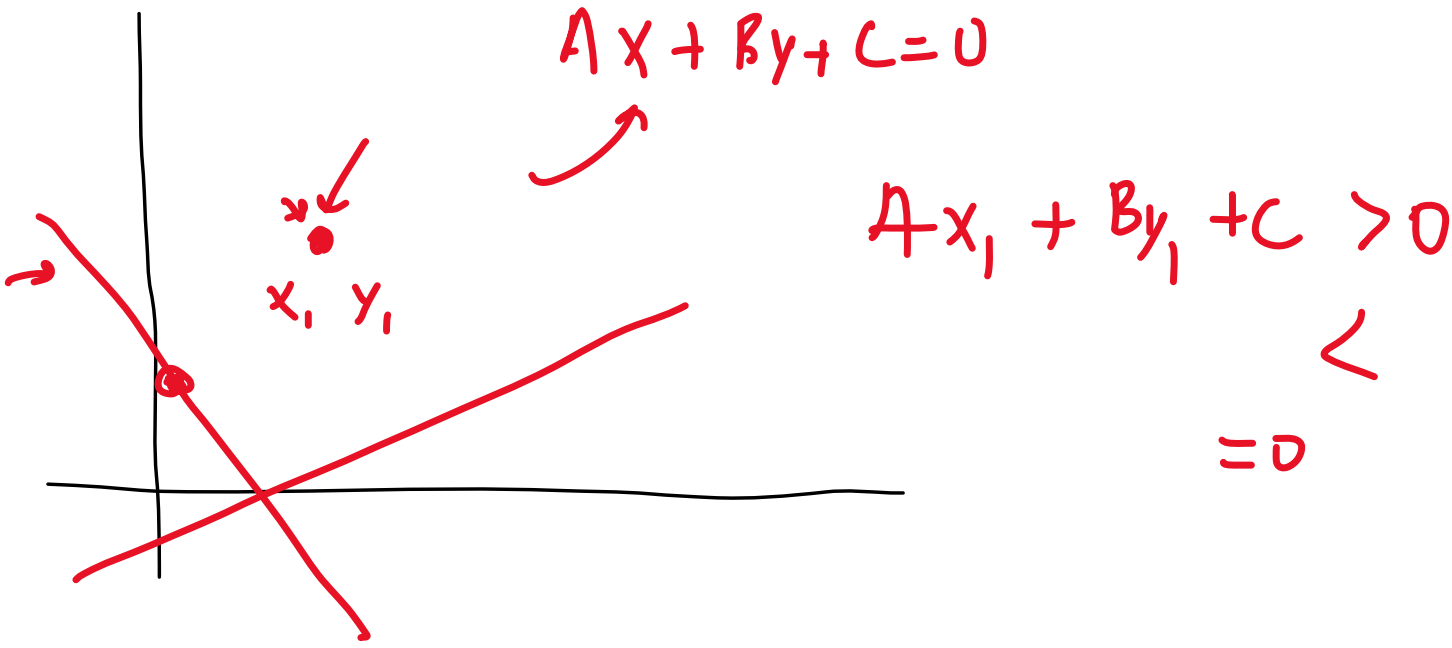
Perceptron Trick

Tuesday, June 15, 2021 12:08 PM



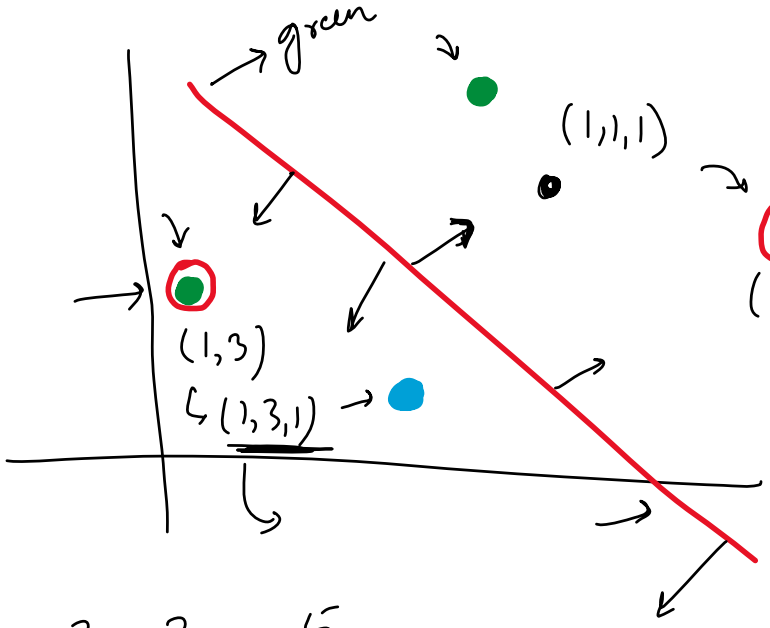
How to label regions?

Tuesday, June 15, 2021 1:12 PM



Transformations

Tuesday, June 15, 2021 1:31 PM



2 3 5

0.01 0.03 0.01

+ − −

3x + 6y + 6 = 0

2 3 5

4 5 1

− − −

−2x − 2y + 4 = 0

learning rate → 0.01

coef = coef − (η coordinate)

Algorithm

Tuesday, June 15, 2021 2:31 PM

x_0	x_1 cgpa	x_2 iq	y placed
1	7.5	61	1
1	8.9	109	1
1	7.0	81	0

$Ax + By + C = 0$

$w_0 + w_1 x_1 + w_2 x_2 = 0$

$\sum_{i=0}^{2} w_i x_i = 0$

$w_0 = C, \; w_1 = A \; w_2 = B$

$w_0 \; w_1 \; w_2$

$w_0 x_0 + w_1 x_1 + w_2 x_2 = 0$

$w_0 x 1 + w_1 x 7.5 + w_2 x 81$

$\sum_{i=0}^{2} w_i x_i = 0$

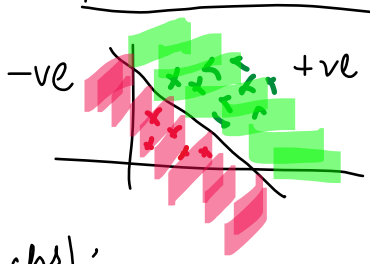
$[w_0 \; w_1 \; w_2] \begin{bmatrix} x_0 \\ x_1 \\ x_2 \end{bmatrix}$

$\geq 0 \rightarrow 1$

$< 0 \rightarrow 0$

$w_0 x_0 + w_1 x_1 + w_2 x_2 \rightarrow 1$

↳ 0



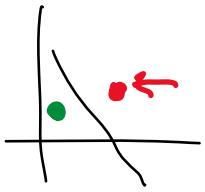
epoch → 1000 ; $\eta = 0.01$

$x_i \in P$

$x_i \in N$

for i in range (epochs):

randomly select a student



$[w_0 \; w_1 \; w_2]$

- if $x_i \in N$ and $\sum_{i=0}^{2} w_i x_i \geq 0$ { (+ve)

  $w_{new} = w_{old} - \eta x_i$

- if $x_i \in P$ and $\sum_{i=0}^{2} w_i x_i < 0$

  $w_{new} = w_{old} + \eta x_i$

$x_i \in N$ $w_i x_i < 0$

$x_i \in P$ $w_i x_i \geq 0$

Simplified Algo

Tuesday, June 15, 2021 2:44 PM

if $x_i \in N$ and $\sum w_i x_i \geq 0$

$$w_n = w_0 - \eta x_i$$

if $x_i \in P$ and $\sum w_i x_i < 0$

$$w_n = w_0 + \eta x_i$$

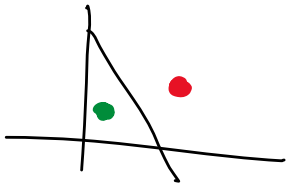
for i in 1000

random student

$$w_n = w_0 + \eta (y_i - \hat{y}_i) x_i$$

$y_i - \hat{y}_i$: 0, $w_n = w_0$; 1, $w_n = w_0 + \eta x_i$; -1, $w_n = w_0 - \eta x_i$

y_i	$\hat{y}_i$	$y_i - \hat{y}_i$
1	1	0
0	0	0
1	0	1
0	1	-1



```
for i in range(epochs):
    select a random student (i)
    W_n = W_0 + η (y_i - ŷ_i) x_i
```

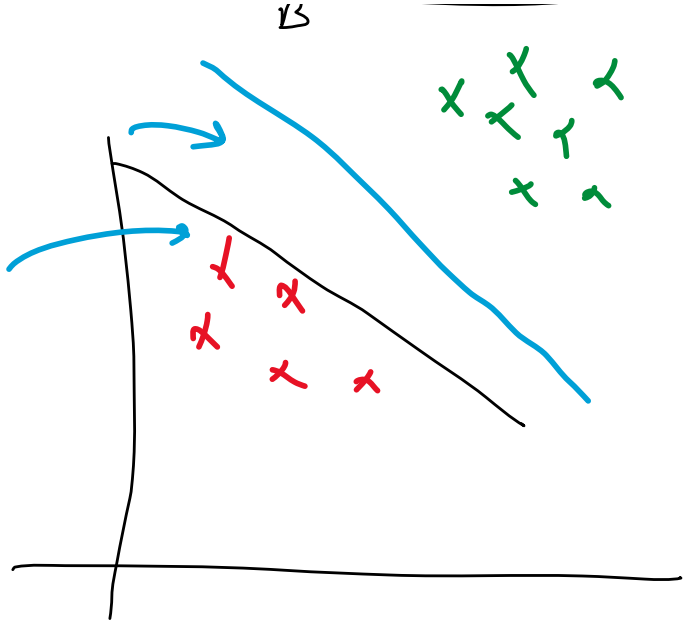
A, B, C

$Ax + By + C = 0$ → $y = mx + b$

$m = -\frac{A}{B}$

$c = -\frac{C}{B}$

def [_ _] [,]

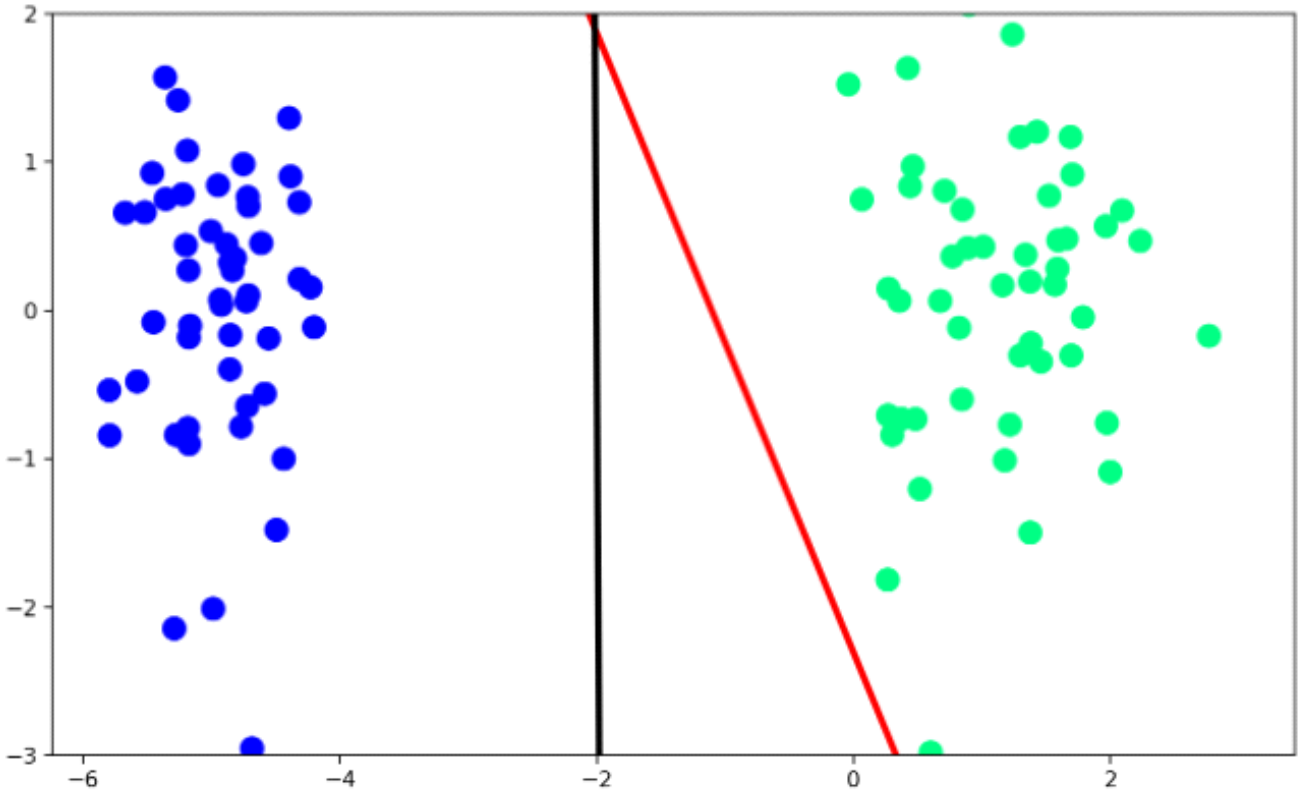


training error

test

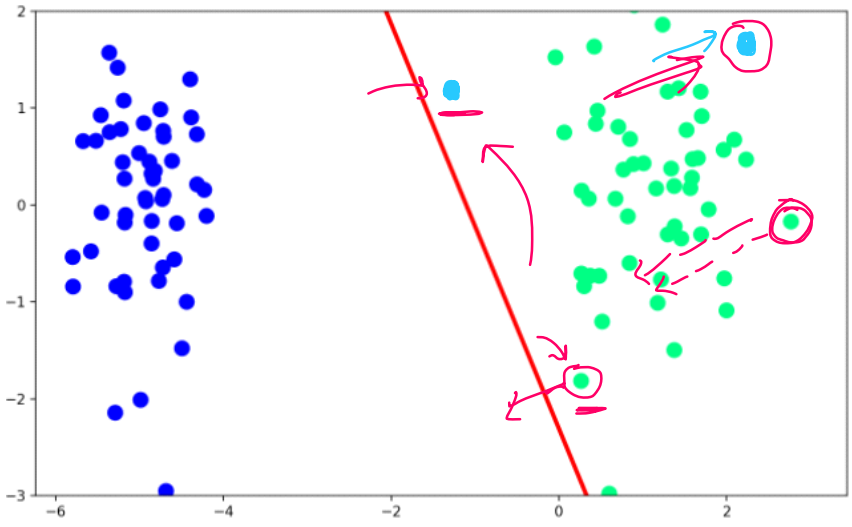
Problem with Perceptron

Thursday, June 17, 2021 12:18 PM



Possible Solution?

Thursday, June 17, 2021 12:18 PM



$$w_{\eta} = w_0 + (y_i - \hat{y}_i) x_i$$

→ misclassified
line – pull

→ correctly
↳ line push
push pull

$$w_{\eta} = w_0 + \eta (y_i - \hat{y}_i) x_i$$

w_0 : $(y_i - \hat{y}_i)$ = 0

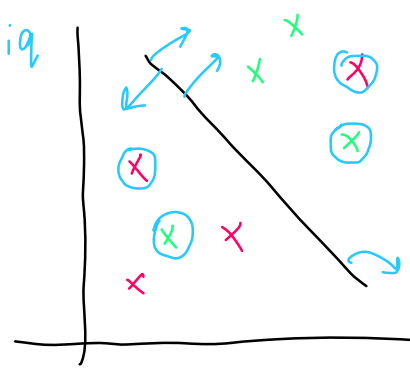
y_i	$\hat{y}_i$	$y_i - \hat{y}_i$
1	1	0
0	0	0
1	0	1
0	1	−1

$(y_i - \hat{y}_i) \neq 0$ model predict

$\Sigma\, w_i x_i = [0, 1]$

$w_1 \times 8.1 + w_2 \times 81 + w_0 =$

≥ 0 (1) < 0 (0) → $\{8.1, 81\}$



$w_1 x_1 + w_2 x_2 + w_0 = 0$

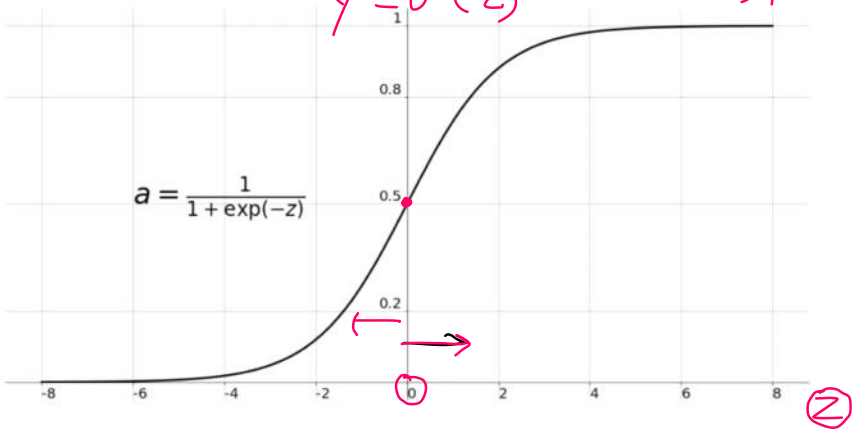
(x_i) placed

cgpa	iq	placed
9	91	0
8.8	78	1
8.1	102	1
7.9	98	1

The Sigmoid Function

Thursday, June 17, 2021 12:19 PM

Sigmoid Function



$$\sigma(z) = \frac{1}{1+e^{-z}}$$

$y = \sigma(z) \rightarrow 1$

$-\infty < z < \infty$

$0 < y < 1$

$\{7.5, 81\} \rightarrow w_1\ w_2\ w_0$

$y_1 = w_1 \times 7.5 + w_2 \times 81 + w_0 = z$

$z \geq 0 \rightarrow 1$

$z < 0 \rightarrow 0$

$\sigma > 0.5 \leftarrow z = \sum w_i x_i = \sigma(z)$

$\sigma(z) < 0.5$

$\{7.5, 81\} \rightarrow z \rightarrow \sigma(z) = 0.5 \geq$

$0.5 <$

if 1 $P(N) = 1 - P(y)$

else 0

$P(y) = 0.7$

placement Y $P(N)$

$\sigma(z)$

$\sum w_i x_i = 0 \quad \sigma = 0.5 \quad P(y) = 0.6$

$\sum w_i x_i > 0 \quad \sigma > 0.5$

$\sum w_i x_i \quad \sigma > 0.6$

$\sigma(\ ) = P(\ )$ $P(y)$ $P(N)$

$w_1\ w_2\ w_0$

$\sigma = 0.5$ $\sigma = 0.5$ $P(0.5) \rightarrow P(y) = 0.5$

$P(y) = 0.1$

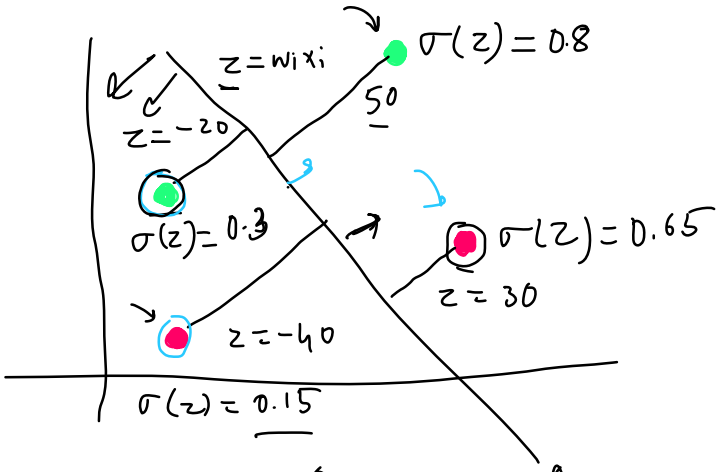
Impact of Sigmoid

Thursday, June 17, 2021 3:27 PM

$$w_\eta = w_0 + \eta (y_i - \hat{y}_i) x_i$$

$$\hat{y}_i = \sigma(z)$$

where $z = \sum w_i x_i$



$$w_\eta = w_0 + \eta \times 0.2 \times x_i$$

$$w_\eta = w_0 - \eta * 0.65 * x_i$$

$$w_\eta = w_0 + \eta * 0.7 * x_i$$

$$w_\eta = w_0 - \eta * 0.15 * x_i$$

y_i	$\hat{y}_i$	$y_i - \hat{y}_i$
1	0.8	0.2
0	0.65	-0.65
1	0.3	0.7
0	0.15	-0.15

$$w_\eta = w_0 + \eta (y_i - \hat{y}_i) x_i$$

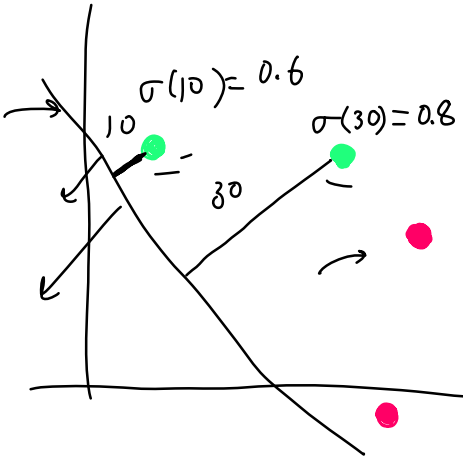
$$\hat{y}_i = \sigma(z)$$

where $z = \sum w_i x_i$

1st

$$w_\eta = w_0 + \boxed{\eta * 0.4 * x_i} = x_1$$

$$w_\eta = w_0 + \boxed{\eta * 0.2 * x_i} = x_2$$

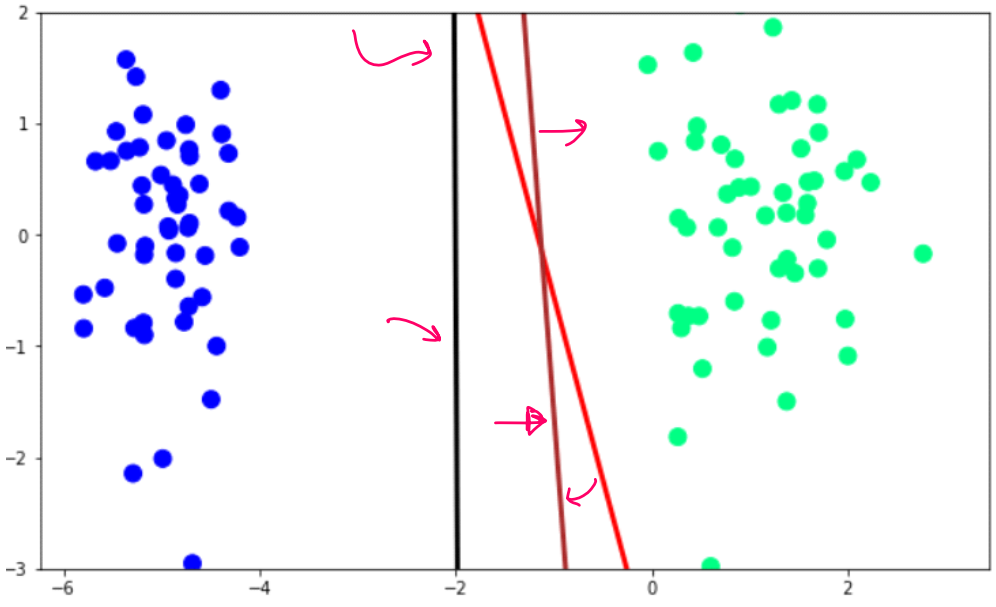


y_i	$\hat{y}_i$	$y_i - \hat{y}_i$
1	0.6	0.4
1	0.8	0.2
0		
0		

$x_1 > x_2$ → code implement

The Problem

Friday, June 18, 2021 2:20 PM



red → step

brown → sigmoid

black → sklearn

for i in 1000

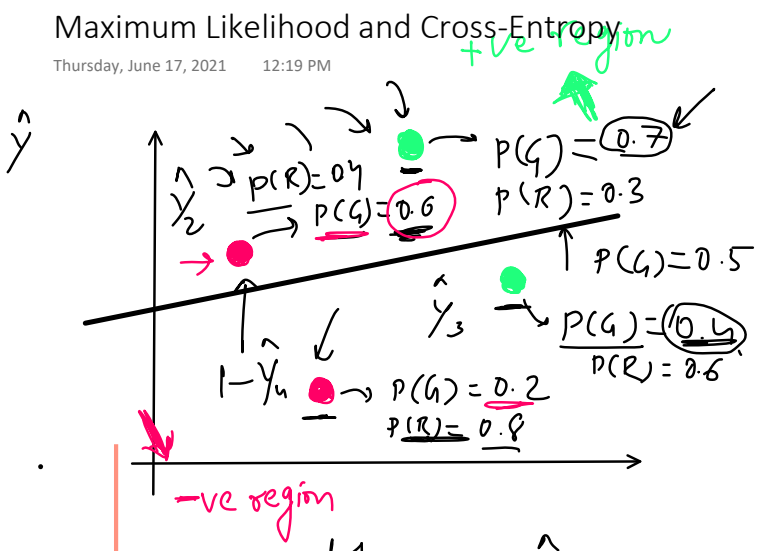
random

Loss function

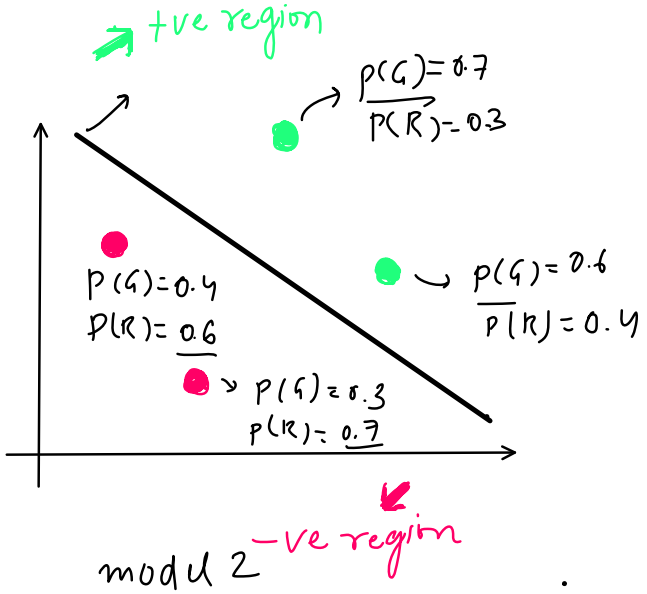
error function

Maximum Likelihood and Cross-Entropy

Thursday, June 17, 2021 12:19 PM



model 1 $\hat{y}_i$



model 2

$\hat{y} = \sigma(z)$

$z = \sum w_i x_i$

model 1 $\rightarrow$ $0.7 \times 0.4 \times 0.4 \times 0.8 = 0.089$

model 2 $\rightarrow$ $0.7 \times 0.6 \times 0.6 \times 0.7$

$= 0.176$

$\log(ab) = \log a + \log b$

$\log(\max) = -\log(0.7) - \log(0.4) - \log(0.4) - \log(0.8)$

0-1 = -ve

cross entropy minimize

$\log(0.1) > \log(0.09)$

$\downarrow \quad \downarrow$

$1 \quad 0.04$

NO

$-\log(\hat{y}_1) - \log(\hat{y}_2) - \log(\hat{y}_3) - \log(\hat{y}_4)$

$(1-\hat{y})$

$y_i = 1$

$-y_i \log(\hat{y}_i) - (1-y_i)\log(1-\hat{y}_i)$

$-y_i \log(\hat{y}_i) = -\log(\hat{y}_1)$

$= -\log(0.7)$

$y_2 = 0 \quad y_3 = 1$

$= -\log(0.7)$

$y_2 = 0 \quad y_3 = 1$

$-y_2 \log(\hat{y}_2) - (1-y_2)\log(1-\hat{y}_2)$

$-\log(1-\hat{y}_2) = -\log 0.4$

$-y_3 \log(\hat{y}_3)$

$-\log(\hat{y}_3) = -\log(0.9)$

$y_4 = 0$

$-(1-y_4)\log(1-\hat{y}_4)$

$-\log(1-\hat{y}_4)$

$-\log(0.8)$

$$L = \sum_{i=1}^{n} -y_i \log(\hat{y}_i) - (1-y_i)\log(1-\hat{y}_i)$$

MSE

$$L = -\frac{1}{n}\sum_{i=1}^{n} y_i \log(\hat{y}_i) + (1-y_i)\log(1-\hat{y}_i)$$

closed form

gradient descent

log-loss error

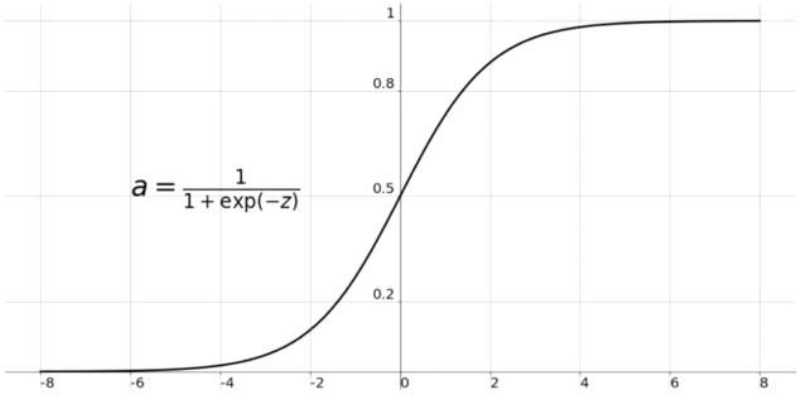
Binary cross entropy

$\min_{w_1, w_2, w_0}$

Derivative of Sigmoid

Thursday, June 17, 2021 12:20 PM

Sigmoid Function



$$\sigma(x) = \frac{1}{1+e^{-x}}$$

$$\sigma'(x) = \frac{d}{dx}\left(\frac{1}{1+e^{-x}}\right)$$

$$\frac{d}{dx}\left(\frac{1}{x}\right) = \frac{d}{dx}(x)^{-1}$$

$$= -x^{-2} = -\frac{1}{x^2}$$

$$\frac{d}{dx}\left[\frac{1}{1+e^{-x}}\right] = \frac{d}{dx}\left[(1+e^{-x})^{-1}\right] = -\frac{1}{(1+e^{-x})^2}\frac{d}{dx}(1+e^{-x})$$

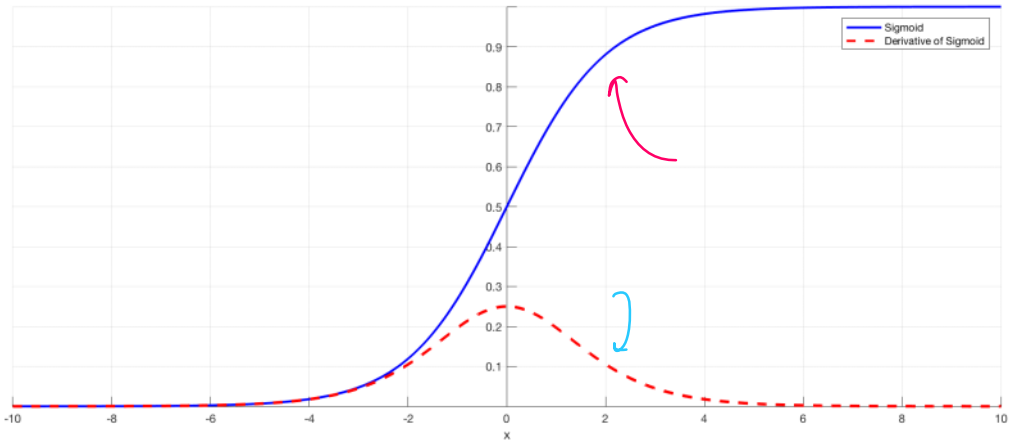
$e^{-x} = e^{-x}$, $-x = -1$

$$= -\frac{1}{(1+e^{-x})^2}\frac{d}{dx}(e^{-x}) = -\frac{e^{-x}}{(1+e^{-x})^2}\frac{d}{dx}(-x) = \frac{e^{-x}}{(1+e^{-x})^2}$$

$$\frac{1\cdot e^{-x}}{(1+e^{-x})(1+e^{-x})} = \frac{1}{1+e^{-x}}\cdot\frac{e^{-x}}{1+e^{-x}} = \sigma(x)\left[\frac{e^{-x}}{1+e^{-x}}\right]$$

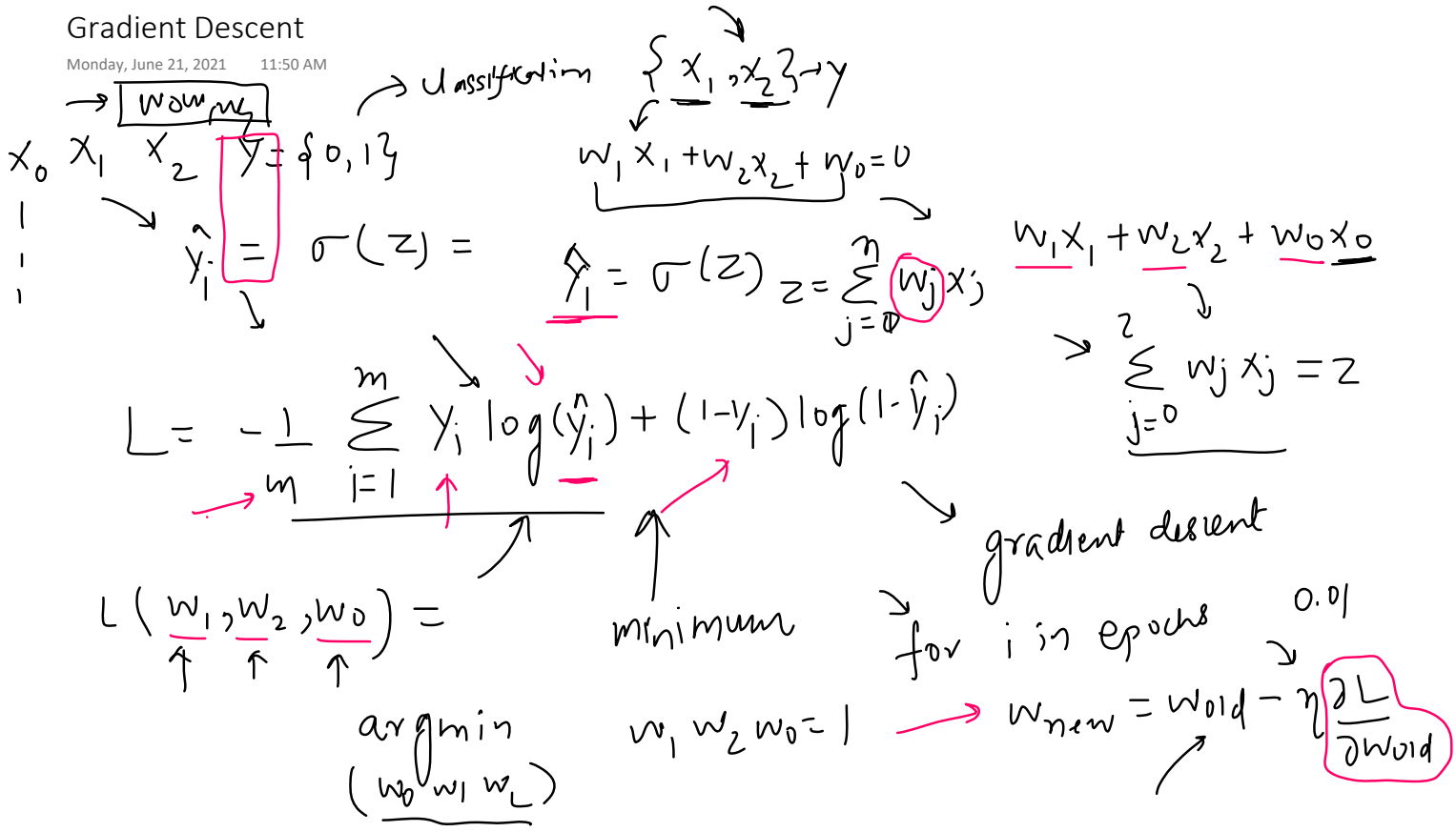
$$= \sigma(x)\left[\frac{1+e^{-x}-1}{1+e^{-x}}\right] = \sigma(x)\left[\frac{1+e^{-x}}{1+e^{-x}} - \frac{1}{1+e^{-x}}\right]$$

$$\sigma(x)\left[1-\sigma(x)\right] \Rightarrow \sigma'(x) = \sigma(x)\left[1-\sigma(x)\right]$$


Sigmoid
Derivative of Sigmoid
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
-10
-8
-6
-4
-2
0
2
4
6
8
10
x

Gradient Descent

Monday, June 21, 2021 11:50 AM



$$w_0 = w_0 - \eta \frac{\partial L}{\partial w_0}, \quad w_1 = w_1 - \eta \frac{\partial L}{\partial w_1}, \quad w_2 = w_2 - \eta \frac{\partial L}{\partial w_2}$$

n cols → n+1 derivatives

$$w_n = w_n - \eta \frac{\partial L}{\partial w_n} \qquad j = 0, 1, 2, \ldots n$$

$$L = -\frac{1}{m} \sum_{i=1}^{m} y_i \log(\hat{y}_i) + (1-y_i) \log(1-\hat{y}_i)$$

$$\frac{\partial L}{\partial w_j} = -\frac{1}{m} \sum_{i=1}^{m} \left[\frac{\partial L}{\partial w_j} \left(y_i \log(\hat{y}_i) + (1-y_i) \log(1-\hat{y}_i) \right) \right]$$

$\hat{y}_i = \sigma(z) \quad z = \sum_{j=0}^{n} w_j x_j$

$$\frac{\partial L}{\partial w_j} y_i \log(\hat{y}_i) = \frac{\partial L}{\partial w_j} y_i \log(\sigma(z))$$

$$- y_i \quad \frac{\partial L}{\partial} \sigma\left(\sum^{n} w_j x_j\right)$$

$$\frac{\partial L}{\partial w_j} = \frac{y_i}{\hat{y}_i} \frac{\partial L}{\partial w_j} \sigma\left(\sum_{j=0}^{n} w_j x_j\right)$$

$$\frac{\partial L}{\partial w_j} \rightarrow y_i \log\left(\sigma\left(\sum_{j=0}^{n} w_j x_j\right)\right) \qquad \hat{y}_i$$

$$= y_i \frac{\partial L}{\partial w_j} \log\left(\sigma\left(\sum_{j=0}^{n} w_j x_j\right)\right) = \frac{y_i}{\hat{y}_i} \hat{y}_i(1-\hat{y}_i) \frac{\partial L}{\partial w_j} \sum_{j=0}^{n} w_j x_j$$

$$y_i(1-\hat{y}_i) \sum_{j=0}^{n} \frac{\partial L}{\partial w_j} w_j x_j = \boxed{y_i(1-\hat{y}_i) \sum_{j=0}^{n} x_j}$$

$$\frac{\partial L}{\partial w_j} (1-y_i) \log(1-\hat{y}_i) \Rightarrow (1-y_i) \frac{\partial L}{\partial w_j} \log(1-\hat{y}_i) \qquad \hat{y}_i = \sigma(z) \quad \sigma\left(\sum_{j=0}^{n} w_j x_j\right)$$

$$\frac{(1-y_i)}{(1-\hat{y}_i)} \frac{\partial L}{\partial w_j} (1-\hat{y}_i) \Rightarrow -\frac{(1-y_i)}{(1-\hat{y}_i)} \frac{\partial L}{\partial w_j} \hat{y}_i \Rightarrow -\frac{(1-y_i)}{(1-\hat{y}_i)} \frac{\partial L}{\partial w_j} \sigma(z)$$

$$\Rightarrow -\frac{(1-y_i)}{(1-\hat{y}_i)} \sigma(z)(1-\sigma(z)) \frac{\partial L}{\partial w_j} \sigma(z) = -\frac{(1-y_i)}{(1-\hat{y}_i)} \hat{y}_i(1-\hat{y}_i) \frac{\partial L}{\partial w_j} \sigma(z)$$

$$\Rightarrow -\hat{y}_i(1-y_i) \frac{\partial L}{\partial w_j} \sum_{j=0}^{n} w_j x_j \Rightarrow -\hat{y}_i(1-y_i) \sum_{j=0}^{n} \frac{\partial L}{\partial w_j} (w_j x_j)$$

$$\Rightarrow \boxed{-\hat{y}_i(1-y_i) \sum_{j=0}^{n} x_j}$$

$$\frac{\partial L}{\partial w_j} = -\frac{1}{m} \sum_{i=1}^{m} \left[y_i(1-\hat{y}_i) \sum_{j=0}^{n} x_j - \hat{y}_i(1-y_i) \sum_{j=0}^{n} x_j \right]$$

$$\frac{\partial L}{\partial w_j} = \frac{1}{m} \sum_{i=1}^{m} \left[\quad \right] \sum_{j=0}^{n} x_j$$

$$\frac{\partial L}{\partial w_j} = -\frac{1}{m}\sum_{i=1}^{m}\left[y_i(1-\hat{y}_i) - \hat{y}_i(1-y_i) \right]\sum_{j=0}^{n} x_j$$

$$= -\frac{1}{m}\sum_{j=1}^{m}\left[y_i - y_i\hat{y}_i - \hat{y}_i + y_i\hat{y}_i \right]\sum_{j=0}^{n} x_j$$

$$\frac{\partial L}{\partial w_j} = -\frac{1}{m}\sum_{i=1}^{m}(y_i - \hat{y}_i)\sum_{j=0}^{n} x_j$$

x_1	x_2	y_i	$\hat{y}_i$
1	2	1	0
3	4	0	0

no. of rows $= m = 2$

no. of cols $= n = 2$

$$\frac{\partial L}{\partial w_0} = -\frac{1}{m}\left[1+0\right]\left[1+1\right]$$

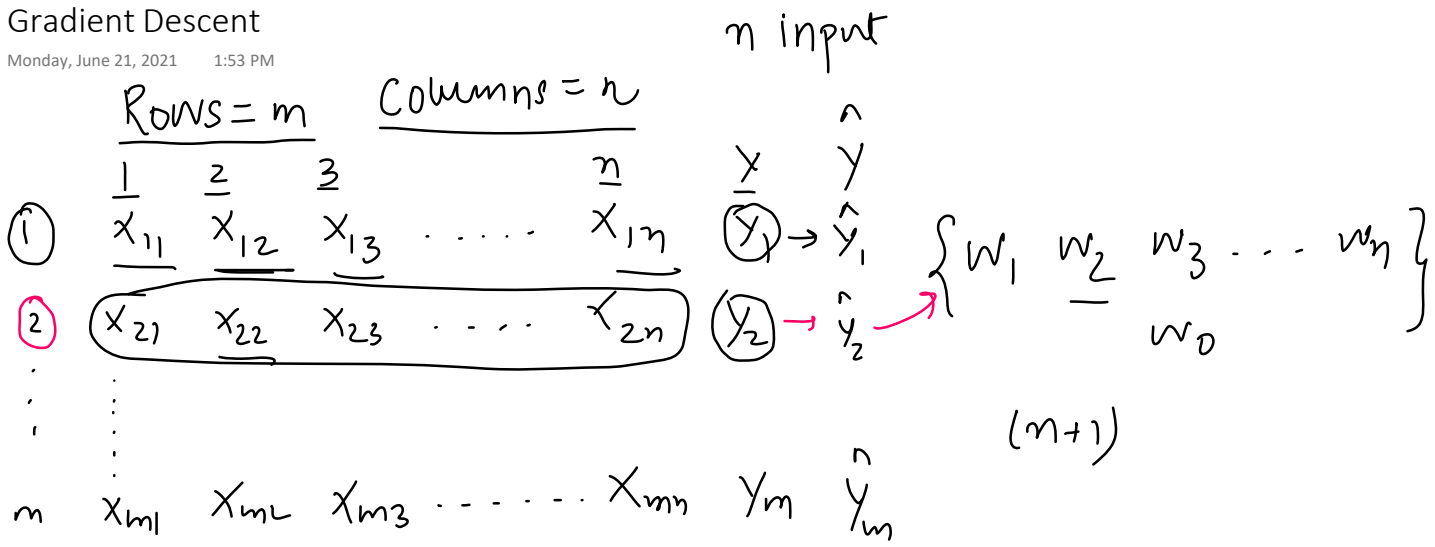
$$= -\frac{1}{2}\left[1\right]\left[2\right] = -1$$

$$= -\frac{1}{2}\begin{bmatrix}1 & 0\end{bmatrix}\begin{bmatrix}1\\1\end{bmatrix}$$

$$= -\frac{1}{2}\left[1\right] = -\frac{1}{2}$$

Gradient Descent

Monday, June 21, 2021 1:53 PM



$$\sigma(w_1 x_{11} + w_2 x_{12} + w_3 x_{13} + \ldots + w_n x_{1n} + w_0) = \hat{y}_1$$

$$\sigma(w_1 x_{21} + w_2 x_{22} + w_3 x_{23} + \ldots + w_n x_{2n} + w_0) = \hat{y}_2$$

$$\hat{Y} = \begin{bmatrix} \hat{y}_1 \\ \hat{y}_2 \\ \vdots \\ \hat{y}_m \end{bmatrix} = \begin{bmatrix} \sigma(w_0 + w_1 x_{11} + w_2 x_{12} + \ldots + w_n x_{1n}) \\ \sigma(w_0 + w_1 x_{21} + w_2 x_{22} + \ldots + w_n x_{2n}) \\ \vdots \\ \sigma(w_0 + w_1 x_{m1} + w_2 x_{m2} + \ldots + w_n x_{mn}) \end{bmatrix}$$

$$\hat{Y} = \sigma\left(\begin{bmatrix} w_0 + w_1 x_{11} + w_2 x_{12} + \ldots + w_n x_{1n} \\ w_0 + w_1 x_{21} + w_2 x_{22} + \ldots + w_n x_{2n} \\ \vdots \\ w_0 + w_1 x_{m1} + w_2 x_{m2} + \ldots + w_n x_{mn} \end{bmatrix} \right)$$

$$\hat{Y} = \sigma\left(\begin{bmatrix} 1 & x_{11} & x_{12} & \ldots & x_{1n} \\ 1 & x_{21} & x_{22} & \ldots & x_{2n} \\ \vdots & & & & \end{bmatrix} \begin{bmatrix} w_0 \\ w_1 \\ \vdots \end{bmatrix} \right)$$

(A: the column of 1s multiplies w_0; X: the data matrix; B: the weight vector W)

$$\hat{y} = \sigma\left(\begin{bmatrix} 1 \\ \vdots \\ 1 \end{bmatrix} \begin{bmatrix} x_{21} & x_{22} & \cdots & x_{2n} \\ & & & \\ x_{m1} & x_{m2} & \cdots & x_{mn} \end{bmatrix} \begin{bmatrix} \vdots \\ w_n \end{bmatrix}\right)$$

$$\hat{y} = \sigma(XW)$$

$$L = -\frac{1}{m}\sum_{i=1}^{m} y_i \log(\hat{y}_i) + (1-y_i)\log(1-\hat{y}_i)$$

$$L = -\frac{1}{m}\left[\sum_{i=1}^{m} y_i \log(\hat{y}_i) + \sum_{i=1}^{m} (1-y_i)\log(1-\hat{y}_i)\right]$$

$(1-y)\log \sigma(1-XW)$

$$\sum_{i=1}^{m} y_i \log(\hat{y}_i) = y_1 \log \hat{y}_1 + y_2 \log \hat{y}_2 + y_3 \log \hat{y}_3 + \cdots + y_m \log \hat{y}_m$$

$$\begin{bmatrix} y_1 & y_2 & y_3 & \cdots & y_m \end{bmatrix} \begin{bmatrix} \log \hat{y}_1 \\ \log \hat{y}_2 \\ \vdots \\ \log \hat{y}_m \end{bmatrix}$$

$$\begin{bmatrix} y_1 & y_2 & y_3 & \cdots & y_m \end{bmatrix} \log\left(\begin{bmatrix} \hat{y}_1 \\ \hat{y}_2 \\ \vdots \\ \hat{y}_m \end{bmatrix}\right)$$

$$y \log \hat{y} = y \log(\sigma(XW))$$

$$L = -\frac{1}{m}\left[y \log \hat{y} + (1-y)\log(1-\hat{y})\right] \quad \text{(min)}$$

where $\hat{y} = \sigma(XW)$ ← GD [W] find

where $\hat{y} = \sigma(XW)$ — GD

Loss function in Matrix form

$$L = -\frac{1}{m}\left[y \log(\sigma(wx)) + (1-y)\log(1-\sigma(wx)) \right]$$

minimum

gradient descent

$W = [\quad]$ $w_0 \rightarrow w_n$

for i in epochs:

$$W = W - \eta \frac{\Delta L}{\Delta W}$$

$$\frac{\Delta L}{\Delta W} = \left[\frac{\partial L}{\partial w_0}, \frac{\partial L}{\partial w_1}, \frac{\partial L}{\partial w_2}, \cdots \frac{\partial L}{\partial w_n} \right] \quad (n+1)$$

$\frac{\Delta L}{\Delta W}$

$$L = -\frac{1}{m}\left[y \log \hat{y} + (1-y)\log(1-\hat{y}) \right]$$

$$\frac{dL}{dw} =$$

$$\frac{d}{dw} y \log \hat{y} \Rightarrow y \frac{d}{dw} \log \hat{y} \Rightarrow \frac{y}{\hat{y}} \frac{d}{dL}(\hat{y})$$

$$\Rightarrow \frac{y}{\hat{y}} \frac{d}{dL} \sigma(wx) \Rightarrow \frac{y}{\hat{y}} \sigma(wx)[1-\sigma(wx)] \frac{d}{dw}(wx)$$

$\Rightarrow \hat{y}$

$$= \frac{y}{\hat{y}} \hat{y}(1-\hat{y}) x = \boxed{y(1-\hat{y})x} \qquad \hat{y} = \sigma(wx)$$

$= \frac{1}{\hat{y}}$

$\frac{d}{dw}(1-y)\log(1-y) \Rightarrow (1-y)\frac{d}{dw}\log(1-\hat{y}) \Rightarrow \frac{(1-y)}{(1-\hat{y})}\frac{d}{dw}[1-\hat{y}]$

$= -\frac{(1-y)}{(1-\hat{y})}\frac{d}{dw}\sigma(wx) \Rightarrow -\frac{(1-y)}{(1-\hat{y})}[\sigma(wx)[1-\sigma(wx)]\frac{d}{dL}(wx)$

$\Rightarrow -\frac{(1-y)}{(1-\hat{y})}\hat{y}(1-\hat{y})x = \boxed{-\hat{y}(1-y)x}$

$\frac{dL}{dw} = -\frac{1}{m}[y(1-\hat{y})x - \hat{y}(1-y)x]$

$= -\frac{1}{m}[y(1-\hat{y}) - \hat{y}(1-y)]x$

$= -\frac{1}{m}[y - y\hat{y} - \hat{y} + y\hat{y}]x$

$\boxed{\frac{\Delta L}{\Delta w} = -\frac{1}{m}(y-\hat{y})x}$

gd

$\boxed{w = w + \eta\frac{1}{m}(y-\hat{y})x}$

$$\underline{W} = \underline{W} + \eta \frac{1}{m} (y - \hat{y})^T$$

$(n+1), 1$

$$W = \begin{bmatrix} w_0 \\ w_1 \\ \vdots \\ w_n \end{bmatrix} \quad X = \begin{bmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1n} \\ 1 & x_{21} & x_{22} & \cdots & x_{2n} \\ \vdots & & & & \\ 1 & x_{m1} & x_{m2} & \cdots & x_{mn} \end{bmatrix} \quad y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_m \end{bmatrix} \rightarrow m, 1$$

$m, n+1$

$$\hat{y} = \begin{bmatrix} \hat{y}_1 \\ \hat{y}_2 \\ \vdots \\ \hat{y}_m \end{bmatrix} \quad (m, 1)$$

$$\underline{W} = \underline{W} + \boxed{\frac{\eta}{m}} (y - \hat{y}) X$$

$(n+1, 1) = (n+1, 1) \quad (1, m)\ m, (n+1) \rightarrow (1, n+1)$

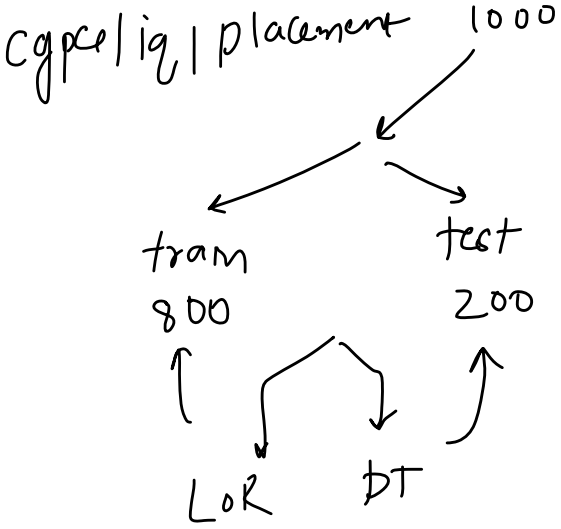
$(n+1), 1$

Accuracy

Tuesday, June 22, 2021 1:12 PM

binary

Actual Label	Logistic Regression Prediction	Decision Tree Prediction
1	1	1
0	0	0
0	0	0
0	0	0
1	1	1
1	1	1
0	1	0
0	0	0
0	0	0
1	1	1



Accuracy = no. of ✓ / total predictions

$\frac{9}{10} = 90\%$

$\frac{8}{10} = 0.8 \rightarrow 80\%$

Accuracy of multi-classification problem

Wednesday, June 23, 2021 8:44 AM

Actual Label	Logistic Regression Prediction	Decision Tree Prediction
0	0	0
0	0	0
0	0	0
2	2	2
0	0	0
2	2	2
0	0	0
2	2	2
1	1	1
1	1	1

iris

setosa, virginica / versicolor

0 1 2

accuracy = # correct prediction / # total

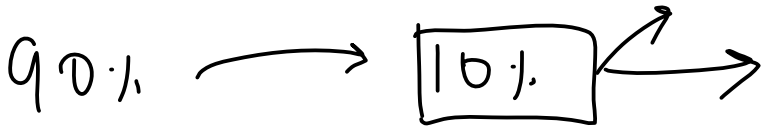
= 10/10 = 1 = 100%

How much accuracy is good?

Wednesday, June 23, 2021 11:14 AM

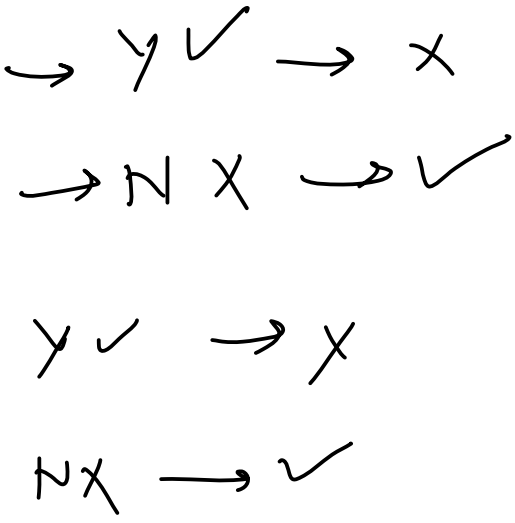
The Problem with Accuracy

Wednesday, June 23, 2021 11:14 AM



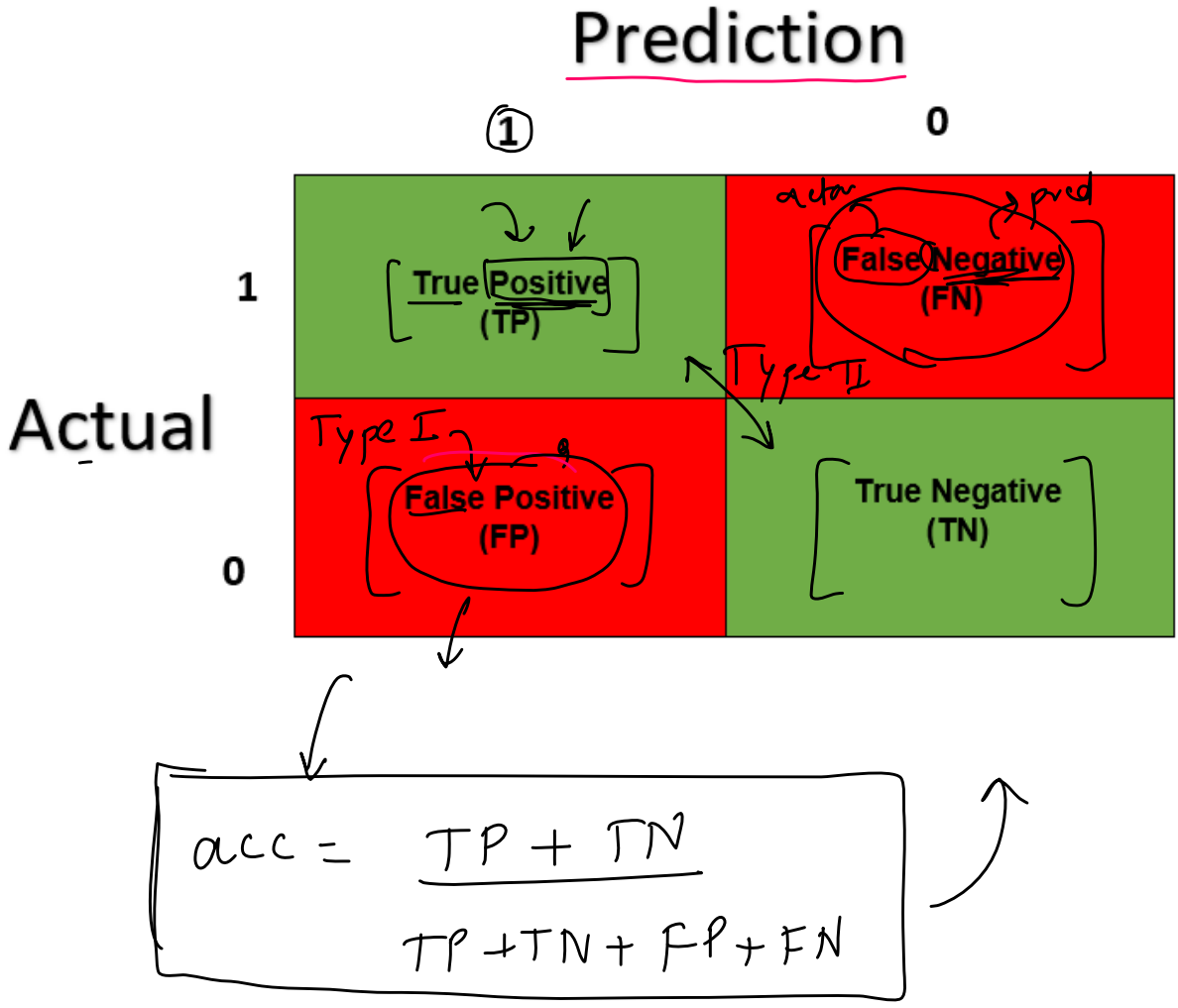
binary

Y | N



Confusion Matrix

Tuesday, June 22, 2021 1:12 PM



Extented

↳ { echo or not }

duplicate

↳ echo

1+2+3=6 → outcome

x y
z y

Type 1 Error

Wednesday, June 23, 2021 8:45 AM

Type 2 Error

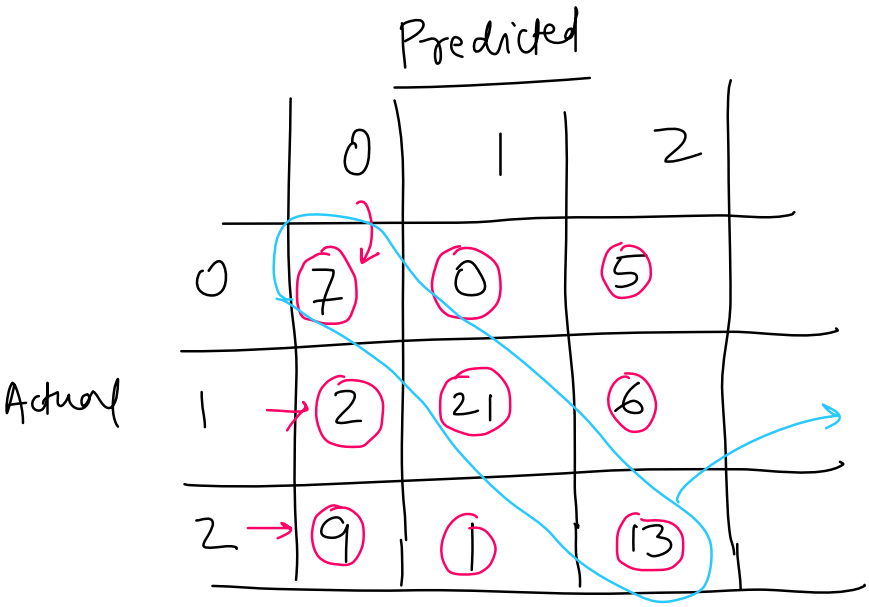
Wednesday, June 23, 2021 8:45 AM

Confusion Matrix for Multi-classification Problem

Wednesday, June 23, 2021 8:45 AM

0, 1, 2 →

10 x 10



3 x 3

binary

2 x 2

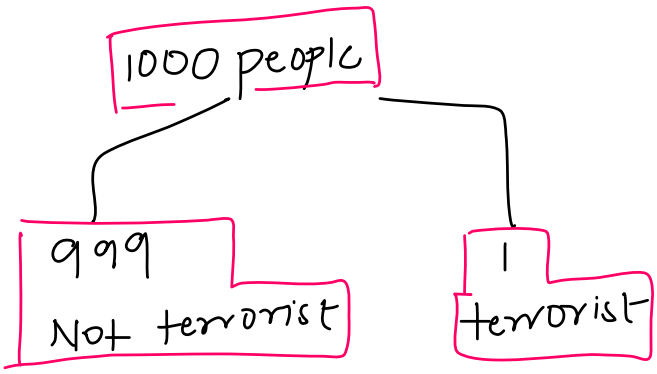
True predn / total = accuracy

3 x 3

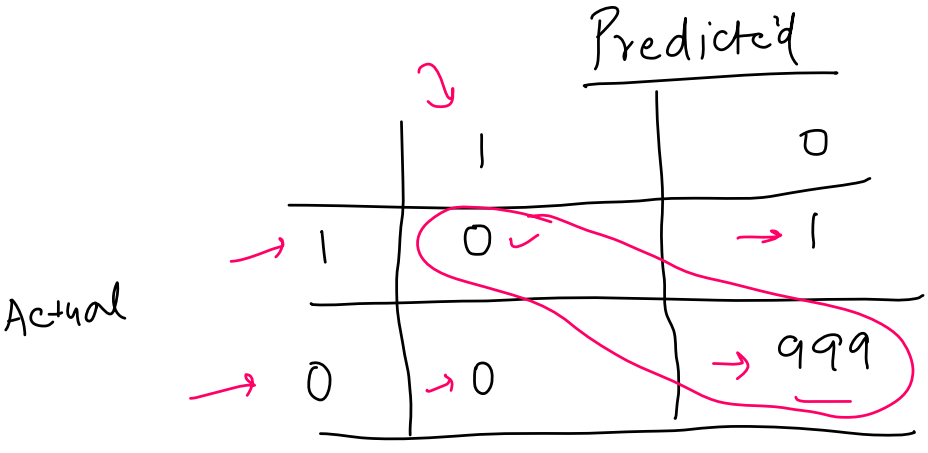
When accuracy is misleading?

Wednesday, June 23, 2021 8:45 AM

Imbalanced Dataset



model → No one is terrorist



Accuracy = $\frac{999}{999+1}$

= 99.9%

Precision

Wednesday, June 23, 2021 8:46 AM

{ spam: 1, not spam: 0 }

Model A

	Sent to Spam	Not sent to spam
Spam	100	170 FN
Not Spam	30 FP	700

Predicted

Model B

	Sent to Spam	Not sent to spam
Spam	100	190
Not Spam	10	700

Actual

$P_A = \frac{100}{100+30}$

$P_B = \frac{100}{100+10}$

$P_A < P_B$

80% FP → Not Spam → Spam

FN → Spam → Not Spam

What proportion of predicted Positives is truly Positive?

1 0

Binary +ve

	Sent to Spam	Not sent to spam
Spam	True Positive	False Negative
Not Spam	False Positive	True Negative

$$\text{Precision} = TP/(TP+FP)$$

Recall

Wednesday, June 23, 2021 8:46 AM

has cancer : 1 , no cancer : 0

Predicted

Actual

	Detected Cancer	Not Detected
Has Cancer	1000	200 (FN)
No Cancer	800 (FP)	8000

A 90%

Recall = $\frac{1000}{1200}$

	Detected Cancer	Not Detected
Has Cancer	1000	500 + B
No Cancer	500	8000

B 90%

$R_B = \frac{1000}{1500}$ (FP)

$R_A > R_B$

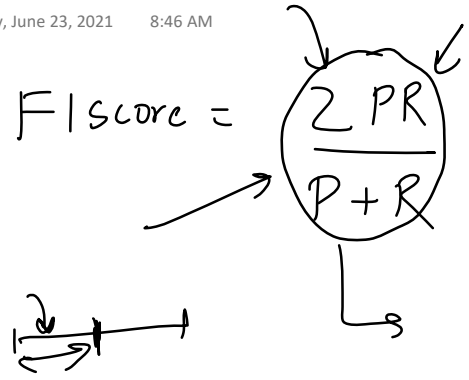
What proportion of actual Positives is correctly classified?

	Detected Cancer	Not detected Cancer
Has Cancer	True Positive	False Negative
No Cancer	False Positive	True Negative

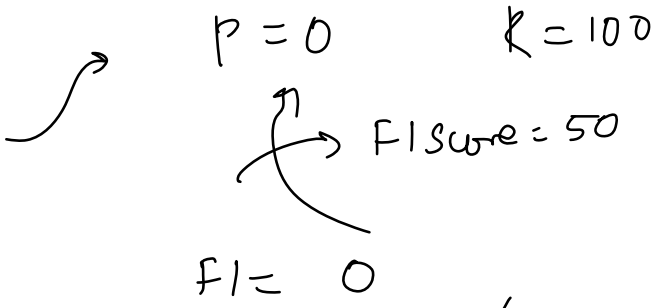
$$\text{Recall} = TP/(TP+FN)$$

F1 Score

Wednesday, June 23, 2021 8:46 AM



F1 score = P+R / 2



(A)

P = R = 80

80 + 80 / 2 = 80

F1 = 80

2 × 80 × 80 / 160 = 80

(B) (75)

P = 60 R = 100

100 + 60 / 2 = 80

= 2 × 60 × 100 / 160

(75)

Multi-class Precision and Recall

Wednesday, June 23, 2021 8:46 AM

Positive (1) ↙ Yes / No

Binary (2) class

Multi (3)

Dog Cat Rabbit

$\frac{2 \times 0.86 \times 0.66}{0.86 + 0.66}$

R_D, R_C, R_R

$F1_D = \frac{2 P_D R_D}{P_D + R_D}$

$F1_C = \frac{2 P_C R_C}{P_C + R_C}$

$F1_R$

Predicted

Actual	Dog	Cat	Rabbit	Total	Recall
Dog	25	5	10	40	0.62
Cat	0	30	4	34	0.88
Rabbit	4	10	20	34	0.58
Total	29	45	34		
Precision	0.86	0.66	0.58		

$R_D = \frac{25}{40} = 0.62$, $R_C = \frac{30}{34} = 0.88$, $R_K = \frac{20}{34} = 0.58$

macro recall ←

weighet recall

$F1_D, F1_C, F1_R$ → macro f1, weighted f1

Multi-class F1 Score

Thursday, June 24, 2021 11:14 AM

Softmax Regression

Monday, June 28, 2021 3:35 PM

cgpa	iq	place?
7.1	71	0
8.5	85	1
9.5	95	2

↑

↘ classified

{ Yes } 1
{ No } 0
Optout 2

{6.5, 65}
↳ Yes / No / Optout

{ Softmax
Multinomial
LR } → Deep Learning

Binary classification

Softmax → general
↑ for 2 class

$0 < \sigma < 1$

Softmax function

K = no. of class {3}

yes → 1
no → 2
opt → 3

$$\sigma(z)_i = \frac{e^{z_i}}{\sum_{j=1}^{K} e^{z_j}}$$

$$\sigma(z)_3 = \frac{e^{z_3}}{e^{z_1} + e^{z_2} + e^{z_3}}$$

(Yes)

$$\sigma(z)_1 = \left\{ \frac{e^{z_1}}{e^{z_1} + e^{z_2} + e^{z_3}} \right\}$$

No

$$\sigma(z)_2 = \frac{e^{z_2}}{e^{z_1} + e^{z_2} + e^{z_3}}$$

Training Intuition

Monday, June 28, 2021 3:35 PM

{Y, N, D} → 0 1 2

cgpa | iq | placed?

–	–	0
–	–	1
–	–	2

→ transform → OHE → 5

cgpa ✓	iq ✓	=0	=1	=2
		1	0	0
		0	1	0
		0	0	1

3 coeff

$w_1\ w_2\ w_0$

D → D_1, D_2, D_3

D_1: cgpa | iq | =0 → m_1 → $w_1^{(0)}, w_2^{(0)}, w_0^{(0)}$

D_2: cgpa | iq | =1 → m_2 → $w_1^{(1)}, w_2^{(1)}, w_0^{(1)}$

D_3: cgpa | iq | =2 → m_3 → $w_1^{(2)}, w_2^{(2)}, w_0^{(2)}$

{ Loss function, Gradient descent }

Loss Function

Monday, June 28, 2021 3:47 PM

$$L = -\frac{1}{m}\sum_{i=1}^{m} y_i \log(\hat{y}_i) + (1-y_i)\log(1-\hat{y}_i)$$

$K = \{1, 2, 3\}$ $i = 1$

$$L = -\frac{1}{m}\sum_{i=1}^{m}\sum_{K=1}^{K} y_K^{(i)} \log(\hat{y}_K^{(i)})$$

$w_1^{(1)}\ w_2^{(1)}\ w_0^{(1)}$

	x_1	x_2	y	$y_{K=1}$	$y_{K=2}$	$y_{K=3}$
1 →	x_{11}	x_{12}	1	1	0	0
2 →	x_{21}	x_{22}	2	0	1	0
3	x_{31}	x_{32}	3	0	0	1

$$y_1^{(1)}\log(\hat{y}_1^{(1)}) + y_2^{(1)}\log(\hat{y}_2^{(1)}) + y_3^{(1)}\log(\hat{y}_3^{(1)}) +$$
$$+ y_1^{(2)}\log(\hat{y}_1^{(2)}) + y_2^{(2)}\log(\hat{y}_2^{(2)}) + y_3^{(2)}\log(\hat{y}_3^{(2)}) +$$
$$+ y_1^{(3)}\log(\hat{y}_1^{(3)}) + y_2^{(3)}\log(\hat{y}_2^{(3)}) + y_3^{(3)}\log(\hat{y}_3^{(3)})$$

$$L = y_1^{(1)}\log(\hat{y}_1^{(1)}) + y_2^{(2)}\log(\hat{y}_2^{(2)}) + y_3^{(3)}\log(\hat{y}_3^{(3)})$$

$\hat{y}_1^{(1)}$, $\hat{y}_2^{(2)}$, $\hat{y}_3^{(3)}$

softmax

$$\hat{y}_1^{(1)} = \sigma(w_1^{(1)}x_{11} + w_2^{(1)}x_{12} + w_0^{(1)})$$

$$y_2^{(2)} = \sigma(w_1^{(2)}x_{21} + w_2^{(2)}x_{22} + w_0^{(2)})$$

$$y_3^{(3)} = \sigma(w_1^{(3)}x_{31} + w_2^{(3)}x_{32} + w_0^{(3)})$$

$$\begin{bmatrix} w_1^{(1)} & w_2^{(1)} & w_0^{(1)} \\ w_1^{(2)} & w_2^{(2)} & w_0^{(2)} \\ w_1^{(3)} & w_2^{(3)} & w_0^{(3)} \end{bmatrix}$$

(L)

$\frac{\partial L}{\partial w_1^{(1)}}, \frac{\partial L}{\partial w_2^{(1)}}, \frac{\partial L}{\partial w_0^{(1)}} \ldots$ gdu

gradient

q-values init=1 $\left[\begin{array}{c}\\ \\ \\ \end{array}\right] =$

loop → 1000 epochs

$$w_1^{(1)} = w_1^{(1)} - \eta \frac{\partial L}{\partial w_1^{(1)}}$$

$$w_2^{(1)} = w_2^{(1)} - \eta \frac{\partial L}{\partial w_2^{(1)}}$$

$\vdots$

Prediction

Monday, June 28, 2021 3:35 PM

$S_x = \{7, 70\} \Rightarrow Y, N, Opt$

m_1 Yes

m_2 No

m_3 Opt-out

$w_1^{(1)}$ $w_2^{(1)}$ $w_0^{(1)}$

$w_1^{(2)}$ $w_2^{(2)}$ $w_0^{(2)}$

$w_1^{(3)}$ $w_2^{(3)}$ $w_0^{(3)}$

$z_1 = 7 \times w_1^{(1)} + 70 \times w_2^{(1)} + w_0^{(1)}$

$z_2 = 7 \times w_1^{(2)} + 70 \times w_2^{(2)} + w_0^{(2)}$

$z_3 = 7 \times w_1^{(3)} + 70 \times w_2^{(3)} + w_0^{(3)}$

$\sigma(y) = \frac{e^{z_1}}{e^{z_1} + e^{z_2} + e^{z_3}}$

$= 0.40$

$\sigma(N) = \frac{e^{z_2}}{e^{z_1} + e^{z_2} + e^{z_3}}$

0.35

$e(0) = \frac{e^{z_3}}{e^{z_1} + e^{z_2} + e^{z_3}}$

0.25

Sigmoid Vs Softmax

Monday, June 28, 2021 3:47 PM

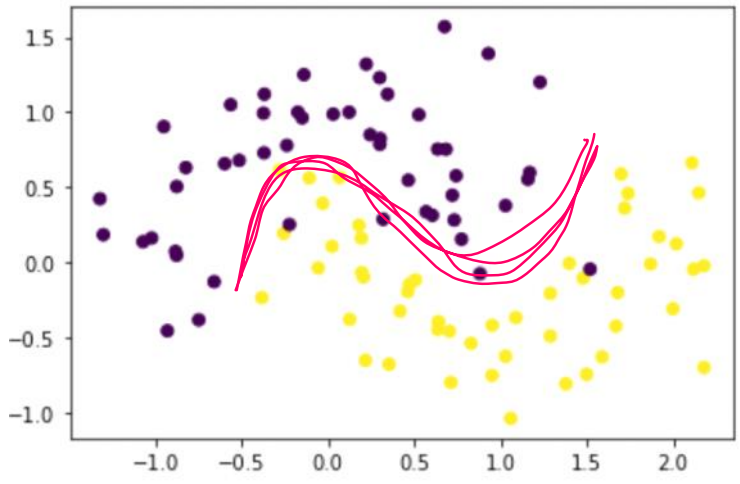
Code Sample

Monday, June 28, 2021 3:36 PM

Polynomial Logistic Regression

Monday, June 28, 2021 6:43 PM

X_1 X_2 y {0,1}



Linear polynomial

degree

x^0 x^1 x^2 x^3

degree = 2

X_1 X_2 → 2 cols

6 cols

X_1^0 X_1^1 X_1^2 X_2^0 X_2^1 X_2^2 y

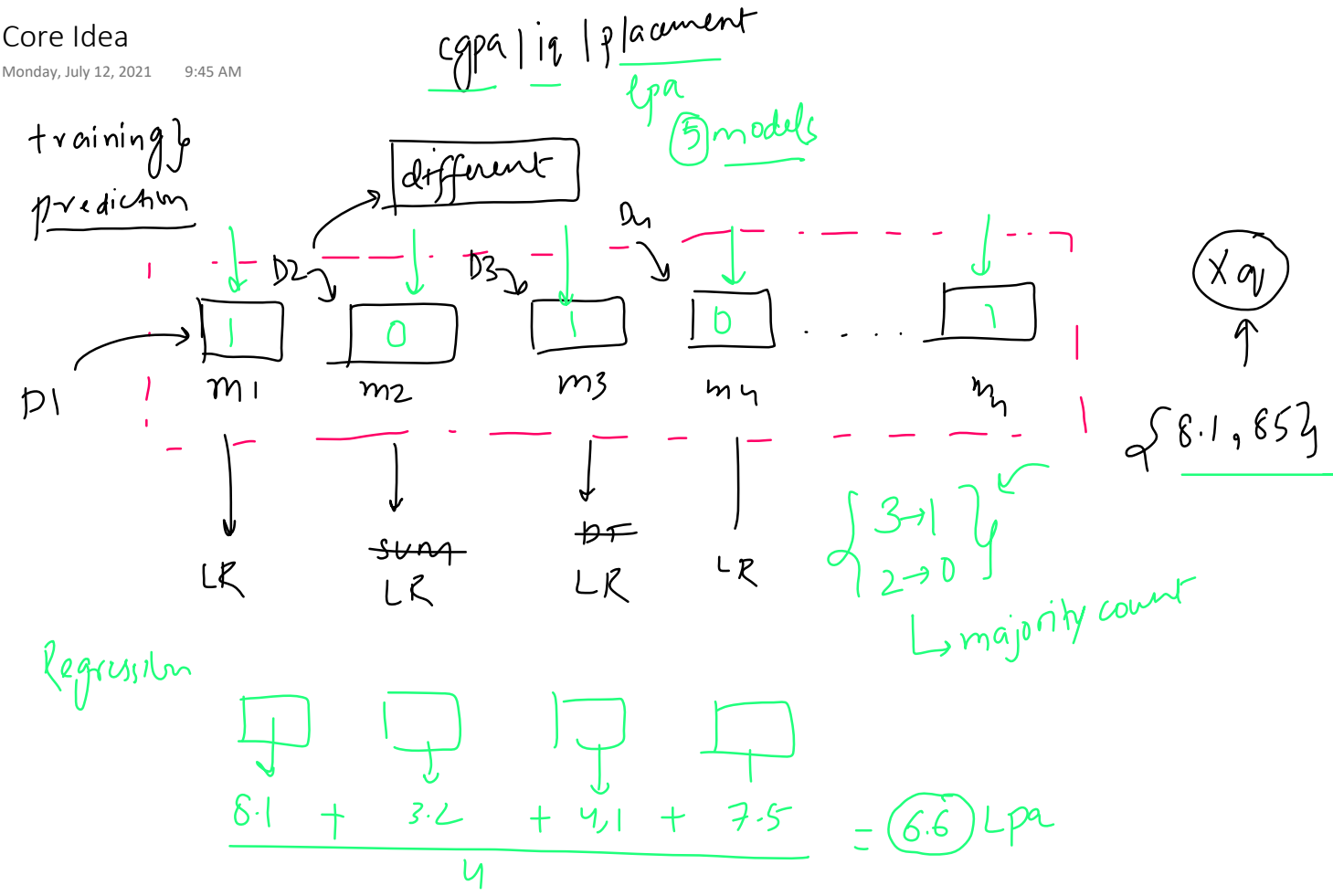
W_0 W_1 W_2 W_3 W_4 W_5

Wisdom of the Crowd

Monday, July 12, 2021 9:56 AM

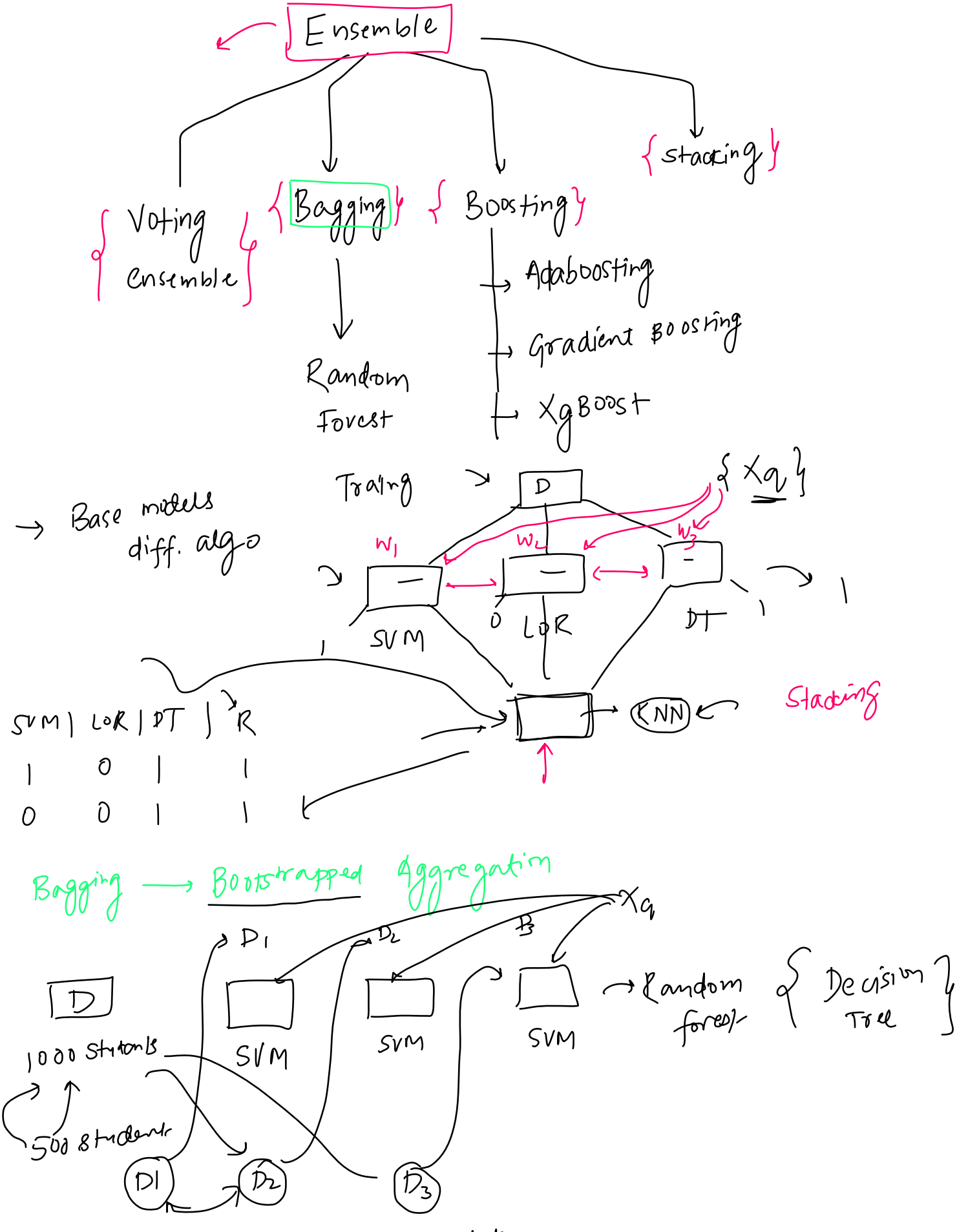
Core Idea

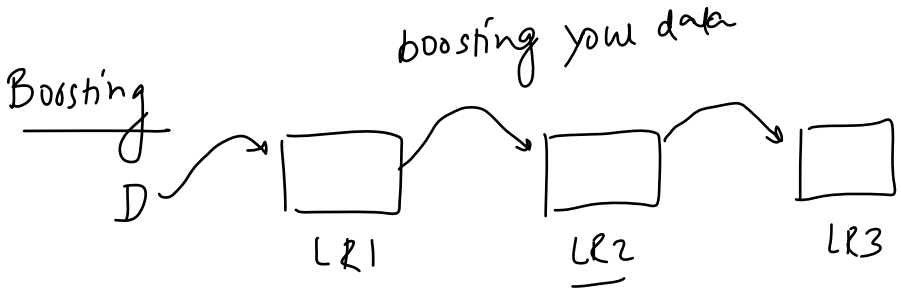
Monday, July 12, 2021 9:45 AM



Type of Ensemble Learning

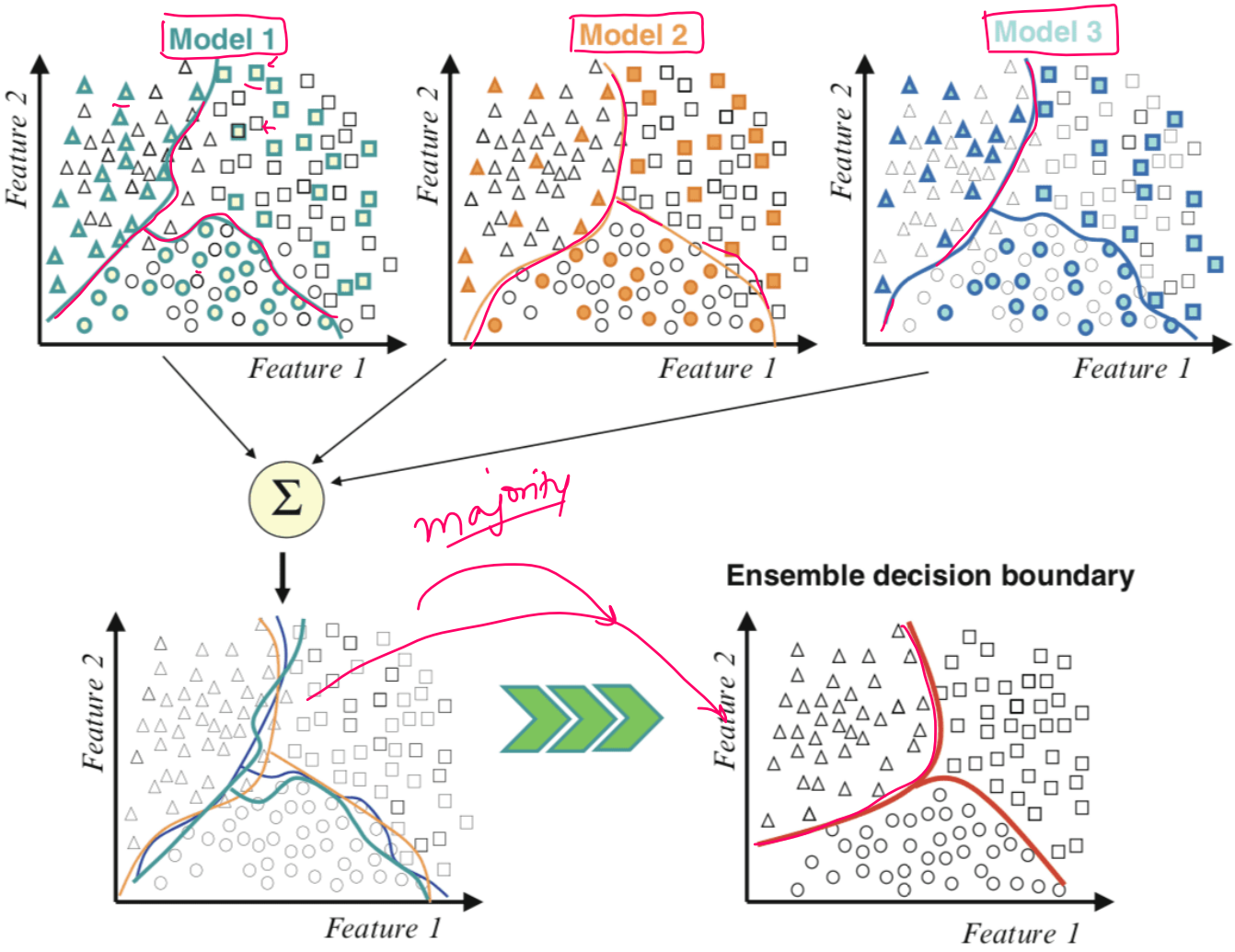
Monday, July 12, 2021 9:45 AM

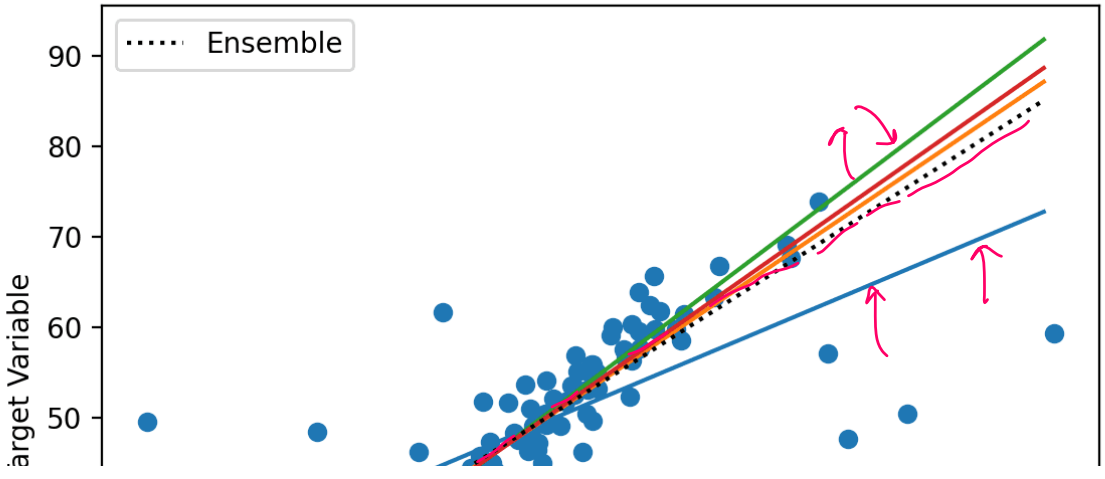


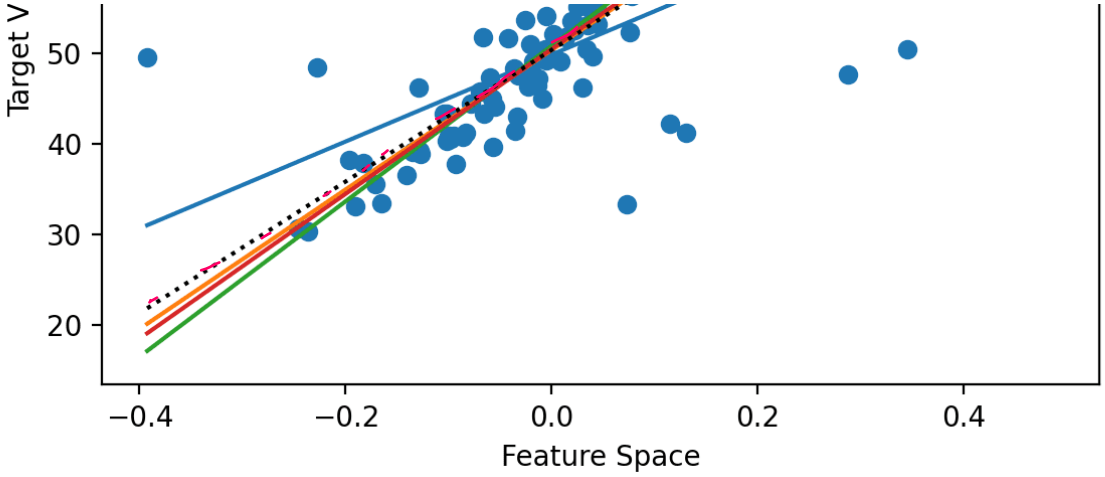

boosting your data
Boosting
D
LR1
LR2
LR3

Why it works?

Monday, July 12, 2021 9:45 AM





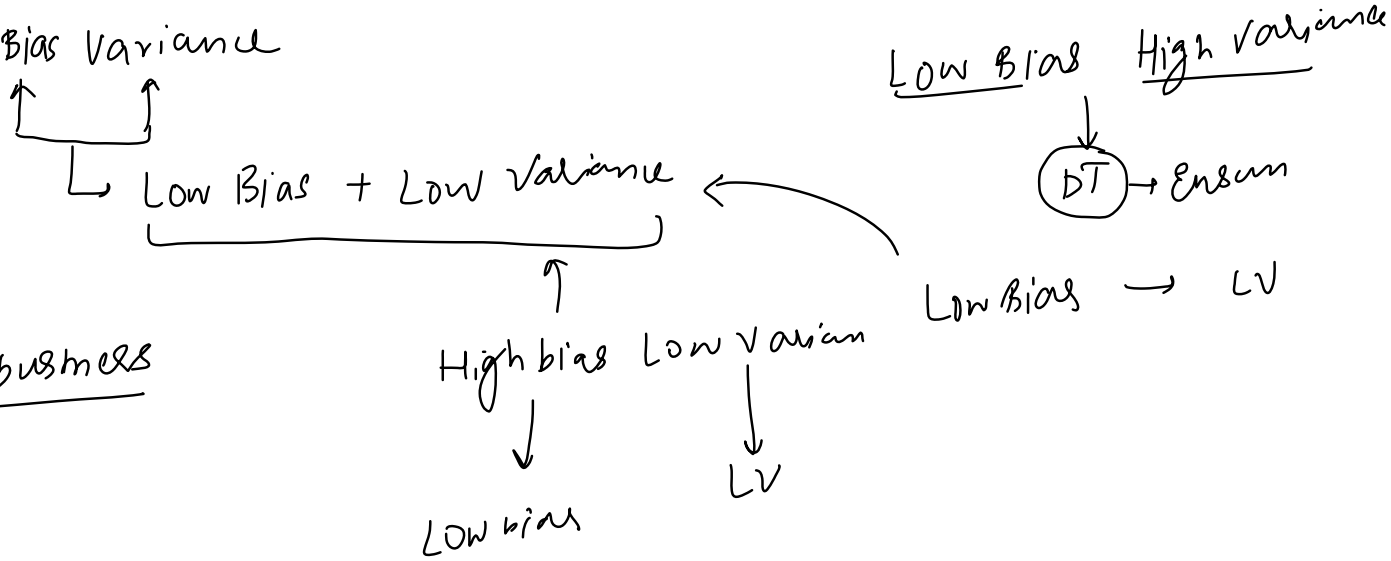

Target V
50
40
30
20
−0.4
−0.2
0.0
0.2
0.4
Feature Space

Benefits

Monday, July 12, 2021 9:45 AM

1) Improvement in performance

2) Bias Variance



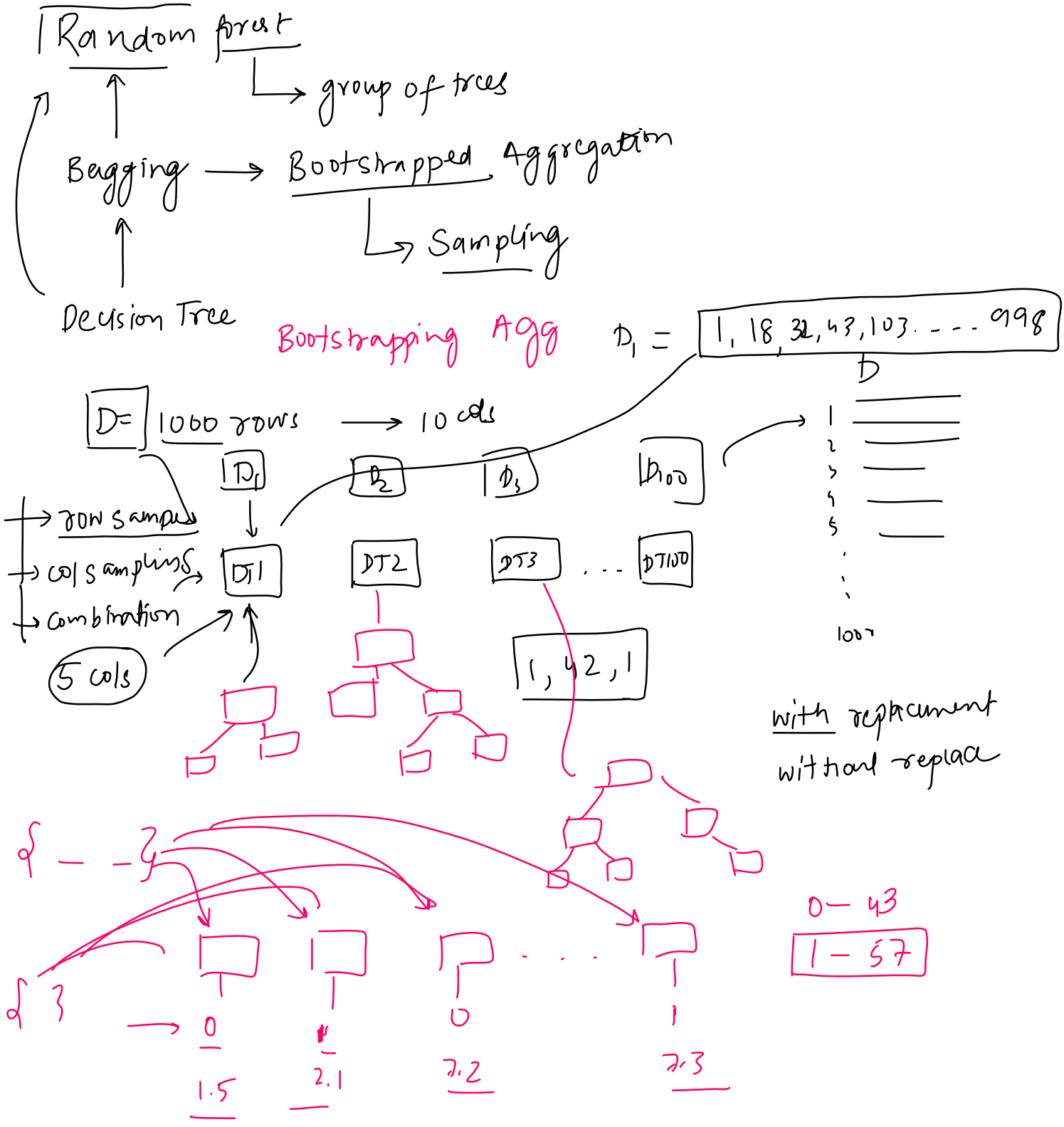
3) Robustness

When to use?

Monday, July 12, 2021 9:46 AM

Intuition

Monday, July 19, 2021 4:53 PM



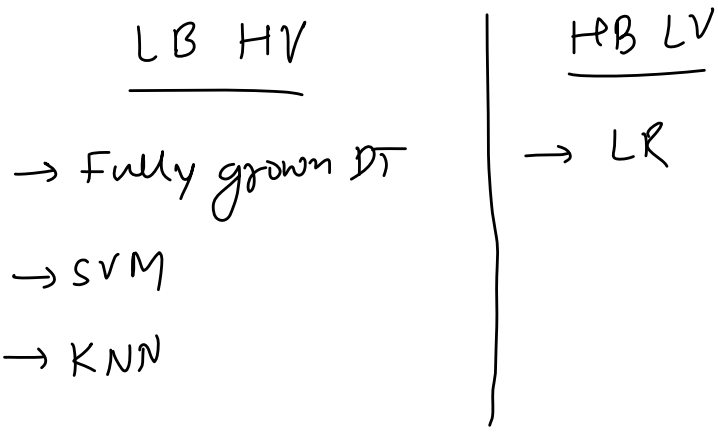
Demo

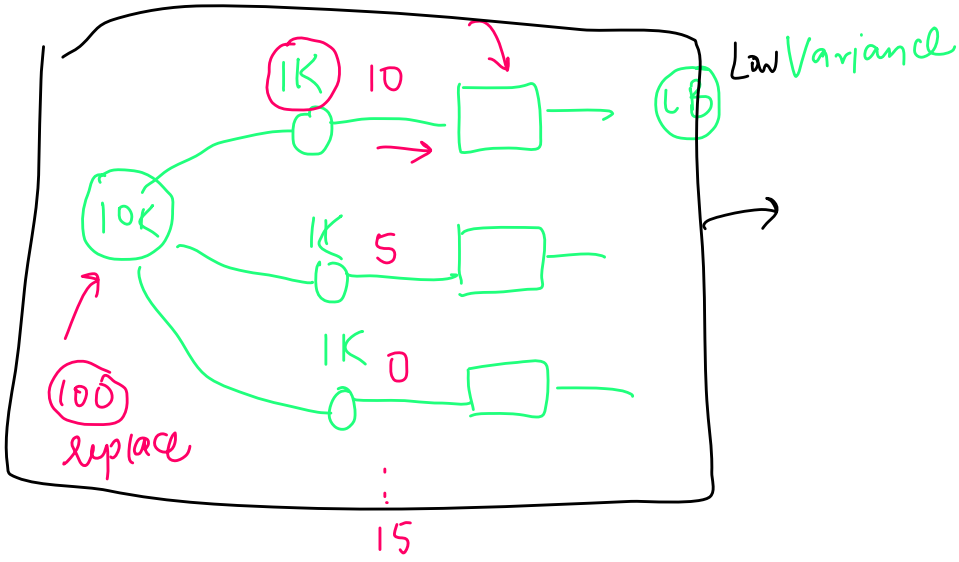
Monday, July 19, 2021 5:05 PM

Bias Variance and Random Forest

Tuesday, July 20, 2021 10:52 AM

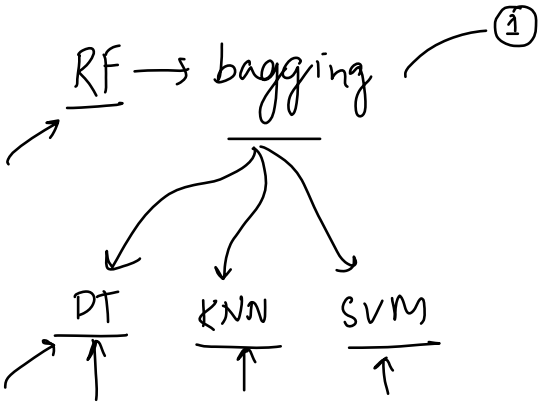
$B \propto \frac{1}{V}$

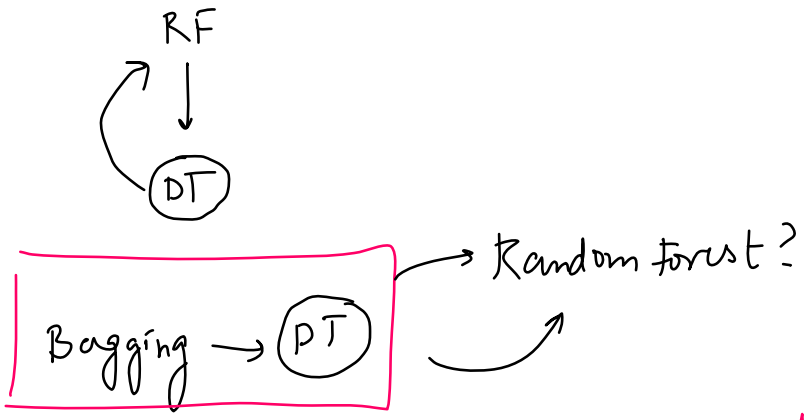


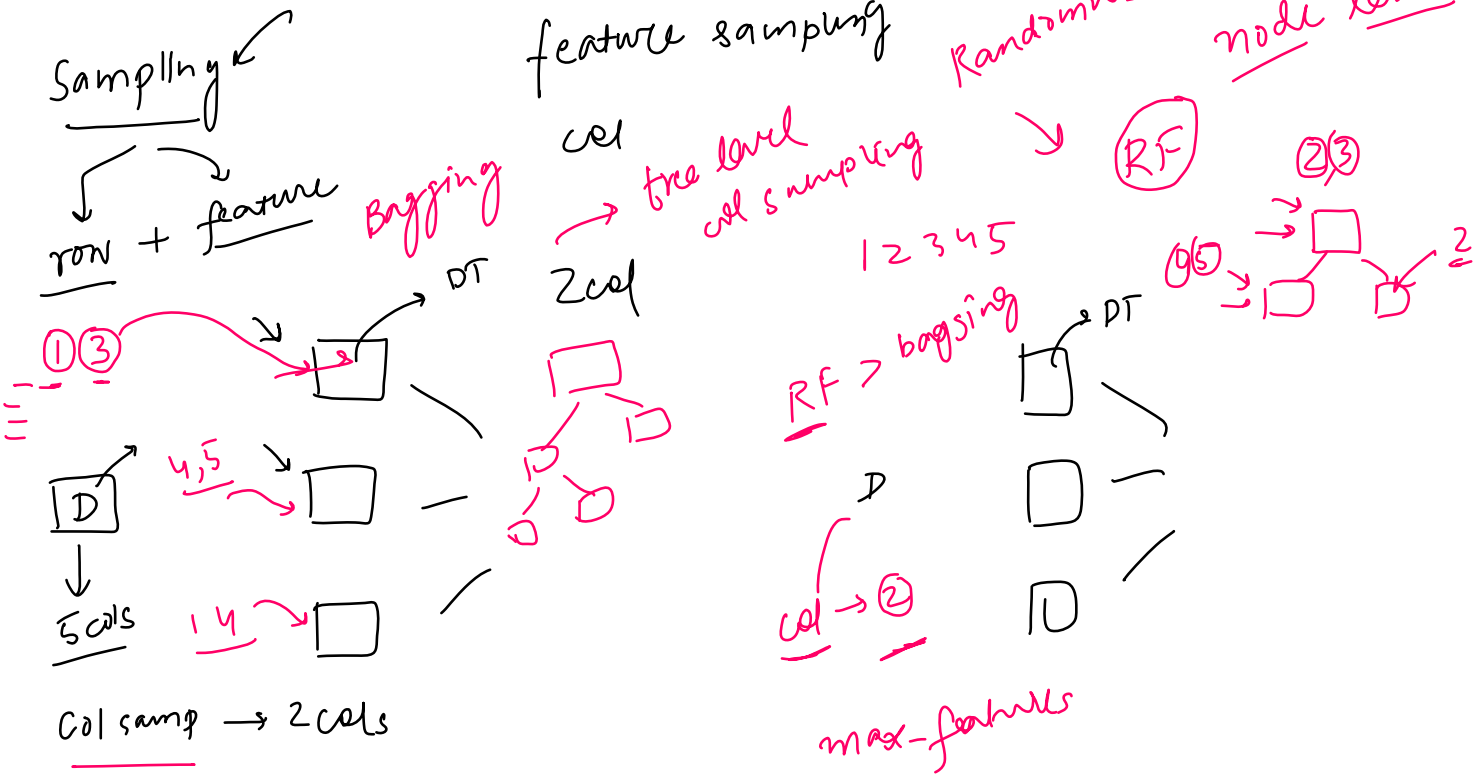


Bagging Vs Random Forest

Monday, July 19, 2021 4:53 PM







Hyperparameters

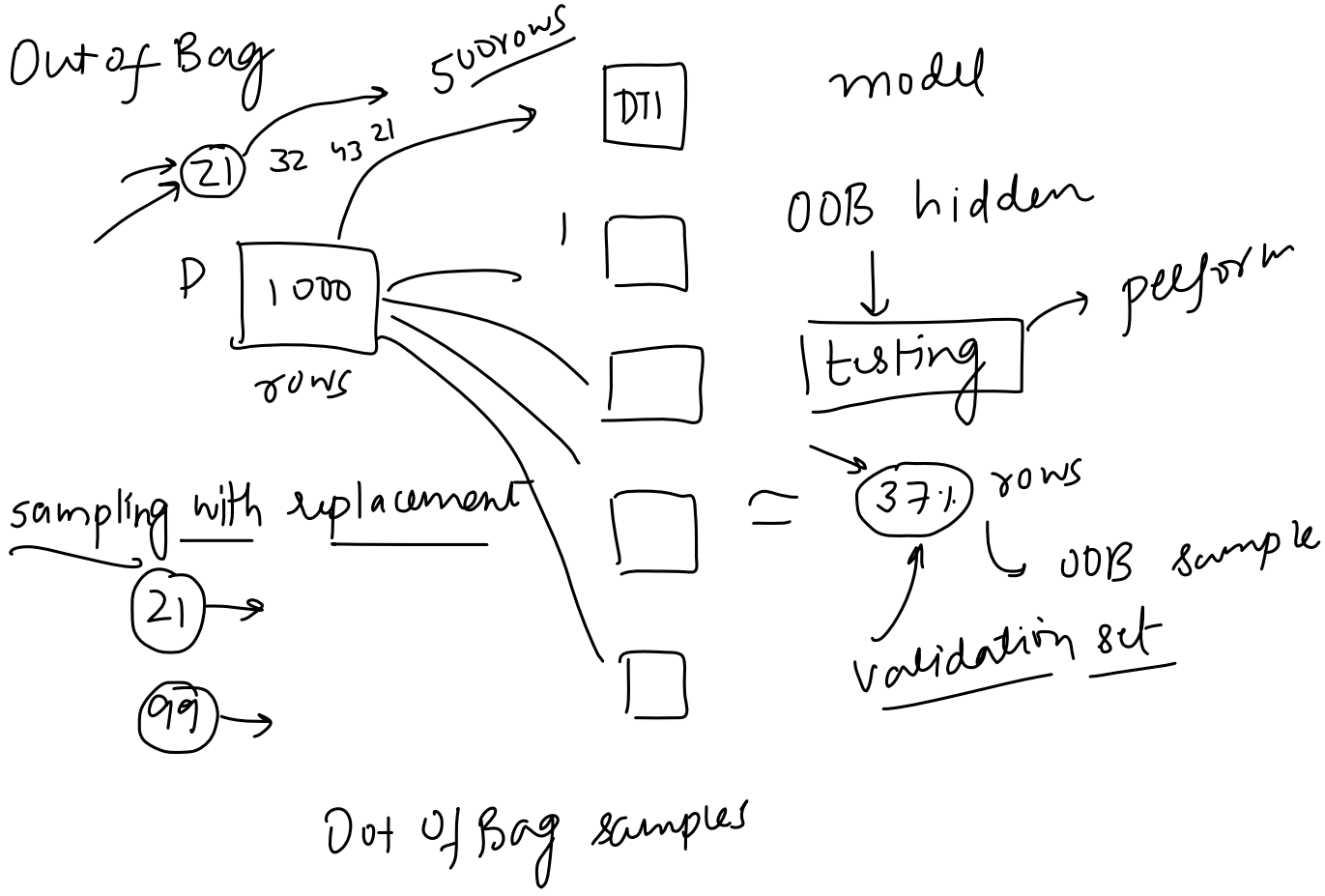
Monday, July 19, 2021 4:56 PM

Code Demo

Monday, July 19, 2021 5:05 PM

OOB Evaluation

Monday, July 19, 2021 5:14 PM



Feature Importance

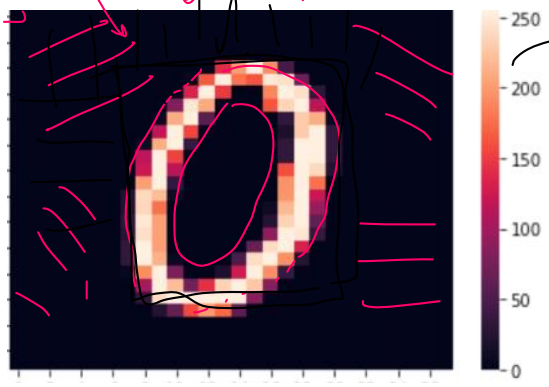
Monday, July 19, 2021 4:56 PM

MNIST

	label	pixel0	pixel1	pixel2	pixel3	pixel4	pixel5	pixel6	pixel7	pixel8	...	pixel774	pixel775	pixel776	pixel777	pixel778
0	1	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0
1	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0
2	1	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0
3	4	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0
4	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0

42000 images

28×28 → 784



feature import

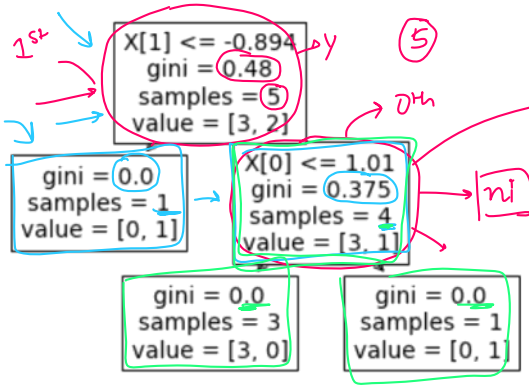
Formula

$$ni = \frac{N\_t}{N}\left[\text{impurity} - \left(\frac{N\_t\_r}{N\_t} \times \text{right\_impurity}\right) - \left(\frac{N\_t\_l}{N\_t} \times \text{left\_impurity}\right)\right]$$

Ntr → 4 5 NTL-1 5

fi [0, 1]

$$fi_K = \frac{\sum_{j \in \text{node split on feature } K} ni}{\sum_{j \in \text{all nodes}} ni}$$



$$= \frac{5}{5}\left[0.48 - \frac{4}{5} \times 0.375 - \frac{1}{5} \times 0.\right] = Y \Rightarrow \left[0.48 - (0.8 \times 0.375)\right]$$

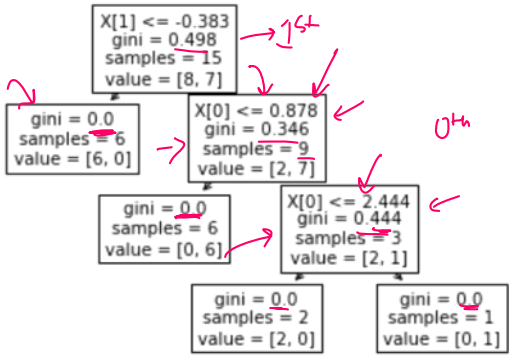
$$\Rightarrow 0.48 - 0.30 = \boxed{0.18}$$

$$\frac{4}{5}\left[0.375\right] = X = 0.8 \times 0.375 = 0.30 = x$$

$$0^{th} = \frac{0.3}{0.3 + 0.18} = 0.625$$

$0^{th} = \frac{0.3}{0.3 + 0.18} = 0.625$

$1^{st} = \frac{0.18}{0.3 + 0.18} = 0.375$



0th Node

$\frac{15}{15}\left[0.49 - \frac{9}{15} \times 0.34\right]$

$= 0.290$

$\frac{9}{15}\left[0.34 - \frac{3}{15} \times 0.44\right]$

$= 0.118$ ←

$\frac{3}{15}\left[0.44\right]$ ←

$= 0.088$

$fi[0] = \frac{0.11 + 0.08}{0.29 + 0.11 + 0.08} = \frac{0.19}{0.48}$

$= 0.39$

$fi[1] = \frac{0.29}{0.48} = 0.60$

Extra Trees

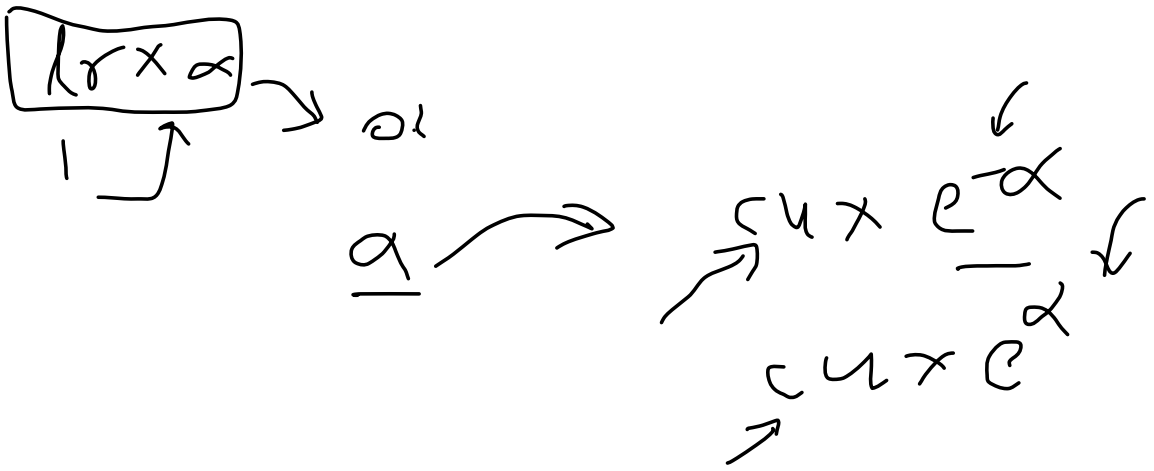
Monday, July 19, 2021 4:57 PM

Before Starting

Monday, August 9, 2021 10:42 AM

1) Weak classifiers (> 50%)

$$\alpha = \frac{1}{2}\log\left(\frac{1-e}{e}\right)$$



The Big Idea

Monday, August 9, 2021 10:43 AM

Step by Step Breakdown

Monday, August 9, 2021 10:43 AM

Points to remember

Monday, August 9, 2021 10:43 AM

Code Walkthrough

Monday, August 9, 2021 10:43 AM

Example and Hyperparameters

Monday, August 9, 2021 10:43 AM

Algorithm

Monday, September 20, 2021 8:23 AM

Input: training set $\{(x_i, y_i)\}_{i=1}^{n}$, a differentiable loss function $L(y, F(x))$, number of iterations M.

1. Initialize $f_0(x) = \arg\min_{\gamma} \sum_{i=1}^{N} L(y_i, \gamma)$.

2. For $m = 1$ to M:

   (a) For $i = 1, 2, \ldots, N$ compute

   $$r_{im} = -\left[\frac{\partial L(y_i, f(x_i))}{\partial f(x_i)}\right]_{f=f_{m-1}}.$$

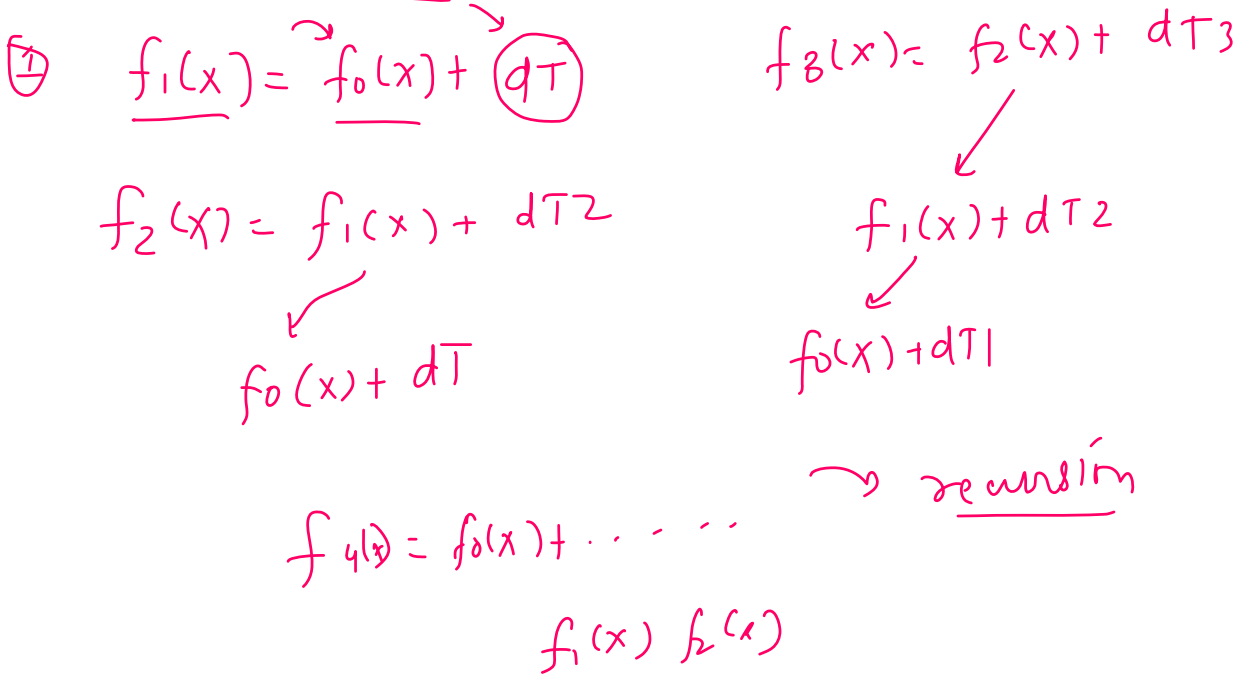
   (b) Fit a regression tree to the targets r_{im} giving terminal regions R_{jm}, $j = 1, 2, \ldots, J_m$.

   (c) For $j = 1, 2, \ldots, J_m$ compute

   $$\gamma_{jm} = \arg\min_{\gamma} \sum_{x_i \in R_{jm}} L\left(y_i, f_{m-1}(x_i) + \gamma\right).$$

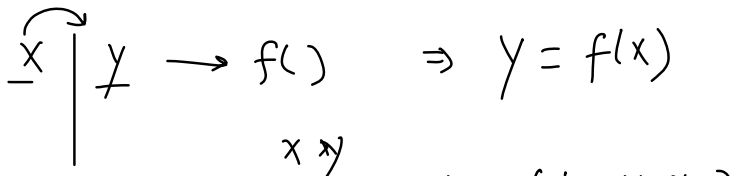
   (d) Update $f_m(x) = f_{m-1}(x) + \sum_{j=1}^{J_m} \gamma_{jm} I(x \in R_{jm})$.

3. Output $\hat{f}(x) = f_M(x)$.



Additive Modelling

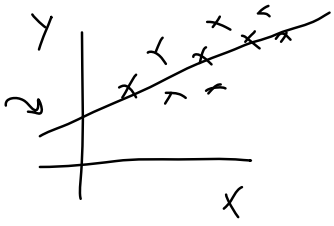
Monday, September 20, 2021 8:28 AM

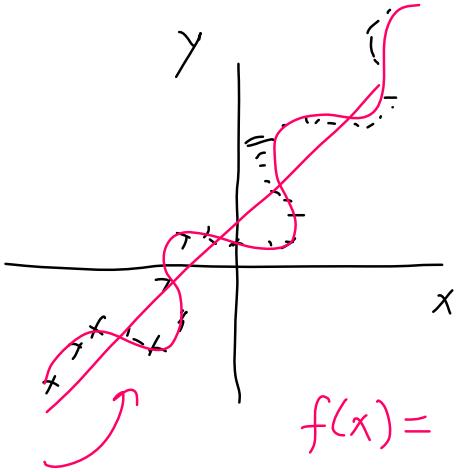


$\Rightarrow y = f(x)$

$y = f(x_1, x_2, x_3)$

$x_1\ x_2\ x_3 \mid y$





$f(x) = x + \sin x$

additive

$F(x) = f_0(x) + f_1(x) + f_2(x) \ldots$

DT PT

$y = x \quad y = \sin(x)$

Explanation

Monday, September 20, 2021 8:25 AM

$$L = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

y_i → actual, $\hat{y}_i$ → pred

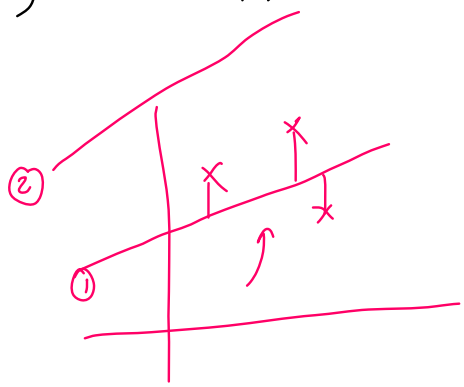
ls

$$L = \frac{1}{2} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \rightarrow \text{diff}$$

$\frac{1}{2}$

$L(y, F(x)) \rightarrow L(y, \hat{y})$

$\hat{y}$



① = 10 = 5 ✓

② = 20 = 10

$y = f(x)$ ↘

DT

$$f(x) = f_0(x) + f_1(x) + f_2(x) + \dots + f_n(x)$$

$$L = \frac{1}{2} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

$$f_0(x) = \underset{\gamma}{\operatorname{argmin}} \sum_{i=1}^{n} L(y_i, \gamma)$$

$$f_0(x) = \underset{\gamma}{\operatorname{argmin}} \frac{1}{2} \sum_{i=1}^{n} (y_i - \gamma)^2$$

γ

$$\frac{d f_0(x)}{d\gamma} = \frac{d}{d\gamma} \frac{1}{2} \sum_{i=1}^{n} (y_i - \gamma)^2 = \frac{1}{2} \sum_{i=1}^{n} \frac{d}{d\gamma} (y_i - \gamma)^2$$

$$\sum_{i=1}^{n} (y_i - \gamma) \frac{d}{d\gamma} (y_i - \gamma) = -\sum_{i=1}^{n} (y_i - \gamma) = 0$$

$$\sum_{i=1}^{n} (\gamma - y_i) = 0$$

$$\sum_{i=1}^{3} (\gamma - y_i) = 0 \Rightarrow (\gamma - 192) + (\gamma - 144) + (\gamma - 91) = 0$$

$$\sum_{i=1} (\gamma - y_i) = 0 \Rightarrow (\gamma - 192) + (\gamma - 144) + (\gamma - 91) = 0$$

$$3\gamma = 192 + 144 + 91$$

mean

For $f_0(x)$ → $\boxed{\gamma = \frac{192 + 144 + 91}{3}}$

mean of output

$$F(x) = \underbrace{f_0(x)}_{\text{mean (leaf)}} + \underbrace{f_1(x)}_{b_1} + f_2(x) + \cdots \rightarrow f_m(x)$$

$\boxed{m = 1}$

$$r_{im} = -\left[\frac{\partial L(y_i, f(x_i))}{\partial f(x_i)}\right]_{f = f_{m-1}}$$

$$r_{i1} = -\left[\frac{\partial L(y_i, f(x_i))}{\partial f(x_i)}\right]_{f = f_0}$$

3 row

r_{11} r_{21} r_{31}

$\hat{y}_i = f(x_i)$

$$r_{i1} = -\left[\frac{\partial}{\partial \hat{y}_i} L(y_i, \hat{y}_i)\right]_{f = f_0}$$

$$L = \frac{1}{2}\sum_{i=1}^{m} (y_i - \hat{y}_i)^2$$

$$r_{i1} = -\left[\frac{\partial}{\partial \hat{y}_i} \frac{1}{2}(y_i - \hat{y}_i)^2\right]_{f = f_0}$$

$$= \left[(y_i - \hat{y}_i)\right]_{f = f_0} = \left[(y_i - f(x_i)\right]_{f = f_0}$$

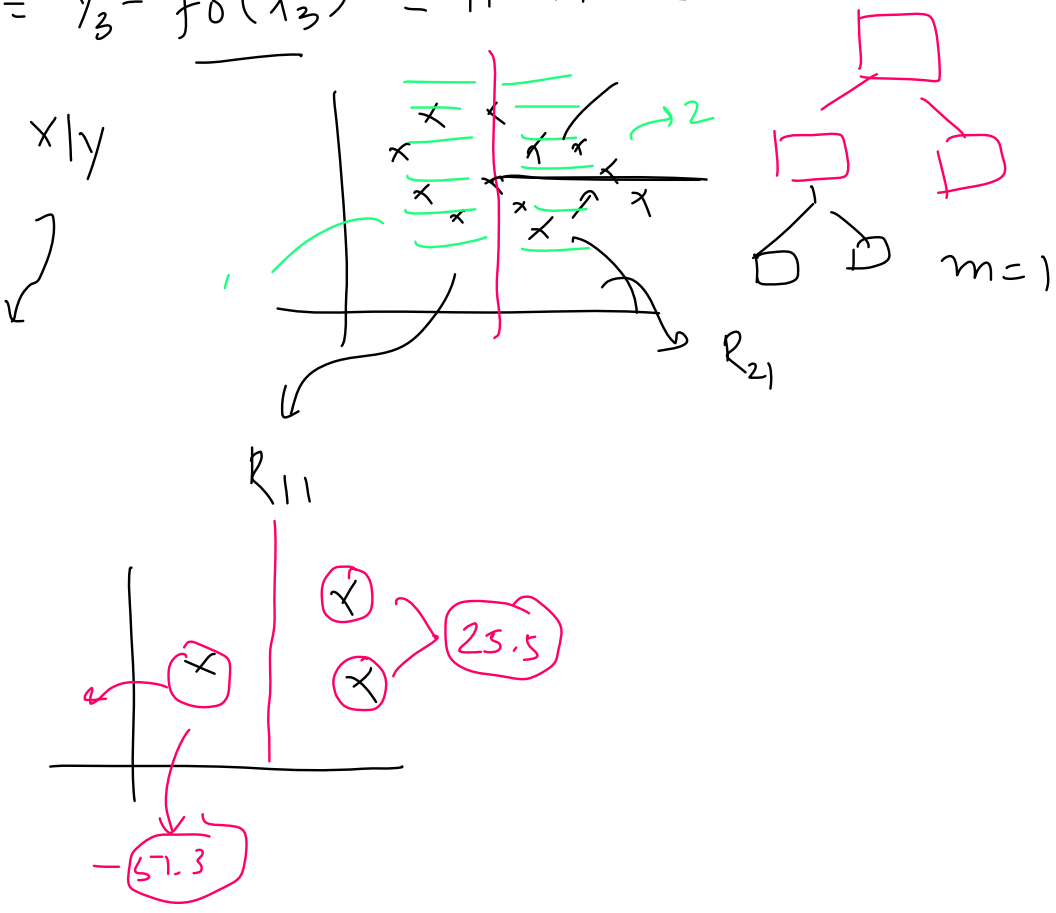
$$= \left[(y_i - y_i) \right]_{f=f_0} = \left[(y_i - I^{(A)}) \right]_{f=f_0}$$

$$r_{i1} = (y_i - f_0(x_i))$$

$$r_{11} = y_1 - f_0(x_1) = 192 - 142 =$$

$$r_{21} = y_2 - f_0(x_2) = 144 - 142 =$$

$$r_{31} = y_3 - f_0(x_3) = 91 - 142 =$$



$$\gamma_{jm} = \underset{\gamma}{\operatorname{argmin}} \sum_{x_i \in R_{jm}} L(y_i, f_{m-1}(x_i) + \gamma)$$

$$\gamma_{j1} \Rightarrow \gamma_{11} = \underset{\gamma}{\operatorname{argmin}} \sum_{x_i \in R_{11}} L(y_i, f_{m-1}(x_i) + \gamma)$$

γ_{11} γ_{21}

γ_{11} γ_{21}

$$\gamma_{11} = \underset{\gamma}{\operatorname{argmin}} \frac{1}{2}(y_i - (f_0(x_i) + \gamma))^2$$

$$\frac{dL}{d\gamma} = \frac{1}{2} \times 2 (y_i - (f_0(x) + \gamma)) \frac{d}{d\gamma} (y_i - f_0(x) - \gamma) = 0$$

$$= -(y_i - f_0(x) - \gamma) = 0$$

$$= y_i - f_0(x) - \gamma = 0$$

$$\gamma_{11} = 91 - 142 - \gamma = 0$$

$$\boxed{\gamma = 91 - 142 = -51}$$

$$\gamma_{21} = \underset{\gamma}{\operatorname{argmin}} \sum_{x_i \in R_{21}} L(y_i, f_0(x_i) + \gamma)$$

$$= \underset{\gamma}{\operatorname{argmin}} \frac{1}{2} \sum_{i=1}^{2} (y_i - (f_0(x_i) + \gamma))^2$$

$$= -\sum_{i=1}^{2} (y_i - f_0(x_i) - \gamma) = 0$$

$$= \sum_{i=1}^{2} (y_i - f_0(x_i) - \gamma) = 0$$

$$= y_1 - f_0(x_1) - \gamma + y_2 - f_0(x_2) - \gamma = 0$$

336
284

$$= 192 - 142 - \gamma + 144 - 142 - \gamma = 0$$

$336 - 284$

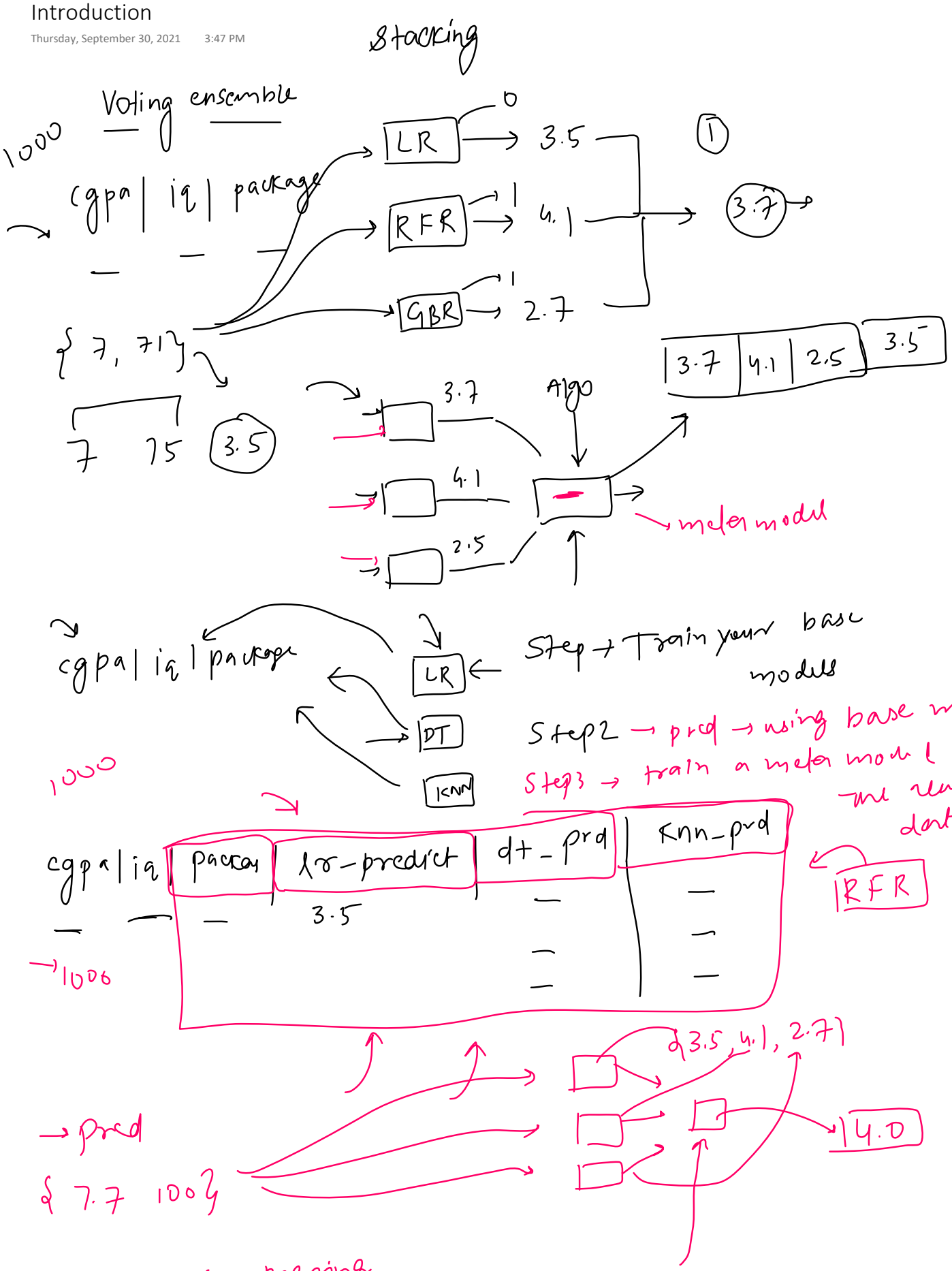
$52 - 2\gamma = 0$

$\gamma = \frac{52}{2} = 26$

Introduction

Thursday, September 30, 2021 3:47 PM

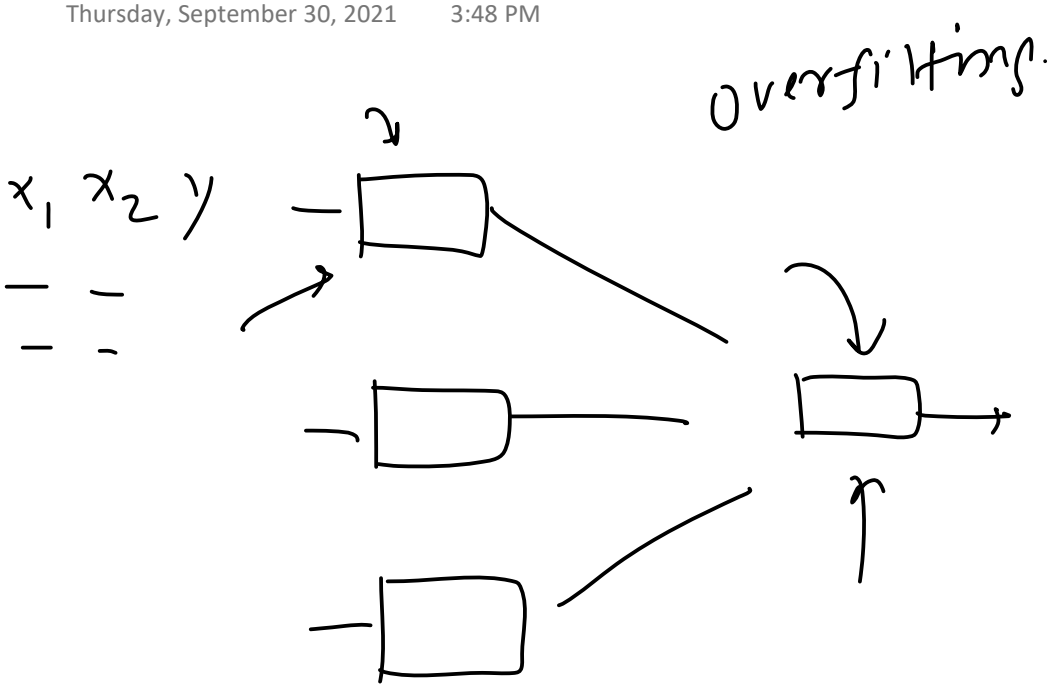
Stacking



→ training

Problem

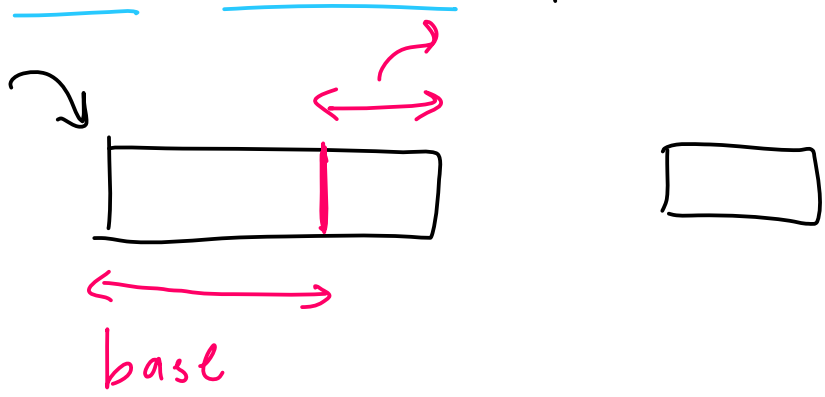
Thursday, September 30, 2021 3:48 PM



Solutions

Thursday, September 30, 2021 3:48 PM

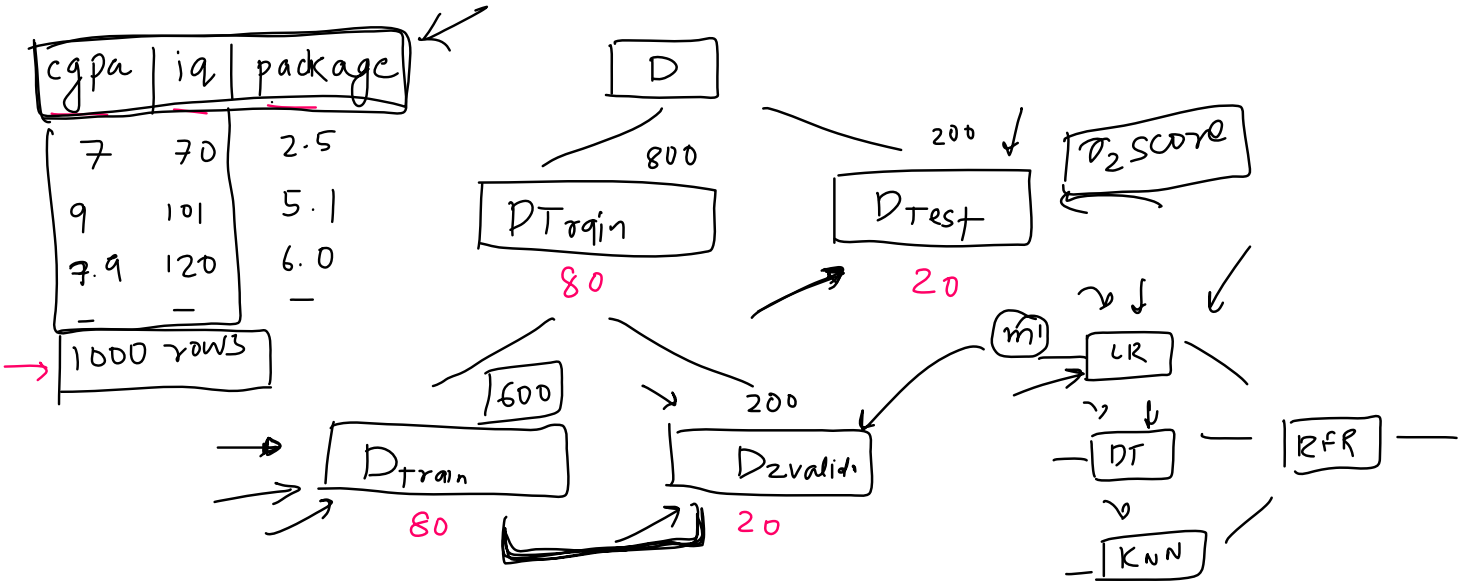
→ Hold out method → Blending



→ K-fold method → Stacking

Hold Out Approach - Blending

Thursday, September 30, 2021 3:48 PM



Step 1 → train 3 base models (3 trained on Dtrain

Step 2 → You form new dataset

Step 3 → Train the meta-model on ↳ meta-models

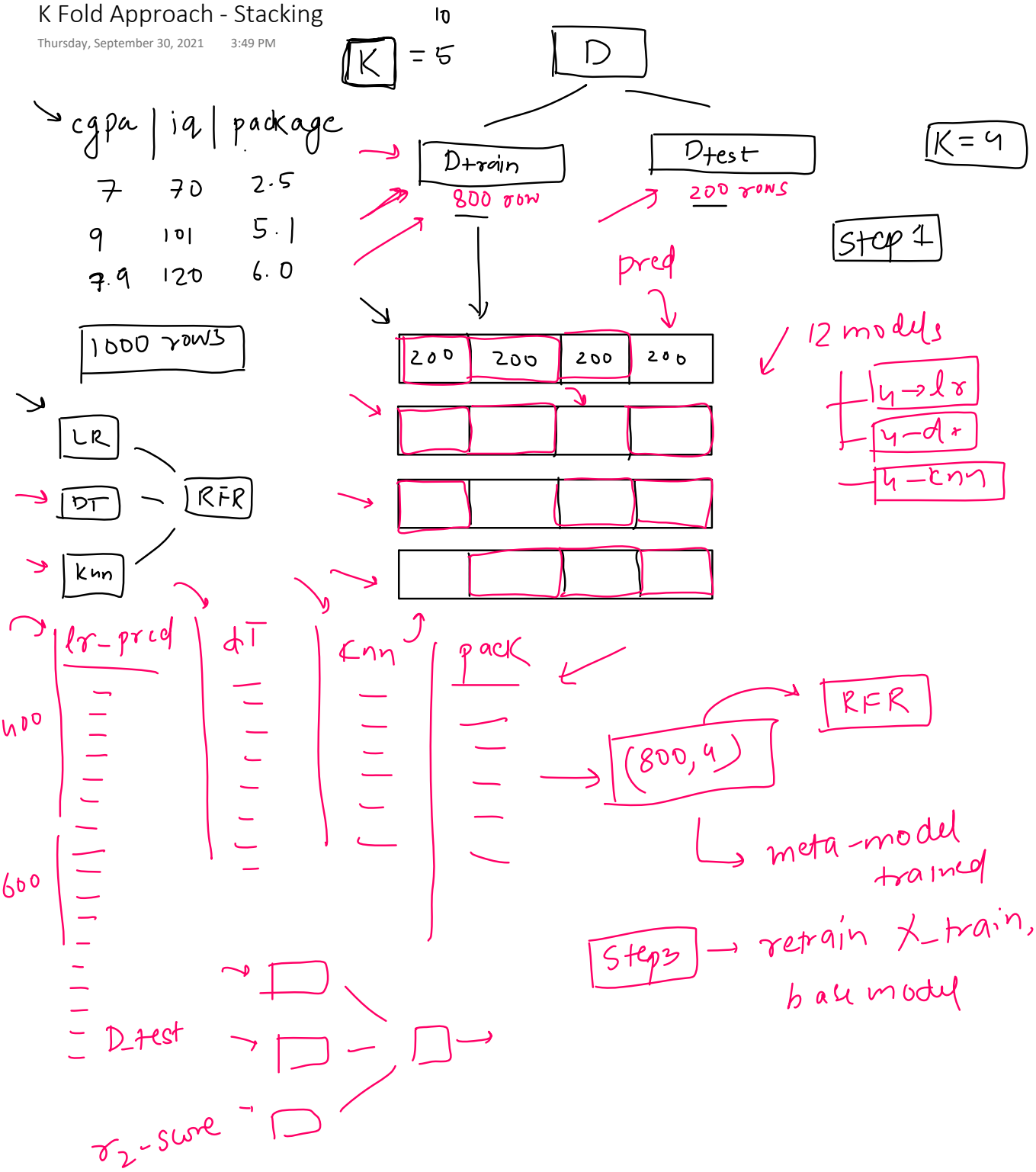
package	lr-pred	dT-pred	Knn-pred

(200, 4)

Blending

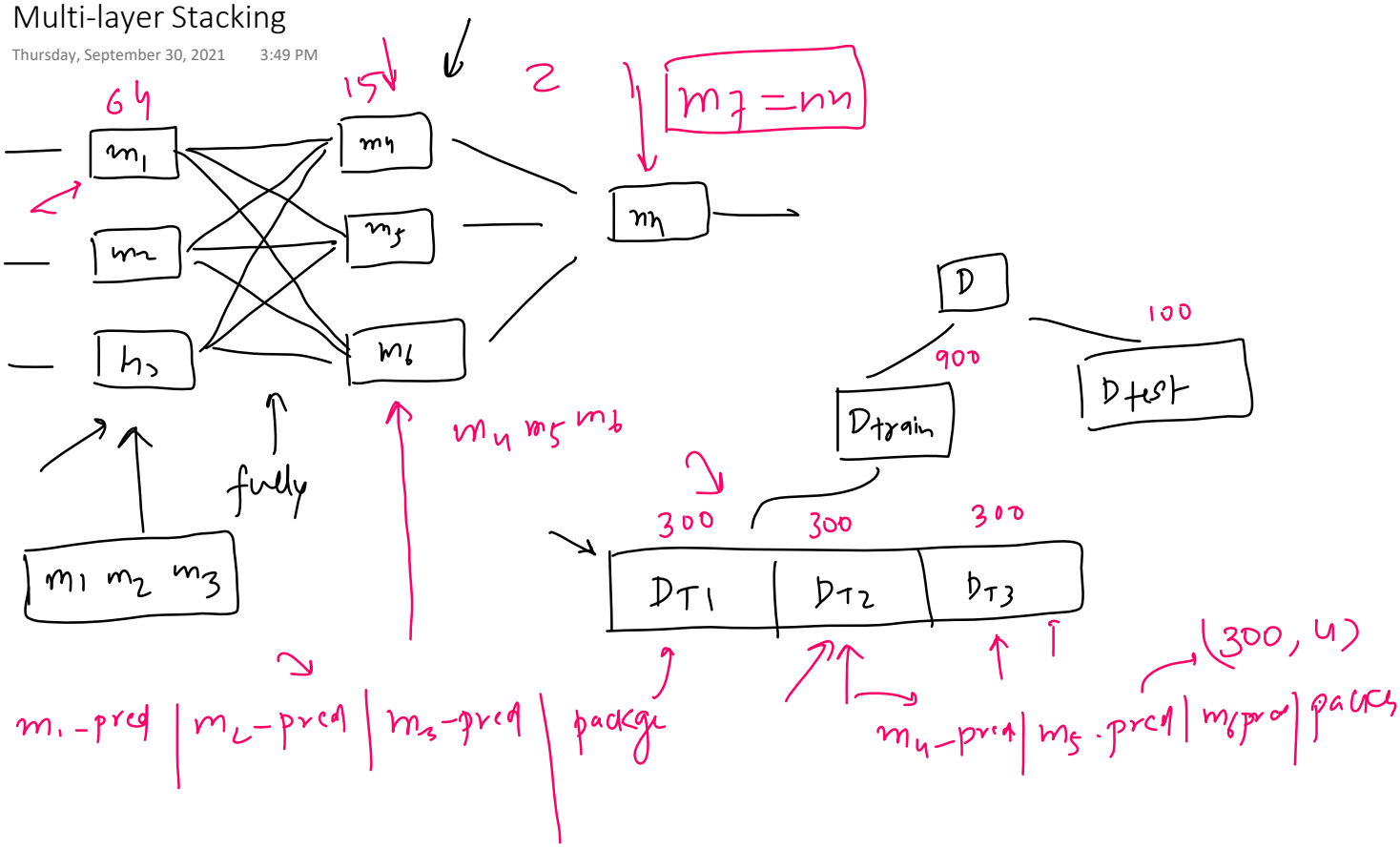
K Fold Approach - Stacking

Thursday, September 30, 2021 3:49 PM



Multi-layer Stacking

Thursday, September 30, 2021 3:49 PM

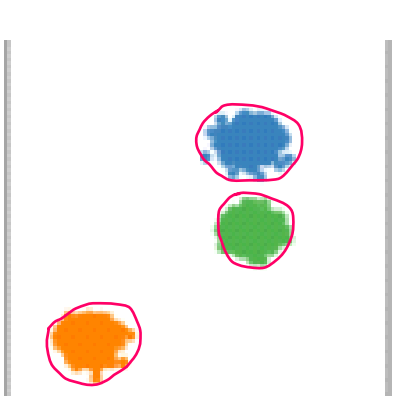


Demo

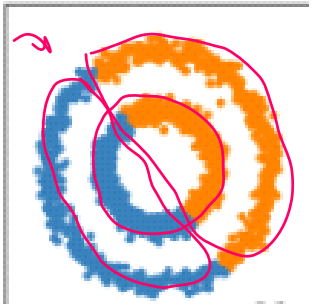
Thursday, September 30, 2021 3:49 PM

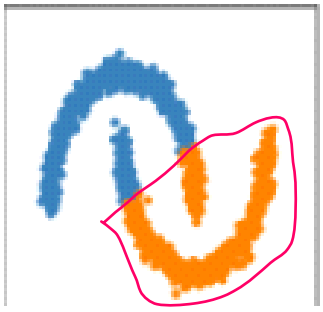
Need of Other Clustering Methods

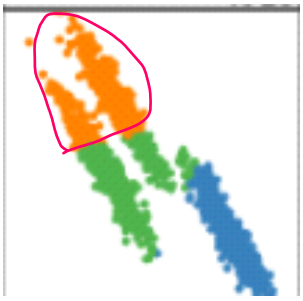
Saturday, November 6, 2021 7:38 AM



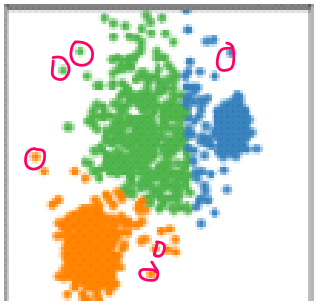
distance to centroid





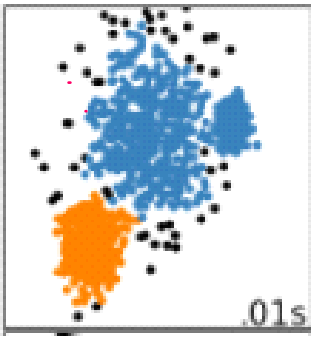


noise



2 More clustering methods:

1. Hierarchical Clustering
2. Density Based Clustering DBSCAN

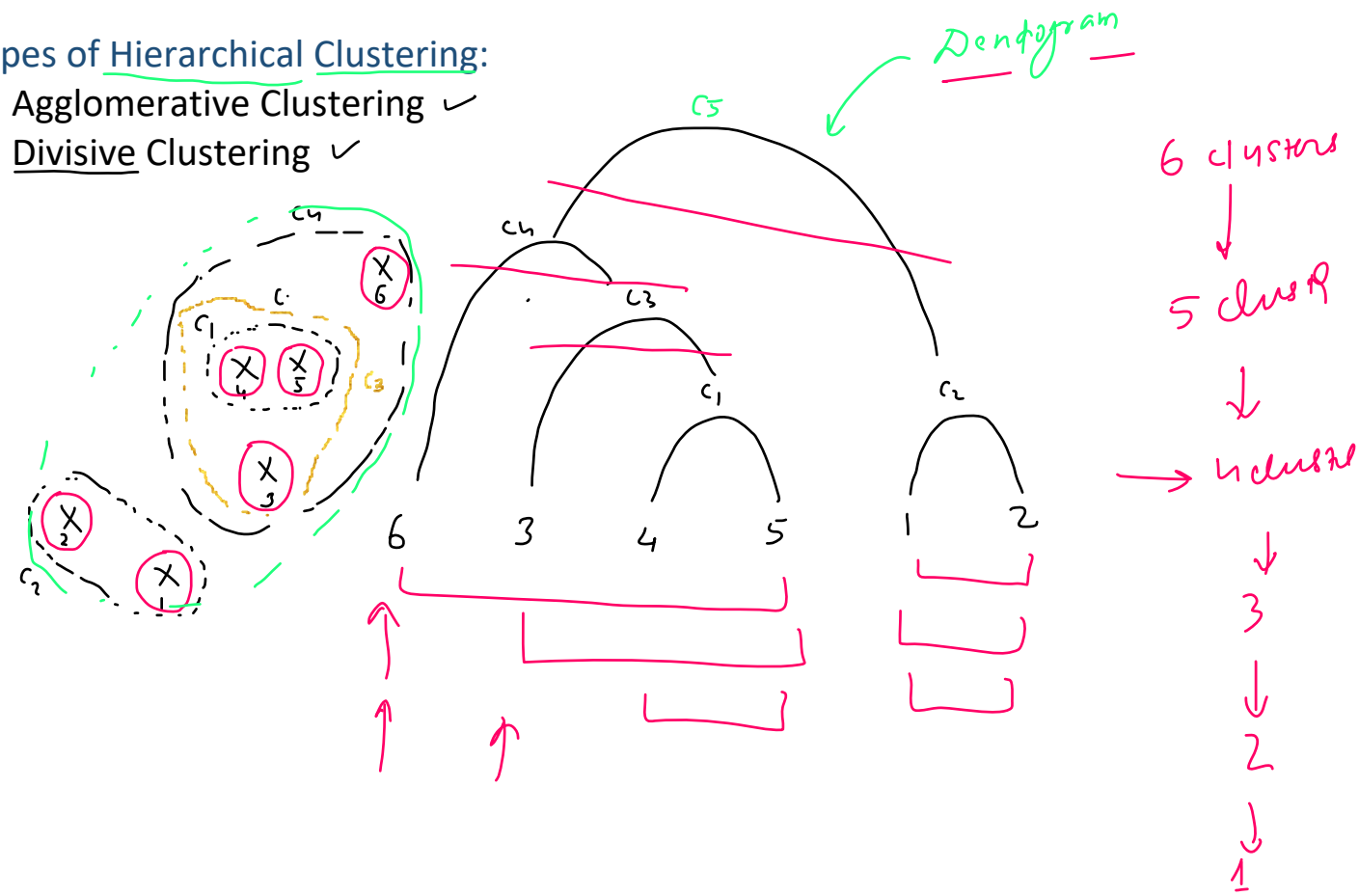

.01s

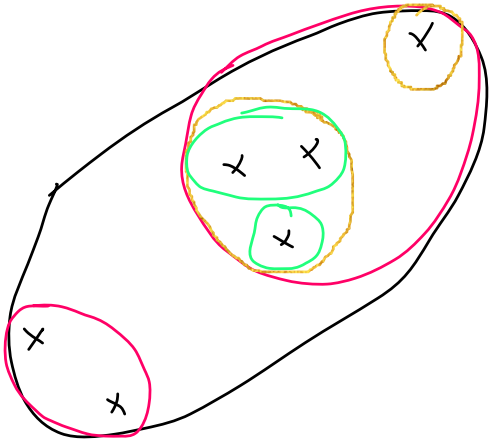
Introduction

Saturday, November 6, 2021 7:39 AM

Types of Hierarchical Clustering:

1. Agglomerative Clustering ✓
2. Divisive Clustering ✓



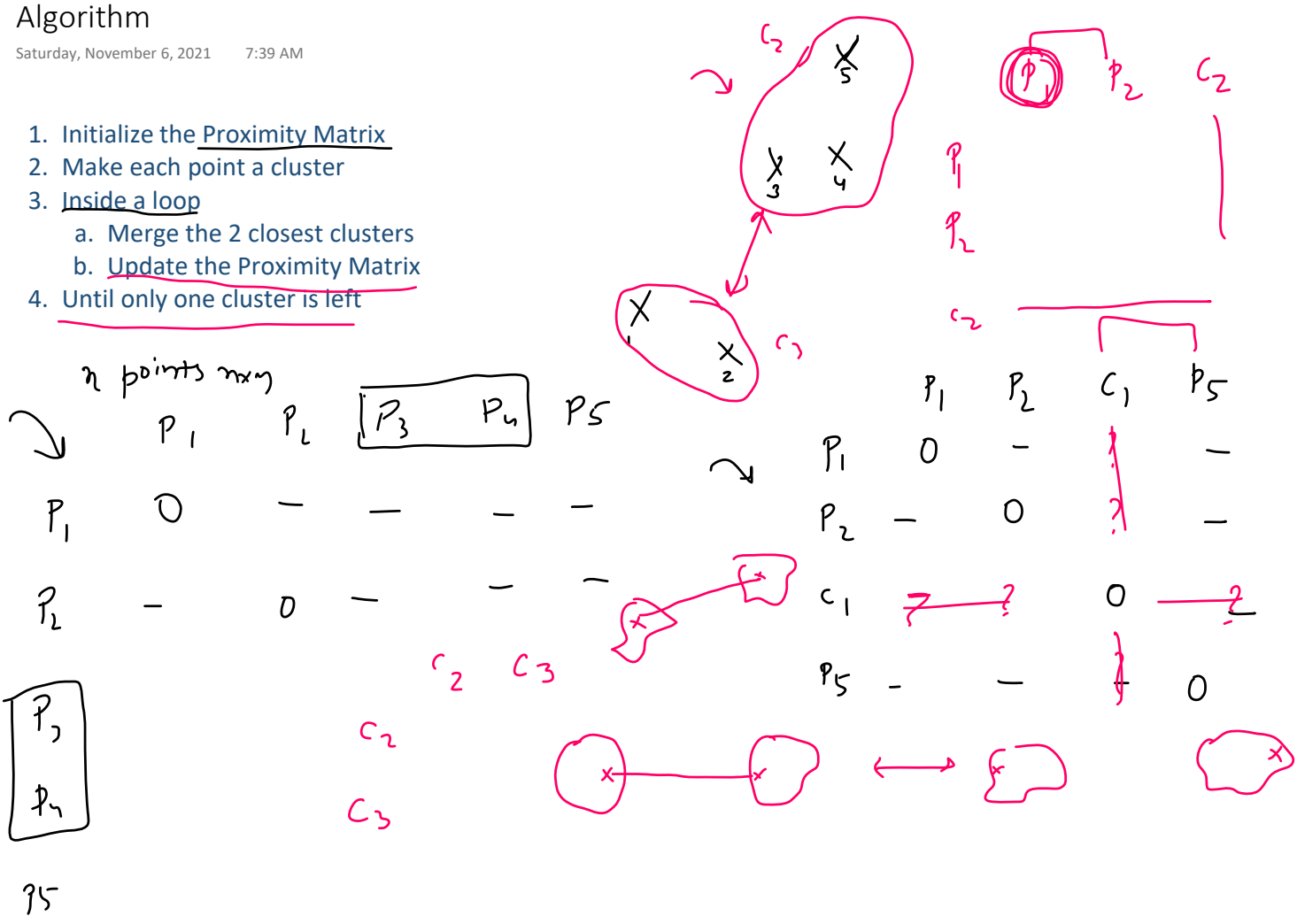


1 cluster
↓
2 clusters
↓
3 clusters
↓
4 clusters
↓
6 clusters

Algorithm

Saturday, November 6, 2021 7:39 AM

1. Initialize the Proximity Matrix
2. Make each point a cluster
3. Inside a loop
   a. Merge the 2 closest clusters
   b. Update the Proximity Matrix
4. Until only one cluster is left



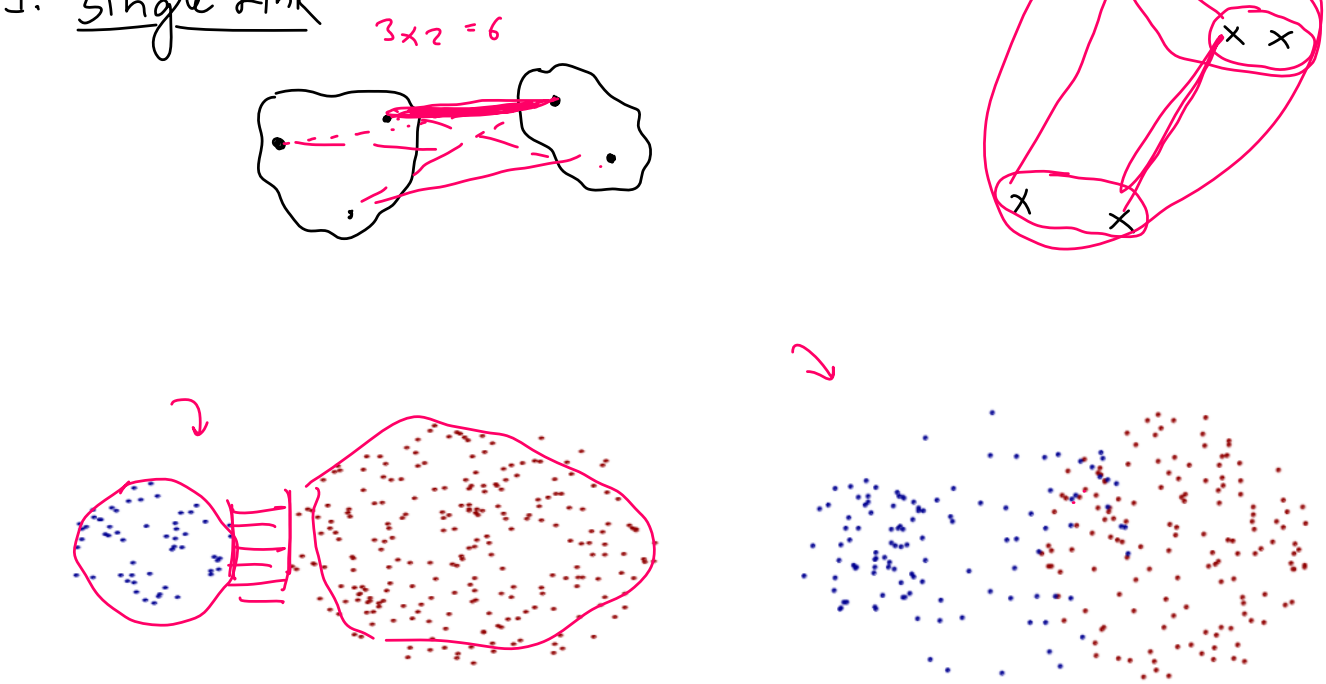
Types

Saturday, November 6, 2021 7:39 AM

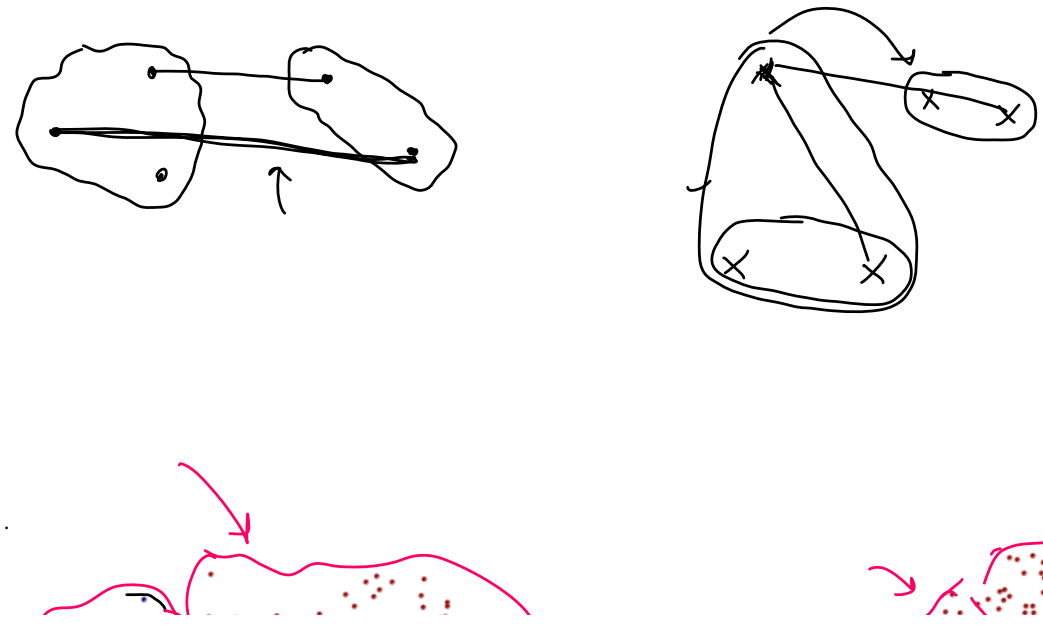
Types of Agglomerative Clustering

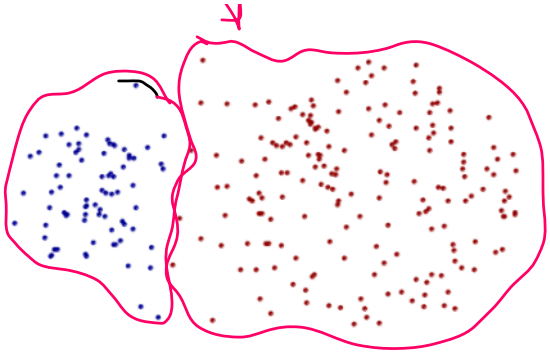
1. Min (Single-link)
2. Max (Complete Link)
3. Average
4. Ward

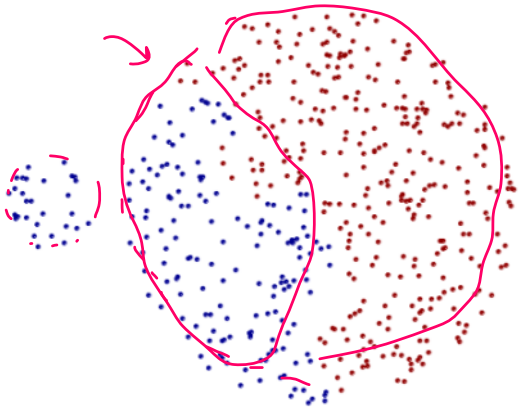
1. Single link



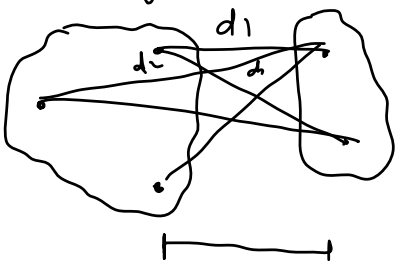
2. Complete link (Max)





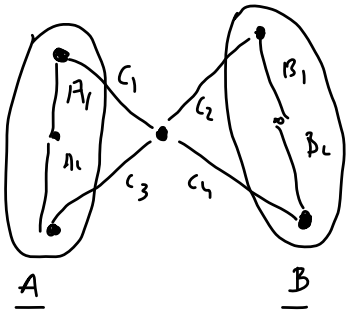


3. Group Average — 2×3



$$\frac{d_1 + d_2 + d_3 + \ldots + d_6}{2 \times 3} =$$

4. Ward → sklearn

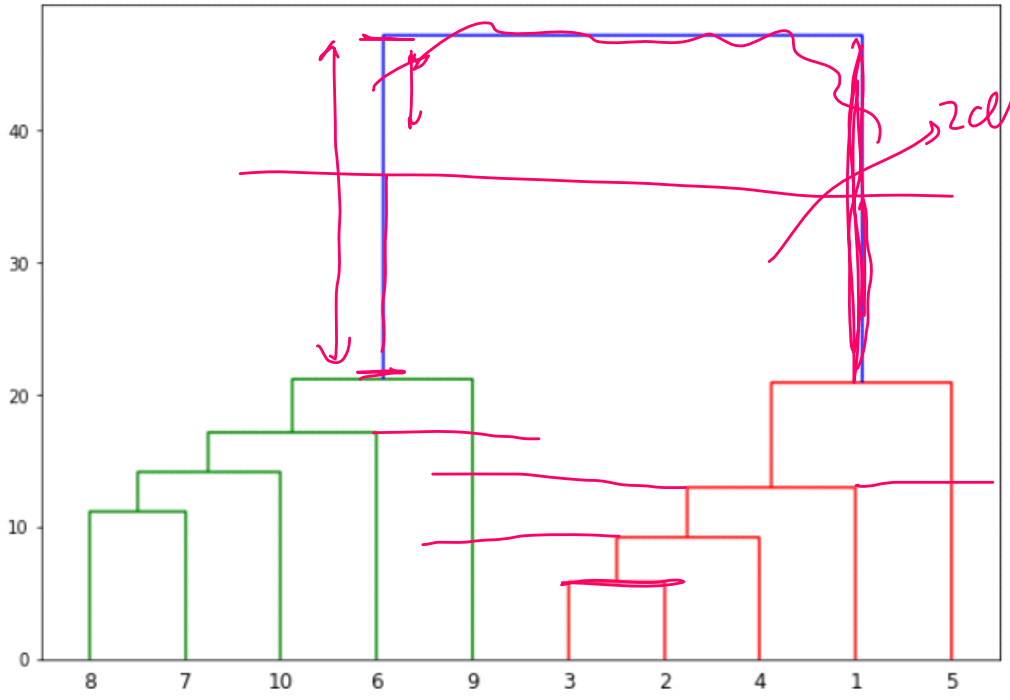


$$\text{dist}_{A \to B} = C_1^2 + C_2^2 + C_3^2 + C_4^2 - A_1^2 - A_2^2 - B_1^2 - B_2^2$$

Variance minimize

How to find the ideal number of clusters

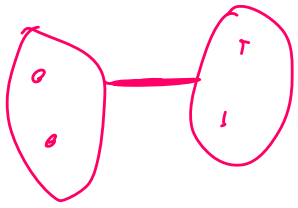
Saturday, November 6, 2021 1:50 PM

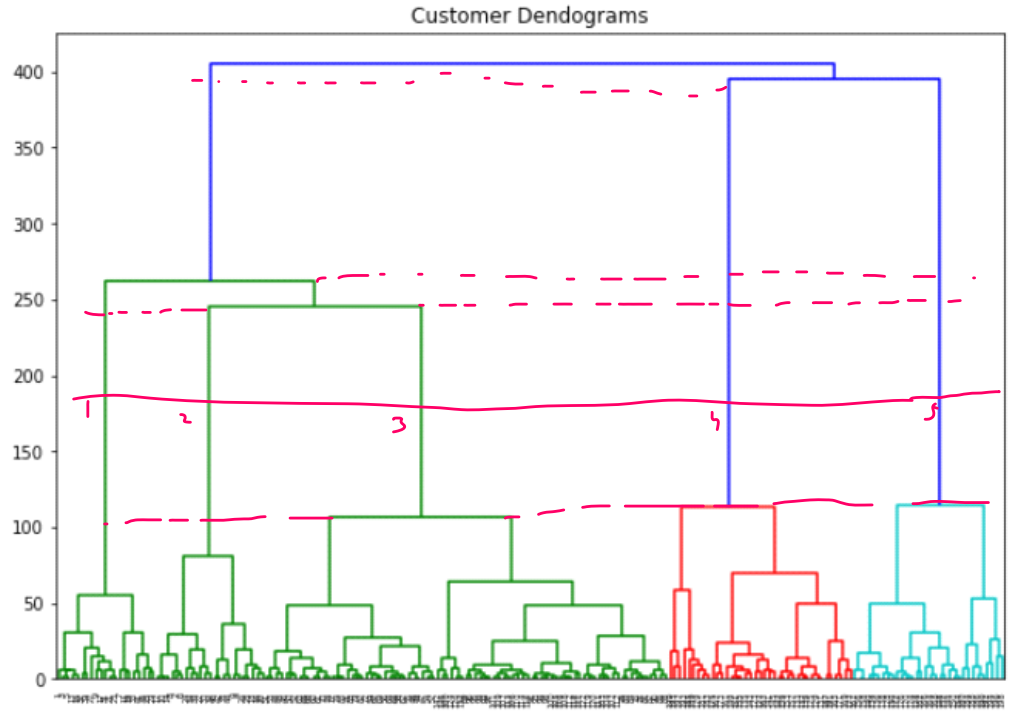


2 clusters

Kmeans → elbow method

inter cluster similarity





5 clusters

Hyperparameter

Saturday, November 6, 2021 7:40 AM

Code Example

Saturday, November 6, 2021 7:40 AM

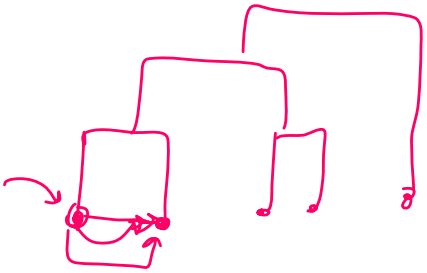
Benefits/Limitations

Saturday, November 6, 2021 7:39 AM

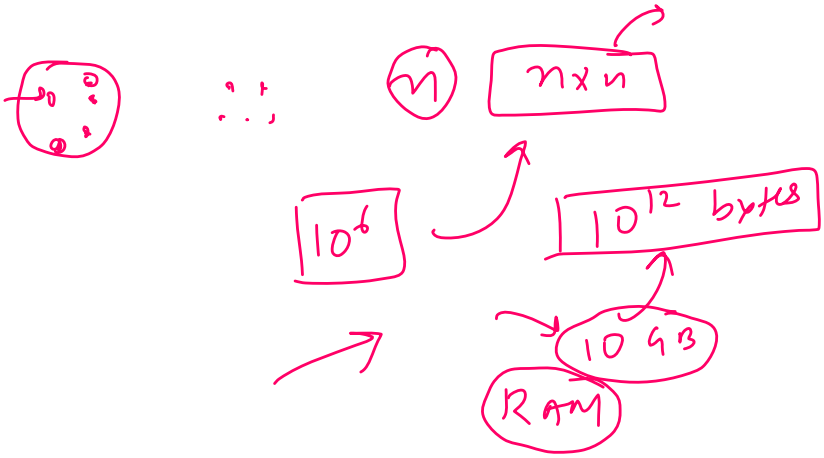
Benefits

1) Widely applicable

2) Dendogram

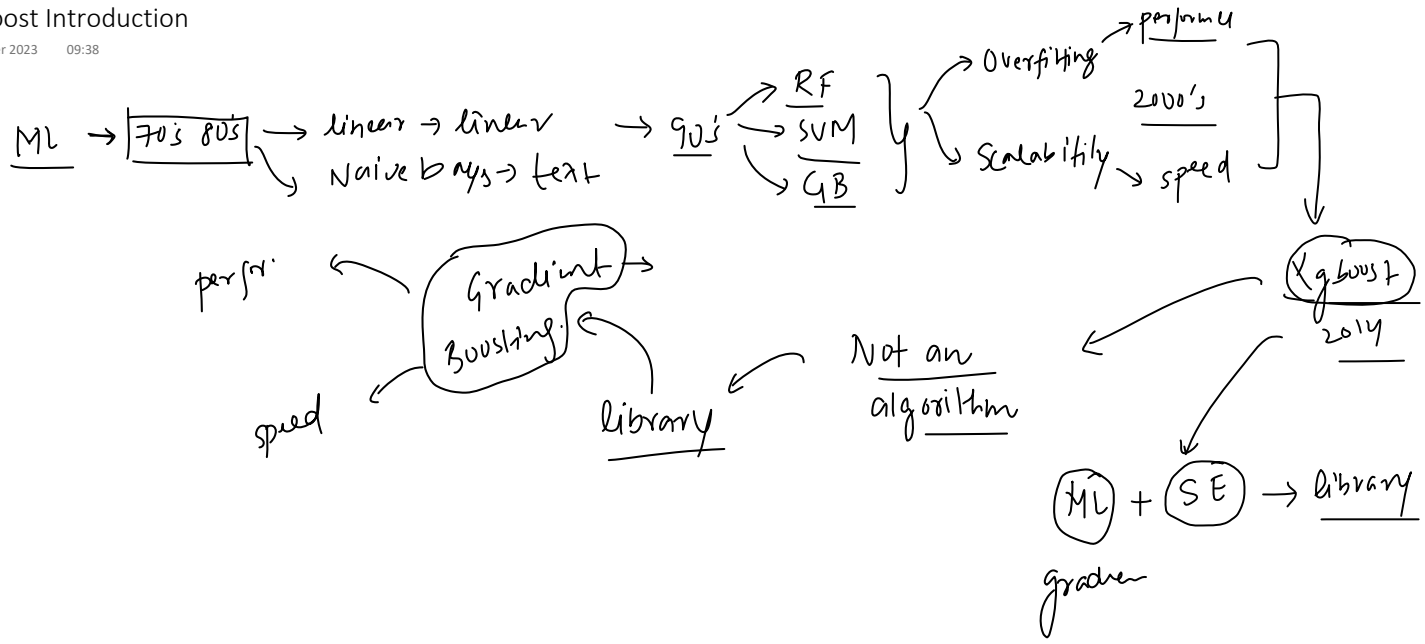


Limitation



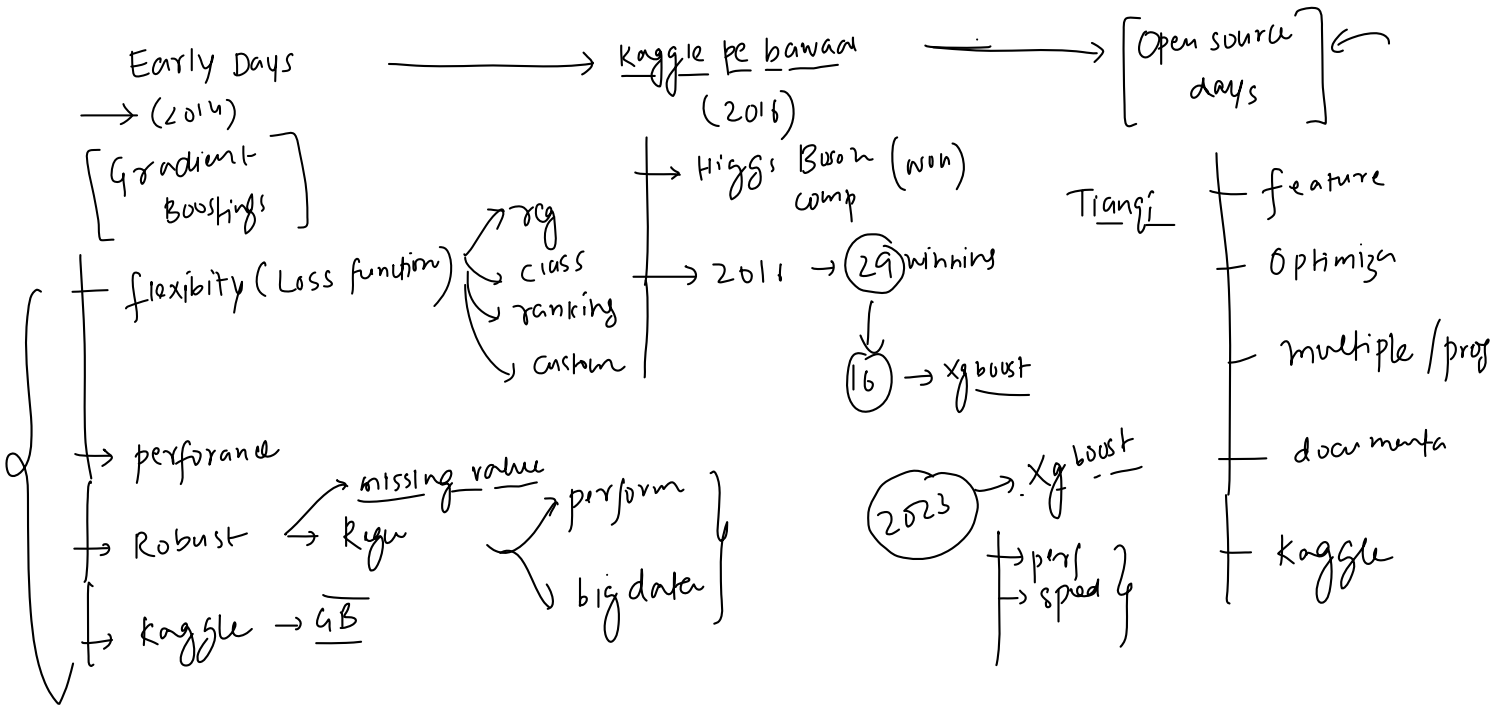
XGBoost Introduction

12 October 2023 09:38



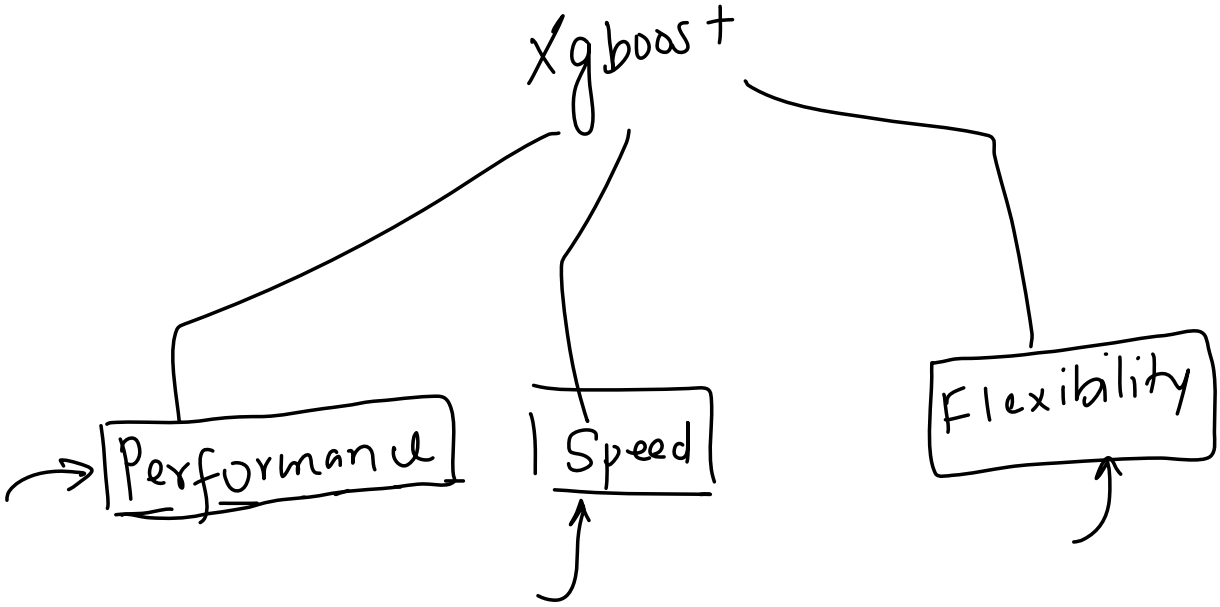
History

12 October 2023 11:01



XGBoost Features

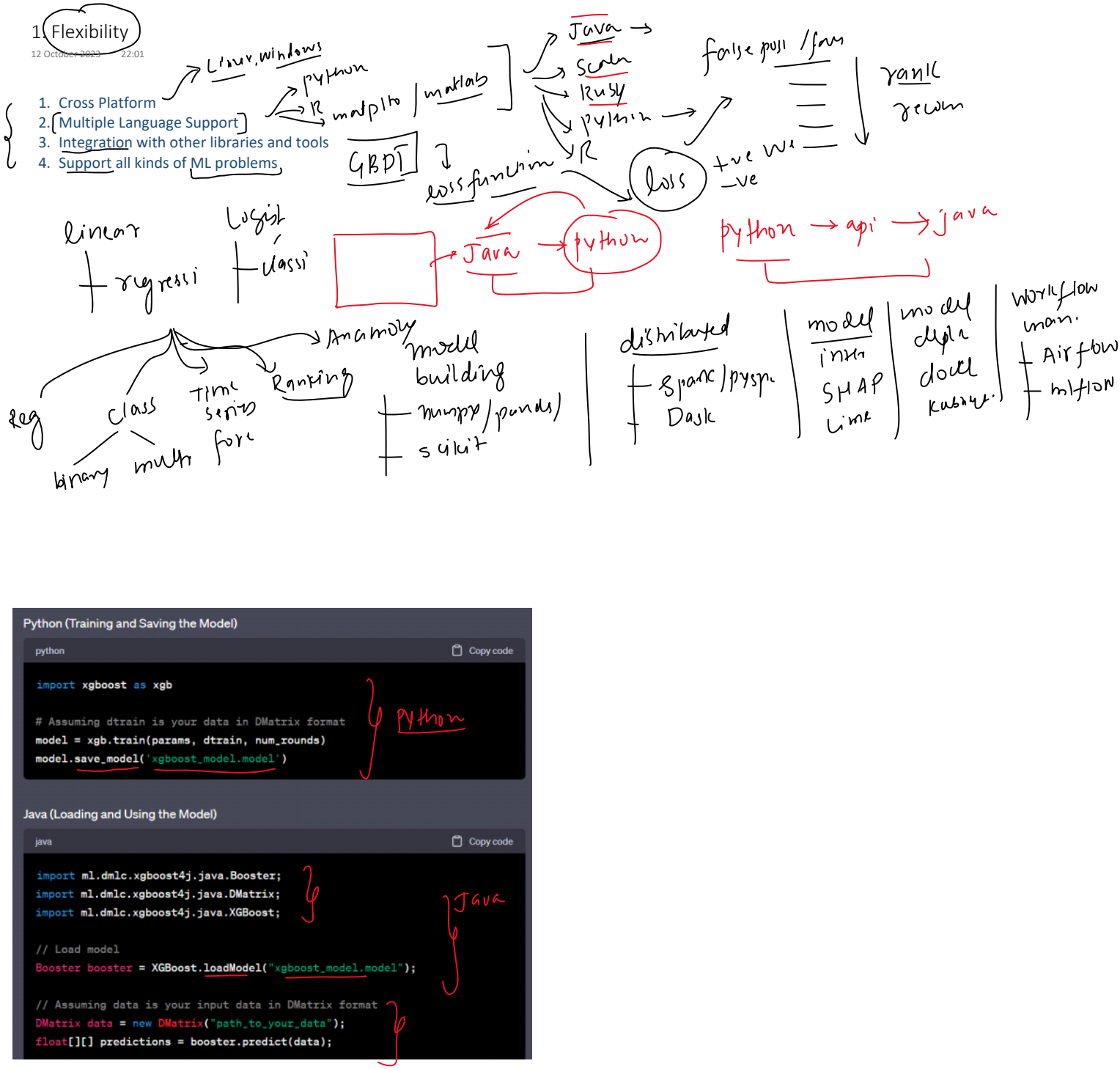
12 October 2023 21:55



1. Flexibility

12 October 2023 22:01

1. Cross Platform
2. Multiple Language Support
3. Integration with other libraries and tools
4. Support all kinds of ML problems



Python (Training and Saving the Model)

python — Copy code

```python
import xgboost as xgb

# Assuming dtrain is your data in DMatrix format
model = xgb.train(params, dtrain, num_rounds)
model.save_model('xgboost_model.model')
```

python

Java (Loading and Using the Model)

java — Copy code

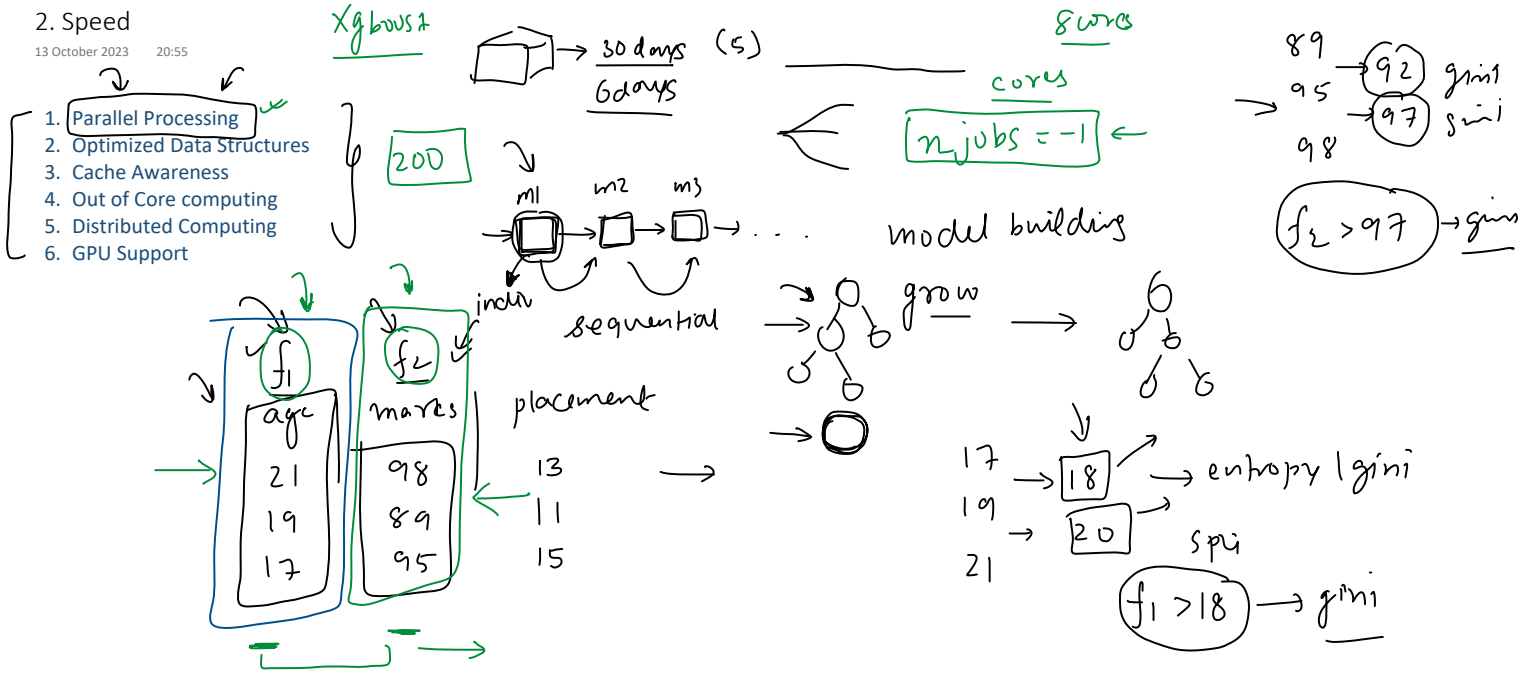
```java
import ml.dmlc.xgboost4j.java.Booster;
import ml.dmlc.xgboost4j.java.DMatrix;
import ml.dmlc.xgboost4j.java.XGBoost;

// Load model
Booster booster = XGBoost.loadModel("xgboost_model.model");

// Assuming data is your input data in DMatrix format
DMatrix data = new DMatrix("path_to_your_data");
float[][] predictions = booster.predict(data);
```

Java

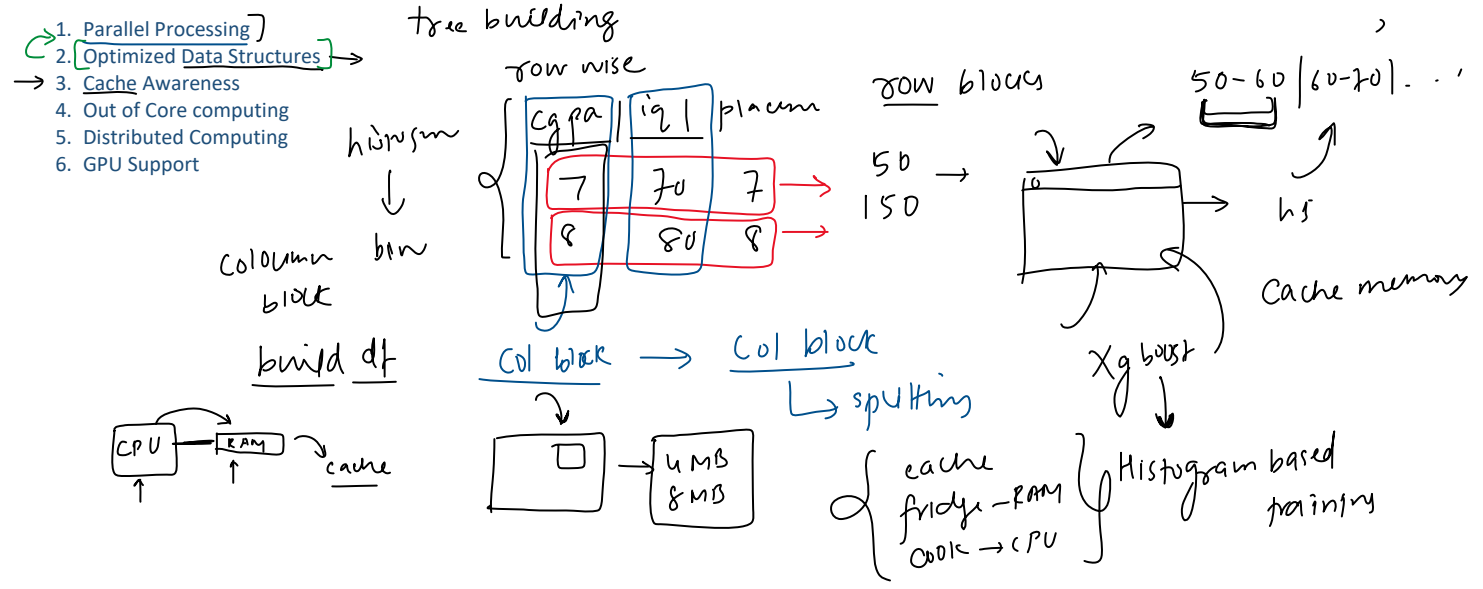
2. Speed

13 October 2023 20:55

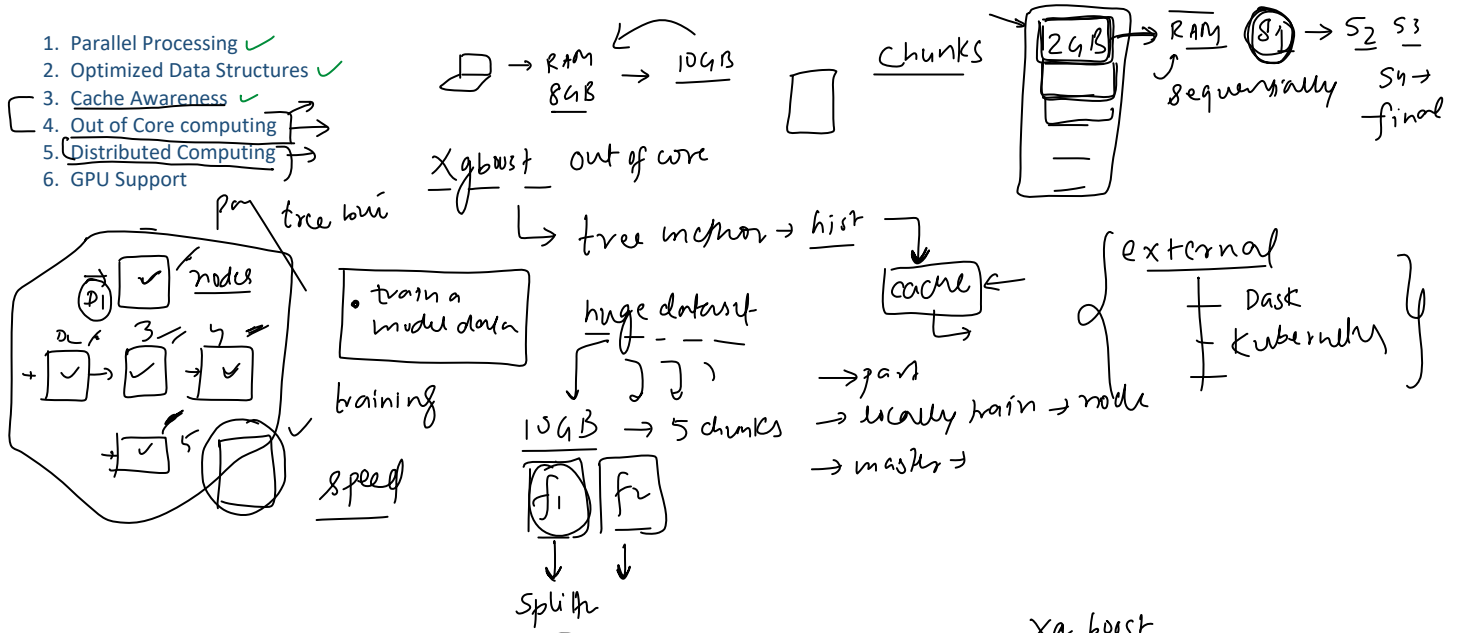
1. Parallel Processing
2. Optimized Data Structures
3. Cache Awareness
4. Out of Core computing
5. Distributed Computing
6. GPU Support



1. Parallel Processing
2. Optimized Data Structures
3. Cache Awareness
4. Out of Core computing
5. Distributed Computing
6. GPU Support



1. Parallel Processing
2. Optimized Data Structures
3. Cache Awareness
4. Out of Core computing
5. Distributed Computing
6. GPU Support



Split

Software → Extreme

Software

Xg boost

Extreme GB

1. Parallel Processing ✓
2. Optimized Data Structures ✓
3. Cache Awareness ✓
4. Out of Core computing ✓
5. Distributed Computing ✓
6. GPU Support →

CPU → data → CPU → GPU

↳ cores → less powerful

more in number

parallel

histogram split finding

tree_method: gpu_hist

Software env

ML → performance

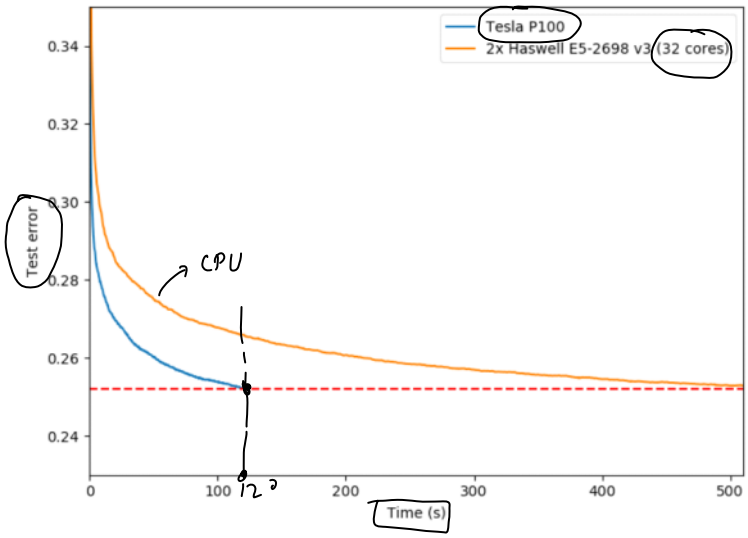


Figure 3. Test error over time for the Higgs dataset, 1000 boosting iterations.

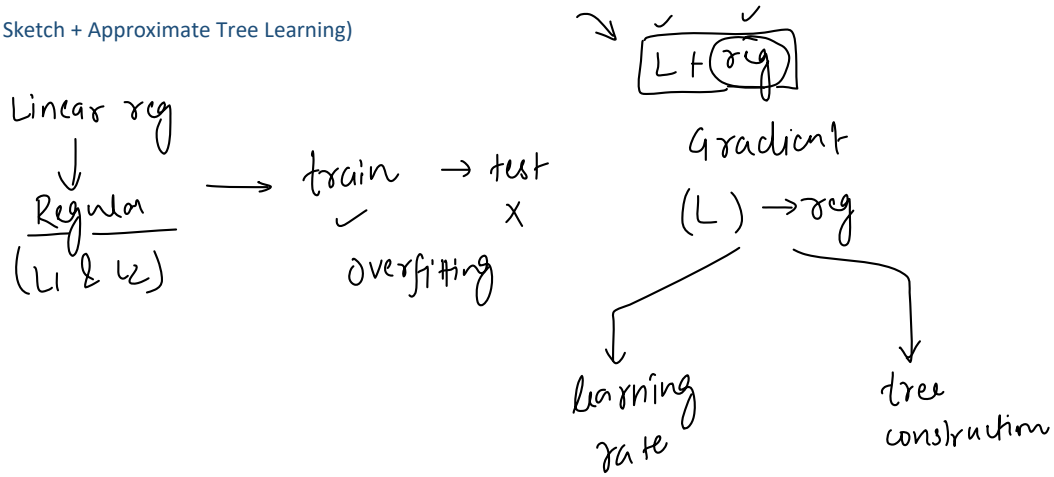
8. Performance →

14 October 2023 10:24

Xgboost

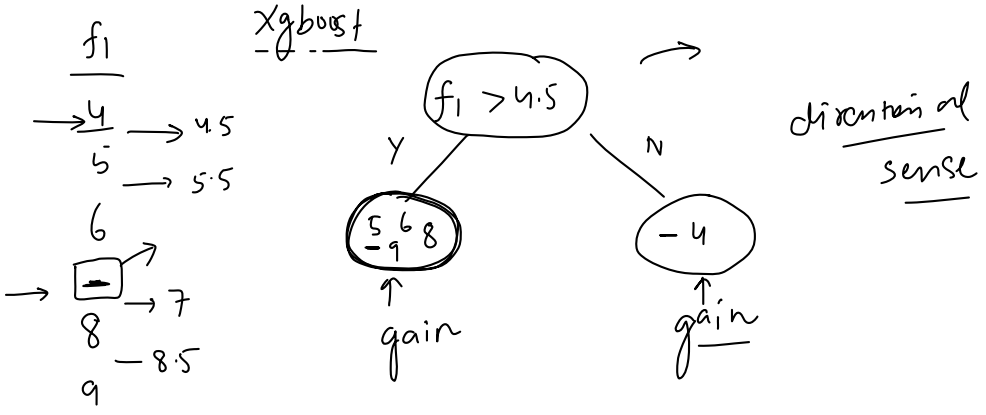
Xgboost → Loss function → min

1. Regularized Learning Objective
2. Handling Missing values
3. Sparsity Aware Split Finding
4. Efficient Split Finding(Weighted Quantile Sketch + Approximate Tree Learning)
5. Tree Pruning

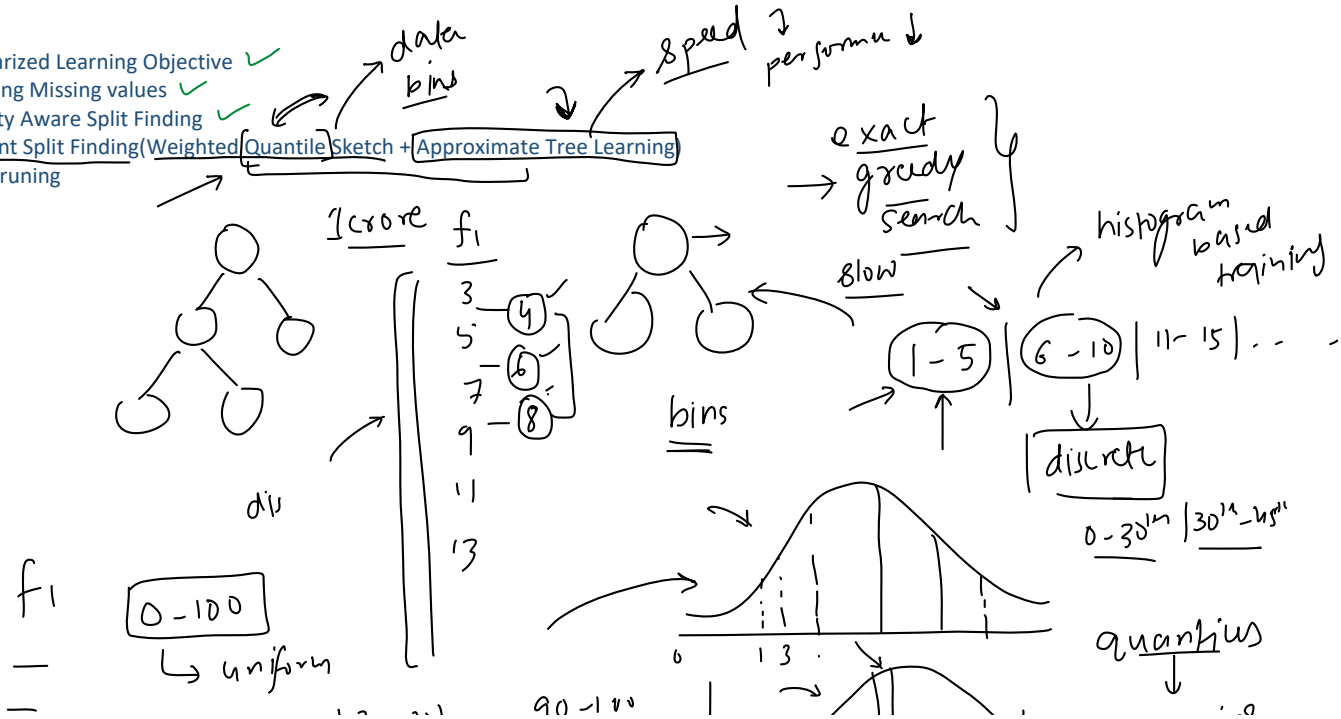


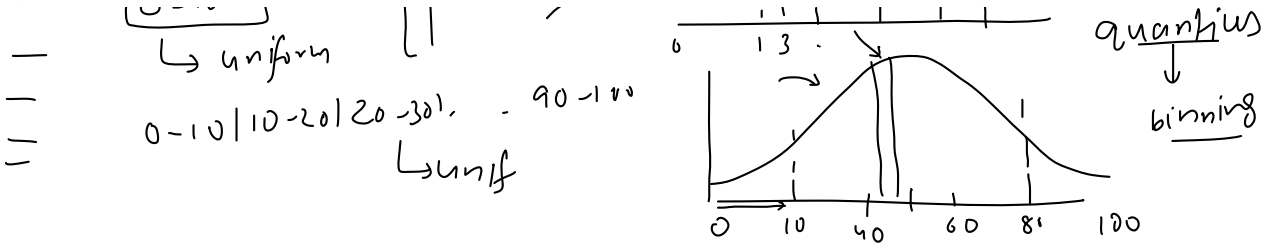
1. Regularized Learning Objective ✓
2. Handling Missing values
3. Sparsity Aware Split Finding →
4. Efficient Split Finding(Weighted Quantile Sketch + Approximate Tree Learning)
5. Tree Pruning

ml → missing
↳ prepro
↳ impute



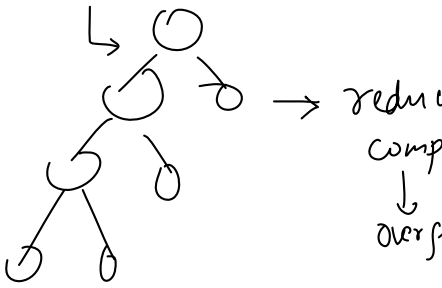
1. Regularized Learning Objective ✓
2. Handling Missing values ✓
3. Sparsity Aware Split Finding ✓
4. Efficient Split Finding(Weighted Quantile Sketch + Approximate Tree Learning)
5. Tree Pruning





GB

1. Regularized Learning Objective ✓
2. Handling Missing values ✓
3. Sparsity Aware Split Finding ✓
4. Efficient Split Finding(Weighted Quantile Sketch + Approximate Tree Learning) ✓
5. Tree Pruning → perform xgboost



→ reduce complexity → overfitting

Post pruning

pre pruning

XgBoost

prepruning

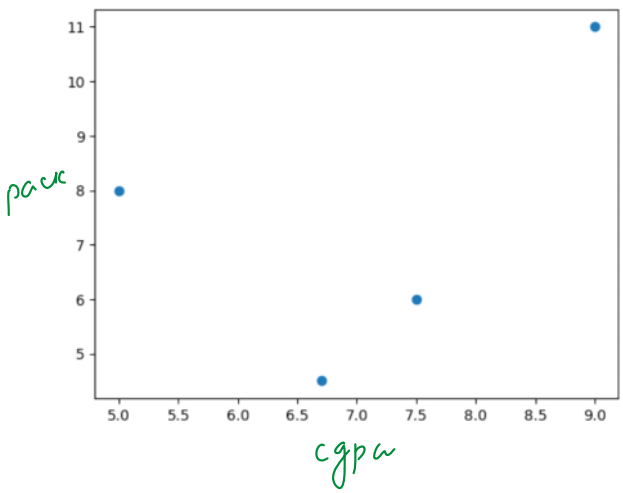
post

gamma value

The Problem Statement

23 October 2023 15:05

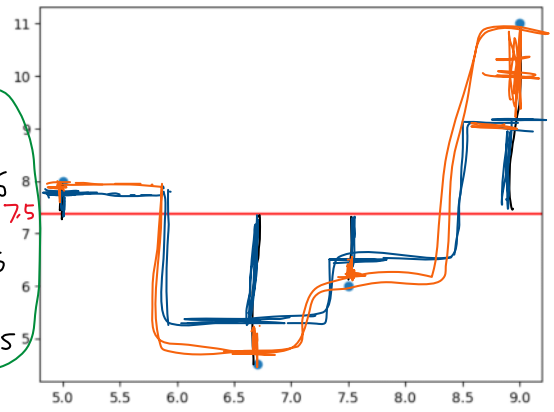
cgpa	package
6.7	4.5
9.0	11.0
7.5	6.0
5.0	8.0

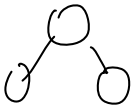


Solution Overview

23 October 2023 15:05

cgpa	package
6.7	4.5
9.0	11.0
7.5	6.0
5.0	8.0





pseudo-residual

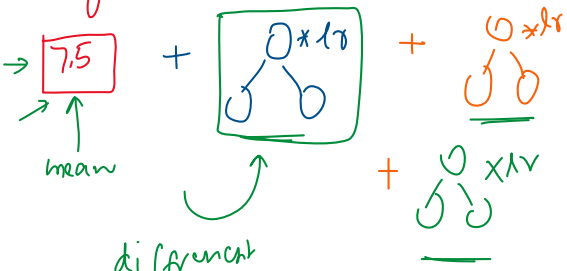
[perf / speed]

4.1 , 9.1

xgboost
↳ gradient boosting

base estimator → mean

stage 1 | stage II



different

GB
↳ vanilla dt → gini / entropy → diff criteria
tree is strong

Step by Step Calculation

23 October 2023 15:07

cgpa	package
6.7	4.5
9.0	11.0
7.5	6.0
5.0	8.0

7.3

stage 1 → mean

model 1	residual 1
7.3	-2.8
7.3	3.7
7.3	-1.3
7.3	0.7

decision tree

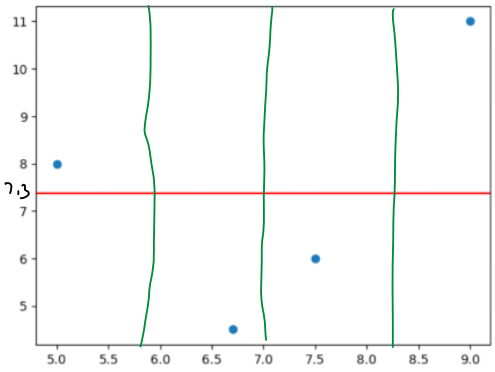
cgpa | residual1

-2.8, 3.7, -1.3, 0.7 → similarity score

0.02

$SS = \frac{(\text{sum of residuals})^2}{\#\text{residuals} + \lambda}$ → reg parameter, $\lambda = 0$

$\frac{(-2.8+3.7-1.3+0.7)^2}{4+0} = \frac{(0.3)^2}{4} = 0.02$



5.0, 6.7 → 5.85

7.5, 9.0 → 7.1, 8.25

cgpa	residual 1
5.0	0.7
6.7	-2.8
7.5	-1.3
9.0	3.7

Splitting criteria ① → 5.85

→ cgpa < 5.85 ✓ | ✗

0.7 | -2.8, -1.3, 3.7

$SS_L = \frac{(0.7)^2}{1+0} = 0.49$

$SS_R = \frac{(-2.8-1.3+3.7)^2}{3} = \frac{0.16}{3} = 0.05$

$gain = (SS_L + SS_R) - SS_{root}$

$= 0.49 + 0.05 - 0.02$

$= 0.52$ ✓

Splitting criteria ② → 7.1

→ cgpa < 7.1 → Similm

0.7, -2.8 | -1.3, 3.7

$SS_L = \frac{(0.7-2.8)^2}{2} = \frac{4.41}{2} = 2.20$

$SS_R = \frac{(-1.3+3.7)^2}{2} = \frac{5.76}{2} = 2.88$

$gain = SS_L + SS_R - SS_{root}$

$= 2.20 + 2.88 - 0.02$

$= 5.06$ ✓

Splitting criteria ③ → 8.25

cgpa < 8.25 ✓

0.7, -2.8, -1.3 | 3.7

$SS_L = \frac{(0.7-2.8-1.3)^2}{3} = \frac{11.56}{3} = 3.85$

$SS_R = \frac{(3.7)^2}{1} = 13.69$

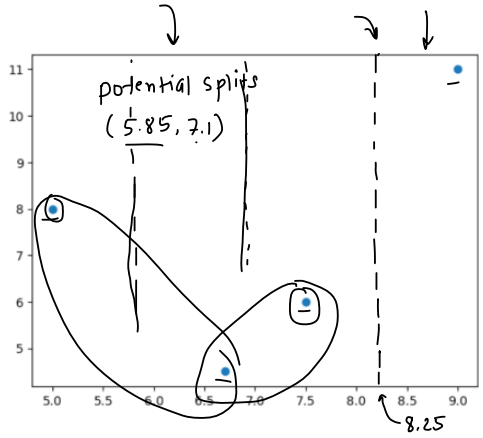
$gain = 3.85 + 13.69 - 0.02$

$= 17.52$

Stage 2 → done

→ cgpa < 8.25

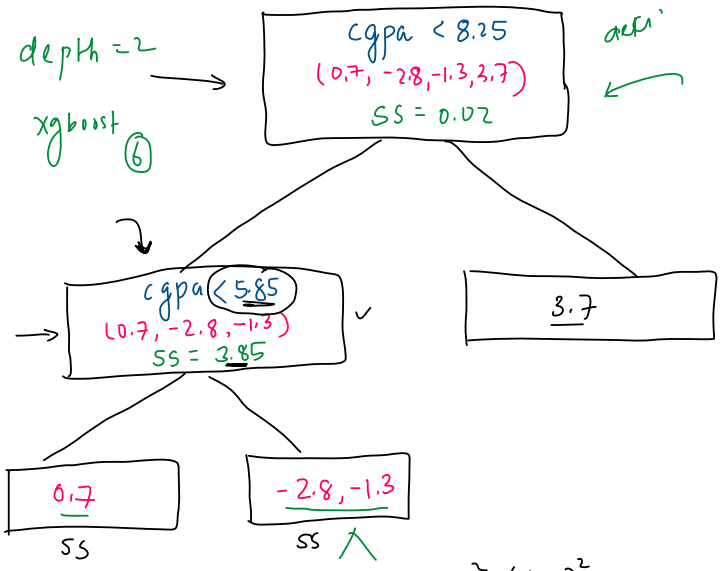
→ 0.7, -2.8, -1.3 | 3.7 ✗ ←



① Splitting Criteria → 5.85

depth = 2

xgboost ⑥



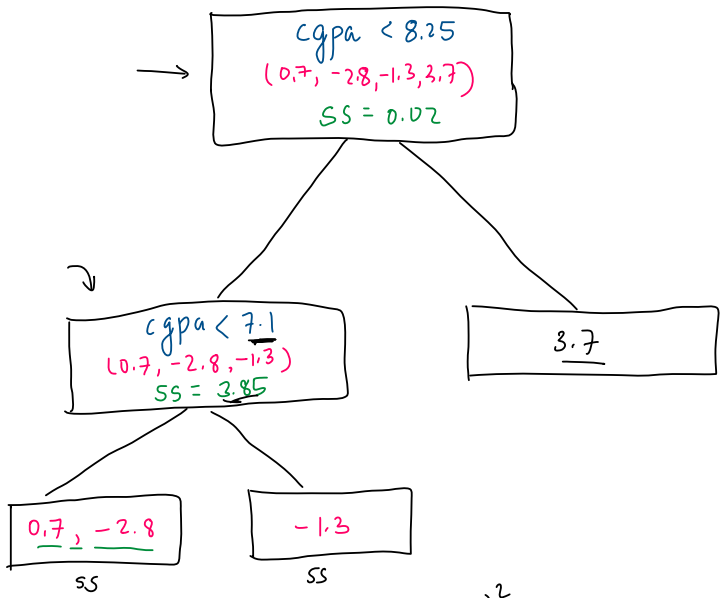
$SS_L = \frac{(0.7)^2}{1} = 0.49$

$SS_R = \frac{(-2.8-1.3)^2}{2} = \frac{(-4.1)^2}{2} = 8.40$

$gain = (SS_L + SS_R - SS_P) = 0.49 + 8.40 - 3.85$

$= \boxed{5.04}$ ✓

② Spitting criteria → 7.1



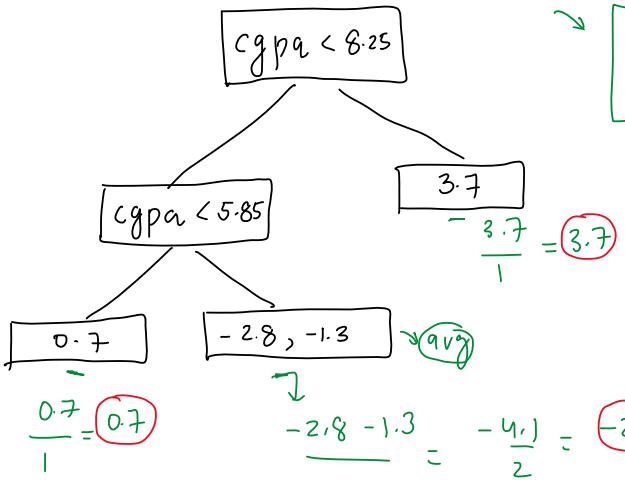
$SS_L = \frac{(0.7-2.8)^2}{2} = \frac{(-2.1)^2}{2} = 2.20$

$SS_R = \frac{(-1.3)^2}{1} = 1.69$

$gain = 2.20 + 1.69 - 3.85$

$= \boxed{0.04}$ — ✗

The final decision tree



$\boxed{\text{Output} = \frac{\text{sum of residuals}}{\#\text{residuals} + \lambda}}$

$\lambda = 0 \rightarrow \lambda = 0$ ↳ avg

$\frac{3.7}{1} = 3.7$

$\frac{0.7}{1} = 0.7$

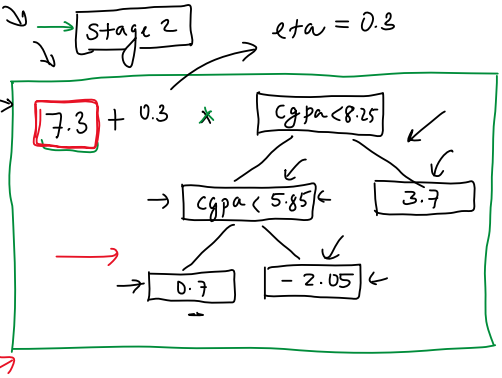
$\frac{-2.8 - 1.3}{2} = \frac{-4.1}{2} = -2.05$

cgpa	package	model 1	residual 1	model 2	residual 2
6.7	4.5	7.3	-2.8	6.69	-2.19
9.0	11.0	7.3	3.7	8.41	2.59
7.5	6.0	7.3	-1.3	6.69	-0.69
5.0	8.0	7.3	0.7	7.51	0.49

→ SS

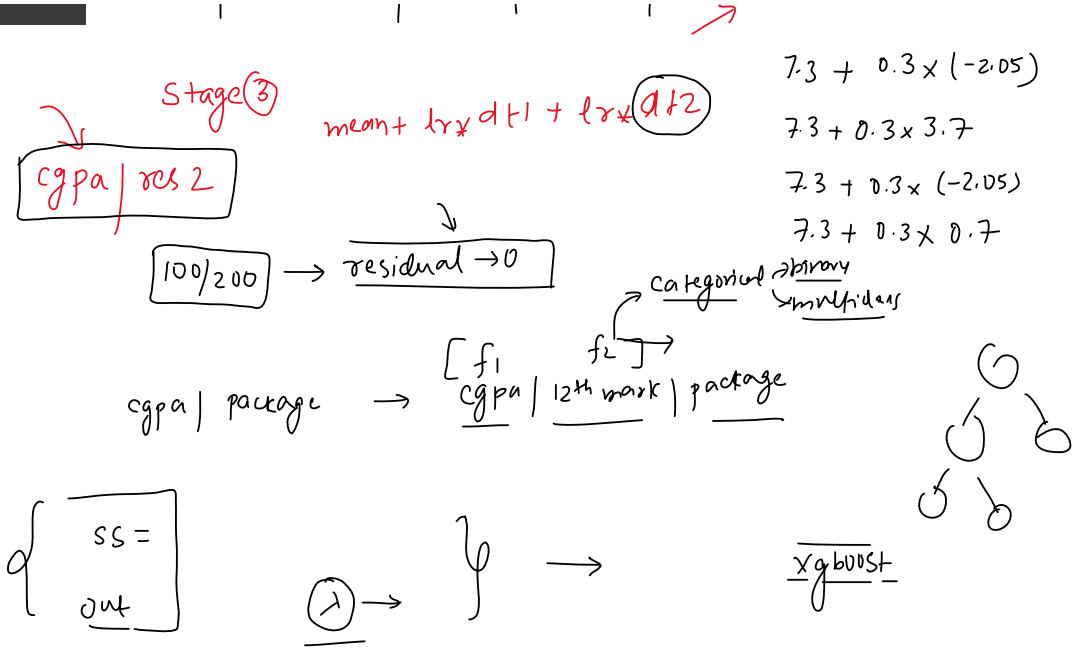
Stage 2 ← 0

eta = 0.3



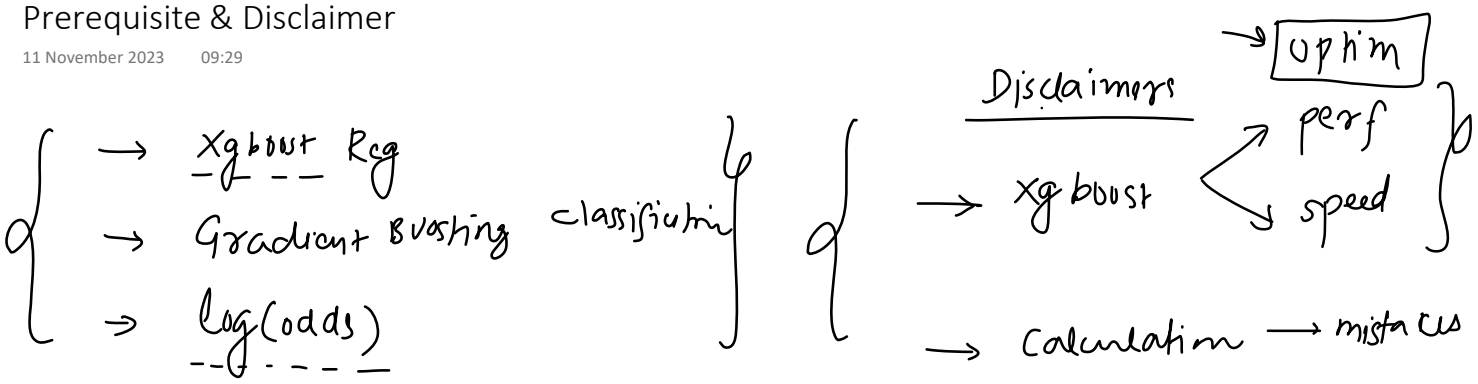
$7.3 + 0.3 \times (-2.05)$

Stage (3)


Stage 3
mean + lr * dt1 + lr * dt2
cgpa | res 2
100/200
residual → 0
7.3 + 0.3 x (-2.05)
7.3 + 0.3 x 3.7
7.3 + 0.3 x (-2.05)
7.3 + 0.3 x 0.7
categorical
binary
multiclass
cgpa | package
[f1 f2]
cgpa | 12th mark | package
SS =
out
λ
XgBoost

Prerequisite & Disclaimer

11 November 2023 09:29



Problem Statement

11 November 2023 09:29

cgpa	placed?
5.70	0
6.25	1
7.10	0
8.15	1
9.60	1

xgboost → classification

6.61 → 1,0

→ overview
→ step by step

Overview

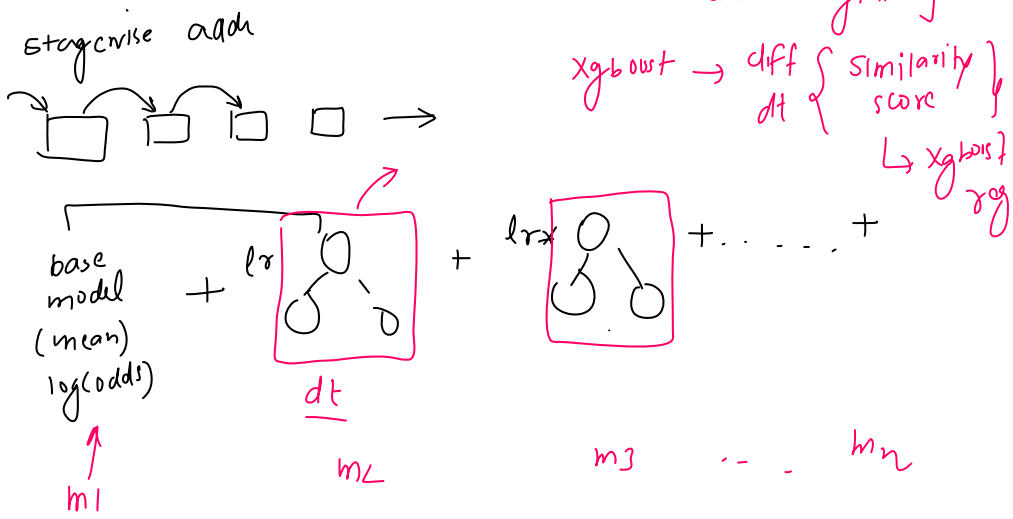
11 November 2023 09:31

cgpa	placed?
5.70	0
6.25	1
7.10	0
8.15	1
9.60	1

xgboost → GBDT →

GBDT → Vanilla dt { entropy / gini }

xgboost → diff dt { similarity score } → xgboost reg



Step by Step Solution

11 November 2023 17:30

Stage 1 — □ → log(odds) → prob

cgpa	placed?	pred1(lo)
5.70	0	0.405 ✓
6.25	1	0.405
7.10	0	0.405
8.15	1	0.405
9.60	1	0.405

base estimator → mean

$\log(\text{odds}) \rightarrow \log\left(\frac{p}{1-p}\right)$ $p \rightarrow$ prob +ve class (1)

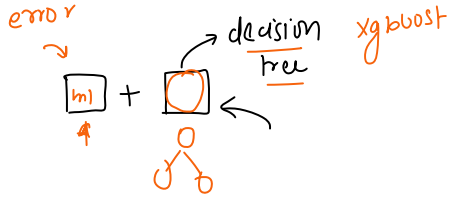
$\log\left(\frac{3/5}{2/5}\right) = \log_e\left(\frac{3}{2}\right) = 0.405$

$p = \frac{e^{\log odds}}{1+e^{\log odds}} = \frac{e^{0.405}}{1+e^{0.405}} = 0.60$

cgpa	placed?	pred1(lo)	pred1(prob)	res1
5.70	0	0.405 ✓	0.6	−0.6
6.25	1	0.405	0.6	0.4
7.10	0	0.405	0.6	−0.6
8.15	1	0.405	0.6	0.4
9.60	1	0.405	0.6	0.4

stage 1 → errors

pseudo residual



split → splitting criterion ←

	cgpa	res1
① 5.97	5.70	−0.6 ←
	6.25	0.4
② 6.67	7.10	−0.6
③ 7.62	8.15	0.4
④ 8.87	9.60	0.4

leaf node → [−0.6, 0.4, −0.6, 0.4, 0.4] → similarity score

SS = 0 ← maximize

$SS = \left(\sum \text{residual}_i\right)^2 \rightarrow$

GBDT Classification

$\frac{\sum \text{prev\_prob}_i (1-\text{prev\_prob}_i) + \lambda}{}$ → reg λ = 0

$= \frac{(-0.6+0.4-0.6+0.4+0.4)^2}{5 \times [0.6(1-0.6)] + 0} = 0$

gain → 1.67

	cgpa	res1
5.97	5.70	−0.6 ←
	6.25	0.4
6.67	7.10	−0.6
7.62	8.15	0.4
8.87	9.60	0.4

Splitting criteria = 5.97

cgpa < 5.97 ✓

−0.6	0.4, −0.6, 0.4, 0.4
SS_L	SS_R
$\frac{(-0.6)^2}{0.6(1-0.6)}$	$\frac{(0.4-0.6+0.4+0.4)^2}{4 \times 0.6 \times 0.4}$

Splitting criteria = 6.67

cgpa < 6.67 ✗

−0.6, 0.4	−0.6, 0.4, 0.4
SS_L	
$\frac{(-0.2)^2}{2[0.6 \times 0.4]}$	$\frac{(0.2)^2}{3 \times 0.24}$

8.15 | 0.4

8.87 { 9.60 | 0.4

$\frac{(-0.6)^2}{0.6(1-0.6)} = \frac{0.36}{0.24} = 1.5$

$\frac{0.36}{4 \times 0.6 \times 0.4} = \frac{1.5}{4} = 0.37$

$\frac{}{2[0.6 \times 0.4]} = \frac{0.04}{2 \times 0.24} = \frac{0.04}{0.48} = 0.08$

$\frac{}{3 \times 0.24} = \frac{0.04}{3 \times 0.24} = 0.05$

gain = (SS_L + SS_R) − SS_root

= (1.5 + 0.37) − 0

= 1.87

gain = 0.08 + 0.05 − 0

= 0.13

Splitting criteria = 7.62

cgpa < 7.62

−0.6, 0.4, −0.6 | 0.4, 0.4

$\frac{(-0.8)^2}{3 \times 0.24} = 0.88$

$\frac{(0.8)^2}{2 \times 0.24} = 1.33$

gain = 0.88 + 1.33 − 0

= 2.22 →

Splitting criteria = 8.87

cgpa < 8.87

−0.6, 0.4, −0.6, 0.4 | 0.4

cgpa	res1
5.70	−0.6 ←
6.25	0.4
7.10	−0.6
→ 7.62	
8.15	0.4
8.87	0.4
9.60	0.4

5.97 = (5.70 + 6.25)/2

6.67 = (6.25 + 7.10)/2

7.62 = (7.10 + 8.15)/2

8.87 = (8.15 + 9.60)/2

Stage 2 model

cgpa < 7.62

−0.6, 0.4, −0.6 (−1.11) | 0.4, 0.4 (1.66)

output →

$\frac{-0.6 + 0.4 - 0.6}{3 \times 0.6 \times 0.4} = \frac{-0.8}{3 \times 0.24} = -1.11$

$\frac{0.4 + 0.4}{2 \times 0.24} = 1.66$

output = $\frac{\sum \text{residual}_i}{\sum \text{prev\_prob}_i (1 - \text{prev\_prob}_i) + \lambda = 0}$

stage 1 → prediction

cgpa	placed?	pred1(lo)	pred1(prob)	res1	pred2(lo)	pred2(prob)	res2
→ 5.70	0	0.405	0.6	−0.6	0.072	0.518	−0.51
→ 6.25	1	0.405	0.6	0.4	0.072	0.518	0.48
→ 7.10	0	0.405	0.6	−0.6	0.072	0.518	−0.51
→ 8.15	1	0.405	0.6	0.4	0.903	0.712	0.28
→ 9.60	1	0.405	0.6	0.4	0.903	0.712	0.28

Stage 2 model

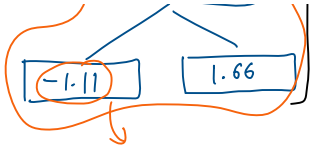
Similar

$\overline{m1}$ + 0.3 × $\overline{m2}$

(log odds) (decision tree)

0.405 + 0.3 × [cgpa < 7.62 → −1.11 | 1.66]

8.15	1	0.405	0.6	0.4	0.903	0.712	0.28
9.60	1	0.405	0.6	0.4	0.903	0.712	0.28



$0.405 + 0.3 \times (-1.11)$

$0.405 + 0.3 \times (-1.11)$

$0.405 + 0.3 \times 1.66$

log(odds) → prob

$\frac{e^{\log(odds)}}{1 + e^{\log(odds)}}$

m3 → decision

same flow

Stage 3

$[m1 + 0.3 \times m2 + 0.3 \times m3]$ → pred → prob → res3

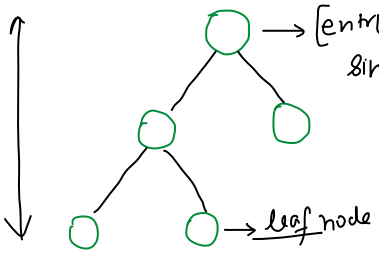
cgpa/res3 → my decision

next video

→ maths → xgboost form

Plan of Attack

30 November 2023 10:42



Regression

Classification

$$\left[\text{Similarity Score} = \frac{\left(\sum \text{Residuals}\right)^2}{N + \lambda}\right] \text{①}$$

$$\left[\text{Similarity Score} = \frac{\left(\sum \text{Residuals}\right)^2}{\sum_{N}[P\,(1 - P)] + \lambda}\right] \text{②}$$

reg

$$\left[\text{Output value} = \frac{\sum \text{Residuals}}{N + \lambda}\right] \text{③}$$

$$\left[\text{Output Value} = \frac{\left(\sum \text{Residuals}\right)}{\sum_{N}[P\,(1 - P)] + \lambda}\right] \text{④}$$

1 intro

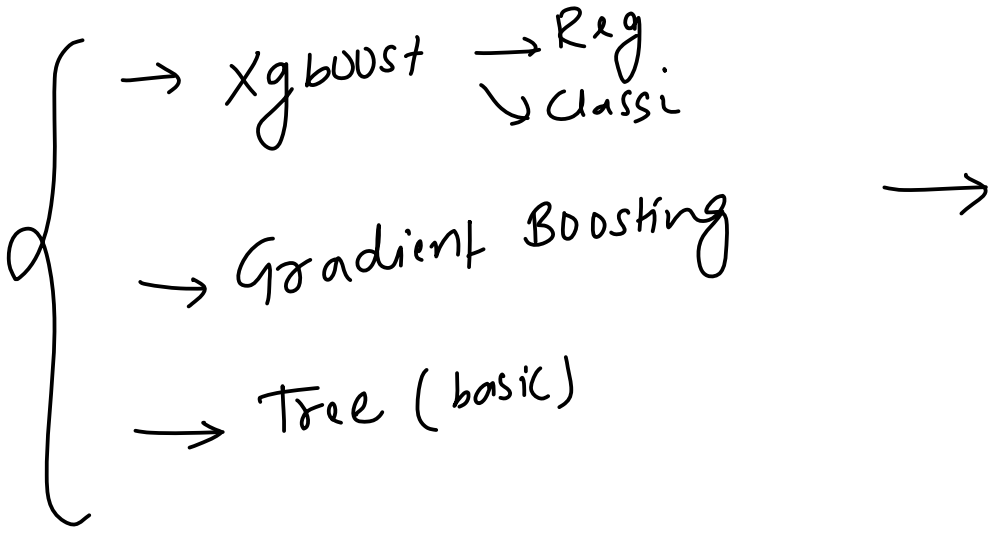
2 regression

3 classification

deep dive → goal

Prerequisite

30 November 2023 10:44

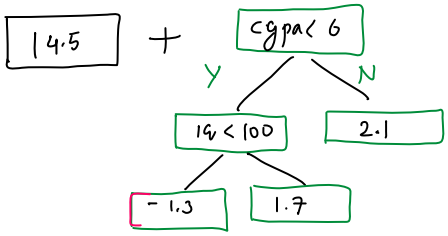


Boosting as an Additive Model

30 November 2023 12:04

Xgboost → 2 model

cgpa	iq	package
6.5	71	13
5	91	10
9.1	120	16



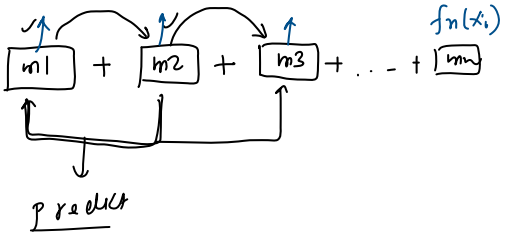
Xgboost

train

sum leaf weight

$m_1 + m_2$

$f_n(x_i)$



[reg]

output = $\frac{\text{sum of residuals}}{N + \lambda}$

$\{g, \log y\}$

$\hat{y}_i = 14.5 + 2.1$

$= 16.6$ lpa

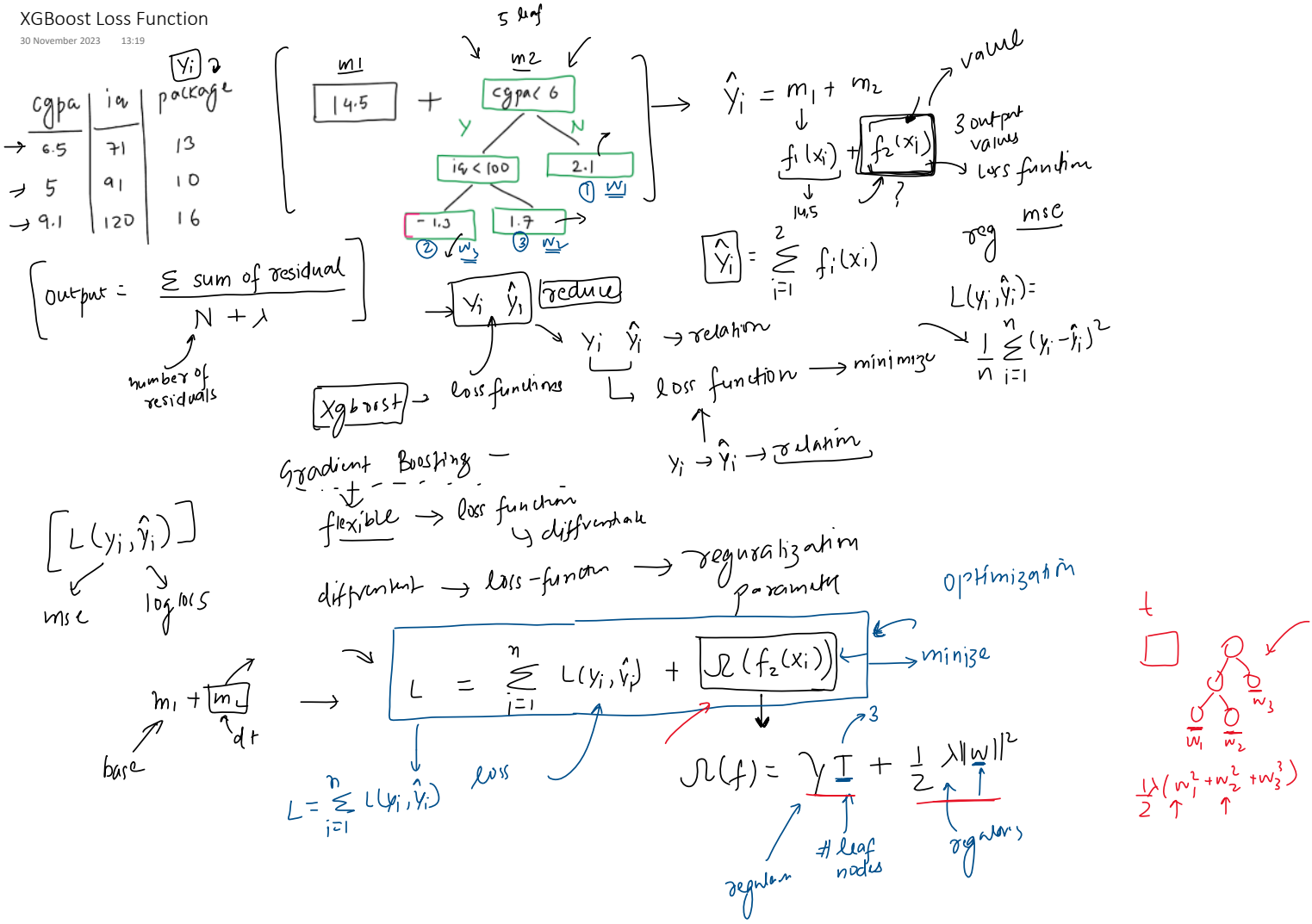
$\hat{y}_i = f_1(x_i) + f_2(x_i) + \ldots + f_n(x_i)$

$\hat{y}_i = \sum_{i=1}^{n} f_i(x_i)$ → Xgboost

1 + 1 — origin story

XGBoost Loss Function

30 November 2023 13:19



The Derivation

30 November 2023 13:57

objective func

$$\mathcal{L} = \underbrace{\sum_{i=1}^{n} L(y_i, \hat{y}_i)}_{\text{loss function}} + \underbrace{\Omega(f_k)}_{\text{regular.}} \leftarrow \gamma T + \frac{1}{2}\lambda \|w\|^2$$

T: leaf, $\|w\|^2$: leaf w

↷ boosting → mean ▢ ▢ ▢

↓

xgboost → 1 model ▢ → objective

Stage I

mean

m1

$f_1(x_i)$ → $\mathcal{L}^{(1)} = \sum_{i=1}^{n} L(y_i, \hat{y}_i)$

Stage II

mean

m1

$f_1(x_i)$

decision

m2

$f_2(x_i)$ → $\mathcal{L}^{(2)} = \sum_{i=1}^{n} L(y_i, f_1(x_i) + f_2(x_i)) + \Omega(f_2(x_i))$ w

f_2 $f_1 X$

$\hat{y}_i = f_1(x_i) + f_2(x_i)$

Stage III

mean

m1

$f_1(x_i)$

m2

$f_2(x_i)$

m3

$f_3(x_i)$ → $\mathcal{L}^{(3)} = \sum_{i=1}^{n} \left(y_i, f_1(x_i) + f_2(x_i) + f_3(x_i) \right) + \Omega(f_3(x_i))$

→ $f_1(x_i) + f_2(x_i) + \dots + f_{t-1}(x_i) \rightarrow \hat{y}_i^{<t-1>}$

Stage t

mean

m1

$f_1(x_i)$

m2

$f_2(x_i)$

. . . .

mt

$f_t(x_i)$ → $\mathcal{L}^{<t>} = \sum_{i=1}^{n} L(y_i, f_1(x_i) + f_2(x_i) + \dots + f_t(x_i)) + \Omega(f_t(x_i))$

last decision

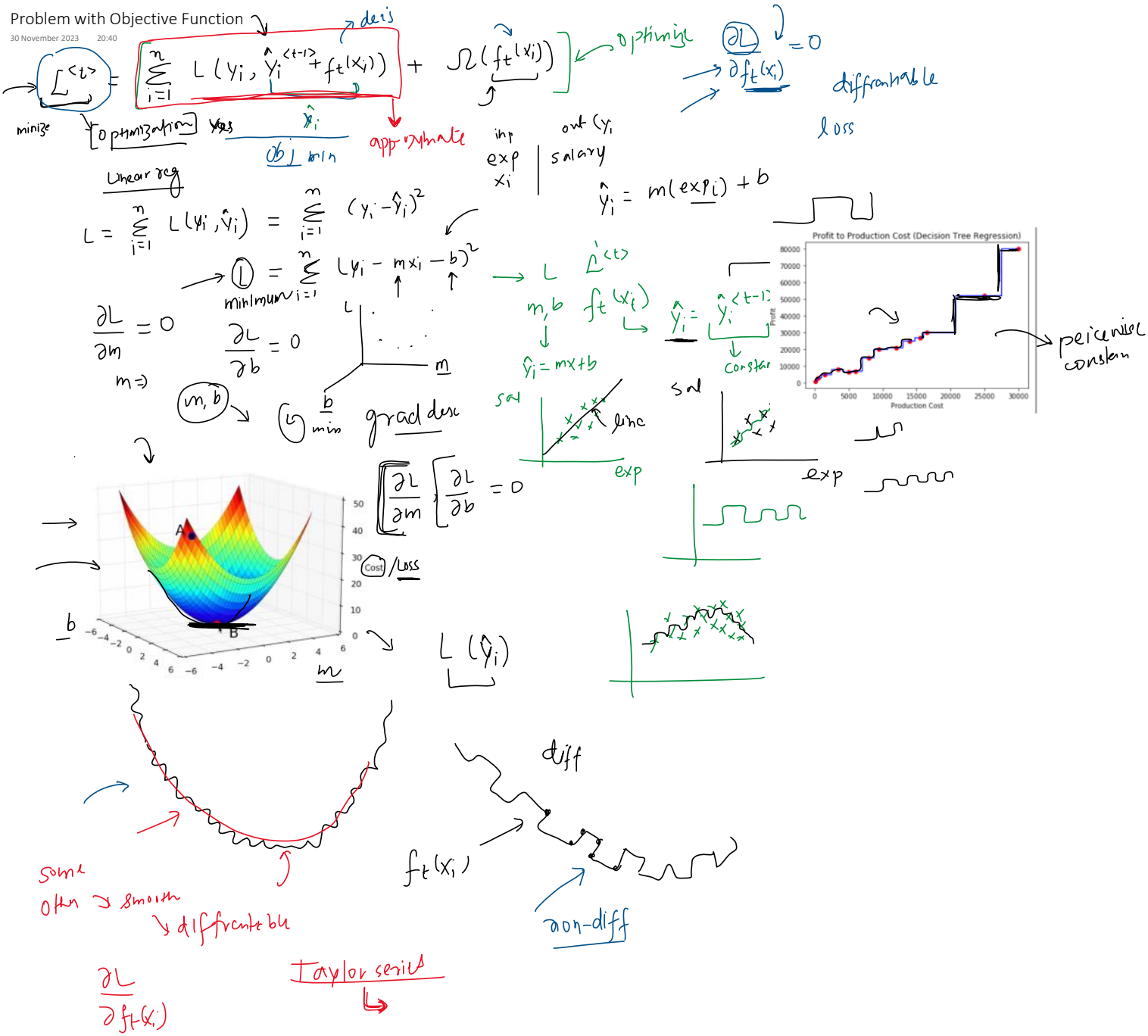
$\mathcal{L}^{<t>} = \sum_{i=1}^{n} L(y_i, \hat{y}_i^{<t-1>} + f_t(x_i)) + \Omega(f_t)$

At any stage t

$$\mathcal{L}^{<t>} = \sum_{i=1}^{n} L(y_i, \hat{y}_i^{<t-1>} + f_t(x_i)) + \Omega(f_t(x_i))$$

Problem with Objective Function

30 November 2023 20:40



The Solution - Taylor Series

01 December 2023 00:02

powerful approx techqi

complex function → approx → polynomial

ex → accurate approx → $3x^2+2$

$3x^2+3x^3+x^4+x^5\ldots$

Objective → usable

↓

Taylor series

How

ex → polynomial

→ derivatives

ex → e^x

$f(x)$ point → a

$$f(x) \simeq f(a) + \frac{f'(a)}{1!}(x-a) + \frac{f''(a)}{2!}(x-a)^2 + \frac{f'''(a)}{3!}(x-a)^3 + \ldots$$

$f(x) = e^x$ $a=0$

$f(a) = e^a = 1$

$\frac{f'(a)}{1!}(x-a) = \frac{e^a}{1}(x-a) = e^0(x-0) = x$

$\frac{f''(a)}{2!}(x-a)^2 = \frac{e^a}{2!}(x-a)^2 = \frac{e^0(x)^2}{2!} = \frac{x^2}{2}$

$\frac{f'''(a)}{3!}(x-a)^3 = \frac{e^a}{3!}(x-a)^3 = \frac{1}{6}x^3 = \frac{x^3}{6}$

$e^x \simeq 1 + x + \frac{x^2}{2} + \frac{x^3}{6} + \ldots$

Applying Taylor Series

01 December 2023 00:51

$$f(x) \simeq f(a) + f'(a)(x-a) + \frac{f''(a)}{2!}(x-a)^2 \ldots\ldots \Big]$$

$\frac{df}{dx} =$ $\quad f(x) = x^2$

$f'(x)$ $\frac{\partial f}{\partial x} = 2x$

$f'(a) = 2a$

$$\rightarrow \mathcal{L}^{<t>} = \left[\sum_{i=1}^{n} L\left(y_i, \hat{y}_i^{<t-1>} + f_t(x_i)\right)\right] + \Big(\Omega(f_t(x_i))\Big) \rightarrow 2^{nd}\ \text{degree}$$

$x \rightarrow \hat{y}_i^{<t-1>} + f_t(x_i) \quad a \rightarrow \hat{y}_i^{<t-1>}$

$$f(a) = \sum_{i=1}^{n} L\left(y_i, \hat{y}_i^{<t-1>}\right) \quad \rightarrow \text{gradient } g_i$$

$$f'(a)(x-a) = \sum_{i=1}^{n} \frac{\partial L\left(y_i, \hat{y}_i^{<t-1>}\right)}{\partial \hat{y}_i^{<t-1>}} f_t(x_i)$$

$\rightarrow$ hessian h_i

$$\frac{f''(a)}{2}(x-a)^2 = \frac{1}{2}\sum_{i=1}^{n} \frac{\partial^2 L\left(y_i, \hat{y}_i^{<t-1>}\right)}{\partial \hat{y}_i^{<t-1>2}} f_t^2(x_i)$$

$$\mathcal{L}^{<t>} = \sum_{i=1}^{n} L\left(y_i, \hat{y}_i^{<t-1>}\right) + g_i f_t(x_i) + \frac{1}{2} h_i f_t^2(x_i) + \Omega(f_t(x_i))$$

Simplification

01 December 2023 01:29

$$\mathcal{L}^{(t)} \simeq \sum_{i=1}^{n}\left[l(y_i, \hat{y}_i^{(t-1)}) + g_i f_t(x_i) + \frac{1}{2} h_i f_t^2(x_i)\right] + \Omega(f_t(x_i))$$

$l(y_i, \hat{y}_i^{(t-1)})$ → 0

$$\mathcal{L}^{(t)} = \sum_{i=1}^{n}\left[g_i f_t(x_i) + \frac{1}{2} h_i f_t^2(x_i)\right] + \Omega(f_t(x_i))$$ → expand

+ leaf nodes

$$\mathcal{L}^{(t)} = \sum_{i=1}^{n}\left[g_i f_t(x_i) + \frac{1}{2} h_i f_t^2(x_i)\right] + \gamma T + \frac{1}{2}\lambda \sum_{j=1}^{T} w_j^2$$

$i = 1 \to n$ #rows

$j = 1 \to T$, $T \to$ #leaf nodes

T = 2, j = 1, 2

$$\mathcal{L}^{(t)} = \sum_{j=1}^{T}\left[\sum_{i\in I_j} g_i w_j + \frac{1}{2}\sum_{i\in I_j} h_i w_j^2\right] + \gamma T + \frac{1}{2}\lambda \sum_{j=1}^{T} w_j^2$$

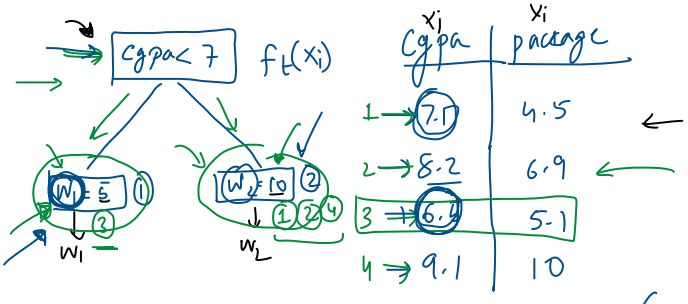
$w_j \to f_t(x_i)$

$\sum_{i=1}^{n} \to \sum_{j=1}^{T}\sum_{i\in I_j}$

$$\mathcal{L}^{(t)} = \sum_{i=1}^{n}\left[g_i f_t(x_i) + \frac{1}{2} h_i f_t^2(x_i)\right] + \gamma T + \frac{1}{2}\lambda \sum_{j=1}^{T} w_j^2$$

100

$$\mathcal{L}^{(t)} = \sum_{j=1}^{T}\left[\sum_{i\in I_j} g_i w_j + \frac{1}{2}\sum_{i\in I_j} h_i w_j^2\right] + \gamma T + \frac{1}{2}\lambda \sum_{j=1}^{T} w_j^2$$



$\sum_{i=1}^{n} \to \sum_{j=1}^{T}\sum_{i\in I_j}$

$i \in I_j$ → instance set of leaf j

2 times

$f_t(x_i) \to w_j$

$f_t(x_3)$ decision

4 terms

$$\to \left(g_1 f_t(x_1) + \frac{1}{2} h_1 f_t^2(x_1)\right) + \left(g_2 f_t(x_2) + \frac{1}{2} h_2 f_t^2(x_2)\right) + \ldots + g_4 f_t(x_4) + \frac{1}{2} h_4 f_t^2(x_4)$$

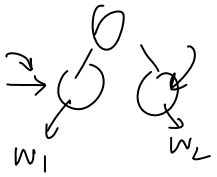
j = 1, j = 2

$$\to \underbrace{g_3 f_t(x_3) + \frac{1}{2} h_3 f_t^2(x_3)}_{g_3 w_1 + \frac{1}{2} h_3 w_1^2} + \underbrace{g_1 f_t(x_1) + \frac{1}{2} h_1 f_t^2(x_1)}_{g_1 w_2 + \frac{1}{2} h_1 w_2^2} + g_2 f_t(x_2) + \frac{1}{2} h_2 f_t^2(x_2) + g_4 f_t(x_4) + \frac{1}{2} h_4 f_t^2(x_4)$$

$$\mathcal{L}^{(t)} = \sum_{j=1}^{T}\left[\sum_{i\in I_j} g_i w_j + \frac{1}{2}\sum_{i\in I_j} h_i w_j^2\right] + \gamma T + \frac{1}{2}\lambda \sum_{j=1}^{T} w_j^2$$

$$\mathcal{L}^{(t)} = \sum_{j=1}^{T}\left[\sum_{i\in I_j} g_i w_j + \frac{1}{2}\sum_{i\in I_j} h_i w_j^2 + \frac{1}{2}\lambda w_j^2\right] + \gamma T$$

$$\mathcal{L}^{(t)} = \sum_{j=1}^{T}\left[\sum_{i\in I_j} g_i w_j + \frac{1}{2}\left(\sum_{i\in I_j} h_i + \lambda\right) w_j^2\right] + \gamma T$$



$$\mathcal{L}^{(t)} = \sum_{j=1}^{T} \left[\sum_{i \in I_j} g_i w_j + \frac{1}{2} \left(\sum_{i \in I_j} h_i + \lambda \right) w_j^2 \right] + \gamma T$$

w_1 w_2

$$\frac{\partial \mathcal{L}}{\partial w_j}$$

$$\frac{\partial \mathcal{L}^{(t)}}{\partial w_j} = \sum_{i \in I_j} g_i + \left(\sum_{i \in I_j} h_i + \lambda \right) w_j = 0$$

$$w_j = - \frac{\sum_{i \in I_j} g_i}{\sum_{i \in I_j} h_i + \lambda}$$

$g_i / h_i \rightarrow$ formula the output

Output Value for Regression

01 December 2023 13:45

$$w_j = \frac{-\sum_{i \in I_j} g_i}{\sum_{i \in I_j} h_i + \lambda}$$ → gradient; ↳ hessian

$$g = \frac{\partial L(y_i, \hat{y}_i^{(t-1)})}{\partial \hat{y}_i^{(t-1)}} \qquad h = \frac{\partial^2 L(y_i, \hat{y}_i^{(t-1)})}{\partial \hat{y}_i^{(t-1)2}}$$

Loss → reg, Classi

Reg → mse → $L = \frac{1}{2}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$

$$g_i = \frac{\partial L}{\partial \hat{y}_i} = -\sum_{i=1}^{n}(y_i - \hat{y}_i) = \sum_{i=1}^{n}(\hat{y}_i - y_i)$$

→ residual for every point; ↗ prev step

$$= \frac{\sum_{i \in I_j}(y_i - \hat{y}_i^{(t-1)})}{\sum_{i \in I_j} 1 + \lambda} \rightarrow \frac{\text{sum of residual}}{\#\text{of residuals} + \lambda}$$

cgpa	iq	pack	y_i ✓	$\hat{y}_i^{(t-1)}$ ✓	residual
8	80	8		8.5	8-8.5
9	90	9		8.5	9-8.5

$$\frac{\partial}{\partial \hat{y}_i}\frac{\partial L}{\partial \hat{y}_i} = \frac{\partial}{\partial \hat{y}_i}(\hat{y}_i - y_i) = 1$$

$$w_j = \frac{\text{sum of residual}}{\#\text{residuals} + \lambda}$$

Output Value for Classification

01 December 2023 13:45

$$w_j = \frac{-\sum_{i \in I_j} g_i}{\sum_{i \in I_j} h_i + \lambda}$$

$$g = \frac{\partial L(y_i, \hat{y}_i^{(t-1)})}{\partial \hat{y}_i^{(t-1)}}$$

$$h = \frac{\partial^2 L(y_i, \hat{y}_i^{(t-1)})}{\partial \hat{y}_i^{(t-1)2}}$$

loss - log loss → gradient boosting

$$\left[w_j = \frac{\text{Sum of residual}}{\sum_{i \in I_j} p_i(1-p_i) + \lambda} \right]$$

p_i → pred prob of prev timestep

Derivation of Similarity score

01 December 2023 13:45

$$\mathcal{L}^{(t)} = \sum_{j=1}^{T}\left[\left(\sum_{i\in I_j} g_i\right) w_j + \frac{1}{2}\left(\sum_{i\in I_j} h_i + \lambda\right) w_j^2 \right] + \gamma T$$

$$w_j = \frac{-\sum_{i\in I_j} g_i}{\sum_{i\in I_j} h_i + \lambda}$$

$$\mathcal{L}^{(t)}(q_i) = \sum_{j=1}^{T}\left[\left(\sum_{i\in I_j} g_i\right) \frac{\left(-\sum_{i\in I_j} g_i\right)}{\sum_{i\in I_j} h_i + \lambda} + \frac{1}{2}\left(\sum_{i\in I_j} h_i + \lambda\right) \frac{\left(\sum_{i\in I_j} g_i\right)^2}{\left(\sum_{i\in I_j} h_i + \lambda\right)^2} \right] + \gamma T$$

$$\mathcal{L}^{(t)}(q_i) = \sum_{j=1}^{T}\left[-\frac{\left(\sum_{i\in I_j} g_i\right)^2}{\sum_{i\in I_j} h_i + \lambda} + \frac{1}{2}\frac{\left(\sum_{i\in I_j} g_i\right)^2}{\sum_{i\in I_j} h_i + \lambda} \right] + \gamma T$$

$$\text{Similarity score} = \sum_{j=1}^{T}\left[-\frac{1}{2}\frac{\left(\sum_{i\in I_j} g_i\right)^2}{\sum_{i\in I_j} h_i + \lambda} \right] + \gamma T$$

iter t

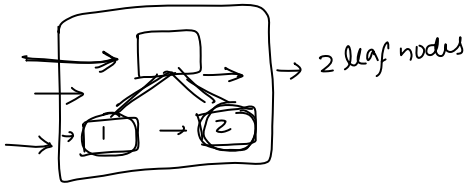
$$\mathcal{L}^{(t)}(q_i) = -\frac{1}{2}\sum_{j=1}^{2 \text{ or } T}\frac{\left(\sum_{i\in I_j} g_i\right)^2}{\sum_{i\in I_j} h_i + \lambda} + \gamma T$$

gini / entropy

t-1 → decision tree

t model

leaf node



$$\mathcal{L}^{(t)}(q_p) = -\frac{1}{2}\frac{\left(\sum_{i\in I_p} g_i\right)^2}{\sum_{i\in I_p} h_i + \lambda} + \gamma$$

or

$$\mathcal{L}^{(t)}(q_c) = -\frac{1}{2}\left[\frac{\left(\sum_{i\in I_L} g_i\right)^2}{\sum_{i\in I_L} h_i + \lambda} + \frac{\left(\sum_{i\in I_R} g_i\right)^2}{\sum_{i\in I_R} h_i + \lambda}\right] + 2\gamma$$

$$\mathcal{L}^{(t)}(q_p) \geq \mathcal{L}^{(t)}(q_c) \qquad -\gamma$$

$$-\frac{1}{2}\frac{\left(\sum_{i\in I_j} g_i\right)^2}{\sum_{i\in I_j} h_i + \lambda} \geq \left(-\frac{1}{2}\left[\frac{\left(\sum_{i\in I_L} g_i\right)^2}{\sum_{i\in I_L} h_i + \lambda} + \frac{\left(\sum_{i\in I_R} g_i\right)^2}{\sum_{i\in I_R} h_i + \lambda}\right] + 2\gamma\right)$$

$$-\frac{1}{2}\frac{\left(\sum_{i\in I_P} g_i\right)^2}{\sum_{i\in I_P} h_i+\lambda} + \gamma - \left(-\frac{1}{2}\left[\frac{\left(\sum_{i\in I_L} g_i\right)^2}{\sum_{i\in I_L} h_i+\lambda} + \frac{\left(\sum_{i\in I_R} g_i\right)^2}{\sum_{i\in I_R} h_i+\lambda}\right] + 2\gamma\right)$$

$$\left[-\frac{1}{2}\frac{\left(\sum_{i\in I_P} g_i\right)^2}{\sum_{i\in I_P} h_i+\lambda} + \frac{1}{2}\frac{\left(\sum_{i\in I_L} g_i\right)^2}{\sum_{i\in I_L} h_i+\lambda} + \frac{1}{2}\frac{\left(\sum_{i\in I_R} g_i\right)^2}{\sum_{i\in I_R} h_i+\lambda}\right] - \gamma$$

differ

$$\frac{1}{2}\left[\frac{\left(\sum_{i\in I_L} g_i\right)^2}{\sum_{i\in I_L} h_i+\lambda} + \frac{\left(\sum_{i\in I_R} g_i\right)^2}{\sum_{i\in I_R} h_i+\lambda} - \frac{\left(\sum_{i\in I_R} g_i\right)^2}{\sum_{i\in I_P} h_i+\lambda}\right] - \gamma$$

Final Calculation of Similarity

02 December 2023 01:27

$$\text{Similarity score} = \sum_{j=1}^{T} \left[-\frac{1}{2} \frac{\left(\sum_{i \in I_j} g_i\right)^2}{\sum_{i \in I_j} h_i + \lambda} \right] + \gamma T$$

reg

$g \rightarrow$ sum of residuals

$h \rightarrow 1 \rightarrow N$

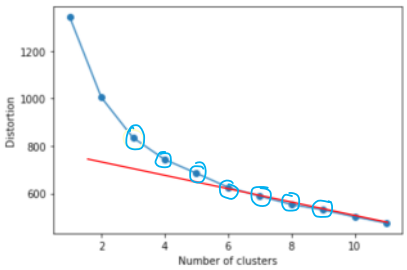
Class

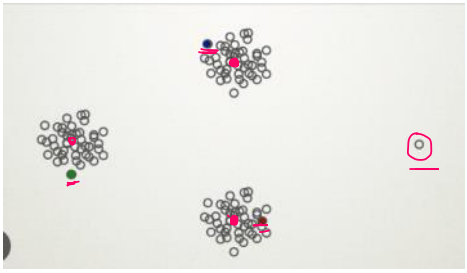
$g \rightarrow$ residual

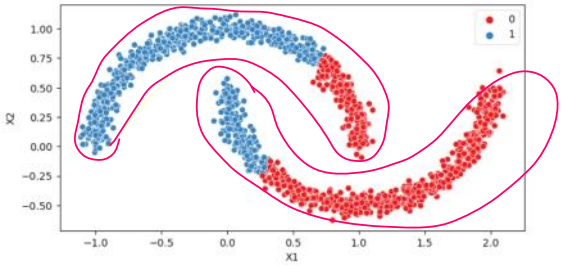
$h \rightarrow p_i(1-p_i)$

Why DBSCAN?

16 December 2023 15:10

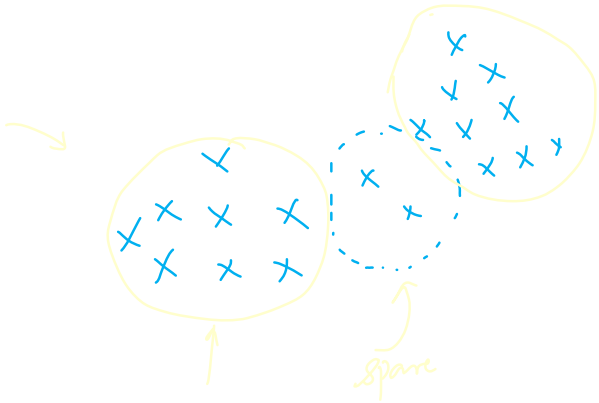






What is Density Based Clustering

16 December 2023 15:12

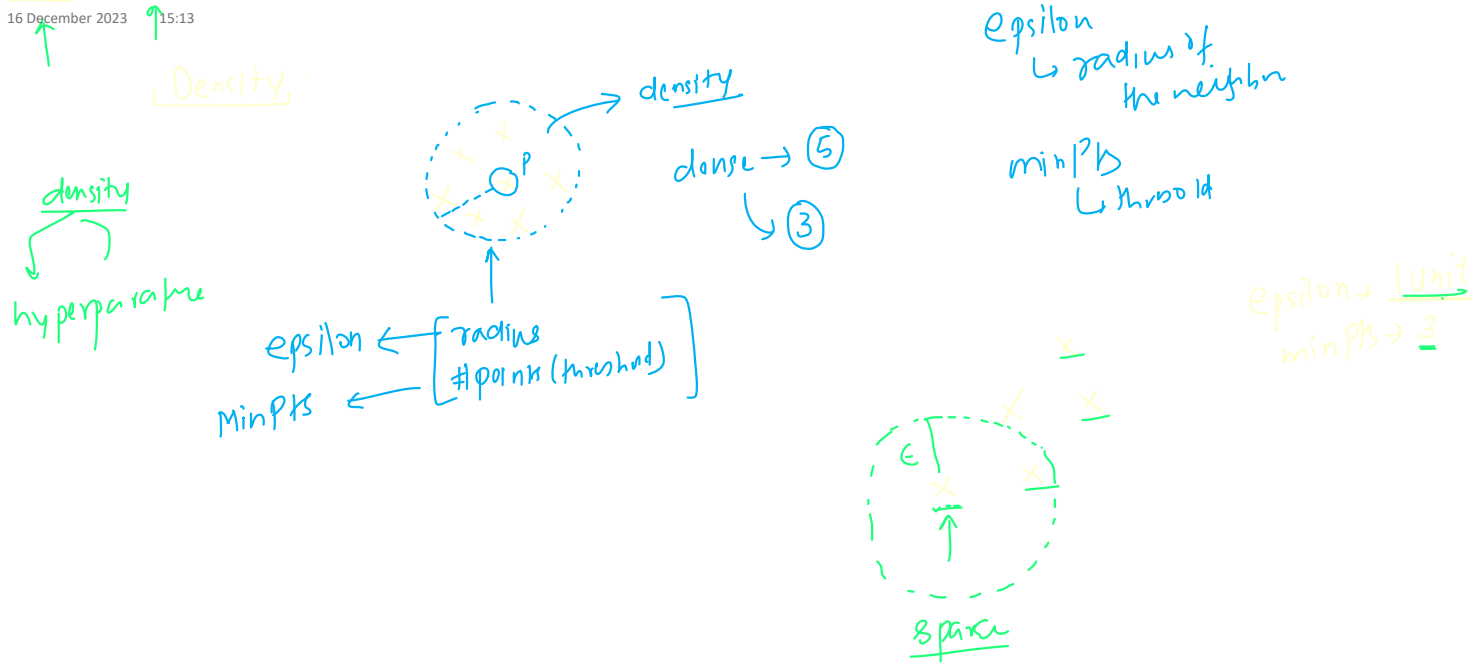


DBSCAN → Density Based Spatial Clustering of Application with Noise

OPTICS

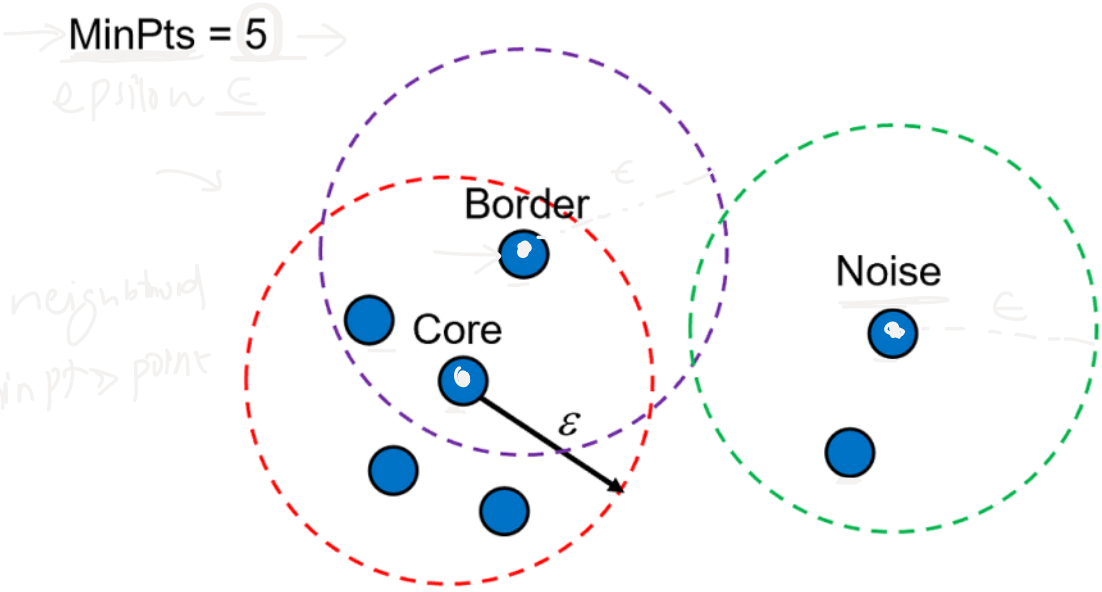
MinPts & Epsilon

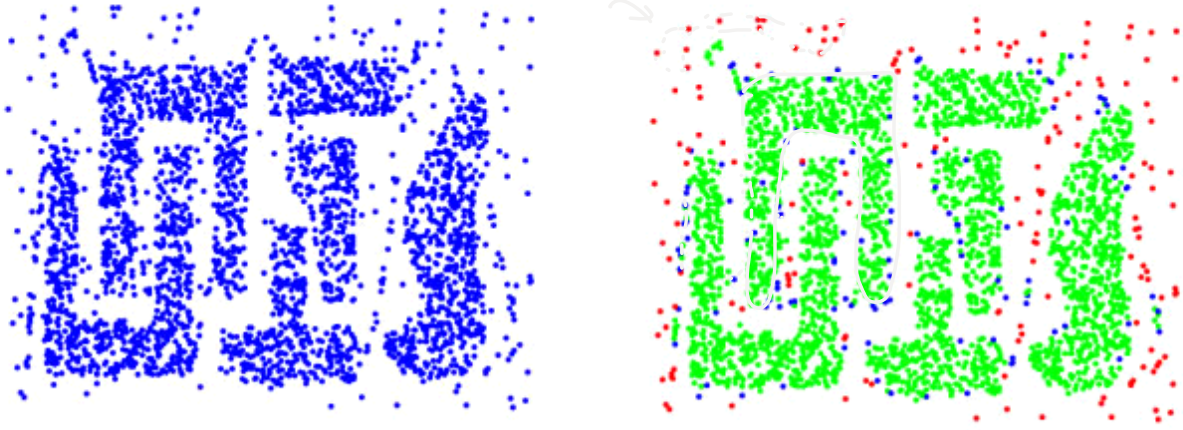
16 December 2023 15:13



Core Points, Border Points & Noise Points

16 December 2023 15:13





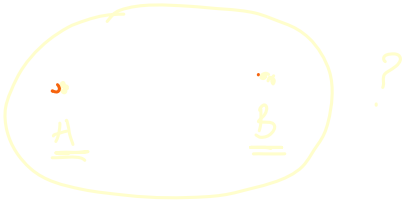
Original Points

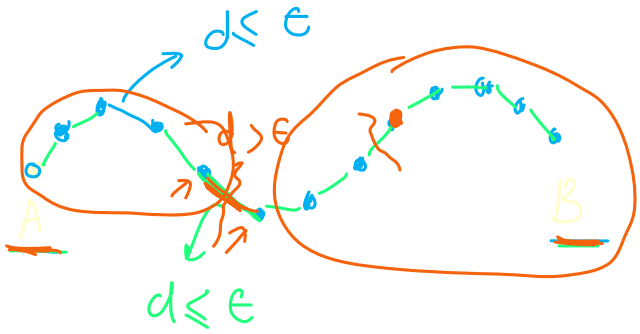
Point types: core, border and noise

Eps = 10, MinPts = 4

Density Connected Points

16 December 2023 15:13





DBSCAN Algorithm

16 December 2023 15:14

Step 1 - Identify all points as either core point, border point or noise point

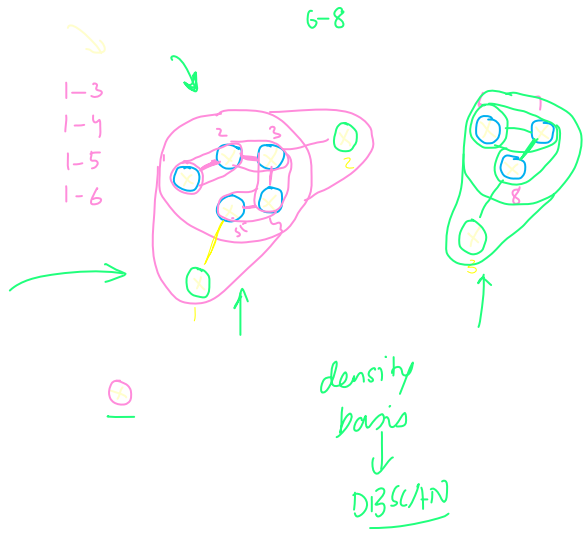
Step 2 - For all of the unclustered core points

Step 2a - Create a new cluster

Step 2b - add all the points that are unclustered and density connected to the current point into this cluster

Step 3 - For each unclustered border point assign it to the cluster of nearest core point

Step 4 - Leave all the noise points as it is.



Code

16 December 2023 15:14

Limitations

16 December 2023 15:14

1. Robust to outliers
2. No need to specify clusters
3. Can find arbitrary shaped clusters
4. Only 2 hyperparameters to tune

1. Sensitivity to hyperparameters
2. Difficulty with varying density clusters
3. Does not predict

Visualization

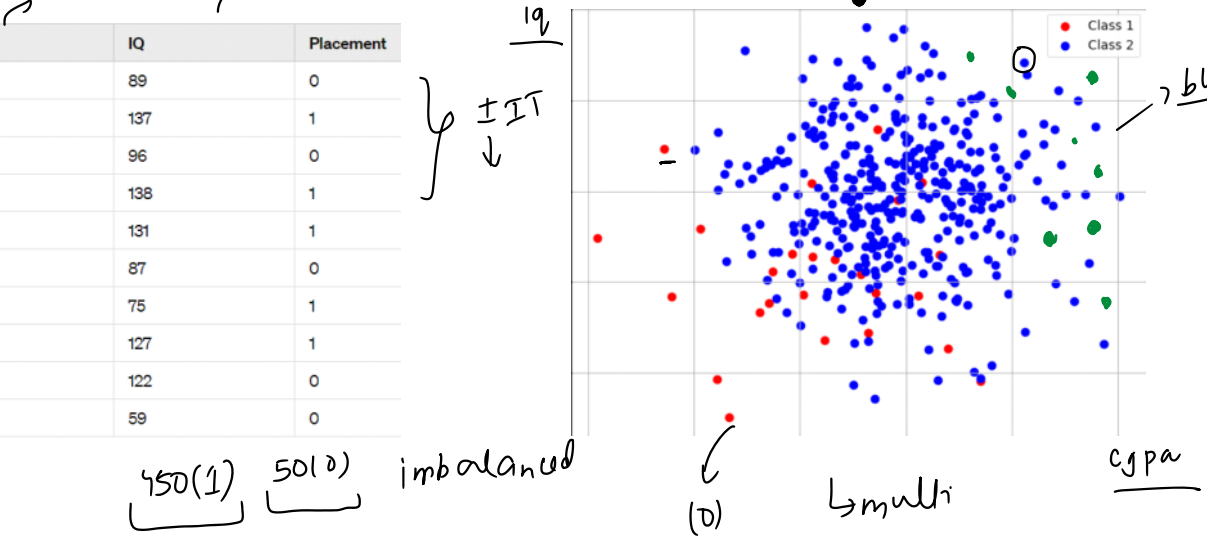
16 December 2023 15:14

What is Imbalanced Data?

02 May 2024 16:12

500 students → model → {cgpa, iq} → 0, 1 (binary)

CGPA	IQ	Placement
5.49	89	0
7.15	137	1
6.03	96	0
5.45	138	1
4.24	131	1
6.46	87	0
4.38	75	1
8.92	127	1
9.64	122	0
3.83	59	0



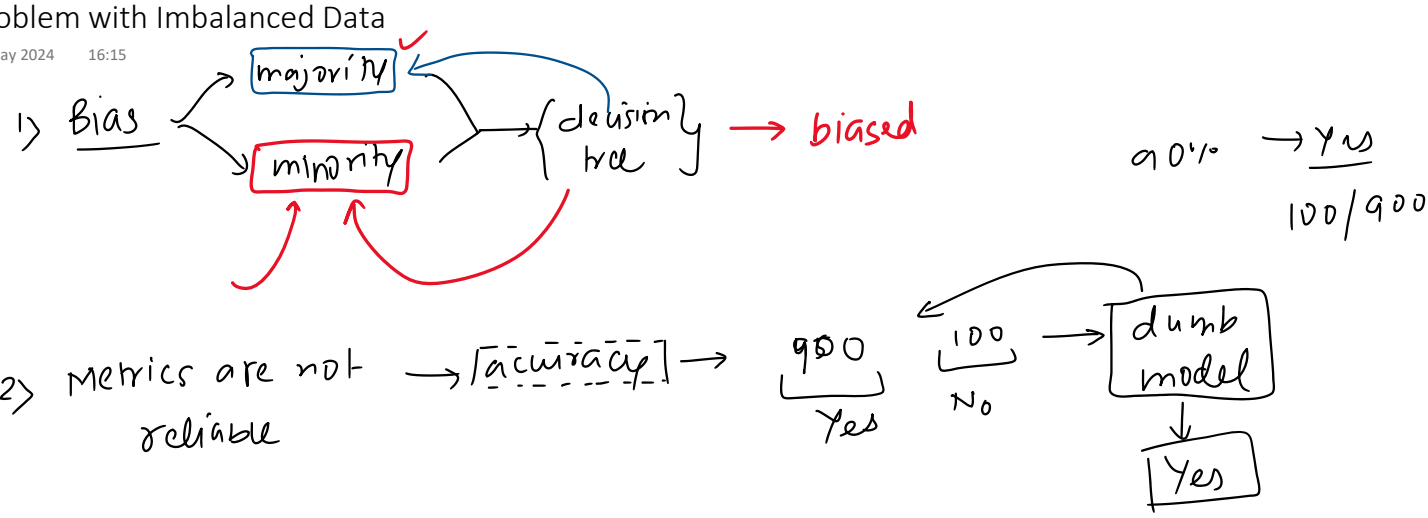
± IT

450(1) 50(0) imbalanced

↳ multi

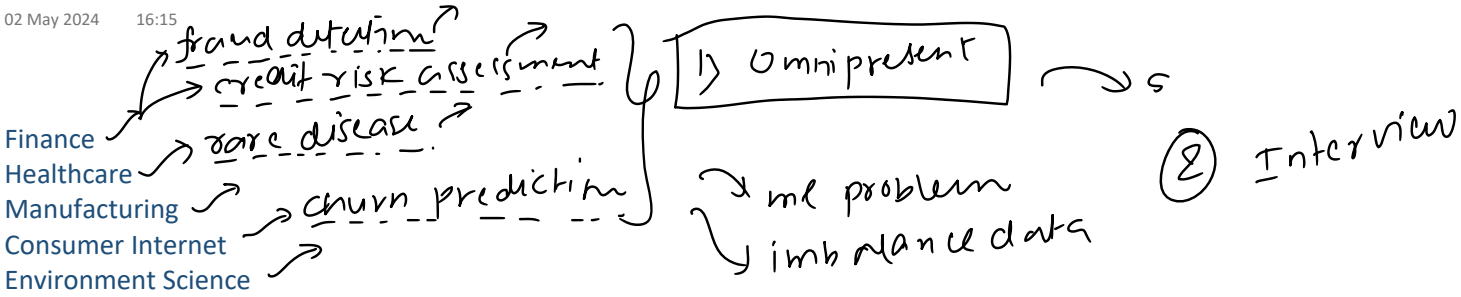
Problem with Imbalanced Data

02 May 2024 16:15



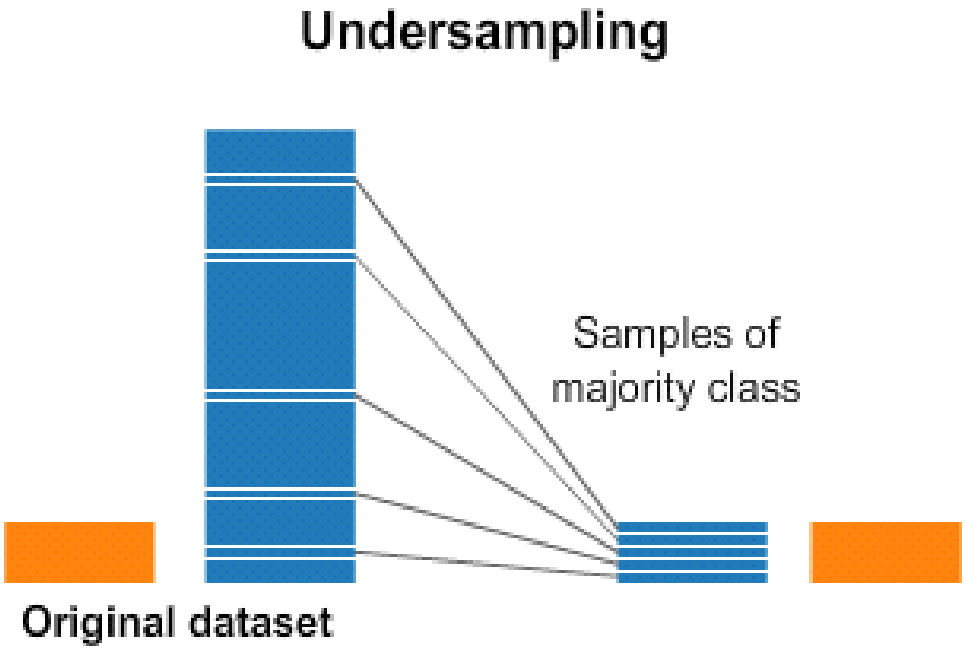
Why studying Imbalanced Data is important?

02 May 2024 16:15



1. Undersampling

02 May 2024 16:17



Advantages

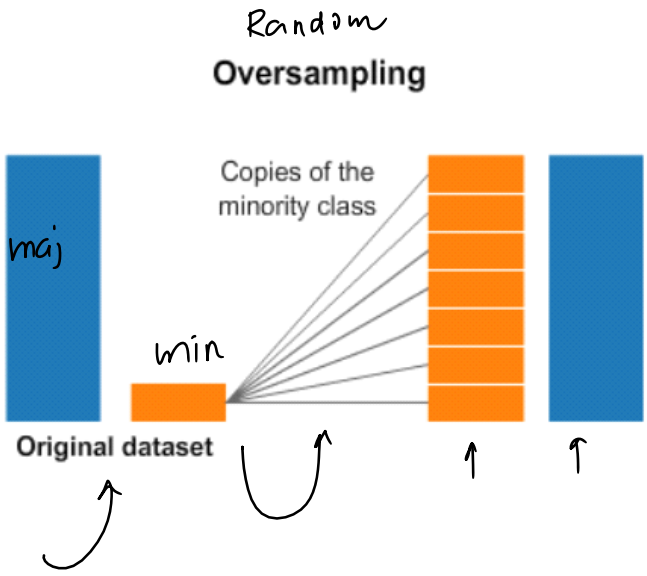
1. Reduction in bias
2. Faster training

Disadvantages

1. Information loss leading to underfitting
2. Sampling Bias

2. Oversampling

03 May 2024 01:13



maj 900 min 100 → balance → 900 (100 duplicate) → 800 → duplication

Advantage

1. Reduced Bias

Disadvantage

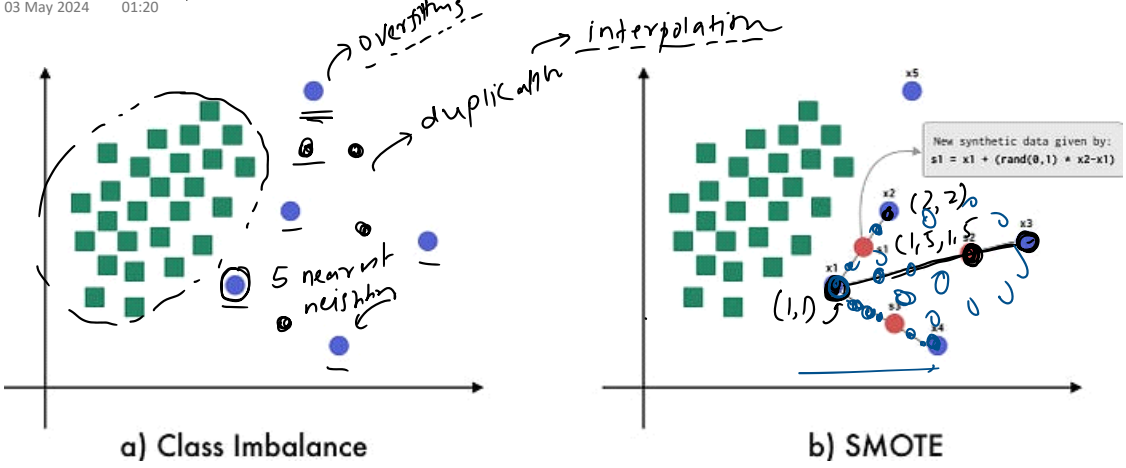
1. Increased size
2. Duplication of data may cause overfitting

ml also → imp → overfitting → decision tree → ROS → overfitting

3. SMOTE

03 May 2024 01:20

popular → Oversampling → min → upsample



a) Class Imbalance

b) SMOTE

5 KNN

- Train a KNN on minority class observations - find each observation's 5 closest neighbours

To create the new synthetic data:

- Select examples from the minority class at random
- Select a neighbour of each example at random (for the interpolation)
- Extract a random number between 0 and 1
- Calculate the new examples as = original sample - factor * (original sample - neighbour)
- The final dataset consists of the original dataset + the newly created examples

0-1 → 0.5

sample − factor * [sample − neighbour]

[1,1] − 0.5 [(1,1) − (2,2)]

(1,1) − 0.5 [(−1,−1)]

(1,1) − (−0.5, −0.5)

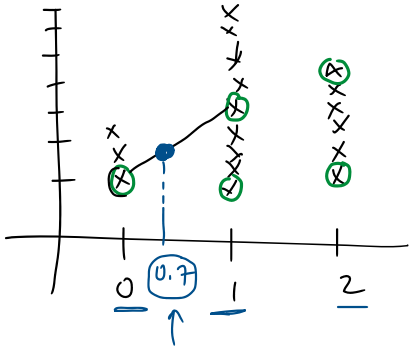
= [1.5, 1.5]

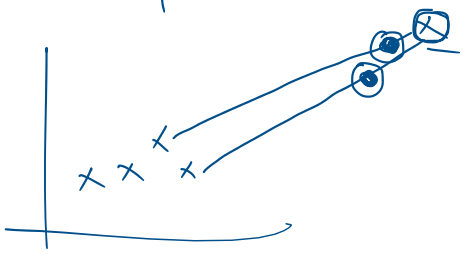
0 → K → K=1

K=5

Disadvantages

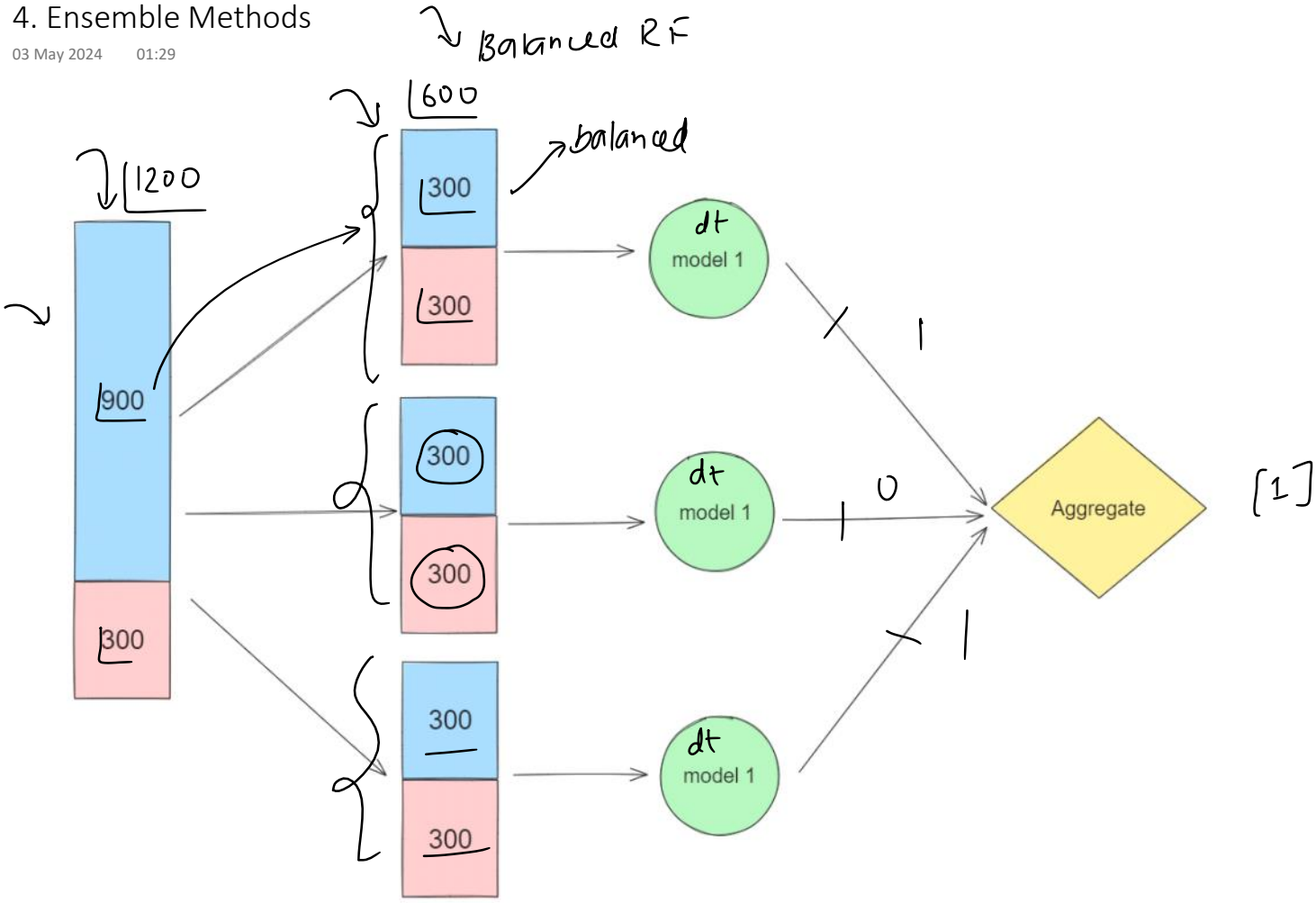
1. Does Not Handle Categorical Data Well
2. Computational Complexity
3. Dependency on the Choice of Neighbors
4. Sensitive to Outliers
5. Balance Achieved May Not Reflect True Nature

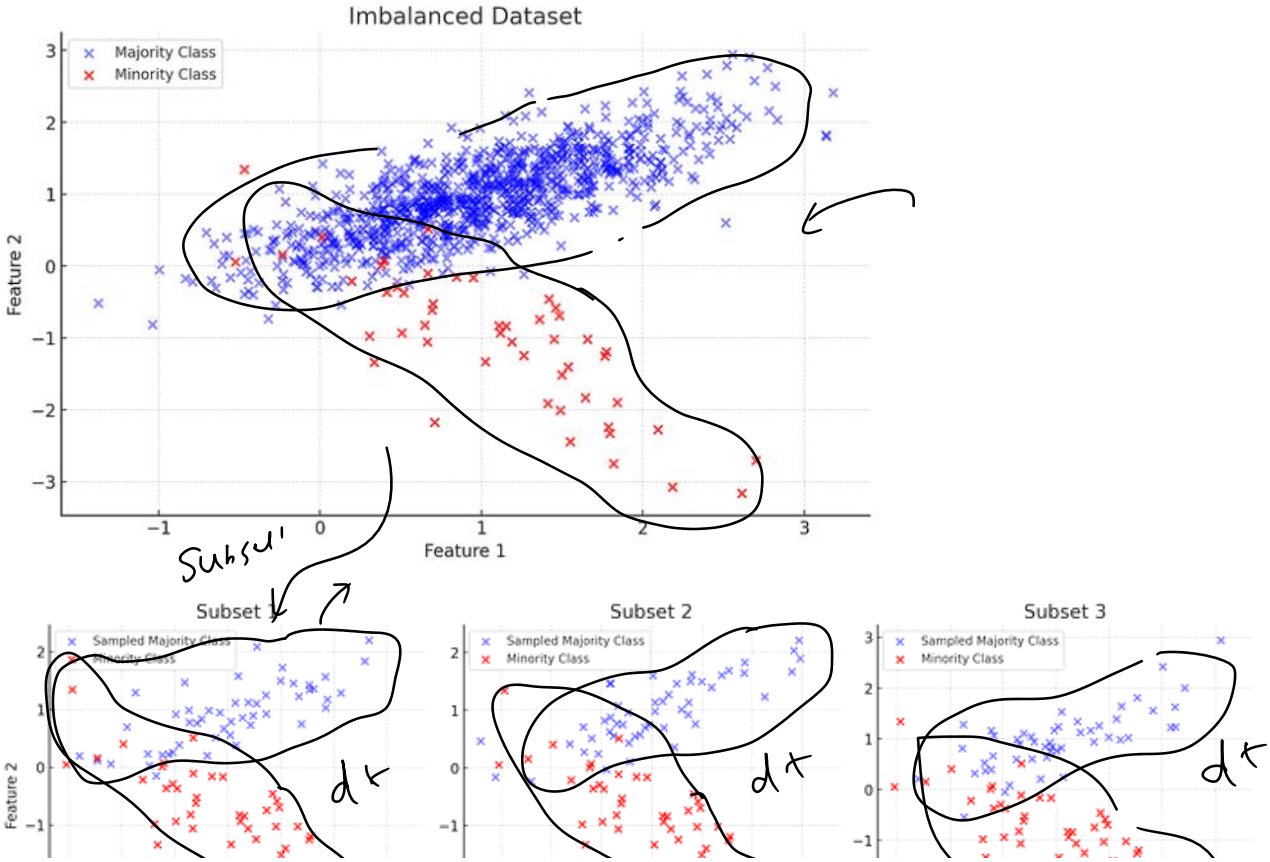


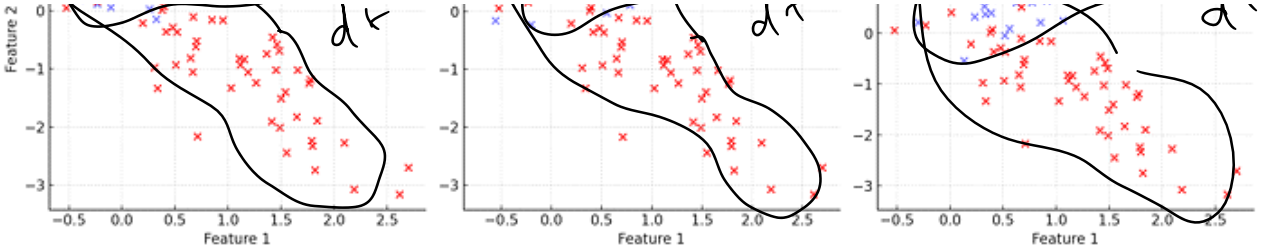


4. Ensemble Methods

03 May 2024 01:29

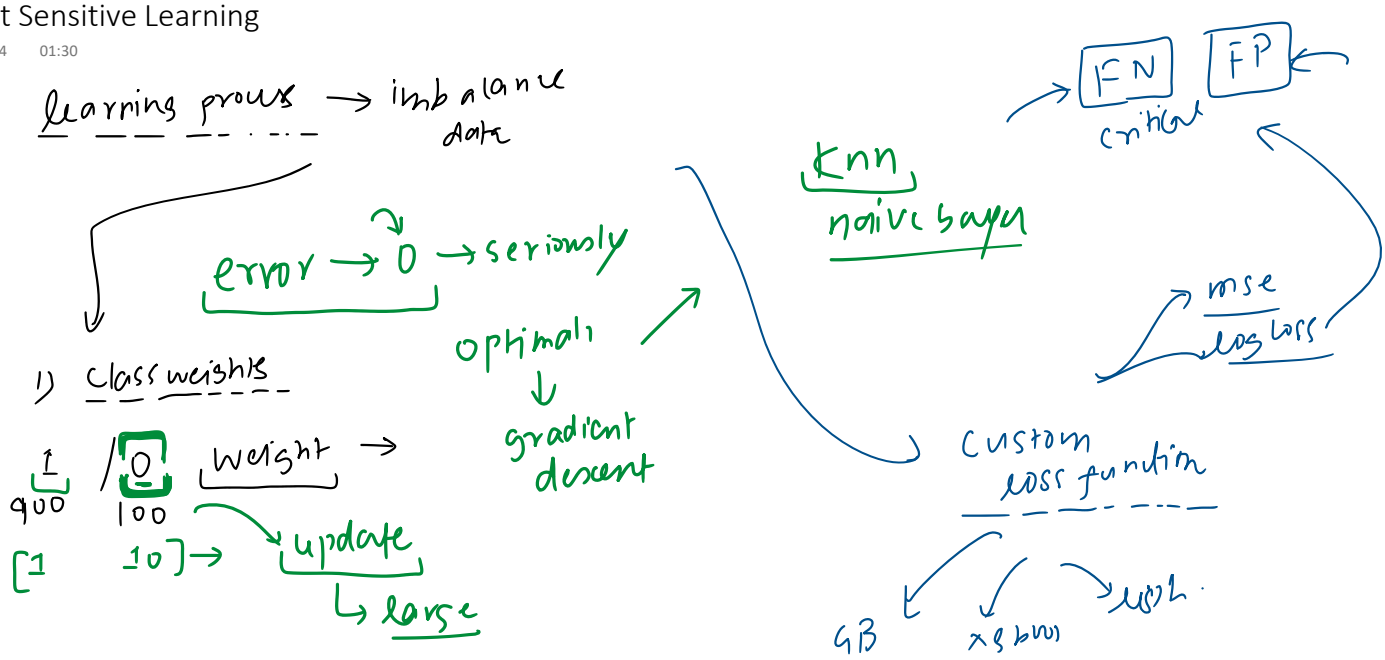





Feature 2
Feature 1
Feature 1
Feature 1

5. Cost Sensitive Learning

03 May 2024 01:30



List of all the techniques

03 May 2024 15:06