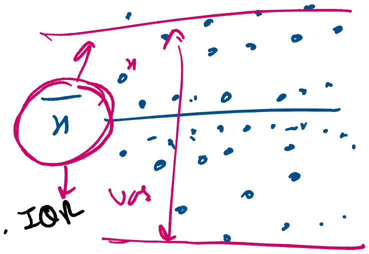
2.StepsToMLProject
Thursday, August 19, 2021 6:12 PM

Central tendency → mean, median, mode



Measure of spreadness → variance σ², std σ, Quantiles Range, IQR

single varia → Sample → mean ± std → majority

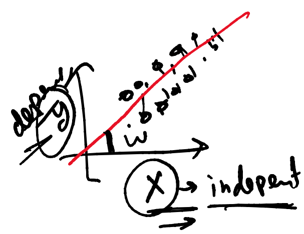
Samp → 170 ± 20

Dependent variable
↓
Price
target

no rooms
Area
populat
educat
point

+ive
−ive
no rel

① Linear Regression



Co-Variance →

it is mathematical measurement of Dependency or relationship of two variable.

X → height
y → weight

Regression →

predicting a continous value of a dependent

... variables.

predicting a continous value ...

Variable on the basis of independ variables.

$$\boxed{Cov(x,y) = \frac{\sum_{i=1}^{N} (\bar{x} - x_i)\cdot(\bar{y} - y_i)}{N-1}}$$

$$-\infty < Cov(x,y) < \infty$$

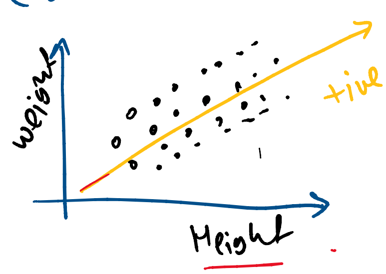
$Cov(x,y) \rightarrow$ +ive x and y both are dependent, and they have a positive relationship

e.g. x = height , y = weight

(160+20) (70+30)

(20 · 30) → 600

$Cov(x,y) \simeq 600 \rightarrow$ +ive → increment in x will result increment in y



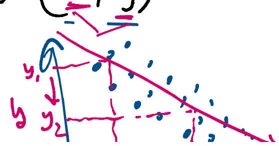
height ↑

weight ↑

x = Age →

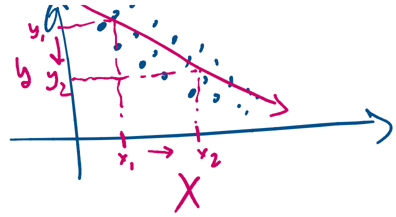
y = strength →

Cov(x,y) = -ive → if we increase x then y will decrease



x ↑ y ↓

x ↓ y ↑



$X \downarrow \quad y \uparrow$

$$Cov(x,y) \approx 0$$

There is no relation in $\underline{X, y}$

$140 \rightarrow P$

$14 \rightarrow S$

Popn

Train

Test

$$\boxed{Cov(x,y) = \frac{\sum_{i=1}^{N} (\bar{x} - x_i)\cdot(\bar{y} - y_i)}{N-1}}$$

→ population → N

→ Sample → $N-1$

target

$y \rightarrow$ price

$20 \approx 5$

feature Importance

features

$X \rightarrow$
- → # rooms x_1
- → Area x_2
- → age x_3

① $Cov(y, x_1) \rightarrow$ +ive 15

② $Cov(y, x_2) \rightarrow$ +ive 500

③ $Cov(y, x_3) \rightarrow$ -ive -75

① $\frac{(\bar{x}_1 - x_1)\cdot(\bar{y} - y)}{N}$ → 15 → 10 → 3 →

② $\frac{(\bar{x}_2 - x_1)\cdot(\bar{y} - y)}{N}$ → 500

③ $\frac{(\bar{x}_3 - x_3)\cdot(\bar{y} - y)}{N}$ → 75

3, 5, 5, 5, 1m, 12, σ

5 – 15 → 3

100

500 → 100 →

250 ± 10

-8 → +8

→ 8 – 50 → 20 ± 12

⇒ **Pearson's Correlation** →

$$\boxed{r = \frac{Cov(x,y)}{\sigma_x \cdot \sigma_y}}$$

$\sigma_x \rightarrow$ std of X

$\sigma_y \rightarrow$ std of y

$$\boxed{r = \frac{\;\;}{\sigma_x \cdot \sigma_y}}$$

$$\boxed{-1 < r < 1}$$

$r \simeq 0$, no relation

$r \simeq 1$ or -1 Both variables are same

$-1, 1$

price (y)

\# Rooms (x_1)

area (x_2)

age (x_3)

$r_{(y, x_1)} = 0.6$

$r(y, x_2) = 0.8$

$r(y, x_3) = -0.7$

$X_2 > x_3 > x_1$

→ 2·D

(100×1)

→ Model

y → target

y is a vector contains possible outcome for given

$\underline{y} \rightarrow$ target

y is a vec... possible outcome for given input (features)

$\underline{X} \rightarrow$ features / Attribute

$y \rightarrow$ actual output

$\hat{y} \rightarrow$ predicted output, estimated

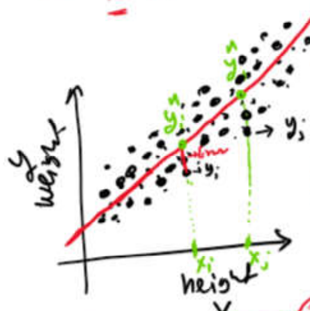
→ Linear Regression

① Hypothesis function

input

$$\hat{y} = H_{\theta}(X) = F(X)$$

$y = mx + c$

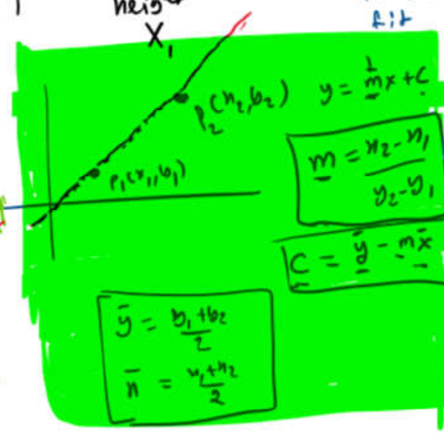


$$\boxed{\hat{y} = \hat{\theta}_0 + \hat{\theta}_1 x_1}$$

$\hat{\theta}_1, \hat{\theta}_0 \rightarrow$ estimated value of parameter for best fit

$$\hat{y} = H_{\theta}(X) = \underbrace{\hat{\theta}_0 + \hat{\theta}_1 X_1}_{\text{Linear}}$$

$$\widehat{weight} = H_{\theta}(\text{Height}) = \hat{\theta}_0 + \hat{\theta}_1 \cdot \text{Height}$$



$y = mx + c$

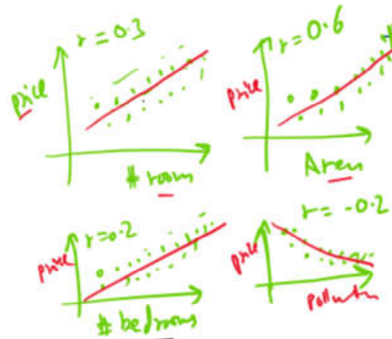
$$\boxed{m = \frac{y_2 - y_1}{x_2 - x_1}}$$

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

$$\boxed{\bar{y} = \frac{y_1 + y_2}{2} \quad \bar{x} = \frac{x_1 + x_2}{2}}$$

$y \rightarrow$ price

$X \rightarrow$
- # rooms $\rightarrow x_1$
- # bedrooms $\rightarrow x_2$
- Area $\rightarrow x_3$
- pollution $\rightarrow x_4$
- ⋮
- x_n



$$\widehat{price} = H_{\theta}(X) = \theta_0 + \theta_1 \cdot \#\text{rooms} + \theta_2 \cdot \#\text{bedrooms} + \theta_3 \cdot \text{Area} + \theta_4 \cdot \text{Pollution}$$

$$\boxed{\hat{y} = H_{\theta}(x)}$$

$$\boxed{H_{\theta}(x) = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \theta_3 x_3 + \cdots \theta_n x_n}$$

$\theta = [\theta_0 \ \theta_1 \ \theta_2 \ \theta_3]_{1\times4}$

$x = [1 \ x_1 \ x_2 \ x_3]_{1\times4}$ $\quad \theta \cdot X^T$

$\theta \cdot X^T \rightarrow \theta_0 + \theta_1 x_1 +$

$$\begin{bmatrix} 2 & 4 \\ 4 & 2 \end{bmatrix}_{2\times 2} \begin{bmatrix} 1 & 2 \\ 2 & 1 \end{bmatrix}_{2\times 2}$$

$$\begin{bmatrix} 2+8 & 4+4 \\ 4+4 & 8+2 \end{bmatrix}_{2\times 2}$$

$x = [1 \; x_1 \; x_2 \; x_3]_{1\times 4}$ $\theta_{1\times 4}$ $\boxed{4\times 1}$

$A = [x_1 \; x_2 \; x_3 \; x_4]$ 1×4

$\theta \cdot x^T \rightarrow \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \theta_3 x_3$ - -

$$B = \begin{bmatrix} a \\ b \\ c \\ d \end{bmatrix}_{4\times 1}$$

$$A_{m\times n} \cdot B_{n\times k} = C_{m\times k}$$

$$A_{1\times 4} \cdot B_{4\times 1} = C_{1\times 1}$$

$$C = [a \cdot x_1 + b \cdot x_2 + c \cdot x_3 + d \cdot x_4]$$

$$H_\theta(x) = [\theta_0 + \theta_1 x_1 + \theta_2 \cdots$$

$X \rightarrow$ 2D matrix which holds all training records

	#rooms	#bedrooms	Area	pollution
	x_1	x_2	x_3	x_4
0	5	11	200	0.2
1	10	2	100	0.1
2	15	5	150	0.5
3	20	2	100	0.3
4	5	1	100	0.4
5	2	0	50	0.5
6	1	0	100	0.6

$\underline{X} =$ the matrix above, $m \times n \rightarrow$ features; $m \rightarrow$ observations

$$y = \begin{bmatrix} 20 \\ 30 \\ 40 \\ 20 \\ 5 \\ 50 \\ 60 \end{bmatrix} \begin{matrix} \rightarrow y_0 \\ \rightarrow y_1 \\ \rightarrow y_2 \\ \rightarrow y_3 \\ \rightarrow y_4 \\ \rightarrow y_5 \\ \rightarrow y_6 \end{matrix}$$

$\theta_0 \rightarrow$ Bias / intercept

$\theta_1, \theta_2, \theta_3 \cdots \theta_n \rightarrow$ coefficients

$$\boxed{\hat{y}_0 = \hat{\theta}_0 + \hat{\theta}_1 x_{1_0} + \hat{\theta}_2 x_{2_0} + \hat{\theta}_3 x_{3_0} + \hat{\theta}_4 x_{4_0}}$$

Error $= |y_i - \hat{y}_i|$

$$\hat{y}_1 = \hat{\theta}_0 + \hat{\theta}_1 x_{1_1} + \hat{\theta}_2 x_{2_1} + \hat{\theta}_3 x_{3_1} + \hat{\theta}_4 x_{4_1}$$

$\vdots$

$$\hat{y}_n = \hat{\theta}_0 + \hat{\theta}_1 x_{1_n} + \hat{\theta}_2 x_{2_n} + \hat{\theta}_3 x_{3_n} + \hat{\theta}_4 x_{4_n}$$

$$\theta = [\theta_0 \; \theta_1 \; \theta_2 \; \theta_3 \; \theta_4]_{1\times 4}$$

$$x_i = [1 \; x_{1_i} \; x_{2_i} \; x_{3_i} \; x_{4_i}]_{1\times 4}$$

$$\Theta = [\theta_0\ \theta_1\ \theta_2\ \theta_3\ \theta_4]_{1\times 4} \qquad X_i = [1\ \ldots]_{1\times 4}$$

$$\hat{y}_i = \Theta \cdot X_i^T \rightarrow \hat{y}_i = \theta_0 + \theta_1 x_{1i} + \theta_2 x_{2i} + \theta_3 x_{3i} \ldots$$

$$\boxed{\hat{y}_i = \Theta \cdot X_i}$$

$$\boxed{\hat{y} = \Theta \cdot X}$$

prediction function

$$\boxed{H_\theta(X) = \Theta \cdot X}$$

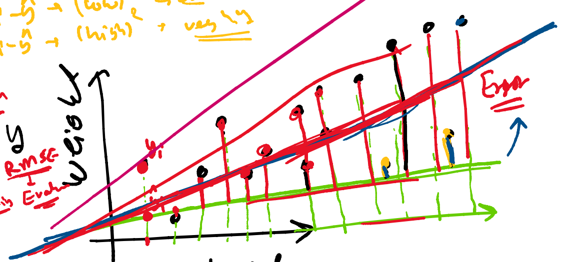
$$\boxed{\hat{y} = H_\theta(X)}$$

→ How to estimate value of θ's

fitting

find unknown value of parameters of $H_\theta(X)$'s using training data

$$\boxed{\hat{y} = H_\theta(X) = \underbrace{\hat{\theta}_0 + \hat{\theta}_1}_{\text{parameter}} x_1}$$



MAE → Evaluation

MSE → RMSE

fitting Evaluation

To estimate Best value for Θ we need to minimize cost function

$$\boxed{J_\theta = \frac{\sum_{i=1}^{n} (\hat{y}_i - y_i)^2}{N}}$$

L_1 penalty

$$\boxed{\underset{\uparrow\ \text{Evaluation}}{MAE} = \frac{\sum_{i=1}^{N} |y_i - \hat{y}_i|}{N}}$$

Linear

L_2 penalty

$$\boxed{MSE = \frac{\sum_{i=1}^{N} (y_i - \hat{y}_i)^2}{N}}$$

Convex

$an^2 + bn + c$

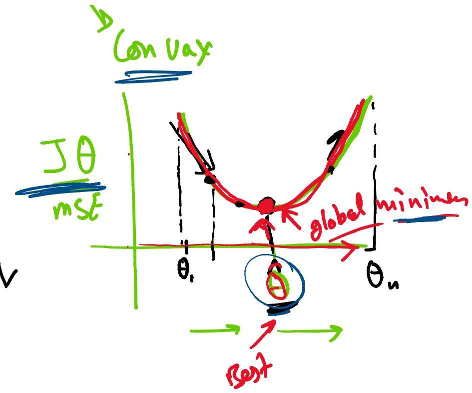
$f(x)$

How to find min or maximum

$f(x)$

How to find min or maximum of a function

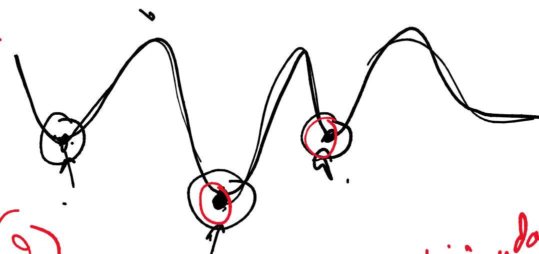
$\rightarrow \quad f'(x) = \underline{0}$, min, max



$f(x) = 2x^3 + 3x + 4$

$\frac{d f(x)}{dx} = 2 \cdot 3\, x^2 + 3 \Rightarrow 6x^2 + 3$

$\hat{y}_i = H_\theta(x) = (\theta_0 + \theta_1 x_i)$ — (1)



$mse(\theta) = \sum_{i=1}^{N} (\hat{y}_i - y_i)^2 \rightarrow$ (2)

$\qquad N$

$y_i \rightarrow$ training data

$N \rightarrow$ no of observation

$\hat{y}_i \rightarrow$ prediction

put eq (1) into eqn (2)

$$\boxed{mse(\hat{\theta}) = \frac{\sum_{i=1}^{N} (\hat{\theta}_0 + \hat{\theta}_1 x_i - y_i)^2}{N}}$$

Training Data

$\rightarrow x_i, y_i \rightarrow$ Constants

$\rightarrow \hat{\theta}_0, \hat{\theta}_1 \rightarrow$ variables

we want to find value of θ such that

$mse(\theta)$ should be minimum

$\rightarrow 2x + 3y = 0$

$\rightarrow 2x - 3y = 1$

To find minimum

$$\boxed{\frac{\partial}{\partial \theta_0} mse(\theta_0) = 0} \rightarrow (i)$$

$$\boxed{\frac{\partial}{\partial \theta_1} mse(\theta_1) = 0} \rightarrow (ii)$$

→ Normal Equation

$m = \frac{y_2 - y_1}{x_2 - x_1}$

$\theta_1 = \frac{\sum^{N} (\bar{y} - y_i) \cdot (\bar{x}_i - x_i)}{\ldots}$

$$\boxed{\hat{y} = \theta_0 + \theta_1 x_1}$$

$$\boxed{\theta_1 = \frac{\sum_{i=1}^{N} (\bar{y} - y_i)\cdot(\bar{x}_i - x_i)}{\sum_{i=1}^{N} (\bar{x}_i - x_i)^2}}$$

price $= \theta_0 + \theta_1 \cdot \#\text{rooms} + \theta_2 \,\text{Area} + \theta_3 \cdots$

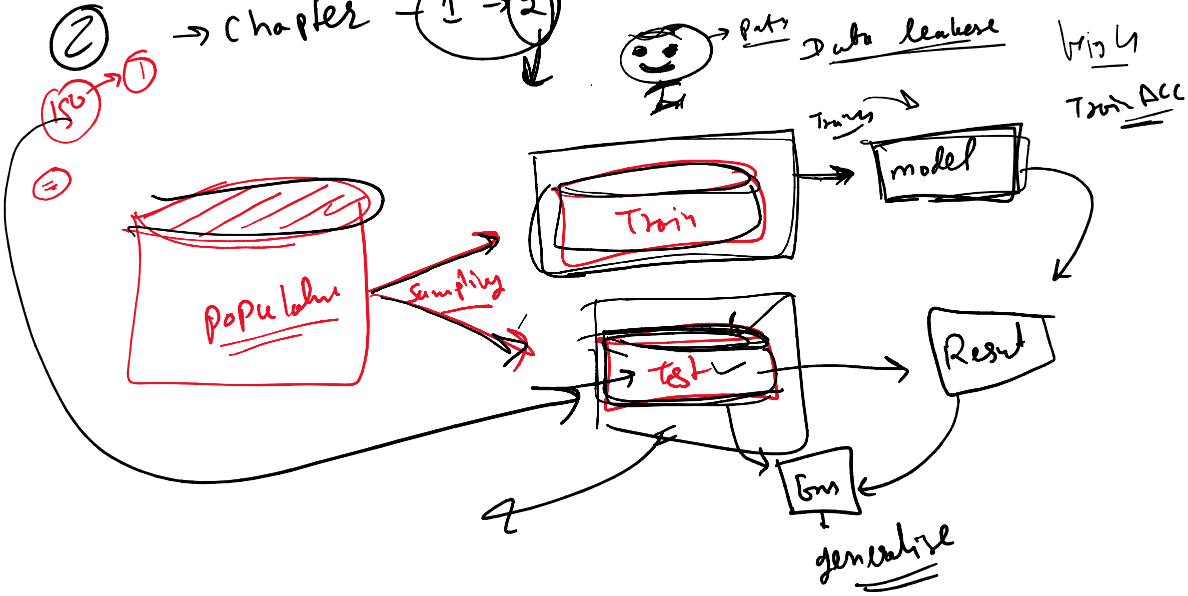
$$\boxed{\theta_0 = \bar{y} - \theta_1 \bar{x}}$$

$$\boxed{\hat{\theta} = (X^T X)^{-1} \cdot X^T \cdot y} \rightarrow \text{normal form}$$

$$\begin{bmatrix} 1 & x_1 & x_2 & x_3 \\ 1 & \vdots & \vdots & \vdots \\ \vdots & \vdots & 1 & 1 \end{bmatrix}$$

① → OOPS

② → chapter 1 → 2



→ Sampling

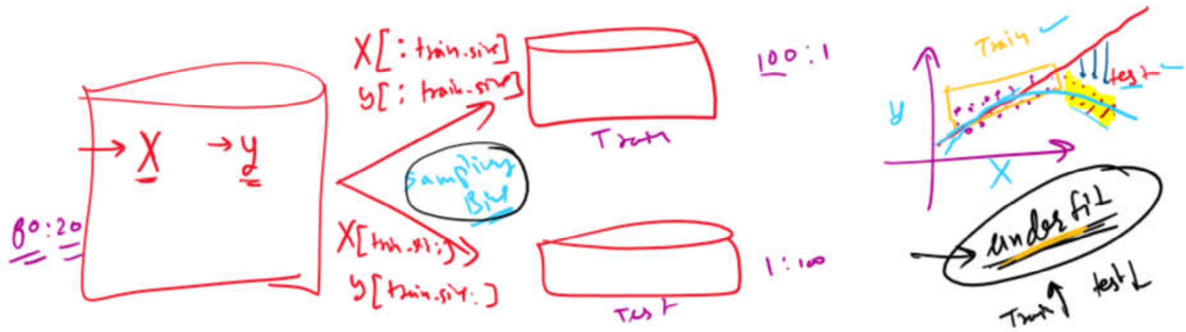
→ train size → test size

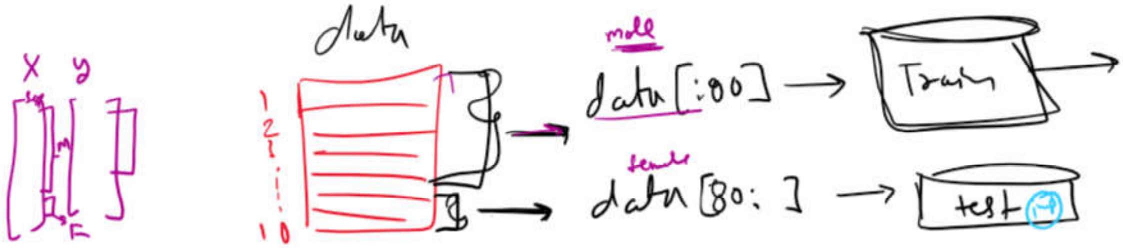
80 20 → Big data set

80 20 → Big data set

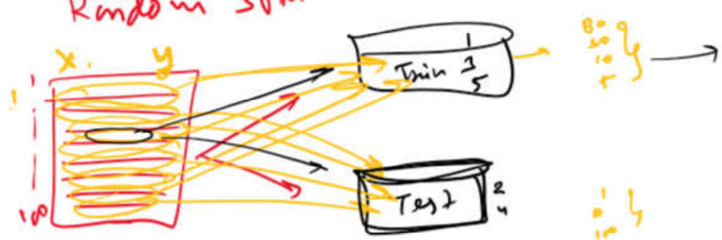
70 30 →

→ Sampling Technique





Random split



→ Stratified Sampling

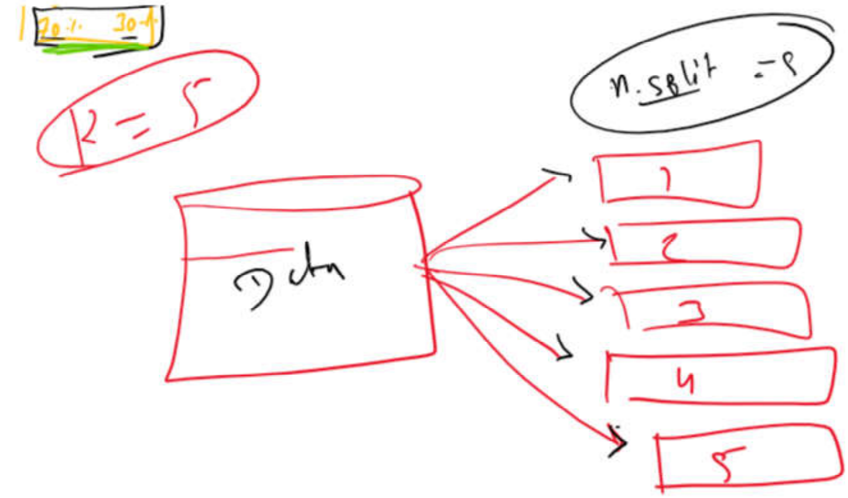
↓

column

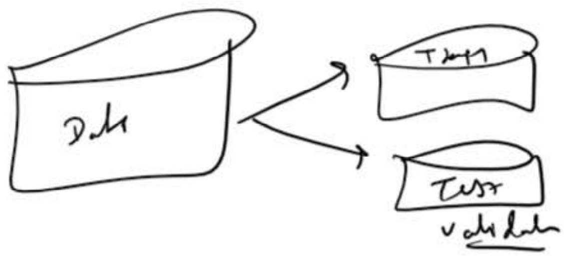
↓

porportion of column will be same

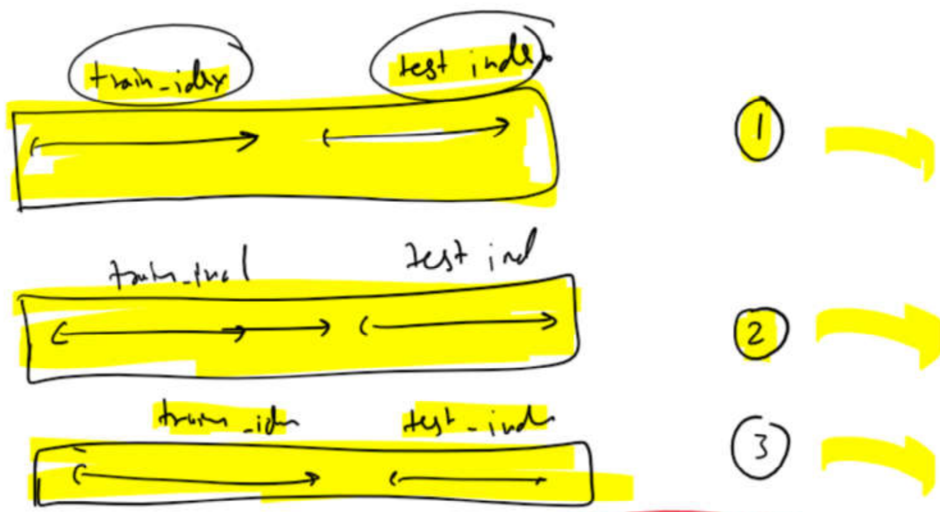
on Train & Test subset



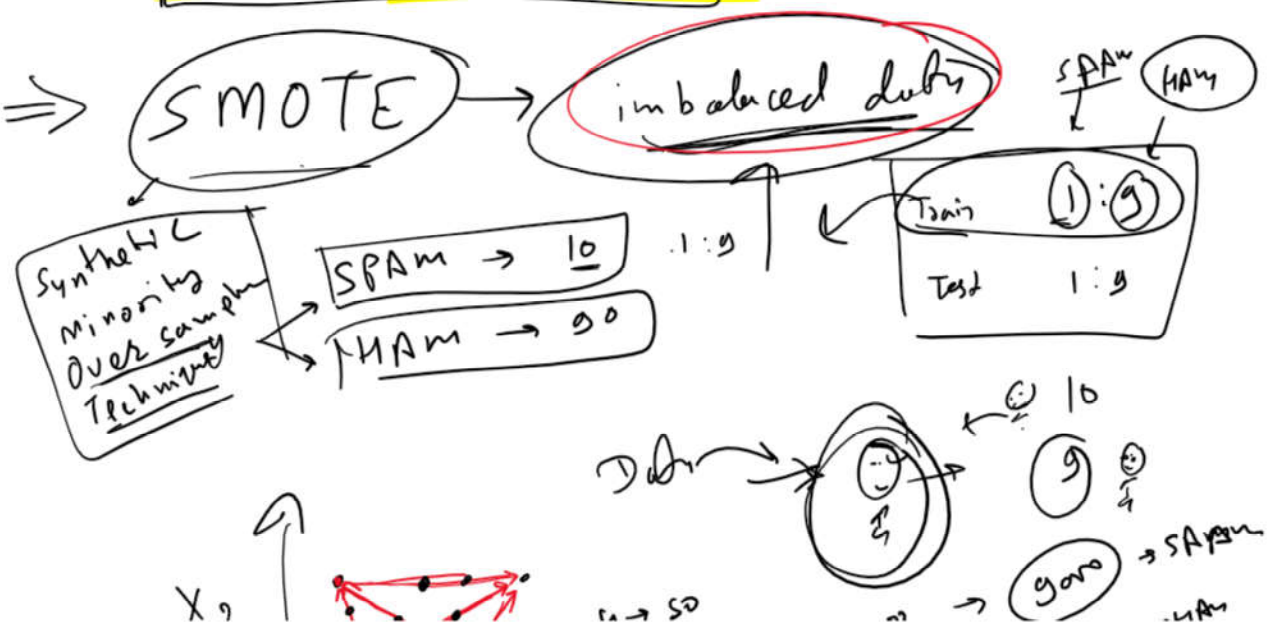
k = 5
n_split = 5
Data
1
2
3
4
5



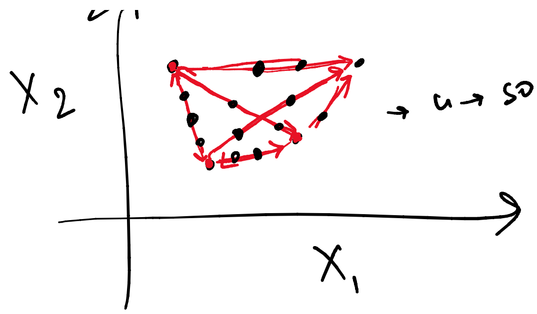
Data
Train
Test
validation

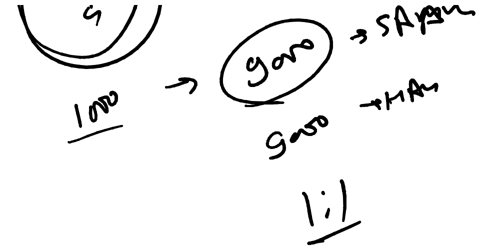


train_idx
test_idx
1
2
3



SMOTE
imbalanced data
SPAM
HAM
Synthetic Minority Over sampling Technique
SPAM → 10
HAM → 90
1 : 9
Train 1 : 9
Test 1 : 9
Data
10
9





prediction function

$H\theta(x)$

$$\rightarrow \hat{y} = \hat{\theta}_0 + \hat{\theta}_1 X_1 + \hat{\theta}_2 X_2 + \hat{\theta}_3 X_3$$

bias

-3 +3   -3 +3   -3 +3

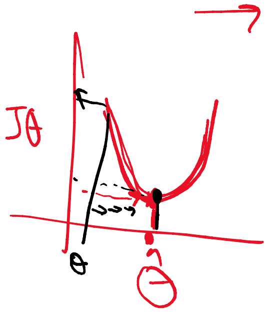
10,0000 – 50,0000

2 – 60   30000 – 250000   0.0 — 1.0

-2 4 2   -2 + 2   -2 +2

Normalize
↓
Scaling

$$\rightarrow J\theta \rightarrow MSE = \frac{\sum_{i=1}^{N} (\hat{y}_i - y_i)^2}{N}$$



sgd is very sensitive to scaling of feature

Preprocessing

① Scaling
Some models are sensitive to feature scaling so it need to be done before you fit your model.

NA

② Deal with missing values

1000 → 700

X_1 X_2 X_3 — X_n

→ ① Drop all rows in which we have missing values

1000 → 700

X_1	X_2	X_3	—	X_4
~~10~~	~~6~~	~~2~~		~~Na~~
5	6	3		4
~~Na~~	~~3~~	~~2~~		~~9~~
~~6~~	~~4~~	~~Na~~		~~9~~

30%

Data loss

→ ① Drop all rows in which we have missing values

→ ② Fix NA values

Continuous
- → mean → outlier
- → median
- → Random for IQR

Category → mode, purpose

male → 70
female → 30

X → ML

③ Feature Selection

→ drop useless column

④ Deal with text or Category data

text → NLP → vector / arrays

ML model or Computer always work on numerical Data

Numbers

$y = b + w_1x_1 + w_2x_2$

ocean_proximity
- NEARBAY
- INLOCEAN
- NE
- ⋮

Category
- ordinal → Label Encoder
- Nominal → One Hot Encoding

⑤ Feature Embedding / Feature Engineering
- drop →
- add →

→ Scaling

→ min - max Scaling → 0 - 1

→ Standard Scaling / Normalization

X

1	2	3	4
3	4	9	0

y

6
10

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

$$\boxed{\hat{x}_i = \frac{(x_i - \bar{x})}{\sigma_x}}$$

→ mean = 0
→ std = 1

Range

2 - 10

2 - 10

60 → 6

$X \rightarrow [10 \; 6 \; 3 \; 9 \; 4 \; 7 \; 2 \; 8 \; 6 \; 5]$ $(2-10)$ $\frac{60}{10} \rightarrow 6$

$X \rightarrow [4 \; 0 \; -3 \; 3 \; -2 \; 1 \; -4 \; 2 \; 0 \; -1]$

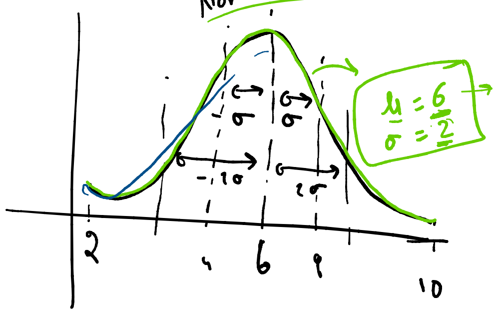
$(x_i - \bar{x}) \rightarrow d$

$(-4, 4)$

min_max

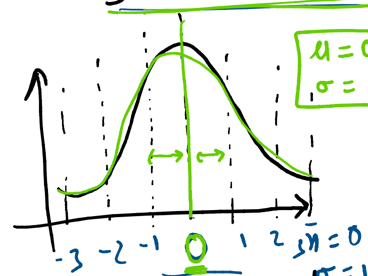
$$\frac{x_i - min}{max - min}$$

Normal Distribution



$\frac{2}{2} \rightarrow 1$

Standard Normal form

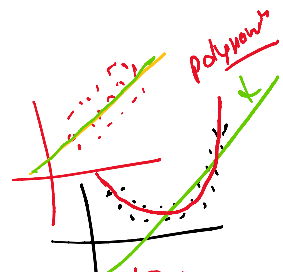


$\bar{x} = 0$, $\sigma = 1$

Normal Dist^n

$$\hat{x}_i = \frac{(x_i - \bar{x})}{\sigma_x}$$

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



Transform / Estimators

fit → parameters

transform

predict

x_i scale → $\mu \rightarrow 2$, $\sigma \rightarrow 2$

x_i (0.3)

Train

Data snooping

Test

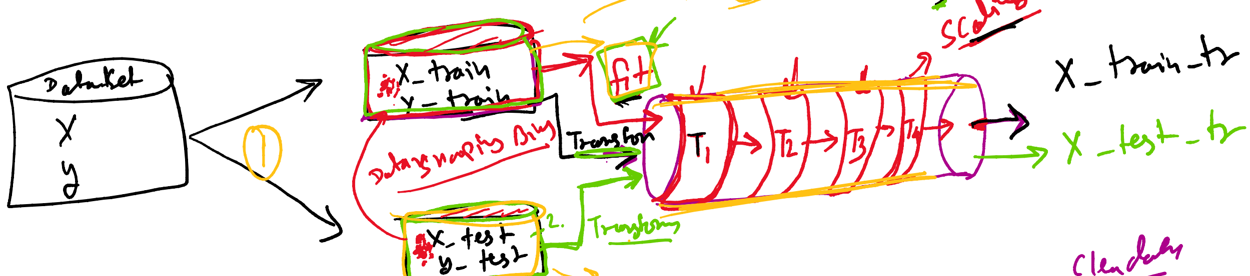
90 % → Data prep

fit transform

$\mu \rightarrow$ mean

$\sigma^2 \rightarrow$ var

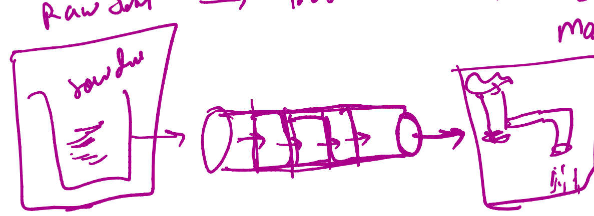
scaling



Acc ← Real world → Test

Final Acc

Raw data → Transform → Clean data → model



→ Category →

① Ordinal Category

① Ordinal Category

ordinal → nominal

perform → Ordinal Codes

fit

good	>	Avg	>	bad
3	>	2	>	1

1 2 0

perform
good
Avg
good
bad
good
bad
avg
good
⋮

Transform →

perf
3
2
3
1
3
1
2
3
⋮

Ordinal Encoding

⇒ Nominal → One Hot encoding / Binary Encoder

	Sex
0	Female
1	Male
1	Male
1	Male
0	Female
1	Male
0	Female
0	Female
0	Female
0	Female
1	Male
1	Male

fit → Transform →

Female (x_1)	Male (x_2)
1	0
0	1
0	1
0	1
1	0
0	1
1	0
1	0
1	0
0	1
0	1

0 0 0 0 0 0 -
0 0 0 1 1 0
0 0 0 1 0 0 1
0

go to →

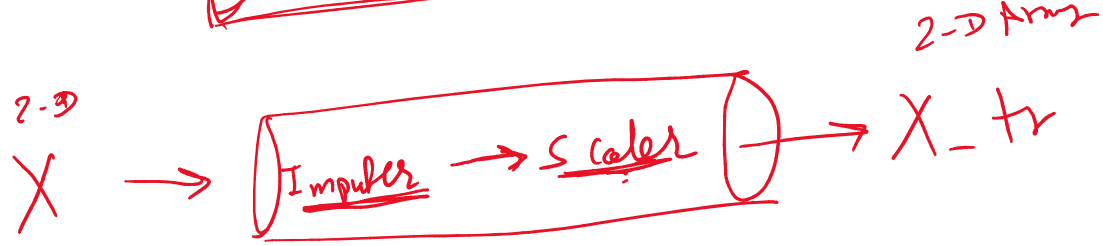
Size

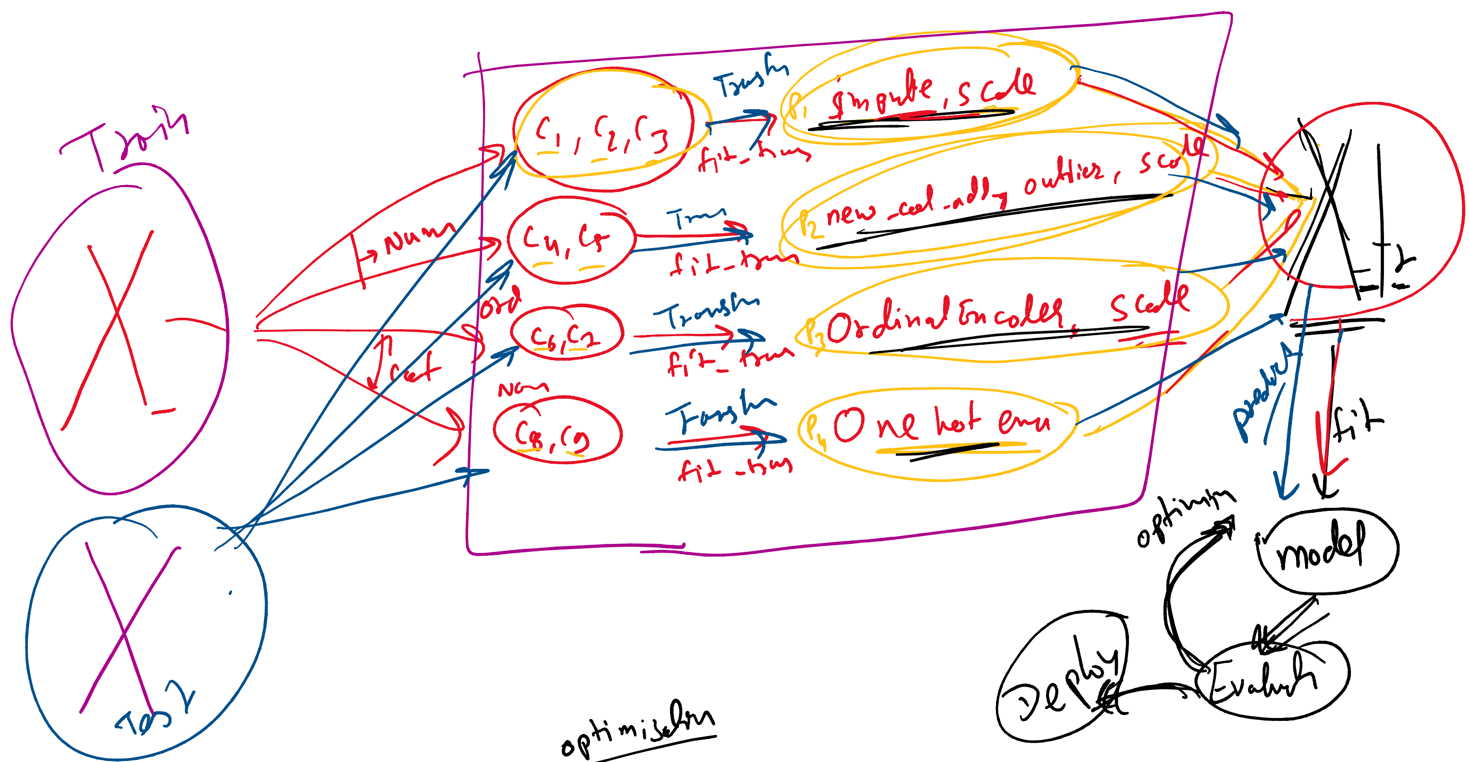
memory

time

Sparse

locations of 1





optimisation

↓

MSE → RMSE (Evaluate)

Error → MSE

Error → MAE

Accuracy → R2-score

$$\boxed{r2\text{-}score = 1 - \frac{RSS}{TSS}}$$

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

$r2$ - score = ... TSS

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

0 - ∞ → Random model

Total Sum of Squared Error $= \sum_{i=1}^{N} (\bar{y} - y_i)^2$ (Train version)

0 - ∞

Residual Sum of Squared Error $= \sum_{i=1}^{N} (\hat{y}_i - y_i)^2$ (prediction Error)

TSS > RSS

r^2 - score → 1 → 100%

r^2 - score → 0 → 100%

r^2 - score → 0.7 → 70%

r^2 score →

99.9992

Error

$\frac{RSS}{TSS}$ → $\frac{0 - \infty}{0 - \infty} = \frac{a}{b}$ → ...

0 - 1

0 - 1 →

0 - 1

$\frac{RSS}{TSS}$ → 0.0001 → Low / Very High

-125×10^9

$r^2 =$

$\frac{a}{b} > 1$

RSS ↑ very high → 1000

TSS → low 1 - 1000 → -999

$\boxed{b > a}$

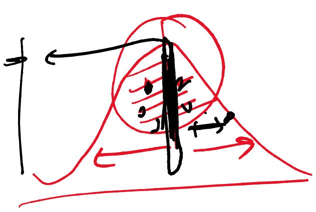
RSS ≃ TSS → 1

$\frac{a}{b} < 1 \to a < b$

RSS < TSS ≮ 1

$$\boxed{R^2 \text{ score} = 1 - \frac{RSS}{TSS}}$$

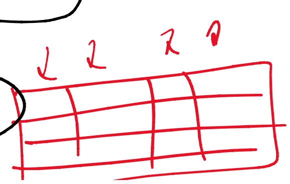
$\frac{0.6}{0.9}$



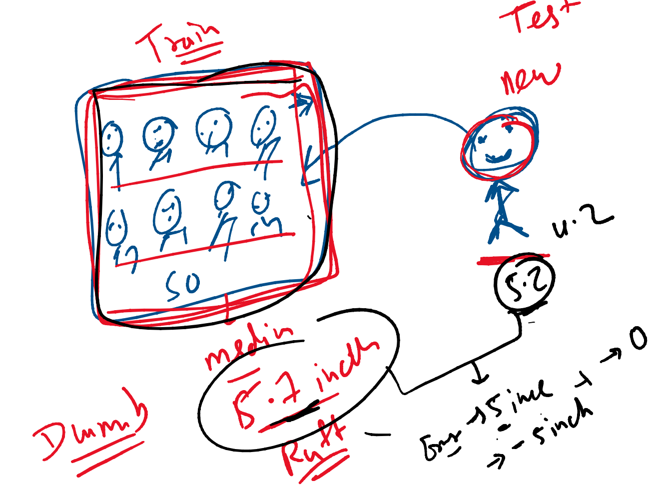
mean ± std → 68

Test

Random



(mean) ± STD



Random

$$TSS_{Error} = \sum_{i=1}^{N} (\bar{y} - y_i)^2$$

pred

your model

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

$$\frac{RSS}{TSS} \rightarrow \frac{\text{your model Error}}{\text{Random model Error}} = \frac{\text{Low}}{\text{High}} \rightarrow 0-1$$

0-1

Error + Acc = 1

Acc = (1 - Error)

-1.79×10^9

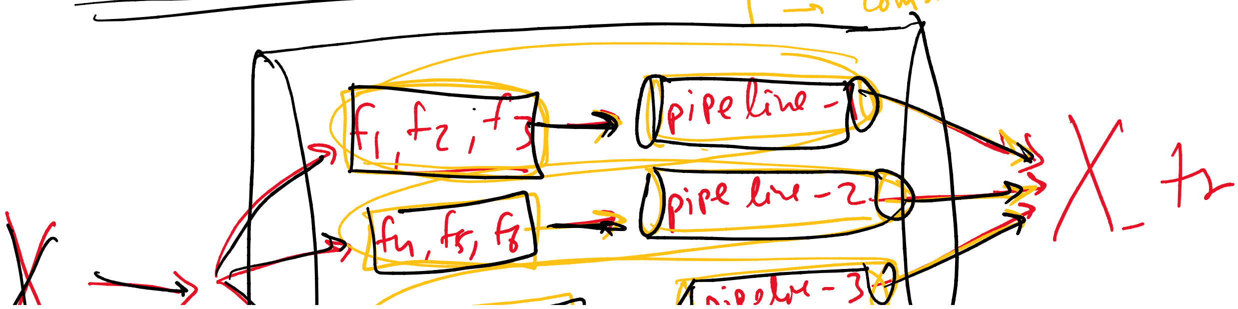
$$\frac{-10^{10}}{10} \rightarrow 10^9 \quad -10^9$$

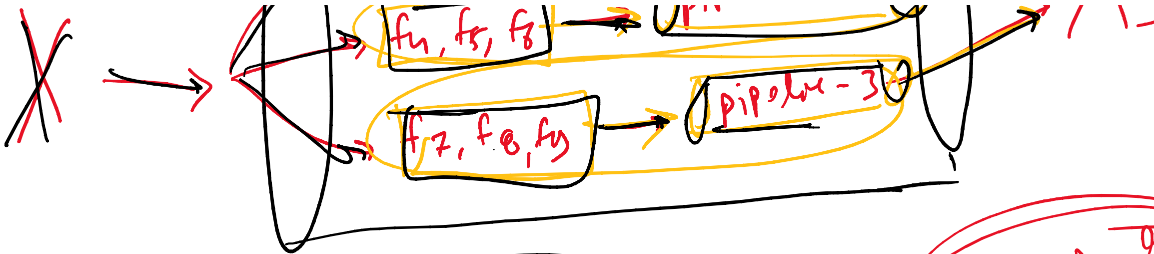
$$\boxed{R^2 = 1 - \frac{RSS}{TSS}} \rightarrow$$

0-1 →

⇒ ColumnTransformer

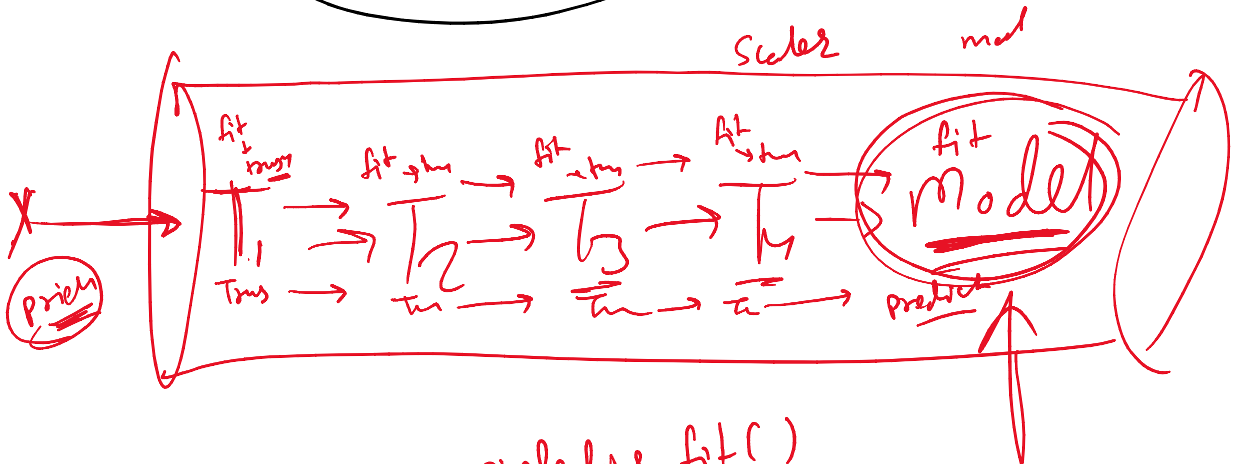
1 → split
2 → build
3 → Att_trn
4 → Combine





1- step

Bag Estimator



→ pipeline.fit()

→ pipeline.predict()

BO → Data pipe

→ Regression →

→ Classification Evalu → metrics

precision
recall
F1
ROC
AUC

→ Algo →

SVR
LR
DT
RF
KNN
Navie Bays

gun
Xm 3-

1000's

3-4

thugs