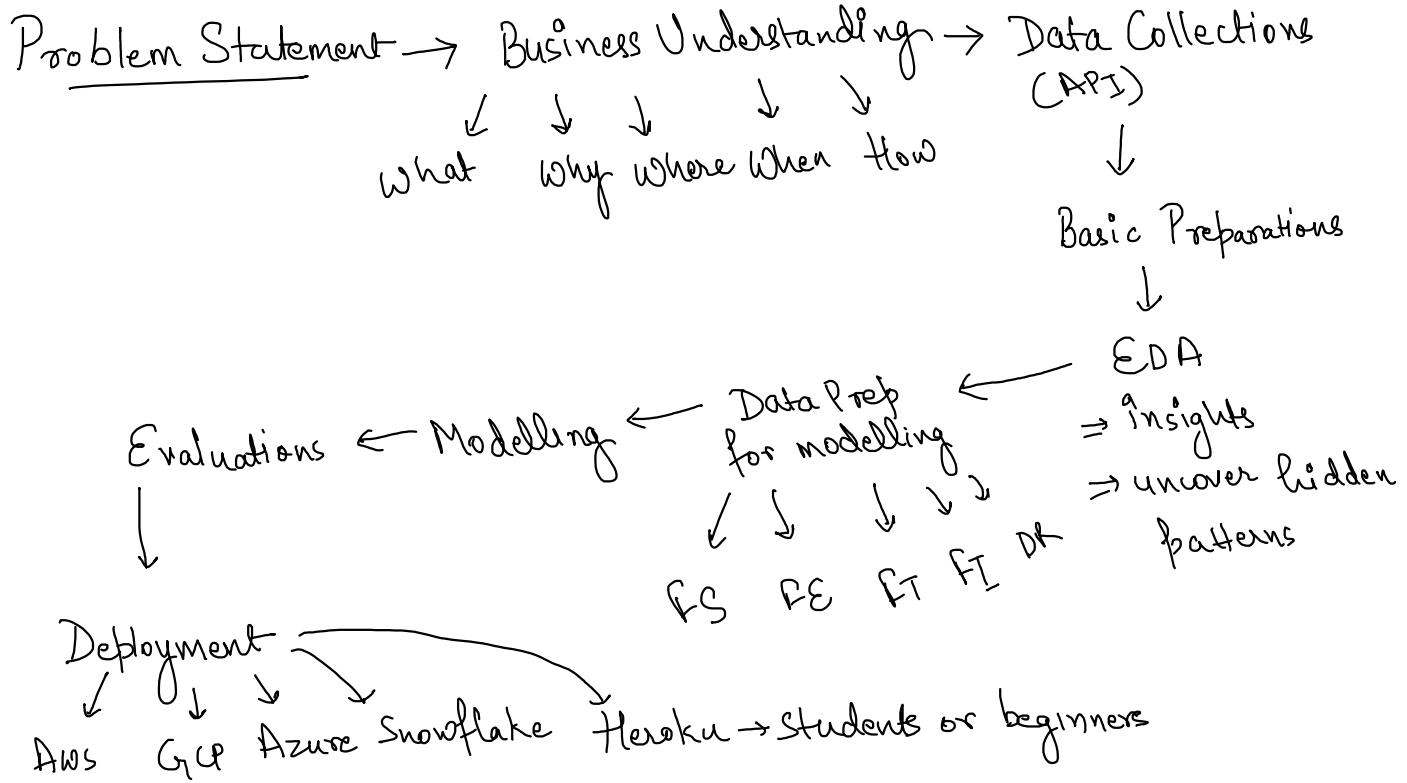


# Lifecycle of Data Science Project

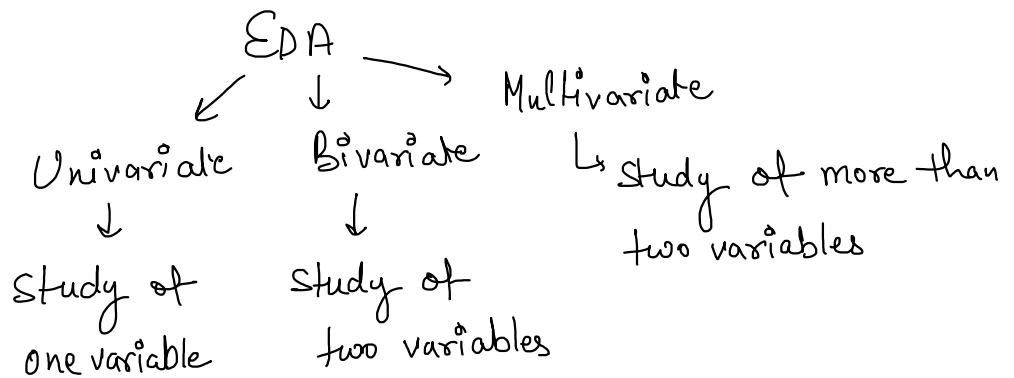


## Starting Project

- Import libraries & load the data
- Basic information  $\Rightarrow df.info()$ 
  - metadata
  - nulls
  - datatype
  - # columns
  - # rows
- Basic description  $\Rightarrow df.describe()$ 
  - count
  - Mean
  - Std dev
  - Quartiles
  - min
  - max

→ Basic Data Preparation ⇒ data quality check / data assessment

→ EDA ⇒ exploratory data analysis

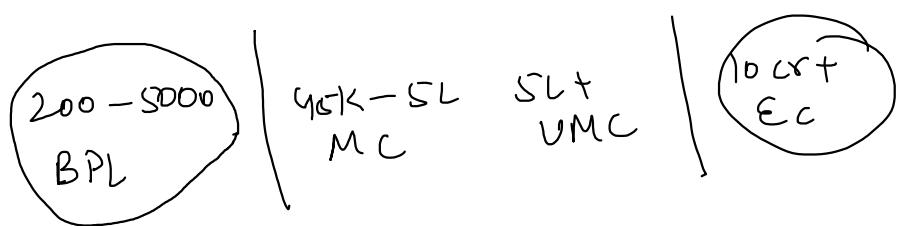


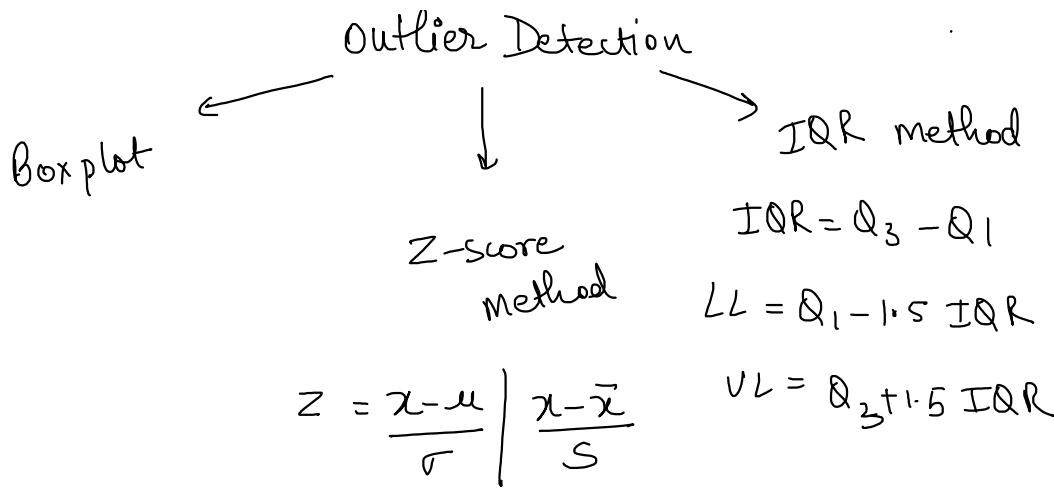
Univariate Analysis ⇒ countplot, distplot, boxplot, histogram, kdeplot

Bivariate Analysis ⇒ Scatter, pie, bar, lineplot, Swarmplot

Multivariate Analysis ⇒ heatmap, Pairplot

## Outliers

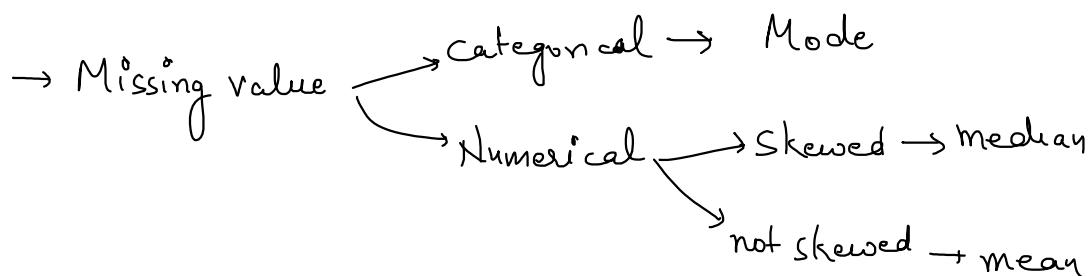
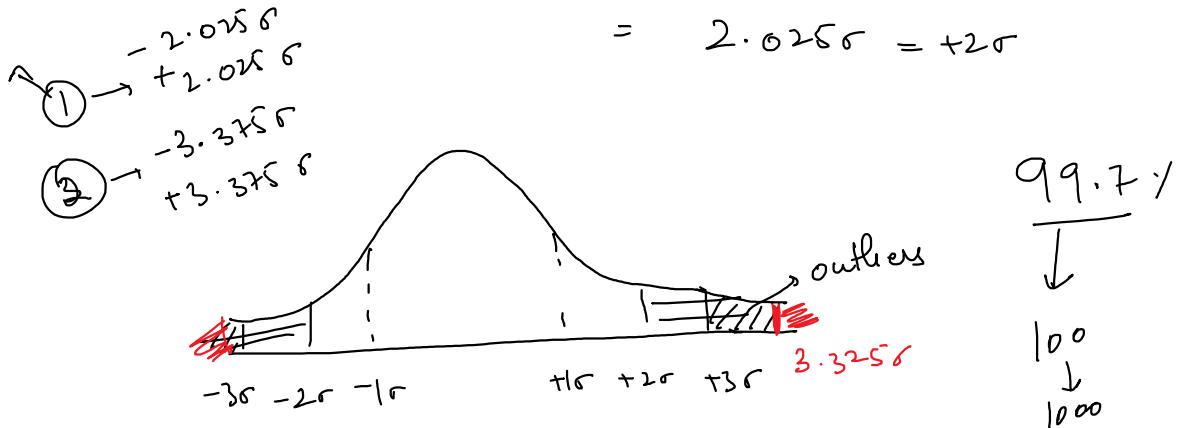


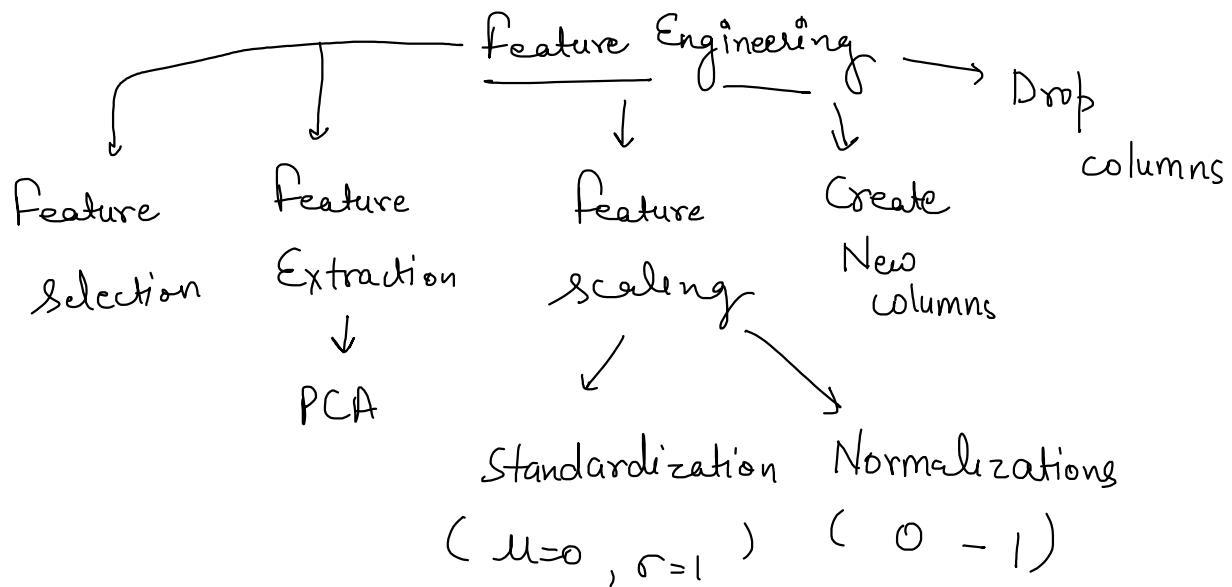


$$Q_1 \Rightarrow -0.675\sigma \quad Q_3 \Rightarrow +0.675\sigma$$

$$\begin{aligned} LL &= Q_1 - 1.5 [Q_3 - Q_1] = -0.675\sigma - [0.675\sigma - (-0.675\sigma)] \\ &= -0.675\sigma - [1.35\sigma] \\ &= -2.025\sigma = -2\sigma \end{aligned}$$

$$UL = Q_3 + 1.5 [Q_3 - Q_1] = +0.675\sigma + [1.35\sigma]$$

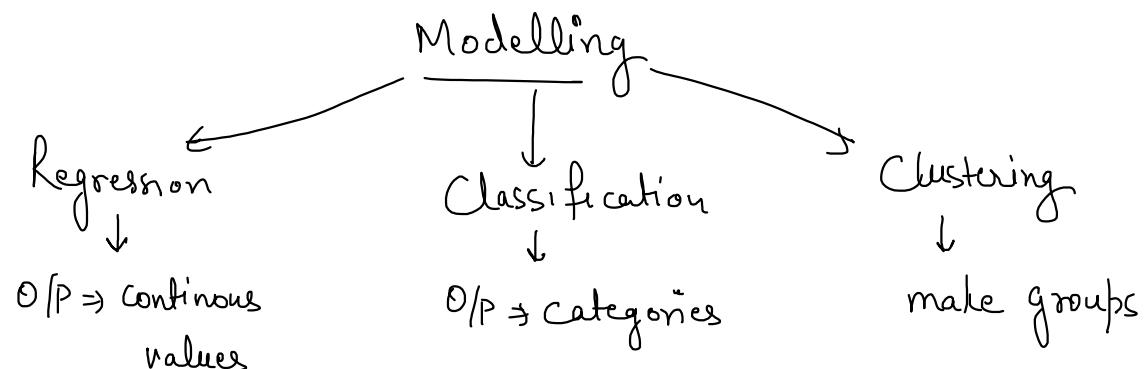




Encoding :  $\rightarrow$  OHE (dummy variable)  $\Rightarrow$  "drop-first=True"

| Category (nominal) |  | A | B | C | D | E |
|--------------------|--|---|---|---|---|---|
| A                  |  | 1 | 0 | 0 | 0 | 0 |
| B                  |  | 0 | 1 | 0 | 0 | 0 |
| C                  |  | 0 | 0 | 1 | 0 | 0 |
| D                  |  | 0 | 0 | 0 | 1 | 0 |
| E                  |  | 0 | 0 | 0 | 0 | 1 |

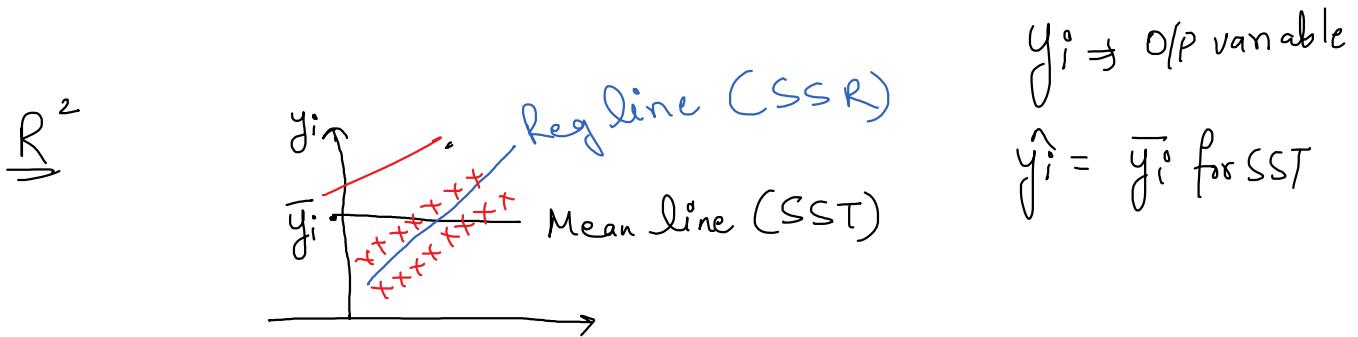
Label encoding / Ordinal encoding  $\Rightarrow$  order is present in category



## Evaluations

→ Regressions       $\begin{cases} \text{fitness based} \Rightarrow R^2, \text{Adj } R^2 \\ \text{error based} \end{cases}$

$\Rightarrow \text{MAE, MSE, RMSE, MAPE}$



$$R^2 = 1 - \frac{SSR}{SST}$$

Case 1 :  $SSR = 0$ ,  $SST = SST$

$$R^2 = 1 - \frac{0}{SST} = 1 - 0 = 1$$

Case 2 :  $SSR = SST$ ,  $SST = SST$

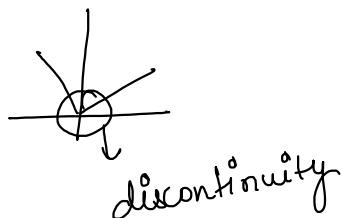
$$R^2 = 1 - \frac{SST}{SST} = 1 - 1 = 0$$

Case 3 :  $SSR > SST$

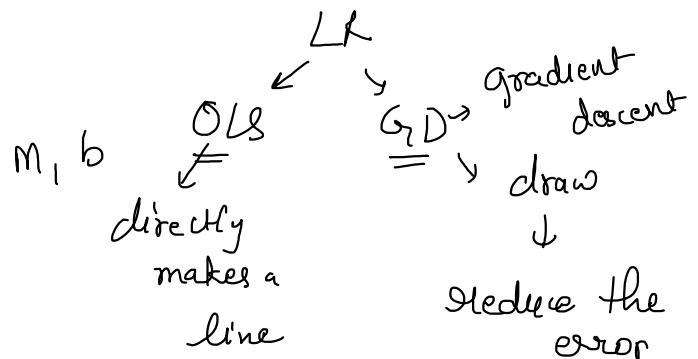
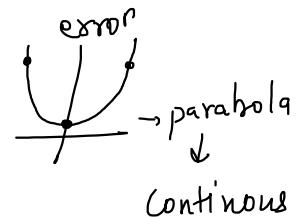
$$\frac{SSR}{SST} > 1$$

$$R^2 = 1 - \frac{SSR}{SST} = -ve$$

$$MAE = \frac{1}{n} \sum (y - \hat{y})$$



$$MSE = \frac{1}{n} \sum (y - \hat{y})^2$$



## Classification

$$\rightarrow \text{Accuracy} = \frac{TP + TN}{TP + FP + TN + FN}$$

## CONFUSION MATRIX

$$\rightarrow \text{Precision} = \frac{TP}{TP + FP}$$

$$\rightarrow \text{Recall or Sensitivity} = \frac{TP}{TP + FN}$$

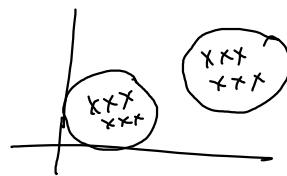
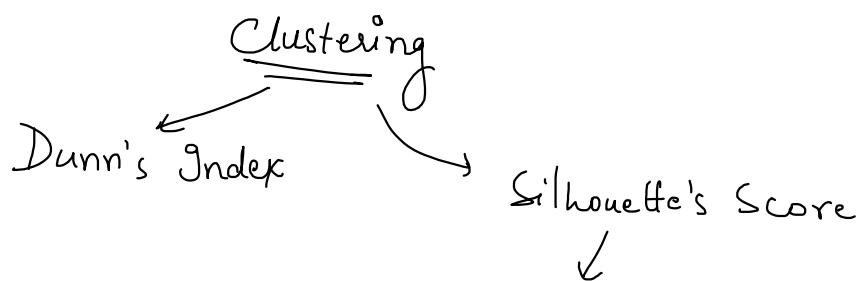
|        |   | Prediction |    |
|--------|---|------------|----|
|        |   | 0          | 1  |
| Actual | 0 | TN         | FP |
|        | 1 | FN         | TP |

$$\rightarrow F1 \text{ score} = \frac{2 \times P \times R}{P + R}$$

$$\rightarrow ROC - AUC \Rightarrow$$

$$FPR = \frac{FP}{TN + FP}$$

$$TNR = \frac{TN}{TN + FP}$$



$$\text{Score} = \frac{b - a}{\max(b, a)}$$

### Indo Pak Relations

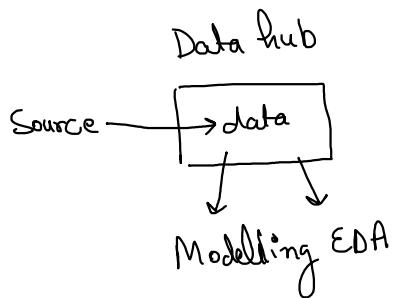
| <u>Urmil</u>                     | <u>Rashmi</u>             | <u>Deepak</u>                   |
|----------------------------------|---------------------------|---------------------------------|
| - GDP of India is greater        | - Area of India is bigger | - religious state / secular     |
| - population of India is greater | - literacy rate is higher | - technology is better in India |

### Project

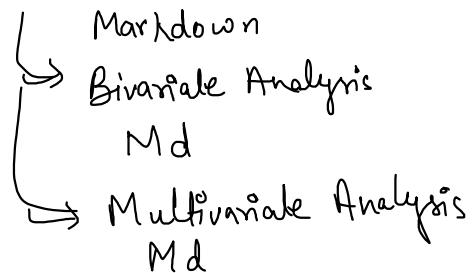
"Air Ticket Price Prediction"

- Data Transformation
- Feature Engineering
- EDA
- Modelling
- Evaluation

↓  
Regression



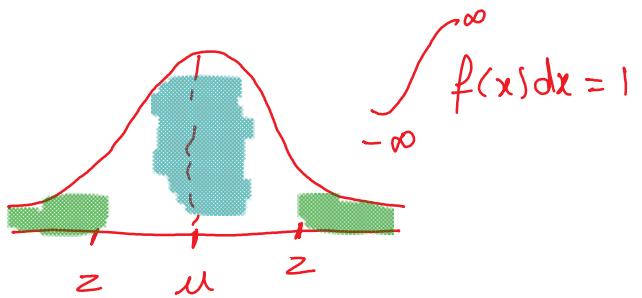
EDA  $\Rightarrow$  Univariate Analysis



$Z$ -score & calculate probability value

$$Z = \frac{x - \mu}{\sigma}$$

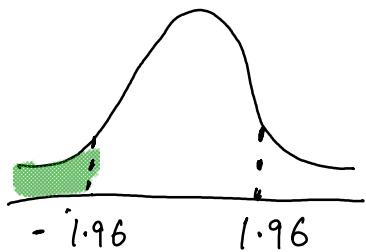
area under curve = probability.



Q  $Z$ -score, can we connect it with probability or can we get probability value for  $Z$ -score?

$$Z\text{-score} = -1.96, \text{ probability} = ?$$

Sol.



$$\Rightarrow \int_{-\infty}^{-1.96} \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} dx = 0.025$$

↓  
probability

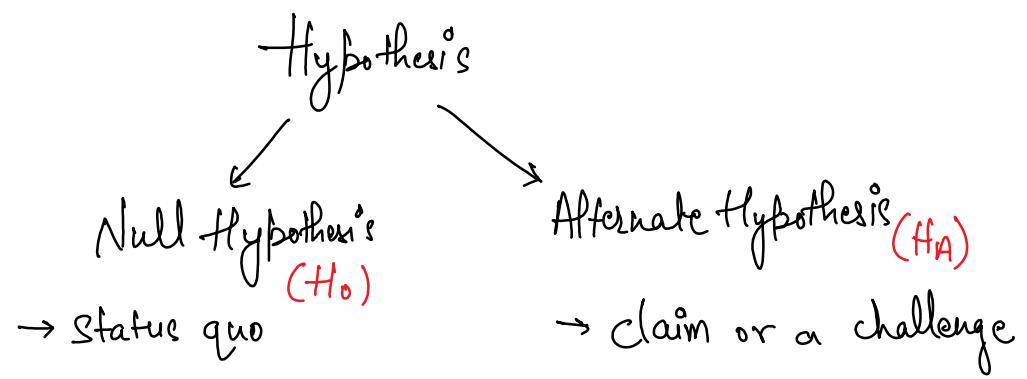
$$Z\text{-Score} \Rightarrow \frac{x - \mu}{\left(\frac{\sigma}{\sqrt{n}}\right)} \rightarrow \text{std error}$$

$$\text{std error} = \frac{\sigma}{\sqrt{n}}$$

## Hypothesis Testing

Hypothesis  $\Rightarrow$  belief

Hypothesis Testing  $\Rightarrow$  Testing Your belief



Q Police claims that a person is criminal?

Sol.  $H_0$  = innocent

$H_1$  = criminal

Q I claim that 90% of this class is above average?

Sol.  $H_0$  : below average or average

$H_1$  : above average

Q RCB is going to win IPL?  $\rightarrow$  Madhukar's claim

Sol.  $H_0$  : Any other team can win

$H_1$  : RCB win

Q Bride claims that groom has taken dowry?

Sol.  $H_0$  : innocent

$H_1$  : has taken dowry / guilty

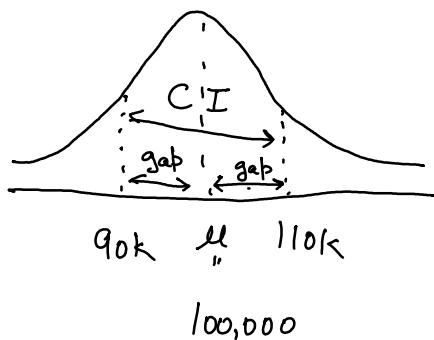
Reality:

$H_0$  : guilty

$H_1$  : innocence

$$H_A : \text{has taken dowry/guilty} \quad \left. \begin{array}{l} \\ \end{array} \right\} H_A : \text{innocence}$$

Build the criteria to test hypothesis :-



$$\begin{array}{l} S_1 = \bar{x}_1 \\ S_2 = \bar{x}_2 \\ \text{population} \end{array}$$

$$n = 100,000$$

In order to build CI,

$$UL = \mu + \text{gap}$$

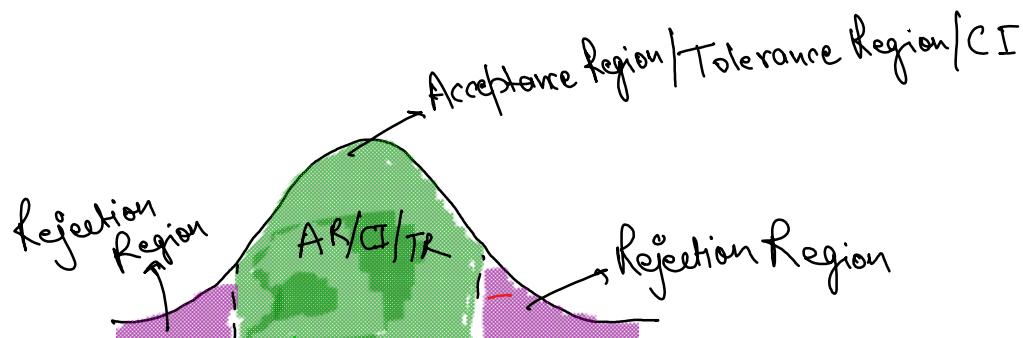
$$LL = \mu - \text{gap}$$

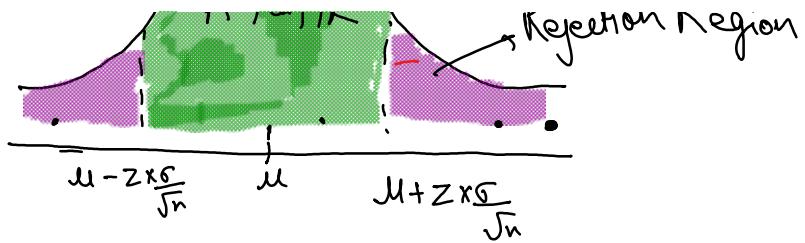
$$\begin{array}{l} \text{gap} \\ \downarrow \\ \text{Std error} \\ \text{Z-Score} \quad (\sigma/\sqrt{n}) \end{array}$$

$$UL = \mu + z \times \frac{\sigma}{\sqrt{n}}$$

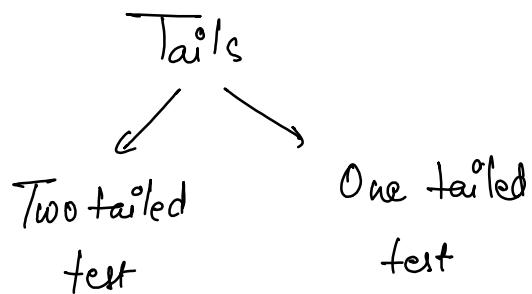
$\frac{\sigma}{\sqrt{n}}$  Margin of error

$\Rightarrow$  Acceptance Region Method

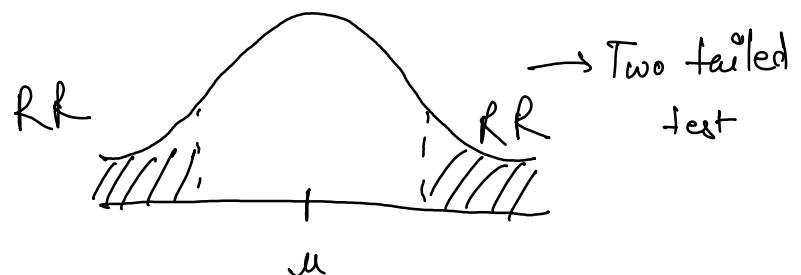




$$AR + RR = 1$$

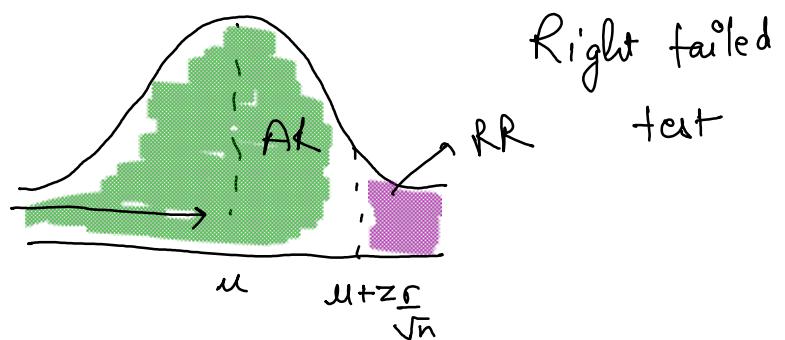


1)  $H_0: \mu = \$100,000$   
 $H_A: \mu \neq \$100,000$



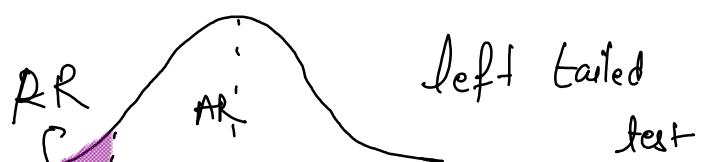
2)  $\mu \leq \$100,000 \Rightarrow H_0$

$\mu > \$100,000 \Rightarrow H_A$

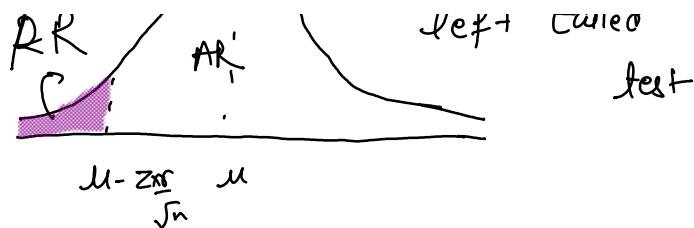


3)  $\mu \geq \$100,000 = H_0$

$\mu < \$100,000 = H_A$



$$\mu < \$100,000 = H_A$$



## Critical Value Method:

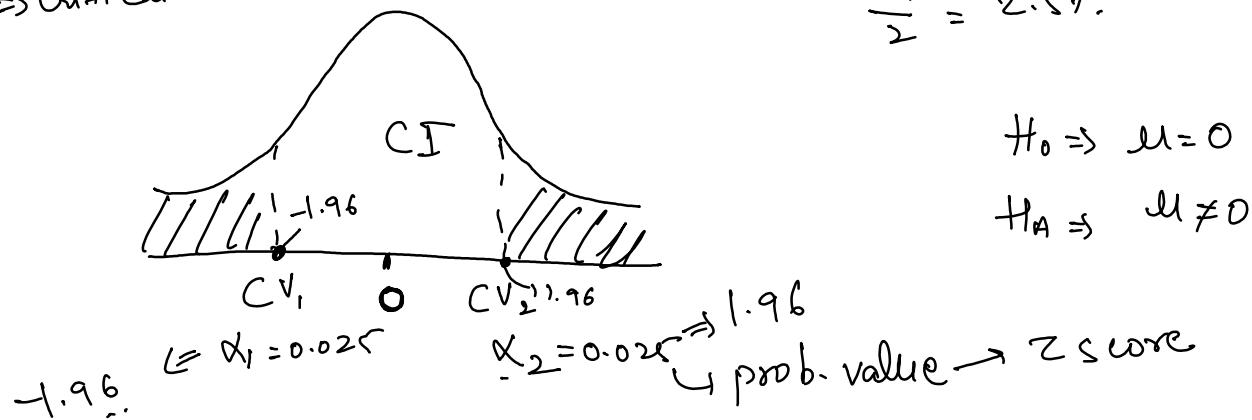
Where to build Rejection Region? → Significance level ( $\alpha$ )

$$CI \Rightarrow 95\%, \quad \underline{\alpha = 5\%}$$

$$CI + SL = 1$$

CV ⇒ critical value

$$\alpha = \frac{5\%}{2} = 2.5\%$$



$$CI + \alpha_1 + \alpha_2 = 1$$

$$\alpha = 0.025 + 0.025 = 0.05$$

$$CI + \alpha = 1$$

$$\alpha = 1 - CI \quad \text{or} \quad CI = 1 - \alpha$$

Thumb rule: If SL is not given, then take  $\alpha = 5\%$  or  $0.05$   
(significance level)

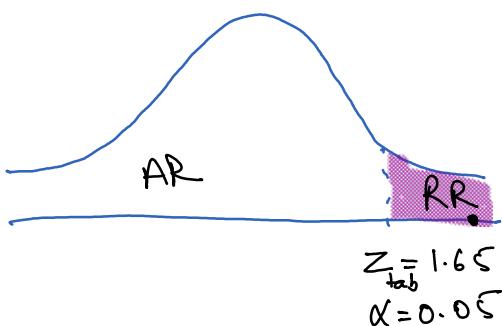
Q1 Q3:

$\alpha$  (significance level)

Steps:

$$\underline{\underline{X = 0.05 \rightarrow \text{Test} \textcircled{1}}}$$

One tailed  $\rightarrow 0.05$   
two tailed  $\rightarrow \frac{0.05}{2} = 0.025$



(II)  $Z_{\text{cal}} \Rightarrow \frac{x - \mu}{\sigma / \sqrt{n}}$

(III) Compare  $Z_{\text{tab}}$  with  $Z_{\text{cal}}$ :

1)  $Z_{\text{cal}} > Z_{\text{tab}}$

$H_0$  Rejected

2)  $Z_{\text{cal}} < Z_{\text{tab}}$

$H_0$  accepted

Q A principal at school claiming that the students in his school has above average IQ. A random sample is taken (30 stds) with average of 112.5. The mean  $\mu$  & std. dev of population is 100 and 15. Test the hypothesis!

$X = 0.05 \rightarrow Z_{\text{tab}} = 1.65$

Sol.  $H_0: \mu \leq 100$



Right tailed

Sol.  
=

$$H_0: \mu \leq 100$$



Right Tailed

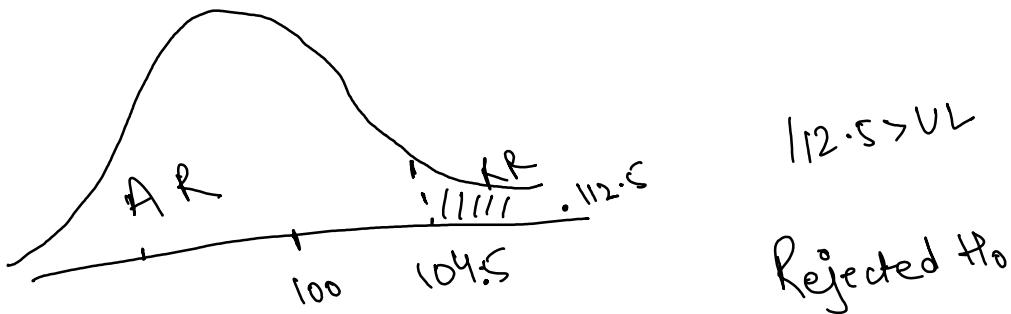
$$H_A: \mu > 100$$

a) AR

$$\mu = 100, \sigma = 15, n = 30, \bar{x} = 112.5$$

$$VL: \mu + Z \times \frac{\sigma}{\sqrt{n}} = 100 + 1.65 \times \frac{15}{\sqrt{30}} = 104.52$$

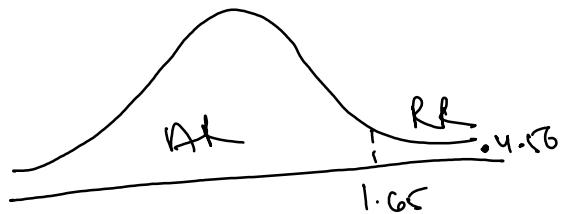
$$LL: \mu - Z \times \frac{\sigma}{\sqrt{n}} = 100 - 1.65 \times \frac{15}{\sqrt{30}} = 95.5$$



## 2) CRITICAL VALUE

$$Z_{cal} = \frac{\bar{x} - \mu}{\sigma / \sqrt{n}} = \frac{112.5 - 100}{15 / \sqrt{30}} = 4.56$$

$$Z_{tab} = 1.65$$



$$Z_{cal} > Z_{tab}$$

Hence, Rejected  $H_0$

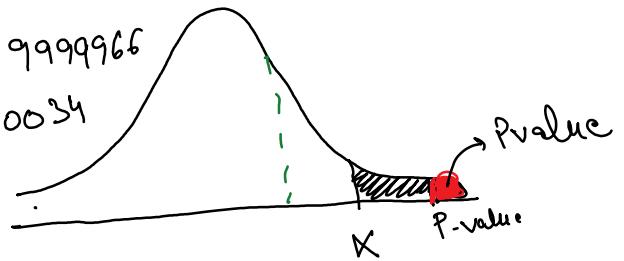
3) P value : probability of your  $H_0$  to be true.

$$z_{\text{cal}} = 4.56$$

$0.9999966 \Rightarrow$  area to the left

$$\begin{aligned} P\text{ value } (z_{\text{cal}} = 4.56) &= 1 - 0.9999966 \\ &= 0.0000034 \end{aligned}$$

$$\alpha = 0.05$$

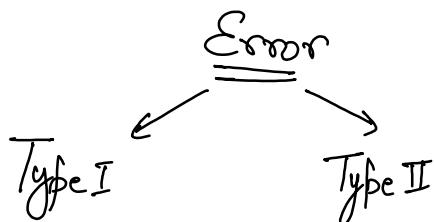


$P\text{ value} < X \Rightarrow$  Reject  $H_0$

$P\text{ value} > X \Rightarrow$  Failed to Reject  $H_0$

$$0.0000034 < 0.05$$

Hence, Rejected  $H_0$



|           |                        | Actual                               |                |
|-----------|------------------------|--------------------------------------|----------------|
|           |                        | $H_0$ is true                        | $H_0$ is false |
| Predicted | Reject $H_0$           | Type I error $\underline{\text{fp}}$ | ✓              |
|           | failed to Reject $H_0$ | Type II error                        |                |

Type I:

→  $H_0$  is true but  $H_A$  is accepted

Type II:

→  $H_A$  is true but  $H_0$  is accepted

$$\text{Type I} \propto \frac{1}{\text{Type II}}$$

- How to Quantify Type I error?  $\rightarrow \alpha =$
- How to Quantify Type II error?  $\rightarrow \beta$

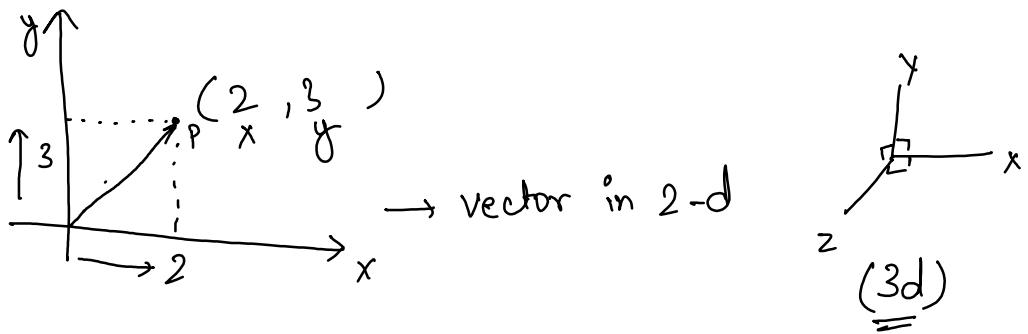
Power of test  $\Rightarrow$  ability of the test to make right decisions  
 ↓  
 influence by no. of samples

$$\begin{aligned} \text{Power} &= 1 - \beta \\ \beta &= 1 - \text{Power} \end{aligned}$$

- Q A researcher has agreed upon data on daily return of portfolio of call option over a recent 250 days period. The mean of daily return is 0.11% std. dev. 0.25%. The researcher believes the mean daily portion is not 0.
- Construct HT of 95% CI.

# Linear Algebra

Sunday, September 24, 2023 8:32 AM



Vector in 2d =  $[2, 3]$

" " 3d =  $[2, 3, 4]$

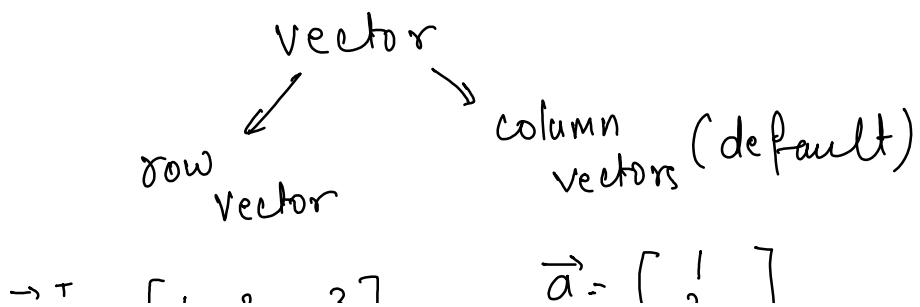
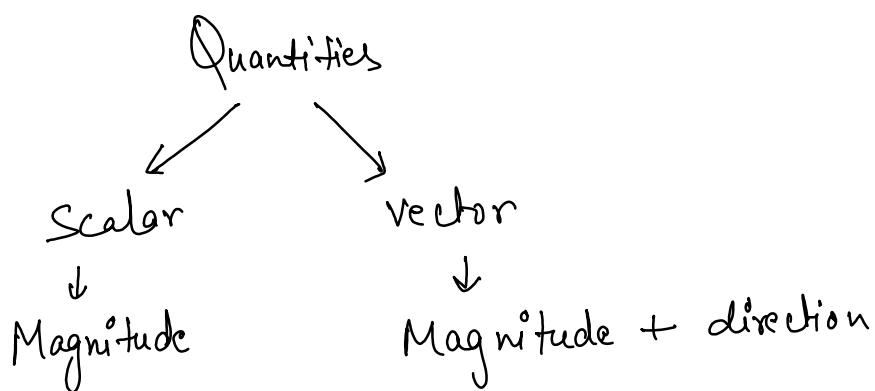
" " 4d =  $[2, 3, 4, 6]$

" " 6d =  $[2, 3, 4, 5, 6, 7]$

" " nd =  $[2, 3, 4, \dots, n]_{1 \times n}$

↳ n-dimensional vector  
(array)

Vector is an n-d array with shape  $(1 \times n)$



Vectors

$$\vec{a}^T = \begin{bmatrix} 1 & 2 & 3 \end{bmatrix} \quad \vec{a} = \begin{bmatrix} 1 \\ 2 \\ 3 \end{bmatrix}$$

## MATRICES

Rows (R)  $\rightarrow \begin{bmatrix} 1 & 4 & 7 \\ 2 & 5 & 8 \\ 3 & 6 & 9 \end{bmatrix}$  (RxC)  $\downarrow$  What is the use of matrix?

Linear Transformation

Square Matrix  $\Rightarrow$  Rows = columns

columns shape values  
 $\Rightarrow 10 \Rightarrow 10 \times 10 \Rightarrow 100 \Rightarrow$  Correlation matrix

$$\begin{bmatrix} 1 & 2 & 5 \\ 0 & 4 & 8 \\ 1 & 3 & 9 \end{bmatrix}_{3 \times 3}$$

Identity Matrix  $\Rightarrow$  np.eye(3)

$$\begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}_{3 \times 3}$$

## MATRIX OPERATIONS

$\Rightarrow$  Addition

$$[a] \quad [b] \quad [c] \quad [d] \quad [a+b]$$

→ Rule :-

$$\begin{bmatrix} a & b \\ c & d \end{bmatrix} + \begin{bmatrix} e & f \\ g & h \end{bmatrix} = \begin{bmatrix} a+e & b+f \\ c+g & d+h \end{bmatrix}$$

⇒ Multiplication

$$\begin{bmatrix} a \\ b \end{bmatrix}_{2 \times 1} \begin{bmatrix} c & d \end{bmatrix}_{1 \times 2} = \begin{bmatrix} ac & ad \\ bc & bd \end{bmatrix}_{2 \times 2}$$

\* Multiplication is only possible when

columns of 1<sup>st</sup> matrix = rows of 2<sup>nd</sup> matrix

\* Size of Matrix after multiplication

$$A_{m \times n} \times B_{n \times p} = C_{m \times p}$$

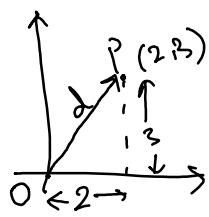
Q

$$\begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix}_{2 \times 2} \times \begin{bmatrix} 7 & 8 \\ 4 & 9 \end{bmatrix}_{2 \times 2}$$

Sol.

$$= \begin{bmatrix} 7+8 & 8+18 \\ 1 \times 7 + 2 \times 4 & 1 \times 8 + 2 \times 9 \\ 21+16 & 24+36 \\ 3 \times 7 + 4 \times 4 & 3 \times 8 + 4 \times 9 \end{bmatrix} \Rightarrow \begin{bmatrix} 15 & 26 \\ 37 & 60 \end{bmatrix}_{2 \times 2}$$

Concepts of Linear Algebra



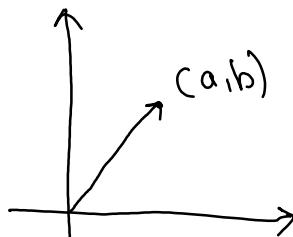
Distance of P from origin?

By Pythagoras theorem

$$d^2 = 3^2 + 2^2$$

$$d = \sqrt{3^2 + 2^2} = \sqrt{13}$$

\* d is the shortest distance b/w O & P



$$d = \sqrt{a^2 + b^2}$$

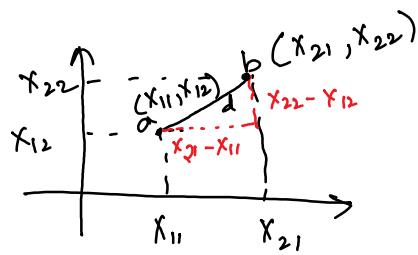
For 3d,

$$d = \sqrt{a^2 + b^2 + c^2}$$

For nd,

$$d = \sqrt{a^2 + b^2 + c^2 + \dots + n^2}$$

Distance between two points:



$$a = (x_{11}, x_{12})$$

$$b = (x_{21}, x_{22})$$

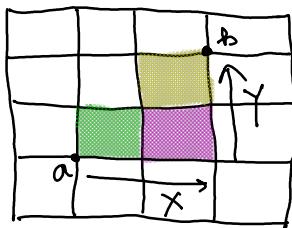
By Pythagoras Theorem,

$$d = \sqrt{(x_{21} - x_{11})^2 + (x_{22} - x_{12})^2}$$

euclidean distance

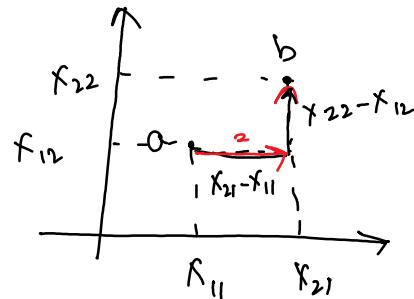
$$\text{Euclidean distance} = \left[ \sum_{i=1}^n |x_{1i} - x_{2i}|^2 \right]^{1/2}$$

Manhattan Distance: When you have large dimensionality (no. of columns) use manhattan distance.



$$X = |x_{21} - x_{11}|$$

$$Y = |x_{22} - x_{12}|$$



$$d = |x_{21} - x_{11}| + |x_{22} - x_{12}|$$

$$\text{Manhattan distance} = \sum_{i=1}^n |x_{1i} - x_{2i}|$$

Minkowski Distance

$$d = \left[ \sum_{i=1}^n |x_{1i} - x_{2i}|^p \right]^{1/p} \quad \text{where } p=1, 2, 3, \dots$$

$$d = \left[ \sum_{i=1}^n |x_{1i} - x_{2i}|^p \right]^{1/p} \quad \text{where } p=1, 2, 3, \dots$$

Lets take  $p=1$ ,

$$d = \sum_{i=1}^n |x_{1i} - x_{2i}| \rightarrow \text{Manhattan distance}$$

Lets take  $p=2$ ,

$$d = \left[ \sum_{i=1}^n |x_{1i} - x_{2i}|^2 \right]^{1/2} \rightarrow \text{Euclidean distance}$$

## Vector Multiplication

dot product

scalar (magnitude or number)

cross product  $\times$

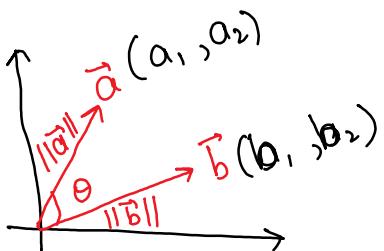
Dot product in linear Algebra:

$$a = \begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_n \end{bmatrix}_{n \times 1} \quad b = \begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_n \end{bmatrix}_{n \times 1}$$

$$C = \underset{1 \times 1}{a \cdot b} \Rightarrow \underset{n \times 1}{a^T \cdot b} = [a_1, a_2, \dots, a_n] \cdot \begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_n \end{bmatrix}$$

$$C = \underset{1 \times 1}{[a_1 b_1 + a_2 b_2 + a_3 b_3 + \dots + a_n b_n]}$$

Angle b/w 2 vectors



Geometric way,  $a \cdot b = \|\vec{a}\| \|\vec{b}\| \cos \theta$  - ①

linear algebra, ,  $a \cdot b = [a_1 b_1 + a_2 b_2]$  - ②

$$\|\vec{a}\| \|\vec{b}\| \cos \theta = [a_1 b_1 + a_2 b_2]$$

$$\cos \theta = \frac{a_1 b_1 + a_2 b_2}{\|\vec{a}\| \|\vec{b}\|}$$

$$\theta = \cos^{-1} \left[ \frac{a_1 b_1 + a_2 b_2}{\|\vec{a}\| \|\vec{b}\|} \right]$$

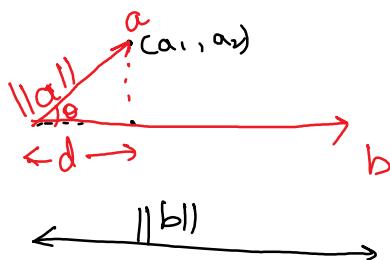
$$\theta = \cos^{-1} \left[ \frac{\sum_{i=1}^n a_i b_i}{\|\vec{a}\| \|\vec{b}\|} \right]$$

where,

$$\|\vec{a}\| = \sqrt{a_1^2 + a_2^2 + \dots + a_n^2}$$

$$\|\vec{b}\| = \sqrt{b_1^2 + b_2^2 + \dots + b_n^2}$$

## Projection :



$$d = \|\vec{a}\| \cos \theta$$

dot product of  $a \cdot b$ ,

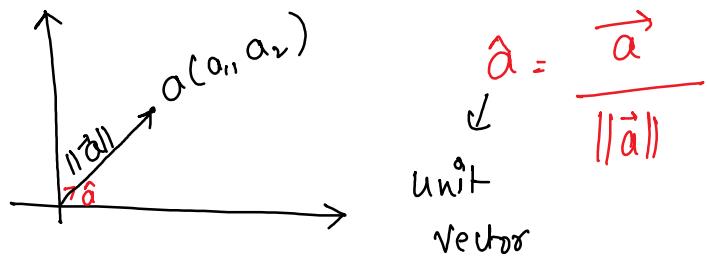
$$a \cdot b = \|\vec{a}\| \|\vec{b}\| \cos \theta$$

$$a \cdot b = d \|\vec{b}\|$$

Projection of  $a$  on  $b$ ,

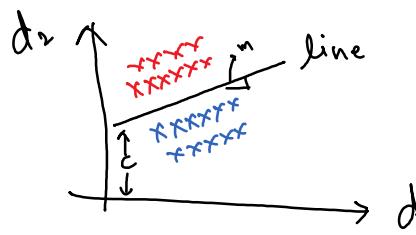
$$\boxed{d = \frac{a \cdot b}{\|\vec{b}\|}}$$

Unit Vector: gives information about direction



## Lines and Planes

Line:  
(2d)

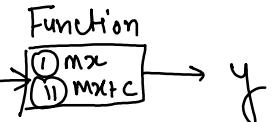


equation of line,  
 $y = mx + c \rightarrow$  y-intercept  
 ↓ slope independent variable / g/p  
 dependent variable / o/p variable

$$m = \frac{y_2 - y_1}{x_2 - x_1} = \tan \theta = \frac{d}{dx} \rightarrow \text{calculus}$$

General Equation of line :

$$ax + by + c = 0 \quad y = mx + c$$



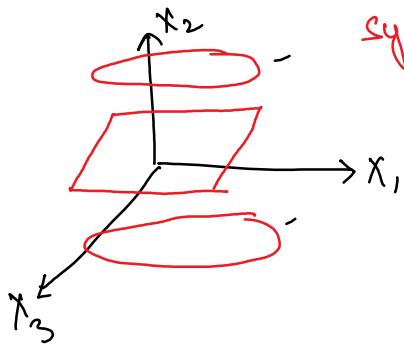
$$by = -c - ax$$

$$y = -\frac{c}{b} - \frac{a}{b}x$$

$$y = \left( -\frac{a}{b} \right)x - \left( \frac{c}{b} \right) = c \rightarrow \begin{matrix} \text{y-intercept} \\ \text{m} \\ (\text{slope}) \end{matrix}$$

Plane

→ divides the 3-d co-ordinate system into 2 spaces



General Equation of plane:

$$ax + by + cz + d = 0$$

↓ change the constant

$$w_1x + w_2y + w_3z + w_0 = 0$$

↓ change the axis name

$$\text{eqn of plane} \Rightarrow w_1x_1 + w_2x_2 + w_3x_3 + w_0 = 0 \quad (3d)$$

for n-dimensions,  $\rightarrow$  hyperplane.

General eqn of hyperplane:

$$w_0 + w_1x_1 + w_2x_2 + w_3x_3 + \dots + w_nx_n = 0$$

Equation of hyperplane:  $w_0 + \sum_{i=1}^n w_i x_i = 0$

Hyperplane equation:  $w_0 + [w_1 x_1 + w_2 x_2 + w_3 x_3 + \dots + w_n x_n] = 0$

$$w_0 + [w_1 \ w_2 \ w_3 \ \dots \ w_n] \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ \vdots \\ x_n \end{bmatrix} = 0$$

$w_0 + w^T x = 0 \Rightarrow$  final equation of hyperplane

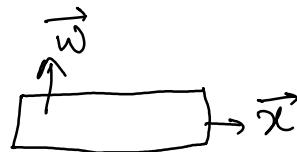
Equation of line passing through origin:

$$y = mx \quad \text{as } C=0$$

Equation of hyperplane passing through origin:

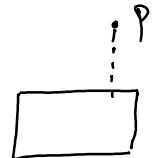
weights

$$\begin{aligned} w^T x &= 0 \\ \rightarrow \text{dot product} \\ w \cdot x &= 0 = \|w\| \|x\| \cos 90^\circ \end{aligned}$$

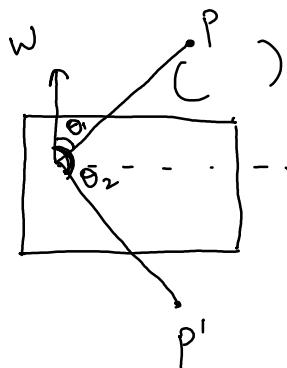


$$\theta = 90^\circ$$

Assignment: find distance of point P from plane?

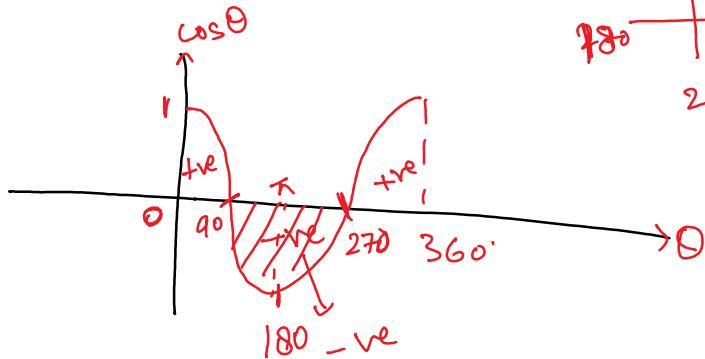
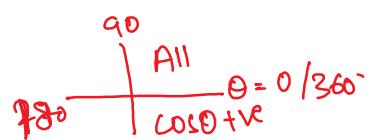


$$\text{Soln} \Rightarrow d \Rightarrow \frac{w^T \cdot P}{\|P\|}$$



$$a) w \cdot P > 0 \quad \|w\| \|P\| \cos \theta \text{ +ve}$$

$$b) w \cdot P' < 0 \quad \|w\| \|P'\| \cos \theta \text{ -ve}$$



## Eigen Value & Eigen Vector

$$A\vec{x} = \lambda \vec{x}$$

Where,    A = matrix  
               $\lambda \Rightarrow$  eigen value

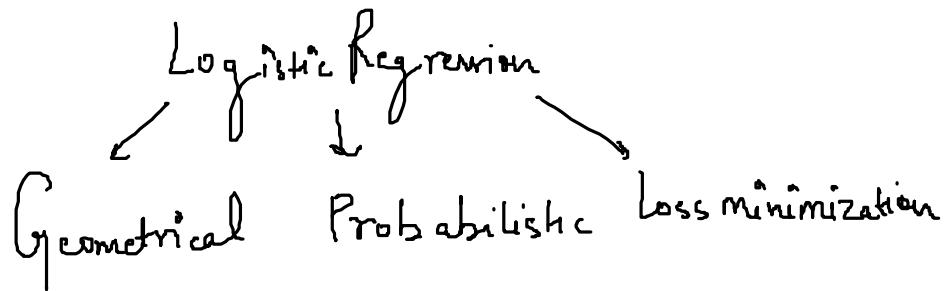
$\lambda$  eigen value for  
scalar (eigenvalue)

# DATA SCIENCE PROJECT LIFE

Saturday, September 23, 2023 3:53 PM

## LOGISTIC REGRESSION

Friday, September 8, 2023 11:13 PM



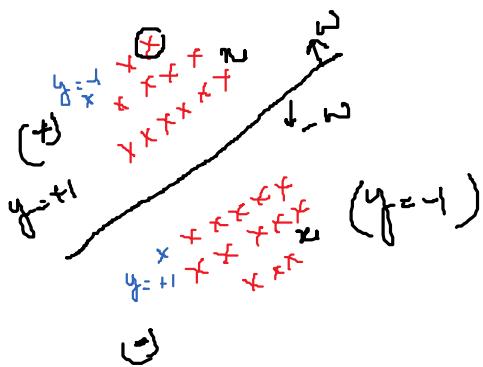
$$y = mx + c$$

$$f(x) = w_0 + w_1 x_1 + w_2 x_2 + \dots + w_n x_n \Rightarrow w^T x + w.$$

$w_0 = 0$ , if line/plane is passing through origin

$$\boxed{f(x) = w^T x}$$

Geometric Intuition:



(i)  $w^T x_i > 0 \rightarrow$  for +ve class

(ii)  $w^T x_i < 0 \rightarrow$  for -ve class

lets multiply the above eqn with  $y_i$ :

a)  $y_i = +1, w^T x_i > 0$

$$y_i w^T x_i > 0$$

b)  $y_i = -1, w^T x_i < 0$

$$y_i w^T x_i > 0$$

$$y_i \cdot w^T x_i > 0$$

$$y_i \cdot w^T x_i > 0$$

c)  $y_i = -1, w^T x_i > 0$

$$y_i \cdot w^T x_i < 0$$

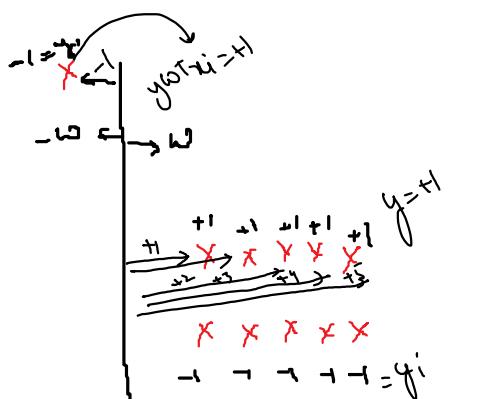
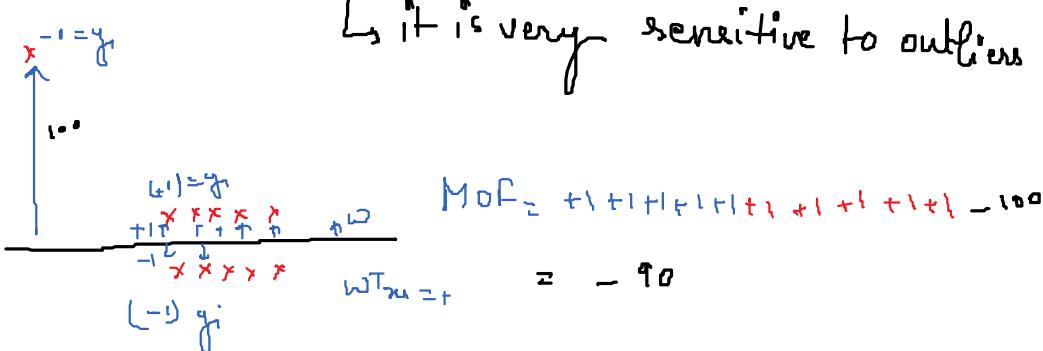
d)  $y_i = +1, w^T x_i < 0$

$$y_i \cdot w^T x_i < 0$$

for correct classifications  $y_i \cdot w^T x_i > 0$

" incorrect "  $y_i \cdot w^T x_i < 0$

$\text{MOF} := \arg \max(y_i \cdot w^T x_i)$  where  $y_i = +, -$



Sigmoid Function

Probabilistic Squeezes      Robust to outliers

b/w 0 & 1

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

$$\sigma(y_i \omega^T x_i) = \frac{1}{1 + e^{-y_i \omega^T x_i}}$$

$$MoF = \arg\max [\sigma(y_i \omega^T x_i)] = \arg\max \left[ \frac{1}{1 + e^{-y_i \omega^T x_i}} \right]$$

$$= \arg\max \left[ \log \left( \frac{1}{1 + e^{-y_i \omega^T x_i}} \right) \right] \quad \log \frac{1}{a} \Rightarrow \log a^{-1} \\ \Rightarrow -\log a$$

$$= \arg\max [-\log (1 + e^{-y_i \omega^T x_i})]$$

$$= \arg\min [\log(1 + e^{-y_i \omega^T x_i})]$$

$$\log 1 = 0$$

$$= \arg\min [-\log e^{-y_i \omega^T x_i}]$$

$$= \arg\min [+y_i \omega^T x_i \cancel{\log e^{-1}}]$$

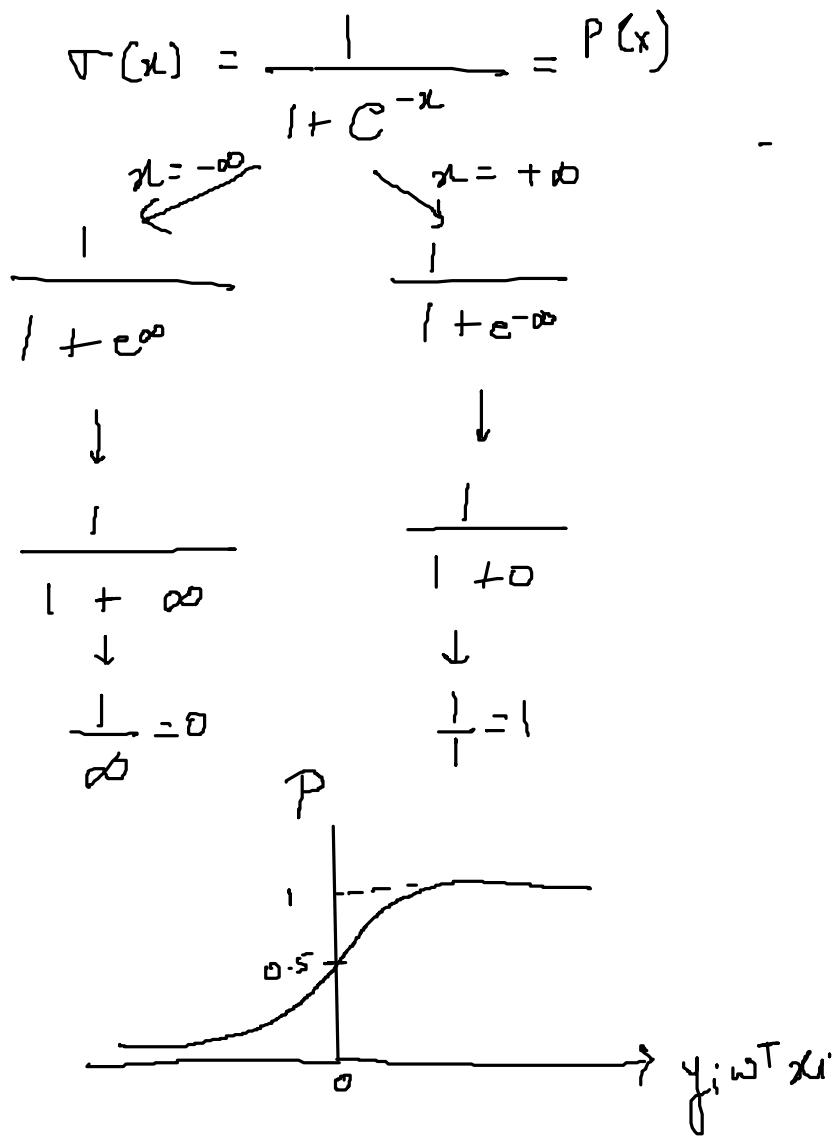
$$= \arg\max \underline{y_i \omega^T x_i}$$

$$\text{Loss } f^n = \arg\min \left[ \log(1 + e^{-y_i \omega^T x_i}) \right]$$

0 - 1

Squares Value b/w 0 and 1:

Squares Value b/w 0 and 1:



Probabilistic Way

$$P = \frac{1}{1 + e^{-y}}$$

$$P(1 + e^{-y}) = 1$$

$$P + Pe^{-y} = 1$$

$$Pe^{-y} = 1 - P$$

$$e^{-t} = \frac{1-p}{p}$$

$$\frac{1}{e^t} = \frac{1-p}{p}$$

$$e^t = \left( \frac{p}{1-p} \right) \rightarrow \text{odd's ratio}$$

Take  $\ln$

$$\ln e^t = \ln \left( \frac{p}{1-p} \right)$$

$$y = \ln \left( \frac{p}{1-p} \right) \rightarrow \text{logit function}$$

class labels

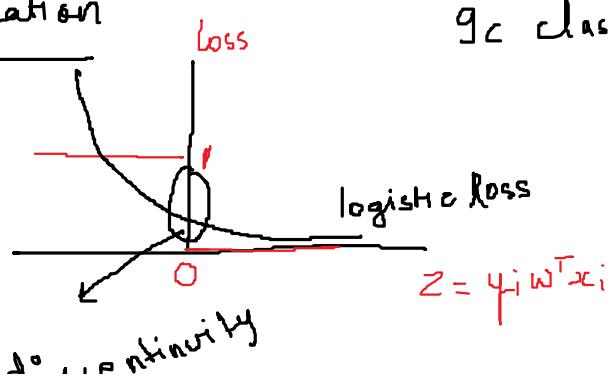
$$(0, 1) = y_i$$

$$\text{Log loss} \Rightarrow y_i \ln(y_i) + (1-y_i) \ln(1-y_i)$$

$y_i = 0$   
 $y_i = +1$

Loss minimization

(0-1 loss)



g\_c classification,

$$z = y_i w^T x_i$$

$$z = y_i w^T x_i$$

## Overfitting & Underfitting:

$$w^* = \underset{w}{\operatorname{argmin}} \sum_{i=1}^n \log(1 + e^{-y_i w^T x_i})$$

$$= \underset{w}{\operatorname{argmin}} \sum_{i=1}^n \log(1 + e^{-z_i})$$

all correct predictions,  $y_i w^T x_i = \infty = z_i$

$$= \underset{w}{\operatorname{argmin}} \sum_{i=1}^n \log(1 + e^{-\infty}) = 0$$

loss = 0  $\rightarrow$  This is overfitting

$$w^* = \underset{w}{\operatorname{argmin}} \left[ \sum_{i=1}^n \log(1 + e^{-y_i w^T x_i}) + \underbrace{\lambda w^T w}_{\text{large}} \right]$$

$w^T$  is variable  
Regularizer

if  $w^T \rightarrow \infty$

The above formulation is known as regularization!

$\lambda \rightarrow$  hyperparameters

$\lambda = 0$  overfitting

$\lambda = \text{large}$  underfitting

## RIDGE REGULARIZATION

$$w^* = \underset{w}{\operatorname{argmin}} \left[ \log(1 + e^{-y_i w^T x_i}) + \lambda w^T w \right]$$

$$\Rightarrow \text{loss function} + \underbrace{\|w\|^2}_{\text{L2 Norm}}$$

## LASSO Regularization

$$w^* = \underset{w}{\operatorname{argmin}} \left[ \log(1 + e^{-y_i w^T x_i}) + \lambda \|w\|_1 \right]$$

Sparse vector: [1, 0, 0, 1, 0, 0, 0, 0, 1, 0, 0]

$w^2 \rightarrow$  it oscillate around 0 but won't go 0.

$w \rightarrow$  converge to 0 fast

## Elastic Net:

$$w^* = \underset{w}{\operatorname{argmin}} \left[ \log(1 + e^{-y_i w^T x_i}) + \lambda_1 w^T w + \lambda_2 \|w\|_1 \right]$$

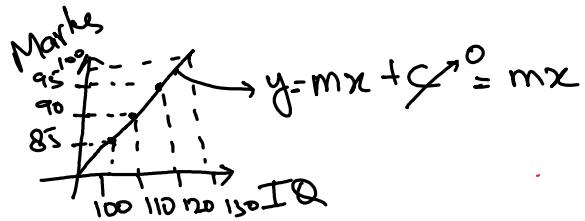
\* Always Standardize

Always Standardize

fit → training → fit(X-train)  
transform → " → .transform(X-train)  
test → " → .transform(X-test)

# Linear Regression

Variance → Co-variance → Correlation → Regression  
 → directional relationship → relationship + strength → relationship + strength → quantify



→ in 2 dimension, we have line  
 → for more than two dimension, we have plane.

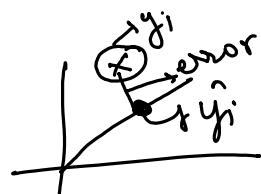
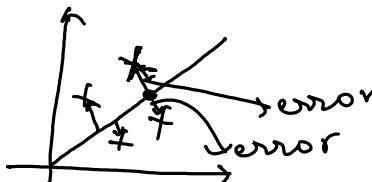
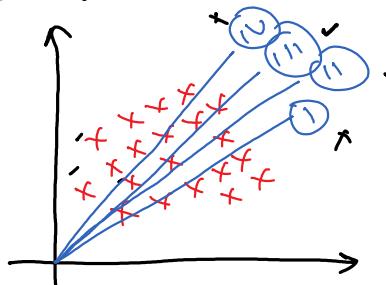
$$\boxed{\text{Marks} = \text{slope} \times \text{IQ}} \rightarrow x_i \text{ (GFP variable)}$$

y (output variable)

$$y = mx + c$$

weights.

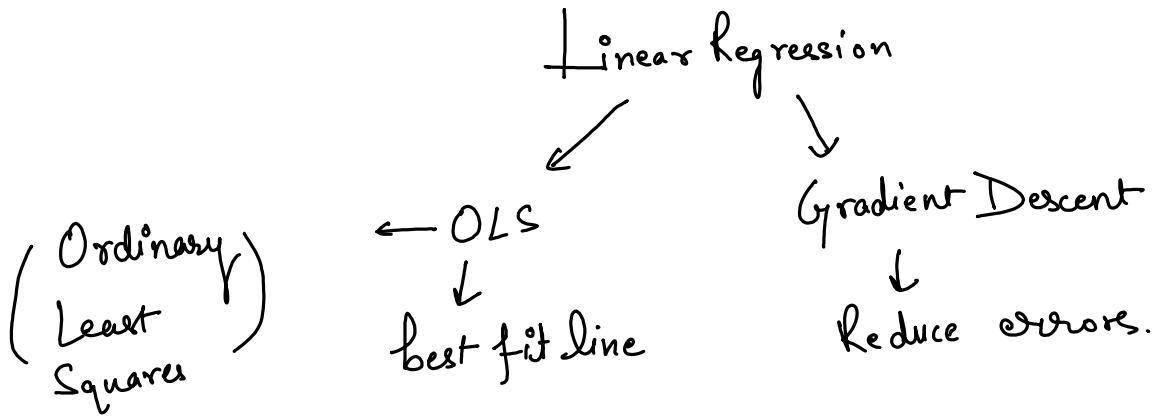
Reality:



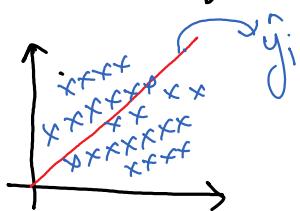
$$\text{error} = y_i - \hat{y}_i$$

Best fit line ⇒ line with minimum errors.

error ⇒ 111 is best fit line

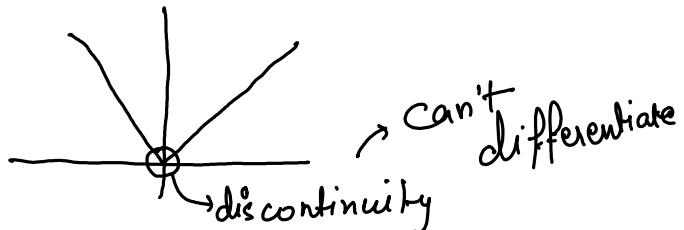


How to create a line?

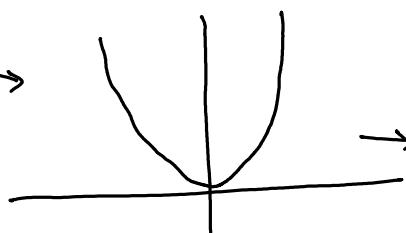


$m, x_i^*, b(c) \rightarrow$  intercept

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



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



→ we have no discontinuity &  
hence we can differentiate

$$E(m, b) = (y_i^* - \hat{y}_i)^2 \Rightarrow [y_i^* - (mx_i + b)]^2 = 0$$

$$E(m, b) = \sum_{i=1}^n [y_i^* - mx_i^* - b]^2$$

$$\frac{\partial x^n}{\partial x} = nx^{n-1}$$

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

$$v = \partial b$$

Properties of differentiation

$$\boxed{\textcircled{1} \frac{d(ax)}{dx} = a \quad \textcircled{11} \frac{d x^n}{dx} = n x^{n-1}}$$

$$= 2 (y_i - mx_i - b) \left( \cancel{\frac{dy_i}{db}} - \cancel{\frac{dmx_i}{db}} - \cancel{\frac{db}{db}} \right)^{-1} = 0$$

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

$$\Rightarrow \sum_{i=1}^n (y_i - mx_i - b) = 0$$

$$\Rightarrow \sum_{i=1}^n y_i - \sum_{i=1}^n mx_i - \sum_{i=1}^n b = 0$$

divide by  $n$

$$\Rightarrow \frac{\sum_{i=1}^n y_i}{n} - \sum_{i=1}^n \frac{mx_i}{n} - \sum_{i=1}^n \frac{b}{n} = \frac{0}{n}$$

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

$$\Rightarrow \boxed{b = \bar{y}_i - m \bar{x}_i} \rightarrow \text{intercept for best fit line}$$

Slope

$$\frac{\partial E}{\partial m} = \frac{\partial \sum_{i=1}^n [y_i - \hat{y}_i]^2}{\partial m} = \frac{\partial \sum_{i=1}^n (y_i - mx_i - b)^2}{\partial m} = 0$$

$$- \sim \sum_{i=1}^n i \cdot \dots \cdot i \cdot \bar{x} \cdot \bar{y}^2$$

$$\frac{\partial E}{\partial m} = \frac{\partial \sum_{i=1}^n [y_i - mx_i - (\bar{y}_i - m\bar{x}_i)]^2}{\partial m} = 0$$

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

Slope  
of best  
fit line  $\Rightarrow$

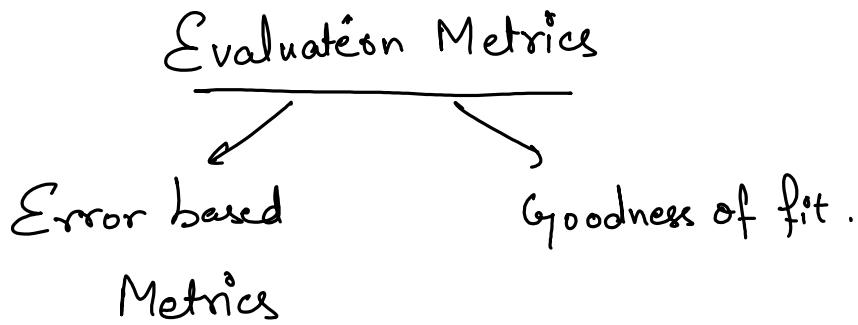
$$m = \frac{\sum (y_i - \bar{y}_i)(\bar{x}_i - x_i)}{\sum (x_i - \bar{x}_i)^2}$$

Assignment! (do the steps)

we got  $m$  &  $b$ .

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

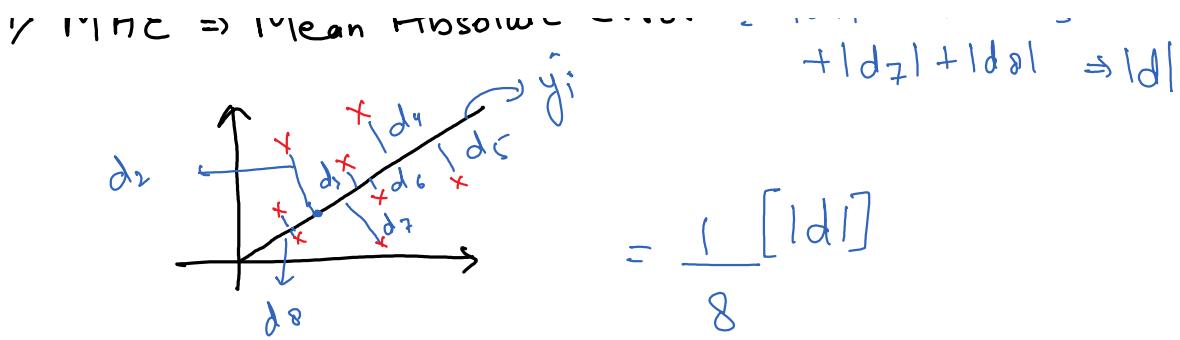
for big datasets don't use it!



### Error based Metrics

$\Rightarrow$  MAE  $\Rightarrow$  Mean Absolute Error  $= |d_1| + |d_2| + |d_3| + |d_4| + |d_5| + |d_6| + |d_7| + |d_8| \Rightarrow |d|$

$\approx \leftarrow \dots \rightarrow \hat{y}_i$



$$MAE = \frac{\sum_{i=1}^n |d_i|}{n}$$

Advantages: 1) Same scale as that of data  
2) Less sensitive to outliers

## 2) Mean Squared Error $\Rightarrow$ MSE

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

LR, this is used as loss function

- Not easily interpretable  $\rightarrow$  demerit
- Sensitive to outliers  $\rightarrow$  demerit

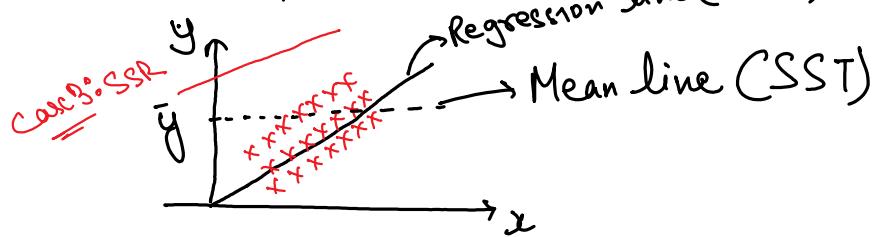
## 3) RMSE $\Rightarrow$ Root Mean Squared Error

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

Advantages:  $\rightarrow$  easily interpretable  
 $\rightarrow$  less sensitive to outliers

Goodness of fit

$\rightarrow R^2$ , R2-Score :  $\rightarrow \underline{\text{coeff. of determination}}$   $\Rightarrow$  tells us how well our model fits the data



$$R^2 = 1 - \frac{SSR}{SST}$$

Case 1:  $SSR = 0$ ,  $SST = SST$

$$R^2 = 1 - \frac{0}{SST} = 1 - 0 = 1 \quad \text{overfitting}$$

Case 2:  $SSR = SST$

$$R^2 = 1 - \frac{SST}{SST} = 1 - 1 = 0 \quad \text{underfitting}$$

\* Case 3:  $SSR > SST$

$$R^2 = 1 - \frac{SSR}{SST} (> 1) = 1 - \left( \begin{array}{l} \text{any value} \\ \text{greater than 1} \end{array} \right) = -ve$$

Problem with  $R^2 \Rightarrow \# \text{ columns} \uparrow \rightarrow R^2 \uparrow$

$\rightarrow$  Adjusted  $R^2$ :  $\text{Adj. } R^2 = 1 - \left[ \frac{(1-R^2)(n-1)}{(n-1-p)} \right]$

where,  
 $n = \# \text{ datapoints}$   
 $p = \# \text{ columns}$

if  $(n-1-p)$  decrease >  $(1-R^2)$  decrease

Adj  $R^2 \downarrow$

if  $(1-R^2)$  decrease >  $(n-1-p)$  decrease

Adj  $R^2 \uparrow$

Note: Adj $R^2$  is not present in sklearn.metrics

Multicollinearity  $\Rightarrow$  Before Modelling  $\Rightarrow$  EDA

↳ One column is highly correlated with other column.

How to deal with it?  $\rightarrow$  dropping any one of the highly correlated columns

Detection  $\rightarrow$  Correlation Matrix  $\Leftarrow$  faster

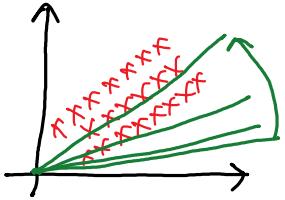
(0.70, -0.70)

↳ Variance Inflation factor (VIF)

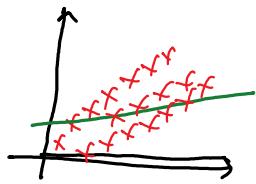
$$VIF = \frac{1}{1-R^2}$$
$$[1, \infty]$$

$$\boxed{VIF > 5}.$$

## Gradient Descent

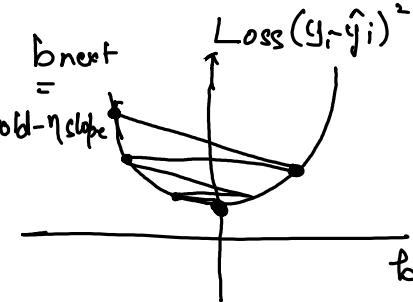


if  $\eta$  change the slope , the line will (m)  
start rotating



In order to move the line,  $\eta$  need to change  
the intercept ( $b$ )

$$L = (y_i - \hat{y}_i)^2, \quad y = mx + b$$



Steps       $m = \text{constant}$ ,  $b = ?$

① take any random value of  $b$

$$\text{② } \frac{dL}{db} = \frac{d(y_i - mx_i - b)^2}{db} = -2(y_i - mx_i - b)$$

↳ equation of slope of  $b$

$$\text{③ } b_{\text{next}} = b_{\text{old}} - \eta \text{slope}$$

↳ learning rate / step size  
↳ hyperparameter

effects of learning rate :  $\rightarrow$

Greater learning rate, descent works faster, but oscillates around minimum

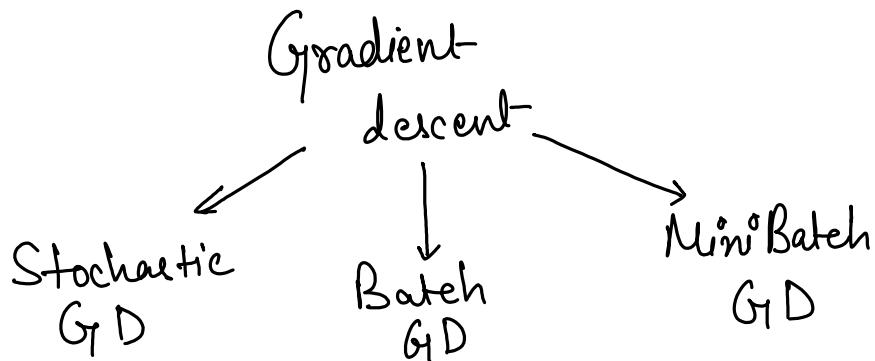
→ Smaller learning rate, descent works slower, but reaches minimum value

Let's work on m & b together.

① take random values of m & b,

$$\textcircled{II} \quad \frac{\partial L}{\partial b} = -2(y_i - mx_i - b) \quad \frac{\partial L}{\partial m} = \frac{\partial (y_i - mx_i - b)^2}{\partial m} \\ = -2(y_i - mx_i - b)x_i$$

$$\textcircled{III} \quad b_{\text{next}} = b_{\text{old}} - \eta \frac{dL}{db} \quad m_{\text{next}} = m_{\text{old}} - \eta \frac{dL}{dm}$$



### Stochastic GD

→ Row wise operation

100 iterations, 100 rows

$$1 \rightarrow \begin{bmatrix} 1 \\ 2 \\ 3 \\ \vdots \\ n \end{bmatrix} \rightarrow b, m$$

### Batch GD

→ 100 iterations, 100 rows

$$1 \rightarrow \begin{bmatrix} 1 \\ 2 \\ 3 \\ \vdots \\ 100 \end{bmatrix} \rightarrow b, m$$

$$2 \rightarrow \begin{bmatrix} 1 \\ 2 \\ \vdots \\ n \end{bmatrix} \rightarrow b, m$$

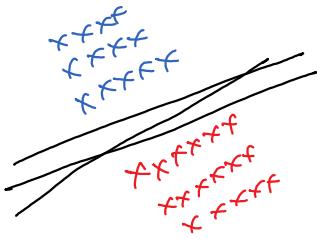
$\begin{matrix} 2 \\ 3 \\ \vdots \\ 100 \end{matrix}$        $b, m$

$Q \rightarrow \begin{matrix} 1 \\ 2 \\ 3 \\ \vdots \\ 100 \end{matrix}$        $b_{\text{next}}, m_{\text{next}}$

$$\begin{aligned} \text{No. of calculations} &= 100 \times 100 \\ &= 10,000 \end{aligned}$$

$$\begin{aligned} \text{No. of calculations} &= 100. \end{aligned}$$

# SVM (Support Vector M/c)



- probability of points closer to hyperplane is 0.5  
So, choose that hyperplane that has good distance from datapoints

SVM

$$\begin{aligned} x^+ &= w^T x_1 + b = 1 \\ x &= w^T x + b = 0 \\ x^- &= w^T x_2 + b = -1 \end{aligned}$$


---


$$w^T(x_1 - x_2) = 2$$

Normalize above eqn ( $\div \|w\|$ )

$$\left( \frac{w^T}{\|w\|} \right) (x_1 - x_2) = \frac{2}{\|w\|}$$

$$x_1 - x_2 = \frac{2}{\|w\|}$$

$$\text{MoF} = f(x) = \underset{w, b}{\arg \max} \frac{2}{\|w\|}$$

for each datapoint in negative zone,

$$w^T x + b < 0$$

for each datapoint in positive zone,

$$w^T x + b > 0$$

Simplifying the equation,  $y_i(w^T x_i + b)$

Case 1:  $y_i = +ve$ ,  $w = +ve$

$$y_i(w^T x_i + b) > 0$$

✓  
Correct  
classifications

Case 2:  $y_i = -ve$ ,  $w = -ve$

$$y_i(w^T x_i + b) > 0$$

Case 3:  $y_i = +ve$ ,  $w = -ve$

$$y_i(w^T x_i + b) < 0$$

✗  
incorrect  
classifications

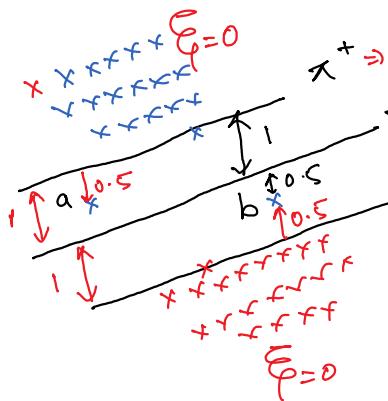
Case 4:  $y_i = -ve$ ,  $w = +ve$

$$y_i(w^T x_i + b) < 0$$

M OF  $f(x) = \underset{w}{\operatorname{argmax}} \frac{2}{\|w\|}$  such that  $\underline{y_i(w^T x_i + b) \geq 1} \Rightarrow \text{Hard Margin SVM}$

Soft Margin

" $\xi$ " → to measure how far a data point is in opp. direction from the right plane.



$$\begin{aligned} & \text{a) } y_i(w^T x_i + b) = -0.5 = 1 - (1.5) \\ & \text{b) } y_i(w^T x_i + b) = 0.5 \end{aligned}$$

$$\text{MoF} \quad f(x) = \underset{w, b}{\operatorname{argmax}} \frac{2}{\|w\|} + C \sum_{i=1}^n \xi_i \rightarrow \text{loss}$$

$$f(x) = \underset{w, b}{\operatorname{argmin}} \left[ \frac{\|w\|}{2} + C \sum_{i=1}^n \xi_i \right]$$

Regularizer                      Loss

Hyperparameter,

$C \uparrow$ , less errors, overfitting

$\lambda \uparrow \rightarrow$  more errors  $\rightarrow$  underfit

$C \downarrow$ , more errors, underfitting

$\lambda \downarrow \rightarrow$  less errors  $\rightarrow$  overfit

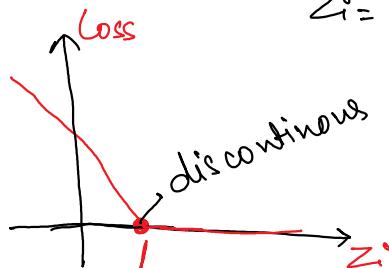
$$f(x) = \underset{w, b}{\operatorname{argmin}} \left[ \sum_{i=1}^n \xi_i + \lambda \frac{\|w\|}{2} \right]$$

$$\lambda C = \frac{1}{\lambda}$$

Loss Minimization (Hinge Loss)

$$y_i(w^\top x_i + b) > 1$$

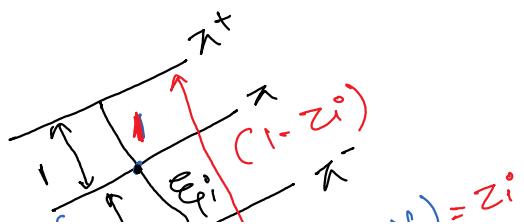
for CC



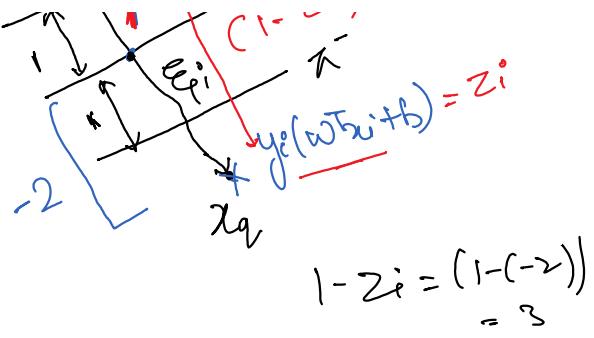
$$z_i = y_i(w^\top x_i + b)$$

$$y_i(w^\top x_i + b) \geq 1$$

$$\text{Hinge loss} \Rightarrow \max(0, 1 - z_i)$$



Hinge loss  $\Rightarrow \max(0, 1 - z_i)$



① for correct classification,

$$y_i(w^T x_i + b) = z_i > 1 \quad , \quad z_i \geq 2 \text{ (assume)}$$

$$\text{hinge loss} = \max(0, 1 - z_i) \Rightarrow \max(0, 1 - 2) = \max(0, -1) = 0$$

② for incorrect classifications,

$$\epsilon_i = 1 - z_i = 1 - (-2) = 3$$

$$\text{hinge loss} = \max(0, 1 - z_i) = \max(0, 3) = 3$$

Dual form of SVM:

$$\text{Primal: } \underset{w, b}{\operatorname{argmin}} \frac{\|w\|^2}{2} + C \sum_{i=1}^n \epsilon_i$$

↑ need not derive it!

$$\text{Dual form: } \max_{\alpha_i} \sum_{i=1}^n \alpha_i - \frac{1}{2} \sum_{i=1}^n \sum_{j=1}^n \alpha_i \alpha_j y_i y_j (x_i^T x_j)$$

↑ similarity

↳ denote support vectors

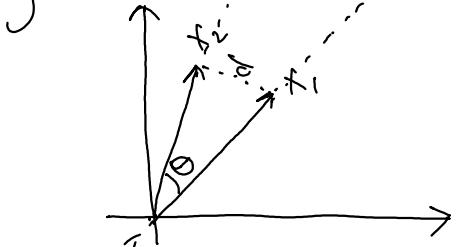
$$x_i \rightarrow x_i$$

$\alpha_i > 0$  (for support vectors)

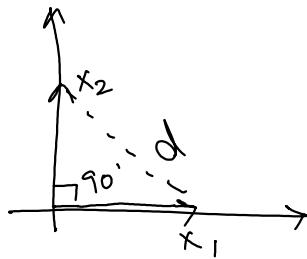
$(x_i^T x_j)$   $\Rightarrow$  Similarity  $\Rightarrow$  how similar two points are.

$\rightarrow$  Cosine Similarity:  $(\cos \theta)$

$\theta$  = angle b/w two vectors.



$\theta \uparrow$

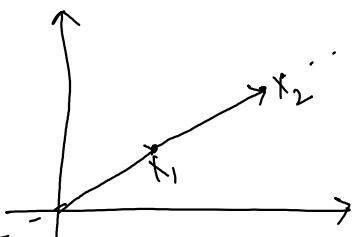


$\theta = 90^\circ$

$$\cos \theta = \cos 90^\circ = 0$$

$d$  is greatest

$$\uparrow \text{Cosine Similarity} \propto \frac{1}{\text{Cosine Distance} \downarrow}$$



$\theta = 0^\circ$  (cosine distance = 0)

$$\cos \theta = \cos 0^\circ = 1$$

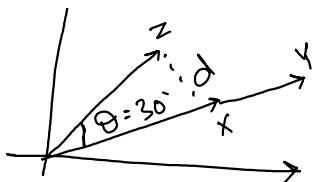
$$\cos \theta = \frac{\vec{x}_1 \cdot \vec{x}_2}{\|\vec{x}_1\| \|\vec{x}_2\|} \Rightarrow \text{lets say}$$

$$\|\vec{x}_1\| \& \|\vec{x}_2\| = 1$$

$$\cos \theta = \vec{x}_1 \cdot \vec{x}_2 \Rightarrow \text{linear algebra} \Rightarrow \underline{x_1^T x_2} \Rightarrow \underline{(x_i^T x_j)}$$



Cosine Similarity ( $x, y$ )  $\Rightarrow$  cosine distance = 0  $\Rightarrow$   $\theta = 0^\circ$



$$0^{\circ} \dots$$

$$\Theta = 0^{\circ}$$

Similarity,  $\cos\theta = \cos 0 = 1$

$x \& z$  are similar.

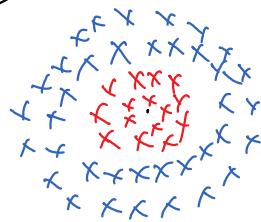
Cosine similarity( $x, z$ )  $\Rightarrow \theta = 30^{\circ}$ , cosine distance  $\neq 0$ ,

Similarity  $\Rightarrow \cos 30^{\circ} = \frac{\sqrt{3}}{2} \rightarrow$  dissimilar

$x \& z$  are dissimilar!

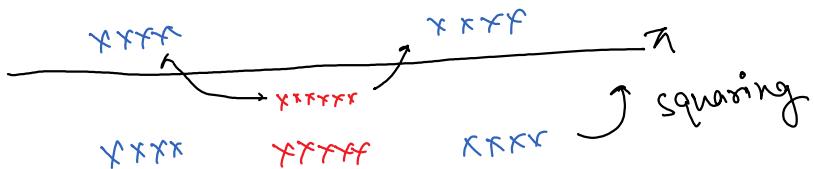
$$x_i^T x_j \rightarrow \begin{cases} \text{Polynomial} \\ (1 + x_i \cdot x_j)^d \\ RBF \\ (e^{-\frac{\|x_i - x_j\|^2}{2\sigma^2}}) \end{cases}$$

KERNELS  $\rightarrow$    
 Polynomial  $\rightarrow c, d$   
 RBF  $\rightarrow c, \sigma$   
 Linear  $\rightarrow c$



→ Higher dimension

↳ linearly  
separable



Polynomial Kernel:  $K(x_1, x_2) = (x_1^T x_2 + c)^d$

Quadratic function,  $d=2$

$$x_1 = \begin{bmatrix} x_{11} \\ x_{12} \end{bmatrix}$$

$$x_2 = \begin{bmatrix} x_{21} \\ x_{22} \end{bmatrix}$$

Applying polynomial kernel with  $d=2$ ,

$$\Rightarrow (1 + \vec{x}_1^\top \vec{x}_2)^2 \Rightarrow \left( 1 + [x_{11} \ x_{12}] \begin{bmatrix} x_{21} \\ x_{22} \end{bmatrix} \right)^2$$

$$\Rightarrow \left[ 1 + (x_{11}x_{21} + x_{12}x_{22}) \right]^2 .$$

$$\Rightarrow \left[ 1^2 + x_{11}^2 x_{21}^2 + x_{12}^2 x_{22}^2 + 2x_{11}x_{21} + 2x_{12}x_{22} + 2x_{11}x_{21}x_{12}x_{22} \right]$$

$$(1, x_{11}^2, x_{12}^2, \sqrt{2}x_{11}, \sqrt{2}x_{12}, \sqrt{2}x_{11}x_{12}) \circ X_1'$$

$$(1, x_{21}^2, x_{22}^2, \sqrt{2}x_{21}, \sqrt{2}x_{22}, \sqrt{2}x_{11}x_{22}) \circ X_2'$$

$$X_1, X_2 \Rightarrow 2d \xrightarrow[\text{Kernel}]{\text{Polynomial}} X_1' \& X_2' \Rightarrow 6d$$

\* Mercer's Theorem: Kernels convert  $d$ -dimension dataset to  $d'$ -dimension dataset such that  $d' > d$ .

Radial Basis Function (RBF):

$$K(x_1, x_2) = e^{-\frac{\|x_1 - x_2\|^2}{2\sigma^2}}$$

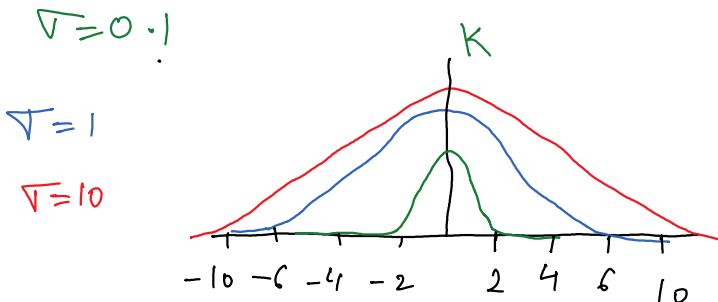
distance  
variance

Case 1:  $d = \|x_1 - x_2\| \Rightarrow \text{distance}$

$$\downarrow K = \frac{1}{e^{d^2/2\sigma^2}} \quad d \uparrow, d^2 \uparrow$$

Case 2:  $\uparrow$

$$\uparrow K = \frac{1}{e^{d^2/2\sigma^2}} \quad \sigma \uparrow, \sigma^2 \uparrow \quad \frac{d^2}{2\sigma^2} \uparrow \quad \downarrow$$

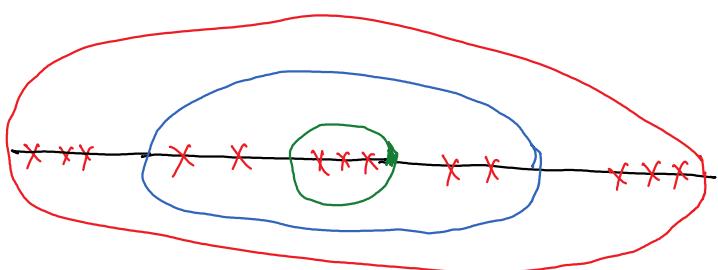


$c, r$

$\sigma \propto \text{similarity range}$

$\sigma \uparrow, \frac{1}{2\sigma^2} \downarrow, \text{similarity} \uparrow$

$\sigma \downarrow, \frac{1}{2\sigma^2} \uparrow, \text{similarity} \downarrow$



$$r \propto \frac{1}{\sigma^2}$$

## Chi square test

Sunday, September 17, 2023 8:40 AM

### Chi Square Test

Non-parametric test  
char-char situations  
(category)

degree of freedom  $\rightarrow$  logically independent values

$$\begin{array}{c} -3 \xrightarrow{2} \\ -2 \xrightarrow{3} \\ \text{avg} = 5, \quad \chi^2 = \frac{7}{2} \\ +1 \xrightarrow{8} \\ +2 \end{array}$$

for 5 values, 4 independent values  
" n values, (n-1) " "

$\rightarrow$  let's say we have n rows,  $df = n-1$

$\rightarrow$  let's say we have  $r$  rows &  $c$  columns  $\Rightarrow df = (R-1)(C-1)$

↳ Chi-square test

Q Whether there is a relation b/w gender & result?

$H_0$  : There is no relationship b/w gender & result

$H_A$  : There is relationship b/w gender & result.

| Result | Pass | Fail | Total |
|--------|------|------|-------|
| Gender |      |      |       |
| Male   | 60   | 40   | 100   |
| Female | 24   | 32   | 56    |
| Total  | 84   | 72   | 156   |

Total male = 100 Total female = 56. total pass = 84. total = 156

Total males = 100, Total females = 56, total pass = 84, total fail = 72

Expectations

$$\text{Expected value} = \frac{\text{Total males} \times \text{Total pass}}{\text{Total no. of people}}$$

| Result  | Pass                               | Fail                               | Total |
|---------|------------------------------------|------------------------------------|-------|
| Gender  |                                    |                                    |       |
| Males   | $\frac{100 \times 84}{156} = 53.8$ | $\frac{100 \times 72}{156} = 46.1$ | 100   |
| Females | $\frac{56 \times 84}{156} = 30.1$  | $\frac{56 \times 72}{156} = 25.8$  | 56    |
| Total   | 84                                 | 72                                 |       |

$$\chi^2 = \frac{(Actual - Expected)^2}{Expected}$$

$$1) \frac{(60 - 53.8)^2}{53.8} = 0.71$$

$$2) \frac{(40 - 46.1)^2}{46.1} = 0.81$$

$$3) \frac{(24 - 30.1)^2}{30.1} = 1.23$$

$$4) \frac{(32 - 25.8)^2}{25.8} = 1.48$$

$$\chi^2_{\text{cal}} = 0.71 + 1.23 + 1.48 + 0.8$$

$$= 4 \cdot 22$$

$$\chi^2 = 0.05 , \quad df = (M-1)(C-1) \\ = (2-1)(2-1) = 1$$

$$\chi^2_{\text{tab}} = 3.841$$

Compare calculated value with tabulated value:

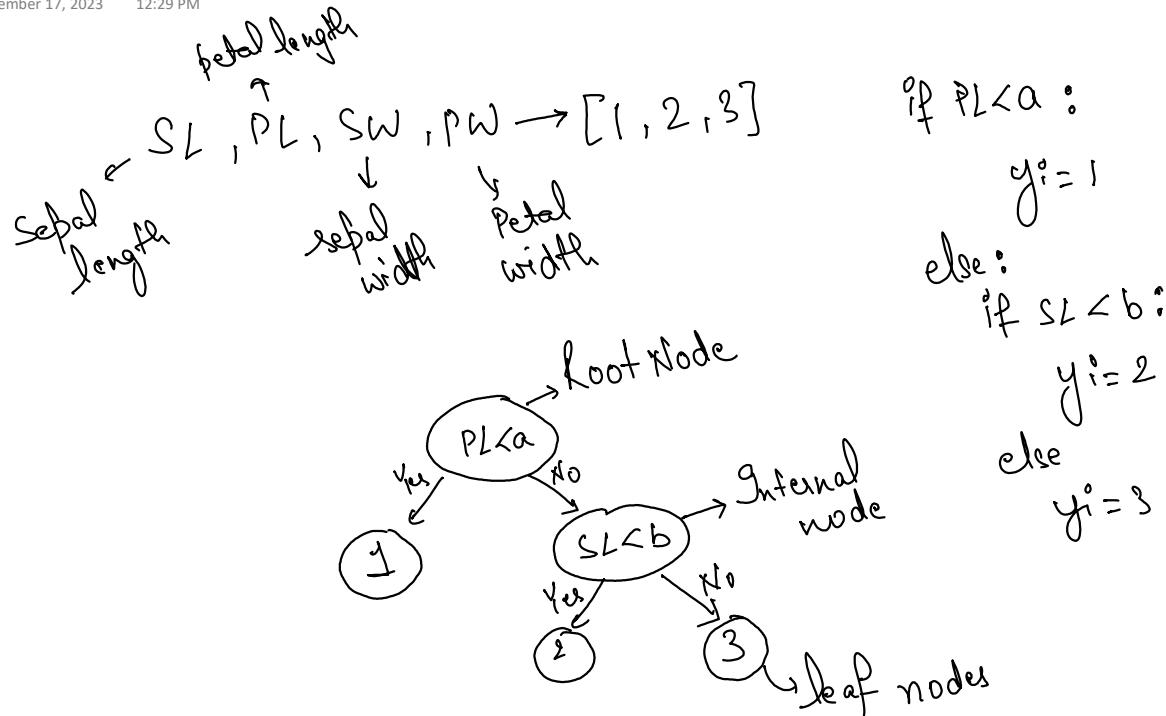
$$4.22 > 3.841$$

$$\chi^2_{\text{cal}} > \chi^2_{\text{tab}}$$

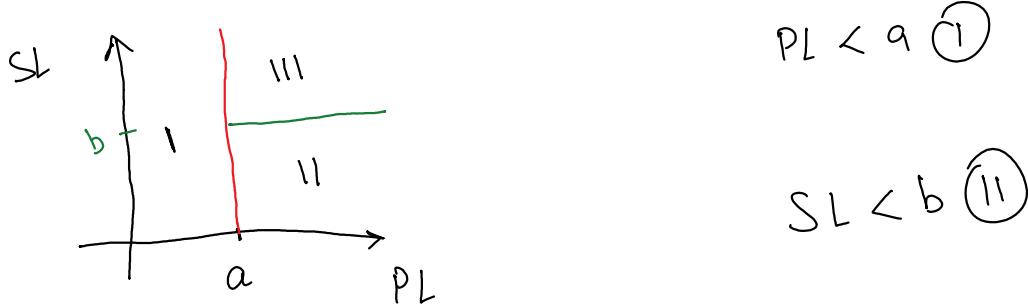
Reject  $H_0$

## Decision Trees

Sunday, September 17, 2023 12:29 PM



### Recursive Partitioning:



DT → Entropy  
 ↳ Gini Impurity  
 ↳ Information Gain

Entropy → Randomness

$$H_D(Y) = - \sum_{i=1}^n P_i^o \lg(P_i^o)$$

| Outlook  | Temperature | Humidity | Windy | PlayTennis |
|----------|-------------|----------|-------|------------|
| Sunny    | Hot         | High     | False | No         |
| Sunny    | Hot         | High     | True  | No         |
| Overcast | Hot         | High     | False | Yes        |
| Rainy    | Mild        | High     | False | Yes        |
| Rainy    | Cool        | Normal   | False | Yes        |
| Rainy    | Cool        | Normal   | True  | No         |
| Overcast | Cool        | Normal   | True  | Yes        |
| Sunny    | Mild        | High     | False | No         |
| Sunny    | Cool        | Normal   | False | Yes        |
| Rainy    | Mild        | Normal   | False | Yes        |
| Sunny    | Mild        | Normal   | True  | Yes        |

$$H_D(Y) = - \sum_{i=1}^n P_i \lg(P_i)$$

$\log_2$

|          |      |        |       |     |
|----------|------|--------|-------|-----|
| Rainy    | Mild | Normal | False | Yes |
| Sunny    | Mild | Normal | True  | Yes |
| Overcast | Mild | High   | True  | Yes |
| Overcast | Hot  | Normal | False | Yes |
| Rainy    | Mild | High   | True  | No  |

$$Y_{\text{Yes}} = 9$$

$$N_{\text{No}} = 5$$

$$P(Y) = \frac{9}{14}$$

$$P(N) = \frac{5}{14}$$

Parent's Entropy  $\Rightarrow$

$$H_D(Y) = - \sum_{i=1}^n P(i) \lg(P_i) = - P(Y) \lg P(Y) - P(N) \lg P(N)$$

$$H_D(Y) \Rightarrow - \frac{9}{14} \lg \left( \frac{9}{14} \right) - \frac{5}{14} \lg \left( \frac{5}{14} \right) = 0.94$$

Properties of entropy:

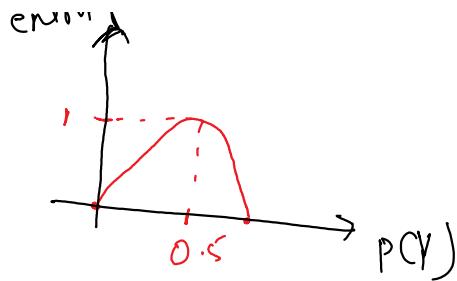
$$H_D(Y) = -P(y_+) \lg P(y_+) - P(y_-) \lg P(y_-)$$

Case 1:  $P(y_+) = 0.99 \quad H_D(Y) = -0.99 \lg 0.99 - 0.01 \lg 0.01$   
 $P(y_-) = 0.01 \quad H_D(Y) = 0.08$

Case 2:  $P(y_+) = 0.5 \quad H_D(Y) = -0.5 \lg 0.5 - 0.5 \lg 0.5$   
 $P(y_-) = 0.5 \quad H_D(Y) = 1$

Case 3:  $P(y_+) = 0 \quad H_D(Y) = -0 \lg 0 - 1 \lg 1$   
 $P(y_-) = 1 \quad = 0$

↑  
Entropy



entropy of column:

$$\begin{array}{c}
 \text{outlook} \rightarrow \begin{cases} \text{sunny } (2Y, 3N) \Rightarrow -\frac{2}{5} \lg\left(\frac{2}{5}\right) - \frac{3}{5} \lg\left(\frac{3}{5}\right) = 0.97 \\ \text{overcast } (4Y, 0N) \Rightarrow -\frac{4}{4} \lg\left(\frac{4}{4}\right) - 0 \lg(0) = 0 \\ \text{rainy } (3Y, 2N) \Rightarrow -\frac{3}{5} \lg\left(\frac{3}{5}\right) - \frac{2}{5} \lg\left(\frac{2}{5}\right) = 0.97 \end{cases}
 \end{array}$$

$$\text{weighted entropy } H_D(Y, \text{outlook}) = \frac{5}{14} \times 0.97 + \frac{4}{14} \times 0 + \frac{5}{14} \times 0.97$$

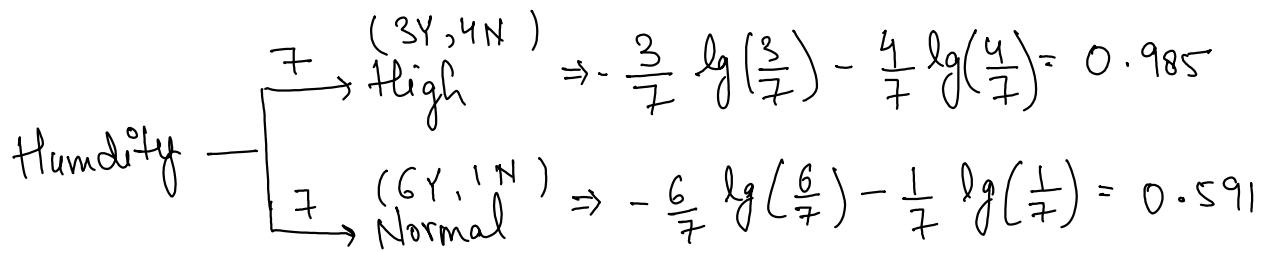
$$= \frac{5}{7} \times 0.97 = 0.69$$

$$\begin{array}{c}
 \text{Temperature} \rightarrow \begin{cases} \text{Hot } (2Y, 2N) \Rightarrow -\frac{2}{4} \lg\left(\frac{2}{4}\right) - \frac{2}{4} \lg\left(\frac{2}{4}\right) = 1 \\ \text{Mild } (4Y, 2N) \Rightarrow -\frac{4}{6} \lg\left(\frac{4}{6}\right) - \frac{2}{6} \lg\left(\frac{2}{6}\right) = 0.918 \\ \text{Cold } (3Y, 1N) \Rightarrow -\frac{3}{4} \lg\left(\frac{3}{4}\right) - \frac{1}{4} \lg\left(\frac{1}{4}\right) = 0.811 \end{cases}
 \end{array}$$

weighted entropy,

$$H_D(Y, \text{temperature}) = \frac{4}{14} \times 1 + \frac{6}{14} \times 0.918 + \frac{4}{14} \times 0.811$$

$$= 0.91$$

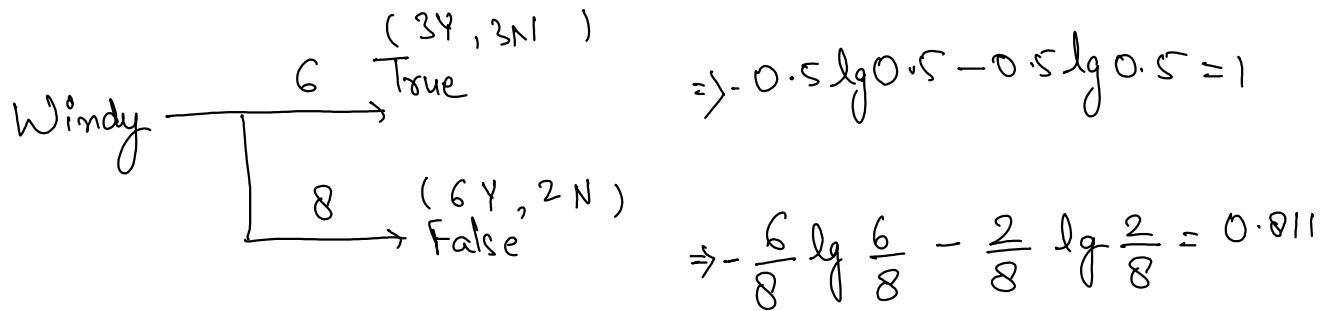


Weighted entropy,

$$H_D(Y, \text{Humidity}) = \frac{7}{14} \times 0.985 + \frac{7}{14} \times 0.591$$

$$= \frac{7}{14} (0.985 + 0.591)$$

$$= 0.788$$



Weighted entropy,

$$H_D(Y, \text{windy}) = \frac{6}{14} \times 1 + \frac{8}{14} \times 0.811$$

$$= 0.892$$

Two ways to choose :

1) Compare weighted entropies & choose the column with smallest entropy

outlook, temp, humidity, windy

choose for first split  
0.69, 0.91, 0.788, 0.892

2) Information Gain  $\Rightarrow IG(Y) = H_D(Y) - \text{weighted entropy}$

$$\text{outlook} \Rightarrow IG(Y) = 0.94 - 0.69 = 0.25$$

$$\text{temperature} \Rightarrow IG(Y) = 0.94 - 0.91 = 0.03$$

$$\text{humidity} \Rightarrow IG(Y) = 0.94 - 0.788 = 0.152$$

$$\text{windy} \Rightarrow IG(Y) = 0.94 - 0.892 = 0.048$$

Since, outlook has highest information gain, we will choose outlook for the split.

Gini Impurity ( $IG$ )

gini impurity  $\leftarrow I_G \neq IG \rightarrow$  information gain

$$IG = 1 - \sum_{i=1}^r (P_i)^2$$

$$IG_1 = 1 - [P(Y)^2 + P(N)^2] \rightarrow \text{Binary class}$$

$$IG_1 = 1 - [P(y_1)^2 + P(y_2)^2 + \dots + P(y_n)^2] \rightarrow \text{multiclass}$$

if you have a large dataset use GINI Impurity.

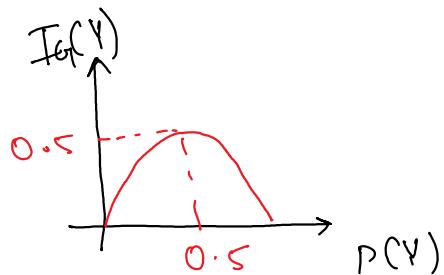
### Properties of Gini Impurity:

Case 1:  $P(y_+) = 0.5$        $IG_1(Y) = 1 - [0.5^2 + 0.5^2] = 1 - [0.25 + 0.25] = 0.5$

                  $P(y_-) = 0.5$

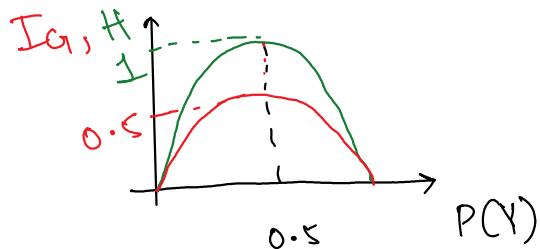
Case 2:  $P(y_+) = 1$        $IG_1(Y) = 1 - [1 + 0] = 0$

                  $P(y_-) = 0$



### Comparison b/w GINI & Entropy:

①



② Computational cost:

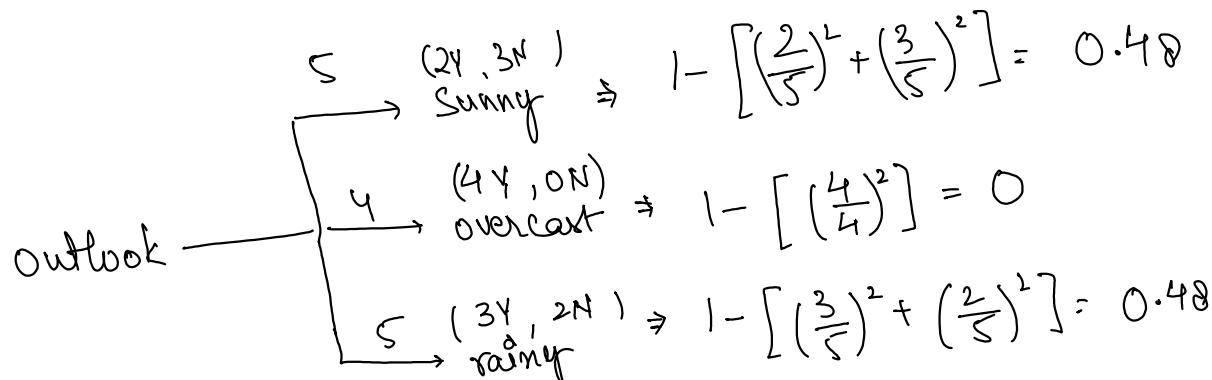
entropy is harder to calculate, higher computational cost

$\text{Gini}$  is easier to calculate, lower computational cost

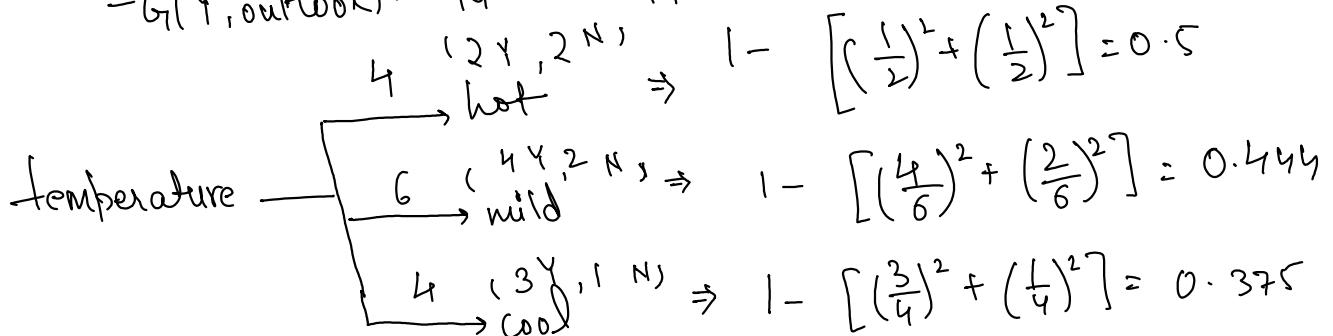
For large datasets  $\rightarrow$  use Gini Impurity

Gini Impurity:

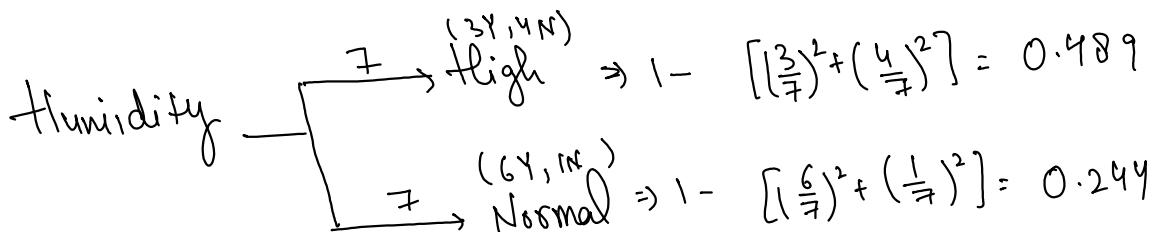
$$I_G(Y) = 1 - \left[ \left( \frac{9}{14} \right)^2 + \left( \frac{5}{14} \right)^2 \right] = 0.459$$



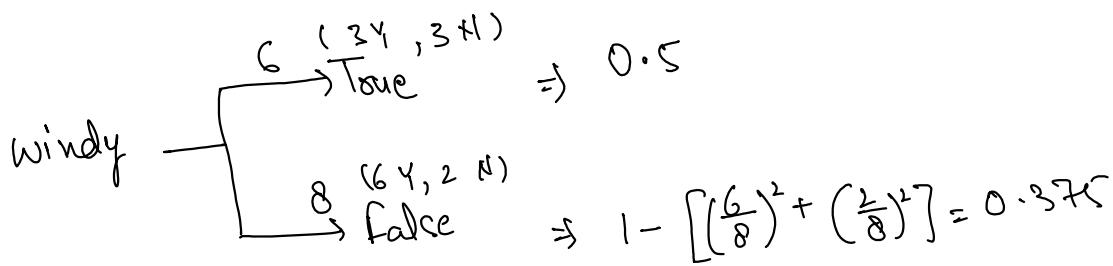
$$I_G(Y, \text{outlook}) = \frac{5}{14} \times 0.48 + \frac{4}{14} \times 0 + \frac{5}{14} \times 0.48 = 0.342$$



$$I_G(Y, \text{temp}) = 0.44$$



$$I_4(Y, \text{humidity}) = \frac{7}{14} \times 0.489 + \frac{7}{14} \times 0.244 = 0.367$$



$$I_4(Y, \text{windy}) = \frac{6}{14} \times 0.367 + \frac{8}{14} \times 0.375 = 0.429$$

Two ways to choose:

▷ choose column with lowest entropy.

| outlook | temp | humidity | windy |
|---------|------|----------|-------|
| 0.342   | 0.44 | 0.367    | 0.429 |

2) Information Gain,

$$\text{outlook} \Rightarrow I_{G_O}(Y) = 0.459 - 0.342 = 0.117$$

$$\text{Temp} \Rightarrow I_{G_T}(Y) = 0.459 - 0.44 = 0.019$$

$$\text{Humidity} \Rightarrow I_{G_H}(Y) = 0.459 - 0.367 = 0.092$$

$$\text{Windy} \Rightarrow IG_{IW}(Y) = 0.459 - 0.429 = 0.030$$

outlook is chosen because it has max IG.

- \* Max\_depth hyperparameter.

# BAGGING & BOOSTING

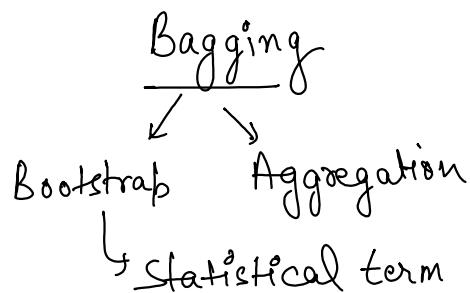
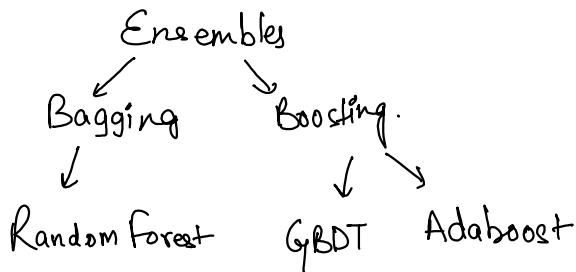
Saturday, September 23, 2023 12:17 PM

## Ensembles

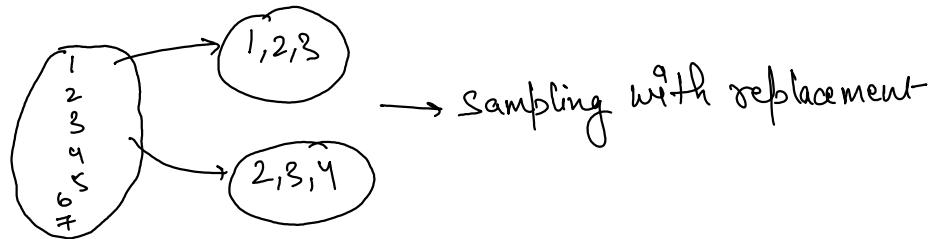
↳ group of musician  
↓

In m/c learning  
↓

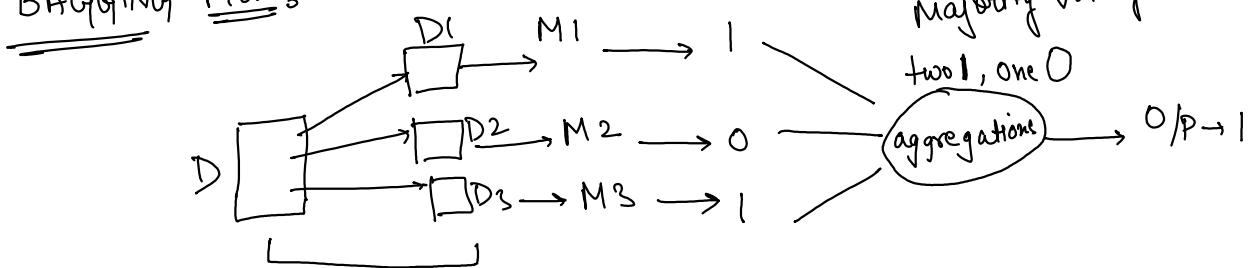
group of models

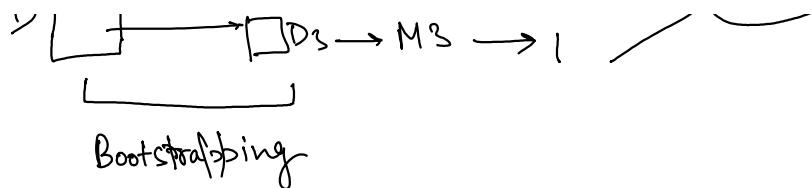


Bootstrap → sampling with replacement



BAGGING Flow: → Parallel working





\* Bagging is used when bias is low & variance is high!

↓                      ↓  
underfitting      overfitting  
why? → assignment

Q How do you create a situation of low bias & high variance with the help of DT?

Sol. Create a DT of reasonable/good depth

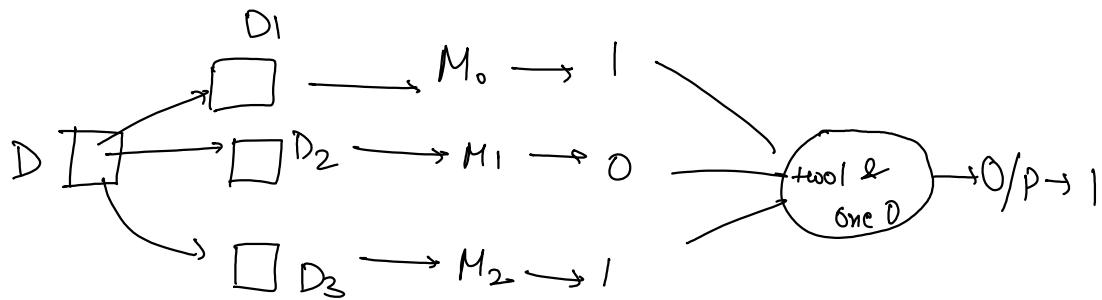
DT → max\_depth↑ → depth of tree↑ → overfitting↑

Bagging = Bootstrapping + Aggregation

Classification      Regression  
↓                      ↓  
Majority voting      mean

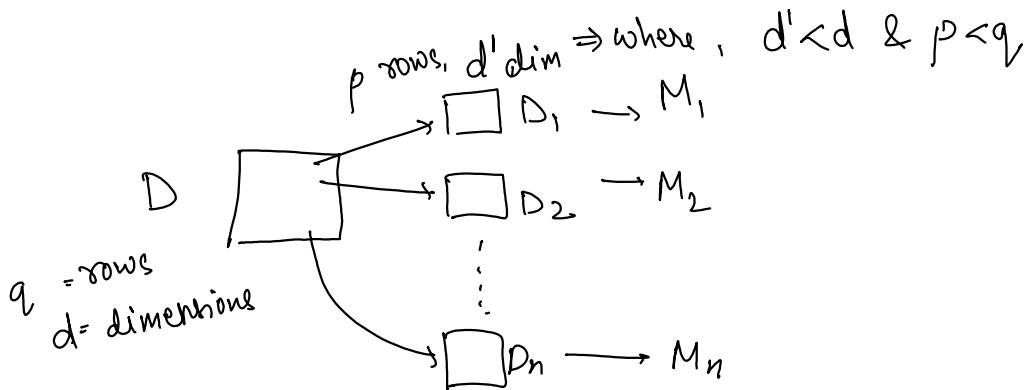
Random forest → group of trees  
↳ decision tree

M = decision tree



\* Models should be different.

RF  $\Rightarrow$  (low bias & high variance) + Row sampling + column sampling  
+  
Aggregation



$M_1, M_2, \dots, M_n$  will be different because of different samples provided to them as input.

Ex

|            | IQ  | CGPA | ECA | CP | Placement |
|------------|-----|------|-----|----|-----------|
| 1) Ankit   | 107 | 8    | 7   | 6  | 1         |
| 2) Poem    | 110 | 9    | 3   | 8  | 1         |
| 3) Vidhya  | 115 | 7    | 4   | 9  | 0         |
| 4) AVIK    | 124 | 6.5  | 8   | 9  | 1         |
| 5) Anupama | 105 | 10   | 1   | 7  | 0         |
|            | /   | -    |     |    |           |

> Anupama 105

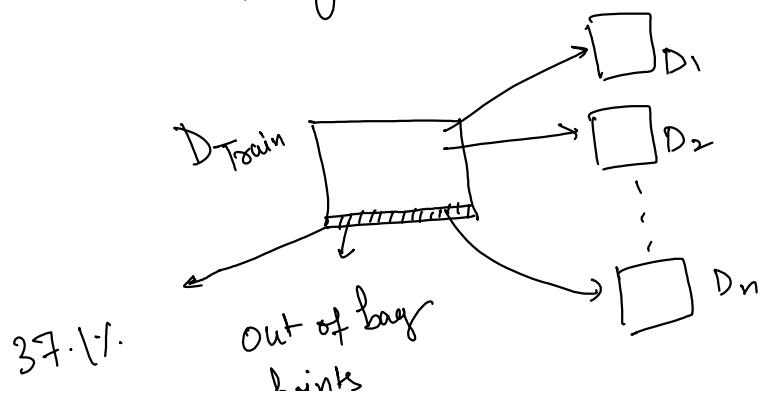
|        | 10    | 1  | 7 | 0      |        |        |
|--------|-------|----|---|--------|--------|--------|
|        | $D_1$ |    |   | $D_2$  |        |        |
| Prem   | IQ    | CP |   | Poem   | CGPA   | ECA    |
| Vidhya | 110   | 8  |   | Vidhya | 9<br>7 | 3<br>4 |

### Hyperparameters:

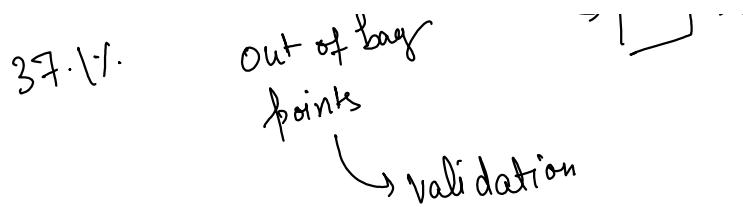
- # models  $\Rightarrow n_{\text{estimators}} \approx [50-2000]$
  - Row sampling rate  $\Rightarrow \frac{p}{q}$
  - Column " "  $\Rightarrow \frac{d'}{d}$
  - max\_depth
  - n\_jobs = -1  $\rightarrow \text{CPU} \rightarrow \text{cores} = 8$
- ↑  $\propto$  Variance ↓

### Oob Score

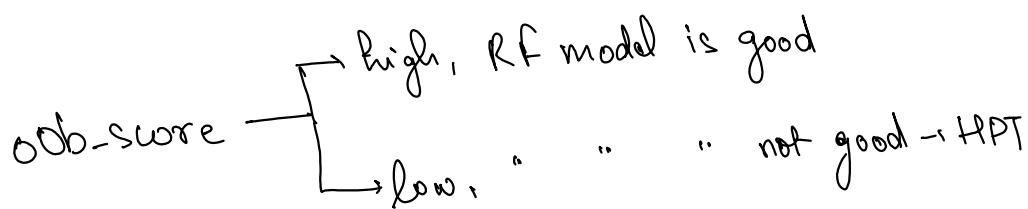
Out of bag



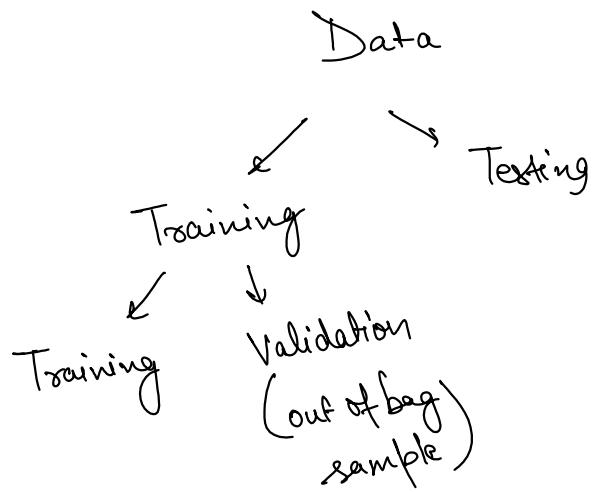
10k  $\rightarrow$  7.5k  
2.5k  
Out of bag  
Validation sample



model = Random forest Classifier(oob-score = True)



oob-Score  $\uparrow$  accuracy



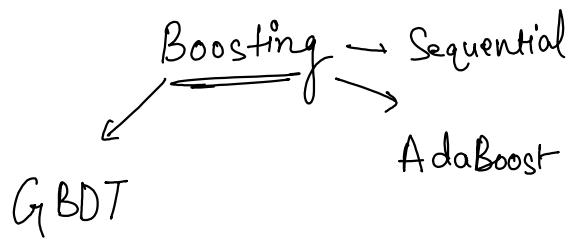
Disadvantage:

- Black box
- no loss function

Advantage:

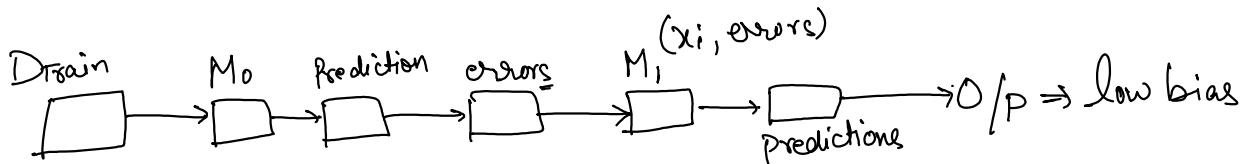
→ feature\_importances: ↓  
100 columns  
↓  
20 to 30 columns  
↓  
80% to 85% of info

Boosting → Sequential



- ① Bagging <sup>(RF)</sup> ⇒ low bias & high variance
- ② Boosting ⇒ high bias & low variance ⇒ decision stumps  
 ↓  
 DT of depth one

Flowchart:



Steps

0)  $D_{Train} = \{x_i^*, y_i^*\}_{i=1}^n \rightarrow M_0 \longrightarrow \text{predictions} \rightarrow \text{errors}$

$\Downarrow$

$y_i^* - \hat{y}_i$

$\Downarrow$

$y_i^* - f_{h_0}(x) = e_i^*$

1)  $M_1 \rightarrow \{x_i^*, e_i^*\}_{i=1}^n \quad e_i^* = y_i^* - h_0(x)$

$\nearrow h_1(x)$

Model at end of stage 1:

$$f_1(x) = \alpha_0 h_0(x) + \alpha_1 h_1(x)$$

\* Boosting  $\Rightarrow$  high bias + additive combinations.

$$2) M_2 \rightarrow \{x_i, e_i\}_{i=1}^n, e_i = \text{error by } f_i$$

$\xrightarrow{h_2(x)}$

$$e_i = y_i - f_i(x)$$

$$f_2(x) = \alpha_0 h_0(x) + \alpha_1 h_1(x) + \alpha_2 h_2(x)$$

for  $k^{\text{th}}$  stage :

$$f_k(x) = \sum_{i=0}^k \alpha_i h_i(x)$$

additive  
weighted  
model

↳ trained to fit on the errors  
at the end of previous stage

$K \Rightarrow$  hyperparameter  $\Rightarrow$  # models

$K \uparrow, \text{bias} \downarrow$

$\rightarrow$  collection of weak learners to make a strong learner.

\* \* \*  
Residual & loss function :

$$L(y_i, f_k(x)) = [y_i - \underbrace{f_k(x_i)}_{z_i}]^2 = [y_i - z_i]^2$$

$$\frac{\partial L(y_i, f_k(x))}{\partial \dots} = \frac{\partial [y_i - z_i]^2}{\partial \dots}$$

$$\frac{\partial L(y_i, f_k(x))}{\partial z_i} = \frac{\partial [y_i - z_i]}{\partial z_i}$$

$$\frac{\partial L}{\partial z_i} = -\frac{2(y_i - z_i)}{\text{ignore it!}}$$

$$-\frac{\partial L}{\partial z_i} = \underbrace{y_i - z_i}_{\text{error (residual)}}$$

negative gradient

pseudo-residual  $\Rightarrow$  makes algorithm very powerful

### Gradient Boosting

g/p  $\Rightarrow \{x_i, y_i\}_{i=1}^n$  + differentiable loss function  $L(y_i, f(x))$

$$0) f_0 = \underset{Y}{\operatorname{arg\min}} \sum_{i=1}^n L(y_i, Y) \rightarrow Y = \bar{y}$$

$\downarrow$  MSE =  $\frac{1}{n} [y_i - \hat{y}_i]^2$

$\downarrow$  constant

1) for  $m=1$  to  $M$

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

for 1 stage,  $m=1$

$$\vartheta_1 = -\lambda \left[ \frac{\partial L(y_i, f_0(x))}{\partial f_0(x_i)} \right]$$

2)  $\hat{h}_m(x)$  that can fit on pseudo-residuals, train it with  $\{x_i^o, y_{im}\}_{i=1}^n$

3)  $y_m = \underset{\gamma}{\operatorname{argmin}} \mathcal{L} \left[ y_i, \underbrace{f_{m-1}(x_i) + \hat{h}_m(x_i)}_{\text{result of previous model}} \right]$

4)  $f_m(x) = f_{m-1}(x) + \gamma_m \hat{h}_m(x)$

New prediction = old prediction + models (additively combined)

→ Application of GB → Xg Boost → Taylor's Series

hyperparameter:  $m \Rightarrow \# \text{models}$

$m \uparrow \propto \text{bias} \downarrow \propto \text{variance} \uparrow$

• Shrinkage:  $f_m(x) = f_{m-1}(x) + \gamma \underbrace{\hat{h}_m(x)}_{\substack{\text{learning rate} \\ 0 < \gamma < 1}}$

→ reduces  $f_m$  which in turn reduces overfitting

GyBDT → drawback → very slow  
 ↳ XgBoost  
 ↳ pip install XgBoost

Example:  $y_i$        $\hat{y}_i = r$       -  
 $12000$

$$\frac{\partial L}{\partial r} = - \sum_{i=0}^n (y_i - r)^2$$

$$16500$$

$$15500$$

$$14000$$

$$L = \frac{1}{2} (12000 - r)^2 + \frac{1}{2} (16500 - r)^2 + \frac{1}{2} (15500 - r)^2 + \frac{1}{2} (14000 - r)^2$$

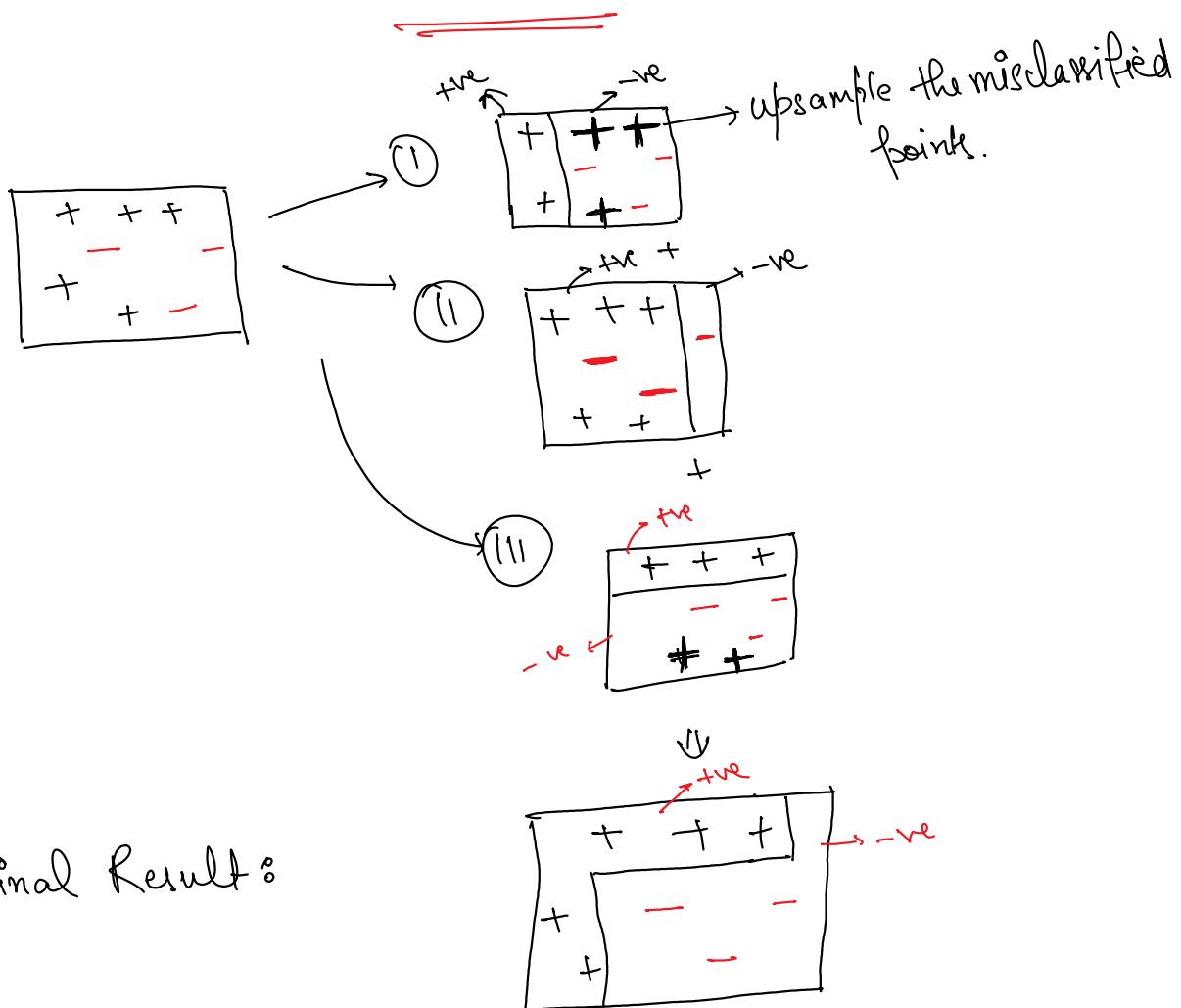
$$0 = \frac{\partial L}{\partial r} = (12000 - r)(-1) + (16500 - r)(-1) + (15500 - r)(-1) + (14000 - r)(-1)$$

$$12000 - r + 16500 - r + 15500 - r + 14000 - r = 0$$

$$58000 - 4r = 0$$

$$r = \frac{58000}{4} = 14500 = \bar{y}_i$$

ADABOOST



$$C = \gamma_1 C_1 + \gamma_2 C_2 + \gamma_3 C_3$$

| $X_1$ | $X_2$ | $Y$ | $\hat{Y}$ | weight = $\frac{1}{n}$ | new weight          | Normalized weight   |
|-------|-------|-----|-----------|------------------------|---------------------|---------------------|
| 3     | 9     | 1   | 1         | $\frac{1}{5} = 0.2$    | 0.16                | $0.16/0.96 = 0.167$ |
| 2     | 4     | 0   | 1*        | $\frac{1}{5} = 0.2$    | 0.24                | $0.24/0.96 = 0.25$  |
| 1     | 5     | 1   | 0*        | $\frac{1}{5} = 0.2$    | 0.24                | $0.24/0.96 = 0.25$  |
| 4     | 6     | 0   | 0         | $\frac{1}{5} = 0.2$    | 0.16                | $0.16/0.96 = 0.167$ |
| 5     | 7     | 0   | 0         | $\frac{1}{5} = 0.2$    | $\frac{0.16}{0.96}$ | $0.16/0.96 = 0.167$ |

$\alpha = \text{error rate}$

error  
 ↳ algebraic sum of weights of  
 misclassified points

$$\alpha = \frac{1}{2} \ln \left( \frac{1 - \text{error}}{\text{error}} \right)$$

misclassified points  
error = 0.2 + 0.2 = 0.4

$$\alpha = \frac{1}{2} \ln \left( \frac{1 - 0.4}{0.4} \right) = \frac{1}{2} \ln \left( \frac{0.6}{0.4} \right) = 0.20$$

$$\text{new weight for correct classified points} = e^{-\alpha} \times \text{old weight} = 0.2 \times e^{-0.2} = 0.16$$

$$\text{new weight for incorrect classified points} = e^{+\alpha} \times \text{old weight} = 0.2 \times e^{0.2} = 0.24$$

|   | $X_1$ | $X_2$ | $Y$ | $\hat{Y}$ | NW    | Range         |
|---|-------|-------|-----|-----------|-------|---------------|
| 0 | 3     | 9     | 1   | 1         | 0.167 | 0 - 0.167     |
| 1 | 2     | 4     | 0   | 0         | 0.25  | 0.167 - 0.417 |
| 2 | 1     | 5     | 1   | 0         | 0.25  | 0.417 - 0.667 |
| 3 | 9     | 6     | 0   | 0         | 0.167 | 0.667 - 0.834 |
| 4 | 5     | 7     | 0   | 0         | 0.167 | 0.834 - 1     |

Randomly choose 5 numbers b/w 0 & 1

0.1, 0.3, 0.5, 0.7, 0.6

[ 0, 1, 2, 3, 2 ]

Adaboost  $\rightarrow$  Image Classification

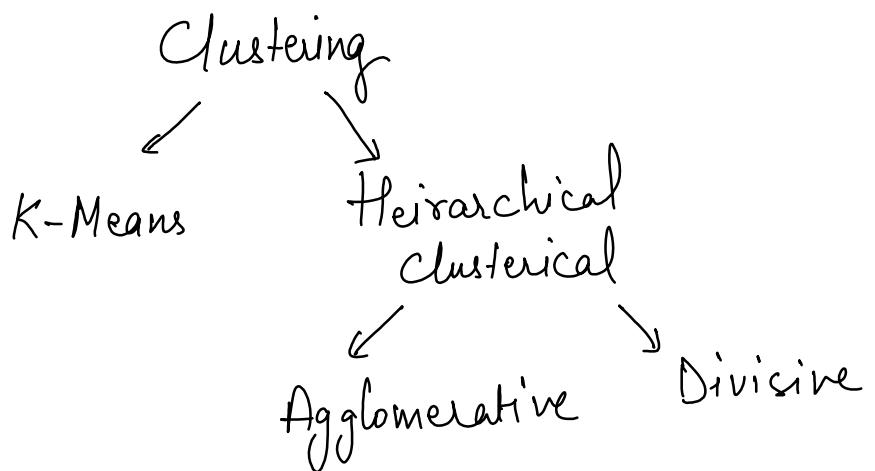
# CLUSTERING

Sunday, September 24, 2023 2:30 PM

## Clustering

$D \Rightarrow \{x_i^o, y_i^o\} \rightarrow y_i^o = f(x) \Rightarrow$  class labels  $\rightarrow$  classification  
 $\downarrow$   
continuous variable  $\rightarrow$  regression

$D = \{x_i^o\} \Rightarrow$  No class labels  $\Rightarrow$  unsupervised learning



unsupervised learning =  $\{x_i^o\}$

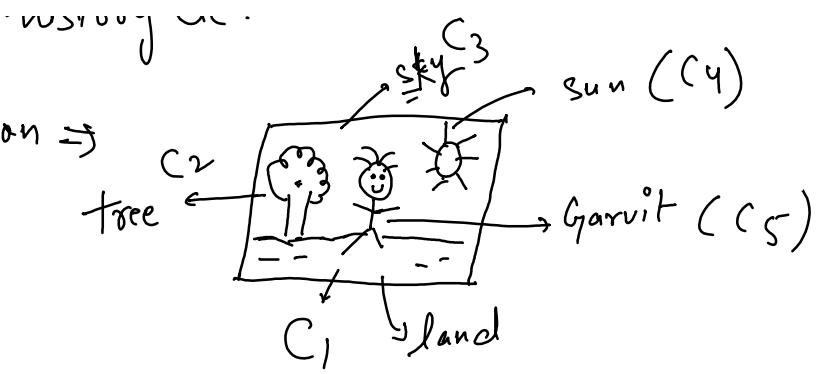
↓  
semi-supervised " =  $\{x_i^o, y_i^o\} \Rightarrow$  smaller portion

$\{x_i^o\} \Rightarrow$  larger portion

Applications: e-commerce: group customers on the basis  
location, gender, income level, product  
history etc.

$\stackrel{\text{sky}}{\longrightarrow} C_3 \longrightarrow \text{sun } (C_4)$

ii) Image segmentation  $\Rightarrow$   
(object detection)



$\rightarrow$  Review Analysis:

Amazon Review

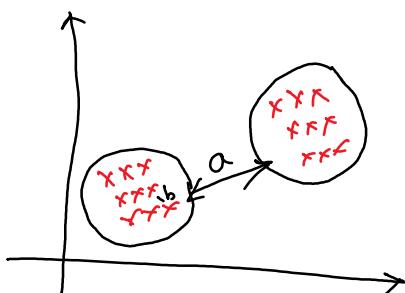
+ve

-ve (NLP)

Review: Product is good!  $\Rightarrow$  +ve

Product is low quality!  $\Rightarrow$  -ve

Metrics



\* Intercluster distance (a)

\* Intraduster distance (b)

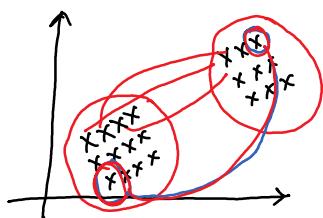
good cluster characteristics:

$\rightarrow$  Intra-cluster distance should be small

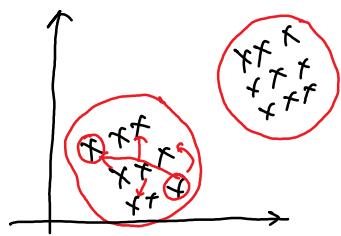
$\rightarrow$  Inter-cluster distance should be large

$$\rightarrow \uparrow \downarrow \text{Dunn's Index} \Rightarrow \frac{\uparrow \max d(i,j)}{\uparrow \max d'(k)} \xrightarrow{\substack{\text{inter-cluster distance} \\ \text{intrachuster distance}}}$$

Range :  $0 \rightarrow \infty$



$\max d(i,j)$  = distance b/w the farthest points in different clusters



$\max d'(k)$  = distance between the farthest within a cluster

$$\rightarrow \text{Silhouette's Score} = \frac{b - a}{\max(b, a)} \quad \Rightarrow \text{sklearn.metrics}$$

$[-1, +1]$  = Range

where,  $b$  = average intercluster distance  
 $a$  = " intrachuster distance

Case 1:  $\frac{b-a}{\max(b,a)}$   $\Rightarrow a$  is min.  $\Rightarrow a=0$   
 $b$  is max  $\Rightarrow b=b$

$$(\text{Silhouette's Score}) \quad S.S = \frac{b - 0}{b} = \frac{b}{b} = 1$$

Case 2:  $b = a$

$$S.S = \frac{b - a}{\max(b, a)} = \frac{0}{\max(a, a)} = 0$$

Case 3:  $b < a$ ,  $b = 0$ ,  $a = a$

$$S.S = \frac{b - a}{\max(b, a)} = \frac{0 - a}{a} = -1$$

hyperparameter  $\xleftarrow{\text{K-Means}}$  Mean  $\xrightarrow{\text{average}} \text{Centroids}$

→ define certain of centroids

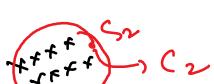


datapoints are assigned → sets (python sets)

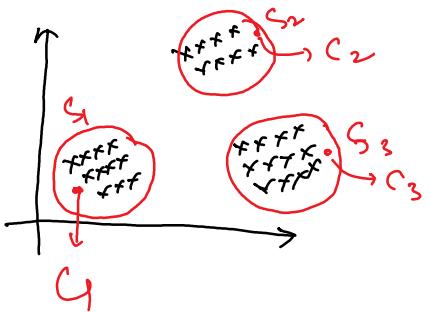


sets should not have anything in common

↑



$k = 3$  (no. of centroids)



$K = 3$  (no. of centroids)

$$S_1 \cap S_2 = \emptyset$$

$$S_2 \cap S_3 = \emptyset$$

$$S_1 \cap S_3 = \emptyset$$



$$\text{distance} \Rightarrow |C_i - x|$$

$$Mof = \underset{C_1, C_2, \dots, C_k}{\operatorname{argmin}} \sum_{i=1}^n \sum_{x \in S_i} \|x - C_i\|^2$$

$$\text{s.t. } x \in S_i$$

$$S_i \cap S_j \neq \emptyset$$

intracenter distance

Complexity theory:  $\rightarrow$  exponential  $\Rightarrow$  NP hard  $\Rightarrow$  very complex problem



Take approximation

### Lloyd's Algorithm:

- 1) Randomly choose  $K$  datapoints from dataset & call them Centroids.

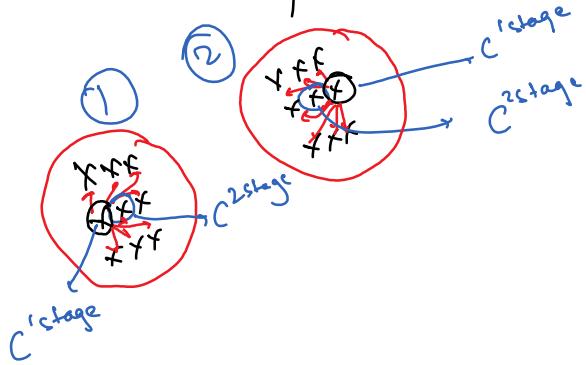
Assignment:

- 2) for each point, select the nearest centroid with help of distance & add the point to the corresponding cluster

3) Recalculate the centroid:

$$C_j = \frac{1}{S_j} \sum_{i=1}^n x_i \quad (x_i \in S_j)$$

4) Repeat Step 2 & 3 until convergence



K-Means++    init = 'kmeans++'

1) You will choose a single datapoint centroid

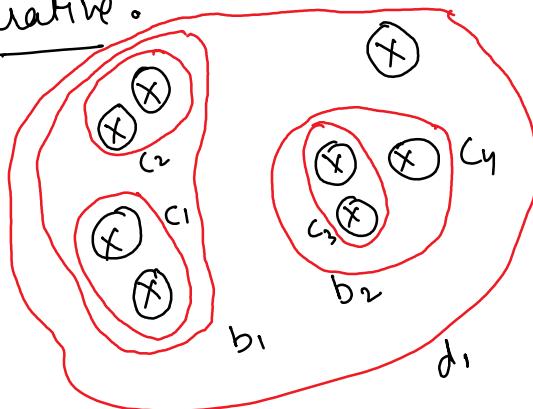
Centroid  
 $C_1$   
datapoints  
 $x_1$   
 $x_2$   
 $x_3$   
 $\vdots$   
 $x_n$

distance  
 $d_1(x_1 - c_1) \rightarrow$  probability  
 $d_2(x_2 - c_1)$   
 $d_3(x_3 - c_1)$   
 $\vdots$   
 $d_n(x_n - c_1)$   
probability of distance

\* points that are farther away from the centroid have higher probability of being selected as the centroid

## Hierarchical Clustering

Agglomerative:



8 clusters

↓  
4 clusters ( $c_1, c_2, c_3, c_4$ )

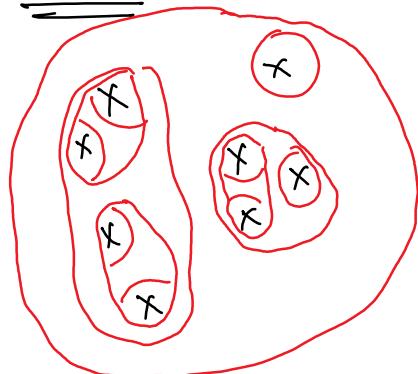
↓  
2 clusters ( $b_1, b_2$ )

↓  
1 cluster ( $d_1$ )

How to do clustering? → on the basis of distance & similarity

"Kernels"

Divisive



1 cluster



2 clusters



4 clusters



8 clusters

Dendograms: A tree like structure that records the merge & split.

# Curse of Dimensionality

binary classification (0, 1)  $\Rightarrow f_1, f_2, f_3$   
 $\Rightarrow 2^3 \Rightarrow 8$

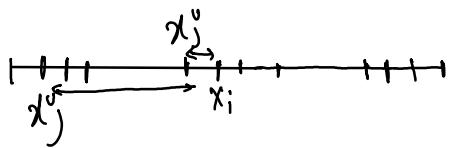
for 10 features,  $2^{10} = 1024$  datapoints

for 100 features,  $2^{100}$

1) Hughes' Phenomenon: Whenever the # dimensions ↑, the model performance ↓

↓  
optimal number

2) Distance functions  $\Rightarrow$  intuition lies only for 3d  
 but not valid for higher dimensions



$$\text{dist-min} = \min_{-} d[x_i^o, x_j^o]$$

$$\text{dist\_max} = \max d[x_i, x_f]$$

$$\frac{\text{dist\_max} - \text{dist\_min}}{\text{dist\_min}} > 0 \text{ for } 1d, 2d, 3d$$

As  $d \uparrow$ ,    It     $\frac{\text{dist\_max}() - \text{dist\_min}()}{\text{dist\_min}} = 0$   
 $d \rightarrow \infty$

hence, all points are equidistant

So, to avoid this, we use similarity.

in  $\downarrow$  NLP  $\rightarrow$  hamming distance

3) As  $d \uparrow$ , chances of overfitting  $\uparrow$

Dimensionality  
Reduction

$\hookrightarrow$  PCA (principal component analysis)

Feature Extraction: Transform your features  $\rightarrow$  new features

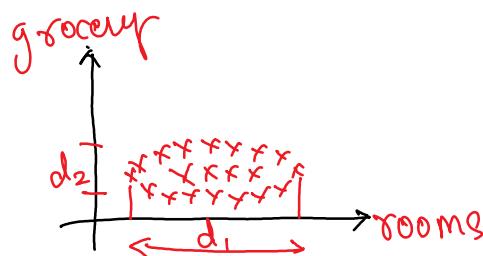
PCA: Reduces the dimensions to the best possible lowest dimension to capture the essence of data.

Benefits:

- faster execution
- visualization (except PCA)

### Basic Intuition

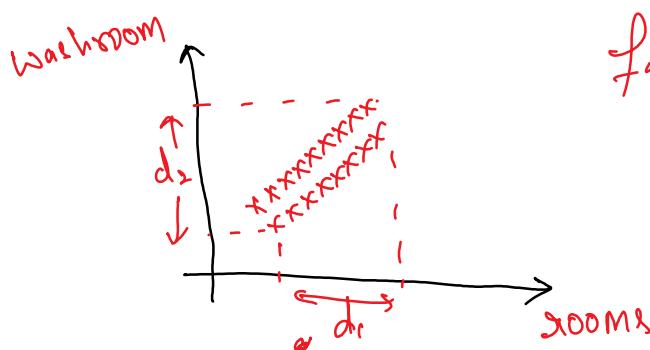
| Rooms | Grocery Shops | Price |
|-------|---------------|-------|
| 3     | 2             | 60    |
| 4     | 0             | 130   |
| 2     | 6             | 170   |
| 5     | 7             | 90    |



axis with higher variance is chosen.

feature selection

# Rooms   # Washrooms   Price



failure case of feature selection

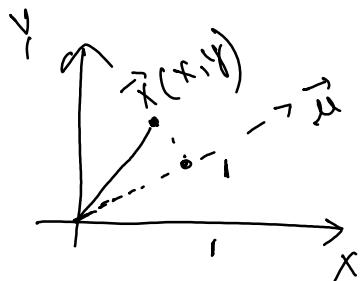
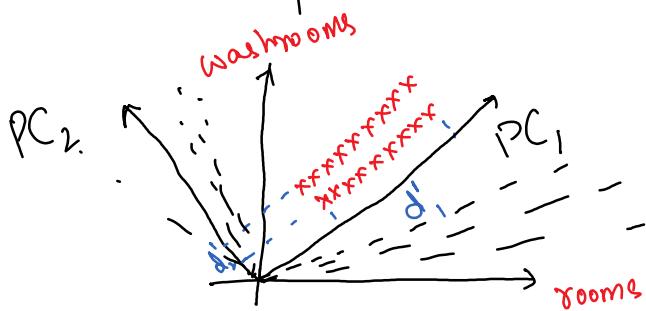
feature extraction

Rooms & Washrooms  $\Rightarrow$  Size of flat

Size      Price      (2d)

feature extraction  $\rightarrow$  creates new features from old features & choose a subset of feature with higher importances.

Geometric Intuition of PCA:



Projection of  $x$  on  $u$   
 $= \frac{\vec{u} \cdot \vec{x}}{\|\vec{u}\|}, \|\vec{u}\|=1$

$$= \vec{u} \cdot \vec{x} \xrightarrow{\text{LA}} (u^T \cdot x)$$

Project each data point on the rotated axis (unit vector)

The unit vector which has the highest variance is chosen as the right axis.

Variance =  $\frac{\sum_{i=1}^n (u^T x_i - u^T \bar{x})^2}{n}$   $\Rightarrow$  Mathematical

$$\text{Variance} = \frac{\sum_{i=1}^n (\bar{x}_i - \bar{x})^2}{n} \Rightarrow \text{Mathematical objective function}$$

Rayleigh Quotient (1920s)

$\Rightarrow$  Covariance  $\Rightarrow$  tell us about the relationship

$\hookrightarrow$  square & symmetric

features ( $x_1$  &  $x_2$ )

$$\begin{matrix} & x_1 & x_2 \\ x_1 & \left[ \begin{matrix} \text{var}(x_1) & \text{cov}(x_1, x_2) \\ \text{cov}(x_1, x_2) & \text{var}(x_2) \end{matrix} \right] \\ x_2 & & \end{matrix}$$

$\Rightarrow$  Matrix  $\rightarrow$  linear transformations

$$A \Rightarrow \text{Matrix}, \lambda = \text{scalar}$$

$A\vec{x} = \underbrace{\vec{x}}_{\text{eigen vector}} \rightarrow \text{eigen value}$

" largest eigenvector of covariance matrix always points in the direction of largest variance".

Steps

1 → Mean Centring  $\rightarrow$  not a mandatory step  $\Rightarrow$  algorithm works very well  
 ↳ Standardization  
 $(\mu=0, \sigma=1)$

2 → Find covariance matrix

$$\begin{matrix} & f_1 & f_2 & f_3 \\ f_1 & \text{Var}(f_1) & \text{Cov}(f_1, f_2) & \text{Cov}(f_1, f_3) \\ f_2 & \text{Cov}(f_1, f_2) & \text{Var}(f_2) & \text{Cov}(f_2, f_3) \\ f_3 & \text{Cov}(f_1, f_3) & \text{Cov}(f_2, f_3) & \text{Var}(f_3) \end{matrix}$$

3 → Eigen decomposition of covariance matrix  
 ↳ eigen values & eigen vector

$$\begin{matrix} f_1 & f_2 & f_3 \\ \text{eigen value} \rightarrow & \lambda_1 & \lambda_2 & \lambda_3 \\ & \downarrow & \downarrow & \downarrow \\ & PC_1 & PC_2 & PC_3 \end{matrix} \quad \begin{matrix} \text{Information comparison} \\ PC_1 > PC_2 > PC_3 \end{matrix}$$

if you chose  $\lambda_1$ ; then it will be 1d  
 " " " or  $\lambda_1 \& \lambda_2$ ; " " " " 2d

" $\lambda_1, \lambda_2, \lambda_3$ " "3d"

How to transform point from 3d to 1d / 3d to 2d?

$\Rightarrow 1d(\text{one PC})$

lets say, dataset has 1000 rows, 3 columns

Shape of unit vector =  $(1, 3)$

$$x \cdot u^T \Rightarrow \begin{bmatrix} \quad \end{bmatrix}_{1000 \times 3} \cdot \begin{bmatrix} \quad \end{bmatrix}_{3 \times 1}$$

$$\Rightarrow \begin{bmatrix} \quad \end{bmatrix}_{1000 \times 1}$$

$\Rightarrow 2d(2PC, PC_1 \& PC_2)$

df.shape =  $(1000, 3)$

Shape of unit vector =  $(2, 3)$

$$u^T \cdot x = \begin{bmatrix} \quad \end{bmatrix}_{1000 \times 3} \cdot \begin{bmatrix} \quad \end{bmatrix}_{3 \times 2}$$

$$= \begin{bmatrix} & \\ & \end{bmatrix}_{1000 \times 2}$$

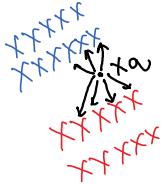
#hyper parameters = n\_components.

K- Nearest Neighbors

You are like your  
neighbors

KNN

$$k = \underline{7}$$



Colour of  $x_q$  = Red

Colour of  $x_q \Rightarrow (2B, 1R) \Rightarrow B \text{ lue} \Rightarrow \text{for } k=3$

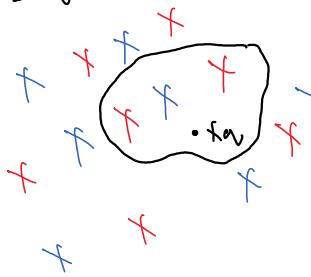
Voting

Colour of  $x_q \Rightarrow (3B, 4R) \Rightarrow R \text{ ed} \Rightarrow \text{for } k=7$

Hyperparameter  $\Rightarrow K = \# \text{ neighbors}$   
 $\downarrow$   
 number of

Reality:

$$K=3$$



$x_q = \text{Red}$

Q Why can't  $g$  have  $K$  as even number?

$$K \neq 2, 4, 6, 8$$

Sol:  $x_q \Rightarrow (4 \text{ Blue}, 4 \text{ Red})$

$$P(B) = \frac{4}{8} = \frac{1}{2} \quad P(R) = \frac{1}{2} = 0.5$$

( $g$  don't know)

That's why, you should always keep  $k$  as odd!

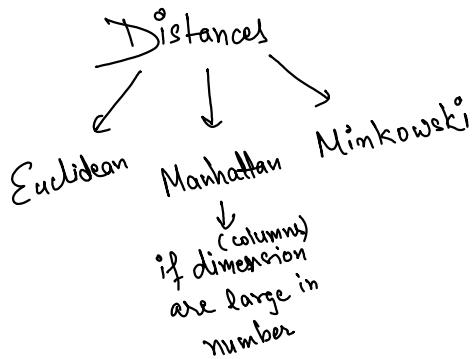
Q  $K=3, K=5, K=6, K=7$ ? Pick wrong value of  $K$ .

$K=6$  is the wrong value

\*  $K = \text{no. of neighbors}$

↳ Tune your model  $\rightarrow k \in \{K\}$  to get best result

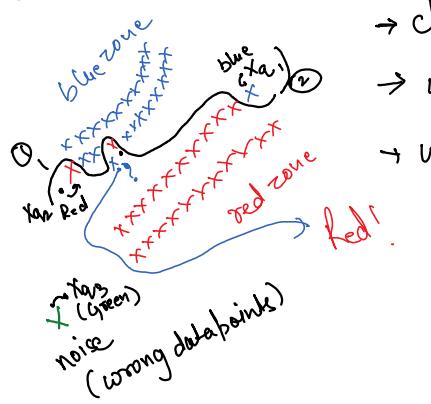
$$K = [3, 5, 7, 9, 11, 13, 15, \dots]$$



\* distance  $\Rightarrow$  KNearestNeighbor (distance=['minkowski'])

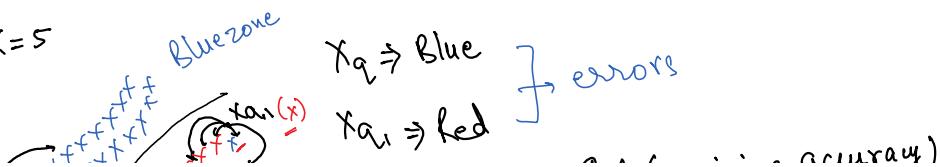
### Effect of $K$ :

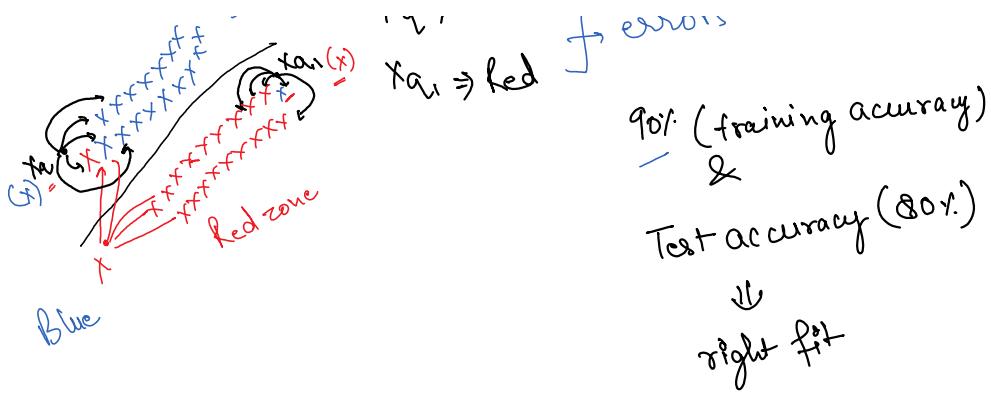
①  $K=1$



- classifier  $\Rightarrow$  non-smooth boundary
  - work very well in training  $\Rightarrow$  train accuracy is 100%
  - won't work well in testing  $\Rightarrow$  test accuracy is low (60%)
- overfitting

•  $K=5$

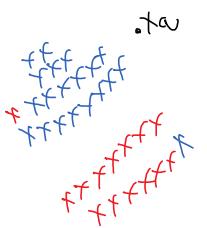




•  $K=n$

$$\text{count(blue)} > \text{count(red)}$$

$\xrightarrow{}$



→ won't work well in training

→ " " " testing

\* gt will always predict datapoints as blue.

overfitting

Underfitting

→ Training accuracy is high

→ Training accuracy is low

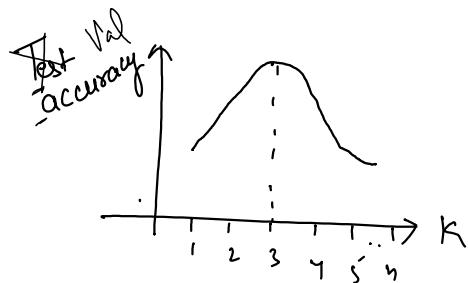
→ Test accuracy is low

→ Test accuracy is low

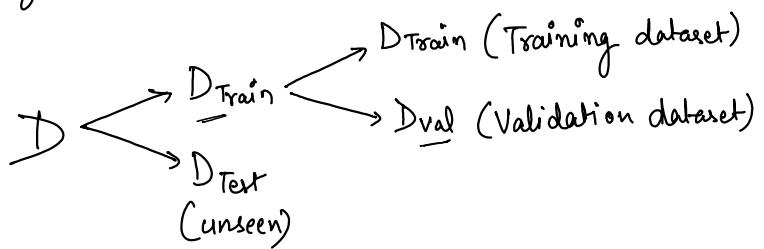
→ Captures all points/memorizes all data points including noise & outliers

→ didn't learn anything

Curve of accuracy with  $K$ :



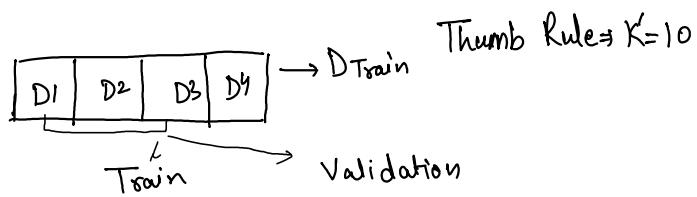
Choose the right value of  $k$ :



Cross-validation → to keep test data as unseen data

( $K'$ -fold CV)       $K' = 4$        $K = \# \text{ neighbors}$

$K = [1, 2, 3]$



$K = 1 \nparallel$

# Neighbors = 1

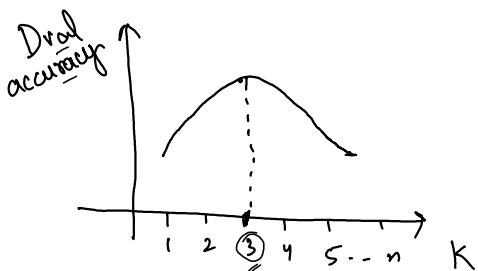
| $K$ | $D_{\text{Train}}$ | $D_{\text{Validation}}$ |
|-----|--------------------|-------------------------|
| 1   | $D_1, D_2, D_3$    | $D_4 \rightarrow a_1$   |
| 1   | $D_2, D_3, D_4$    | $D_1 \rightarrow a_2$   |
| 1   | $D_1, D_3, D_4$    | $D_2 \rightarrow a_3$   |
| 1   | $D_1, D_2, D_4$    | $D_3 \rightarrow a_4$   |

# Neighbors = 2

$K=2$

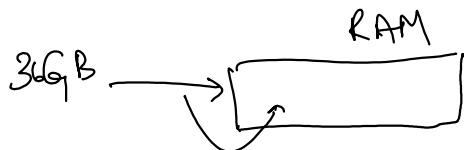
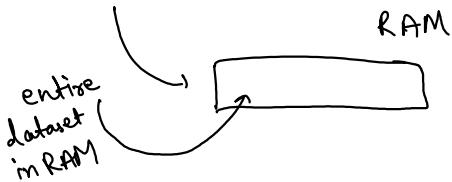
|   |                 |                        |
|---|-----------------|------------------------|
|   | $D_1, D_2, D_3$ | $D_4 \rightarrow a'_1$ |
| 2 | $D_2, D_3, D_4$ | $D_1 \rightarrow a'_2$ |
| 2 | $D_1, D_3, D_4$ | $D_2 \rightarrow a'_3$ |
| 2 | $D_1, D_2, D_4$ | $D_3 \rightarrow a'_4$ |

Choose the value of  $k$  which has highest accuracy.  
 $\downarrow$   
 $\# \text{ neighbors}$



## Disadvantages:

- 1) Lazy learners  $\rightarrow$  calculations at time of execution.
- 2) sensitive to outliers/noise
- 3) Space problems/time complexity



Application:  $\rightarrow$  heavily used in healthcare.

## Evaluation Metrics (All classification algorithm)

### Confusion Matrix

|           |           | Actual   |           |  |
|-----------|-----------|----------|-----------|--|
|           |           | $\oplus$ | $\ominus$ |  |
| Predicted | $\oplus$  | TP       | FP        |  |
|           | $\ominus$ | FN       | TN        |  |

Definitions:

- FP = false positive
- TP = True "
- FN = False Negative
- TN = True "

1) Accuracy  $\Rightarrow \frac{TP + TN}{TP + TN + FP + FN}$

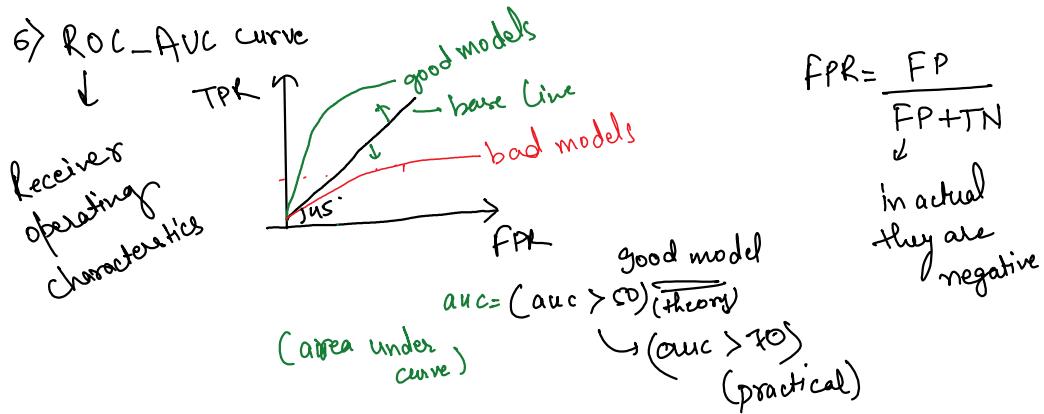
2) Precision  $\Rightarrow \frac{TP}{TP + FP} \Rightarrow \text{Precision} = \frac{\text{True Positive}}{\text{True Positive} + \text{False Positive}}$

3) Recall  $\Rightarrow \frac{TP}{TP + FN}$

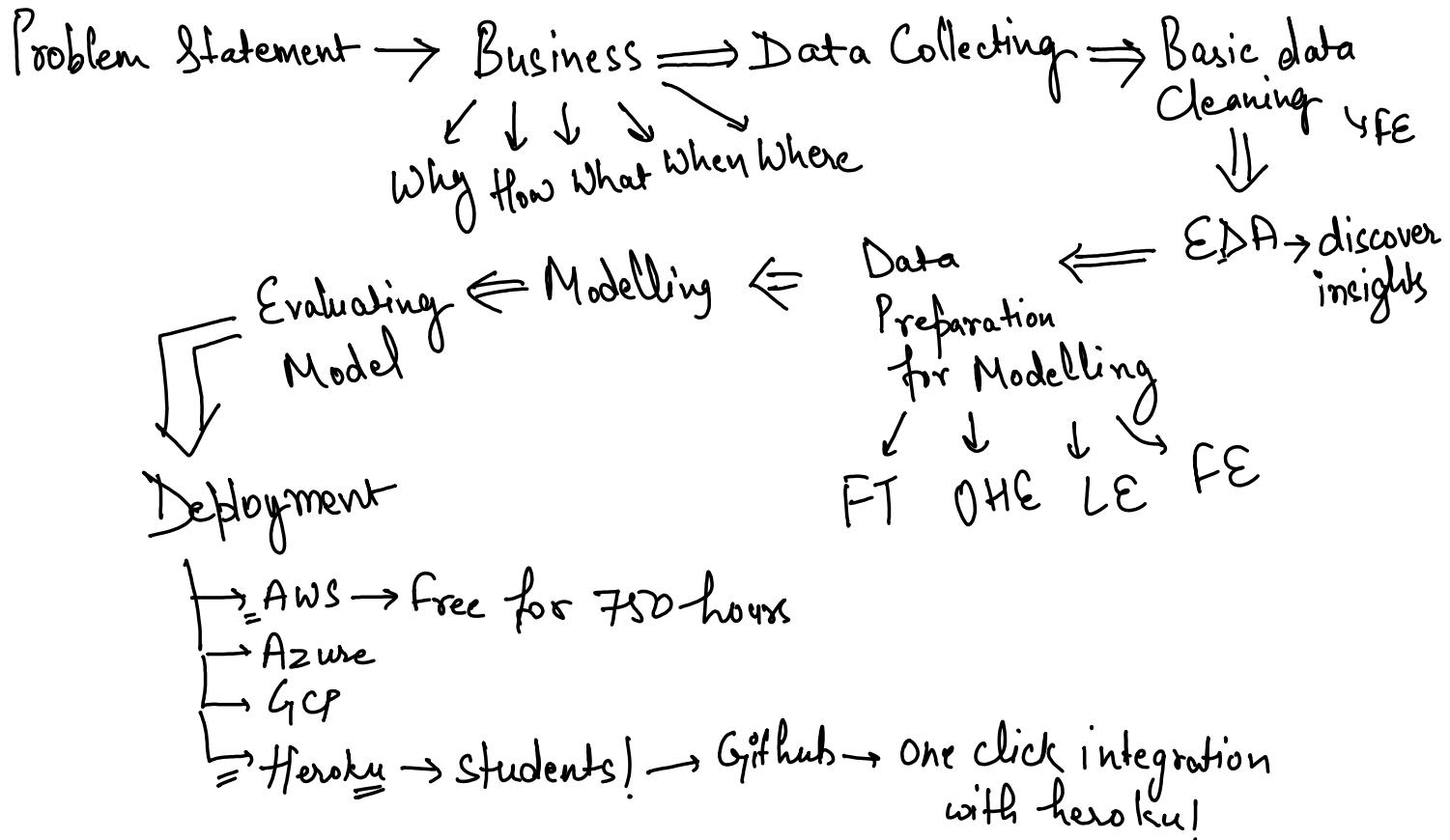
4) Recall  $\Rightarrow \frac{TP}{TP+FN} \Rightarrow$  Sensitivity (true positive rate)

4) Specificity  $\Rightarrow \frac{TN}{TN+FP}$

5) F1-score =  $\frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$



# Data Science Project Life Cycle



## Project

- import data
- Basic information regarding data

[`df.info()`]

- datatypes of columns
- metadata → data about data
- no. of columns
- no. of rows
- ...
- ...

↗ no. of rows  
 ↗ Non-nulls

- Basic description of data (`df.describe()`)

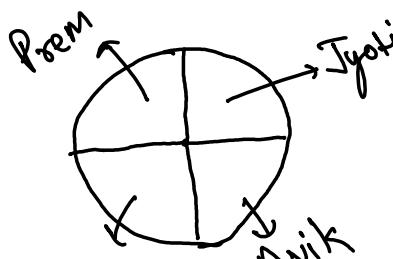
|        |         |          |
|--------|---------|----------|
| → min  | → 25%   | → median |
| → max  | → 50%   |          |
| → mean | → 75%   |          |
| → σ    | → count |          |

- Data study → `unique()`, `nunique()`, `value_counts()`
  - ↓ list of unique values
  - ↓ number of unique value
  - ↓ count of unique values()

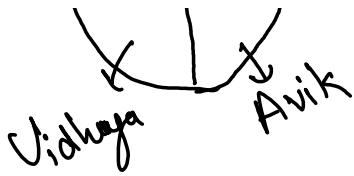
- Basic data cleaning → FE

- EDA
  - Univariate → Pie chart, histogram, kde plot
  - Bivariate → Piechart, scatter, boxplot, linechart
  - Multivariate → Heatmap, pairplot

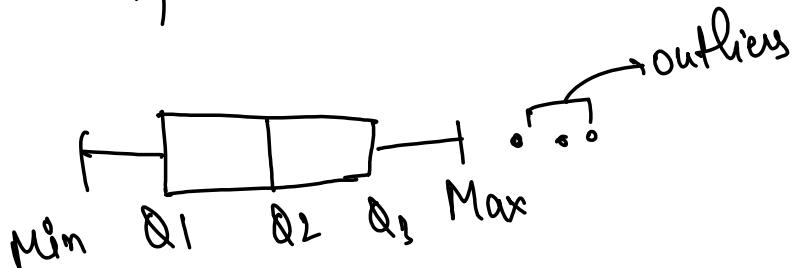
Q Pie or Bar?  $\Rightarrow$  4 categories  $\Rightarrow$  count



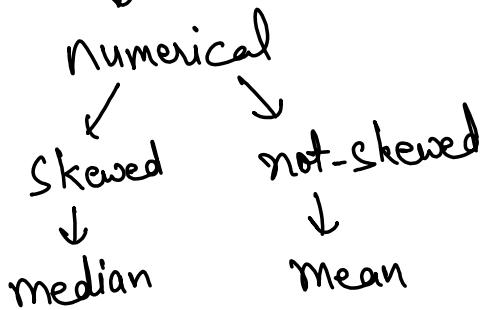
Bar has more precision chart



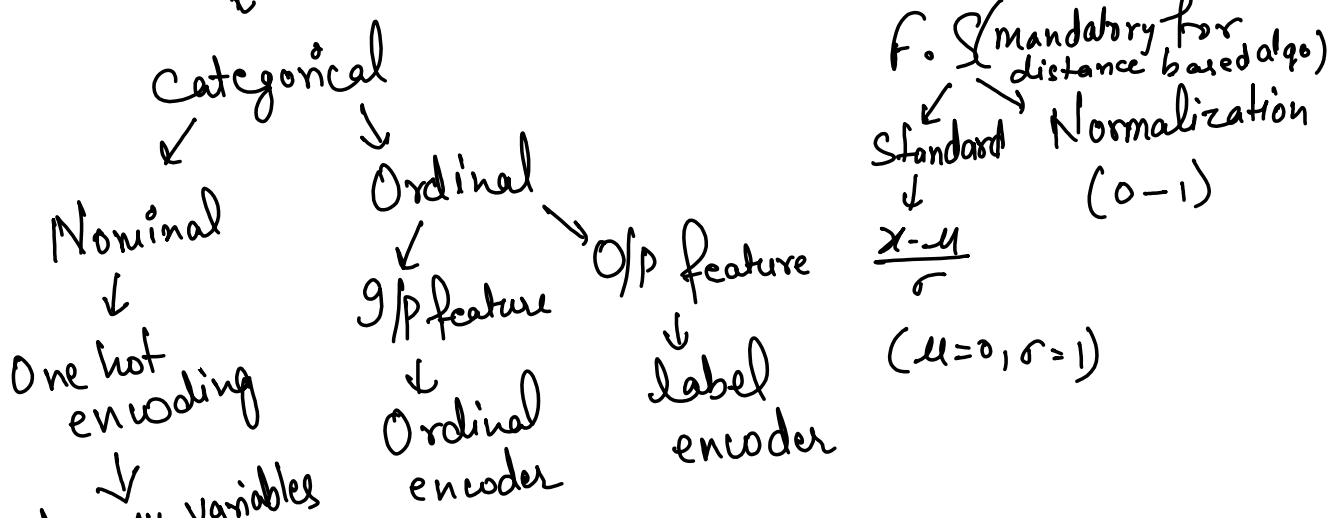
→ Outlier Analysis → Boxplot



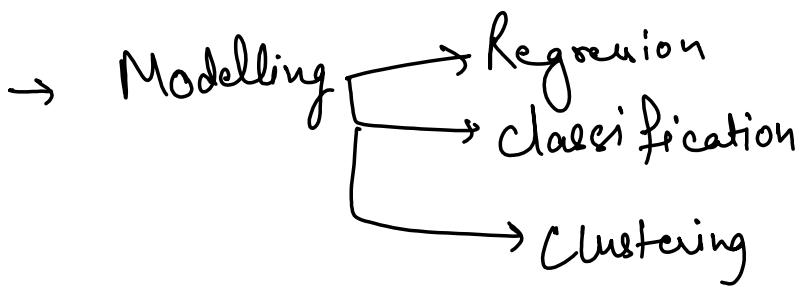
→ Missing values → Category → mode



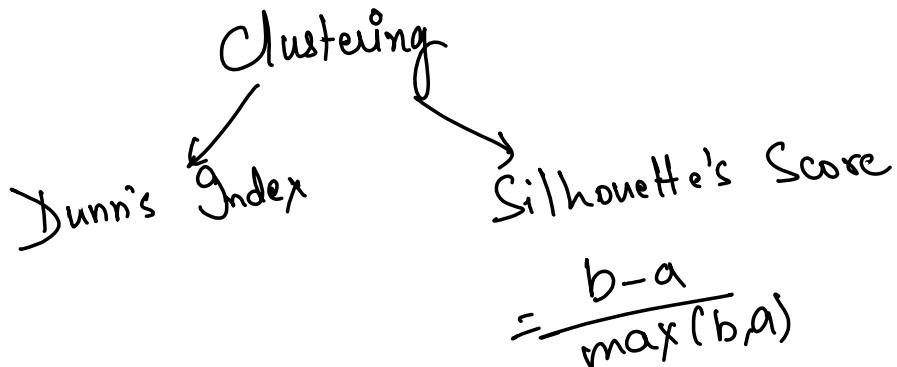
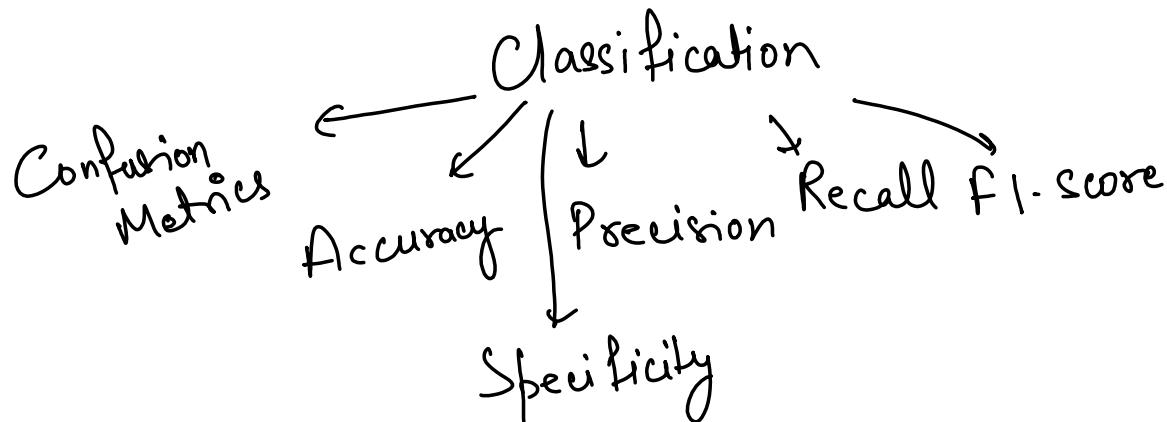
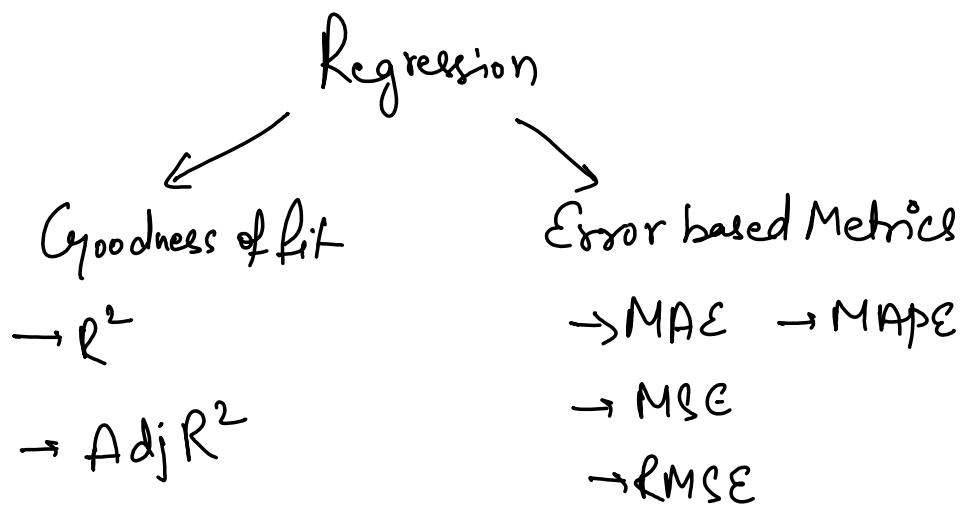
→ Data Preparation for Modelling → Numerical → f.E



dummy variables  $\downarrow$  encoder encoder



→ Evaluation



Deployment → different module

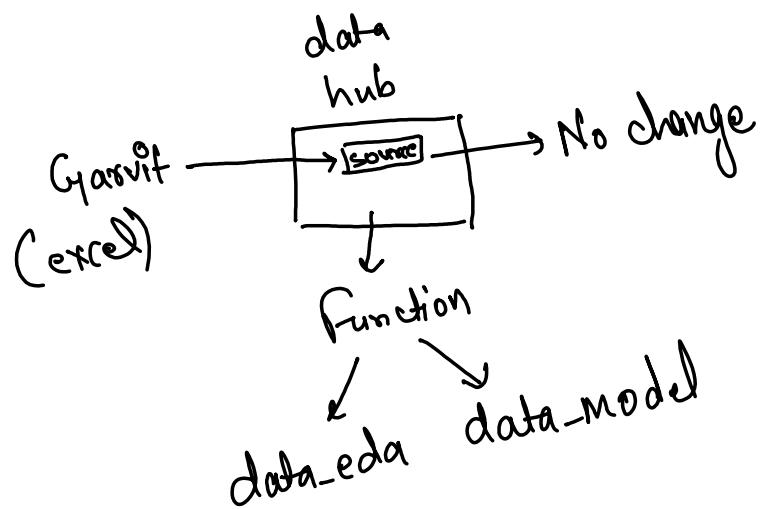
Regression

Air ticket Price Prediction



Data Transformation

What is the requirement? → Data integrity



# Naïve Bayes

Sunday, October 1, 2023 8:36 AM

## Bayes Theorem

$$P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{P(B|A) \times P(A)}{P(B)}$$

$$P(A|B) = \frac{P(B|A) \times P(A)}{P(B)}$$

Posterior =  $\frac{\text{likelihood} \times \text{Prior}}{\text{evidence}}$

\*IQ

Probability & likelihood  $\Rightarrow$  What is the difference?

Let's  $h = \text{height}$

probability  $P(H > 170\text{cm} | \mu=160, \sigma=1)$   $\Rightarrow$  quantification of something happening

likelihood  
 $P(\mu=160, \sigma=1 | H > 170\text{cm})$   $\Rightarrow$  getting best data distribution  
distribution for your observation.

Naïve Bayes  
( ignorant )

Assumptions:

$\Rightarrow$  All features are independent of each other.

⇒ all features have equal contribution in predicting the O/p variable.

$$P(C_x/x_i) = \frac{P(X_i/C_x) P(C_x)}{P(X_i)}$$

$C_x$  → class label

$X_i$  → G/P variables

$$P(\text{Yes}/O, T, H, W) = ?$$

$$P(\text{No}/O, T, H, W) = ?$$

| Outlook  | Temperature | Humidity | Windy | PlayTennis |
|----------|-------------|----------|-------|------------|
| Sunny    | Hot         | High     | False | No         |
| Sunny    | Hot         | High     | True  | No         |
| Overcast | Hot         | High     | False | Yes        |
| Rainy    | Mild        | High     | False | Yes        |
| Rainy    | Cool        | Normal   | False | Yes        |
| Rainy    | Cool        | Normal   | True  | No         |
| Overcast | Cool        | Normal   | True  | Yes        |
| Sunny    | Mild        | High     | False | No         |
| Sunny    | Cool        | Normal   | False | Yes        |
| Rainy    | Mild        | Normal   | False | Yes        |
| Sunny    | Mild        | Normal   | True  | Yes        |
| Overcast | Mild        | High     | True  | Yes        |
| Overcast | Hot         | Normal   | False | Yes        |
| Rainy    | Mild        | High     | True  | No         |

if  $P(\text{Yes}/O, T, H, W) > P(\text{No}/O, T, H, W)$ :

play tennis

else:

Not tennis

$$\# P(\text{Yes}/\text{outlook}) = \frac{P(\text{outlook}/\text{Yes}) P(\text{Yes})}{P(\text{outlook})}$$

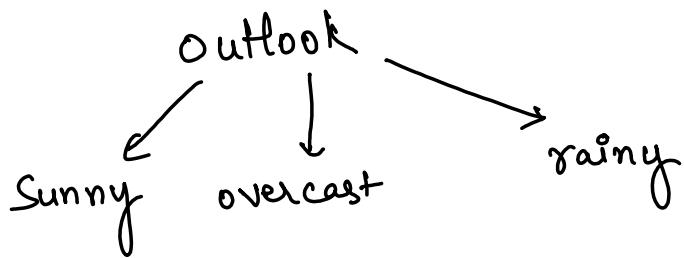
$$P(\text{No}/\text{outlook}) = \frac{P(\text{outlook}/\text{No}) P(\text{No})}{P(\text{outlook})}$$

\* ignore denominator because it is same for all probabilities

$$P(\text{outlook}/\text{Yes}) = ? \quad \& \quad P(\text{Yes}) \& P(\text{No})$$

Working:

$$P(\text{Yes}) = \frac{9}{14} \quad P(\text{No}) = \frac{5}{14}$$



$$P(\text{Sunny}/\text{Yes}) = \frac{2}{9}$$

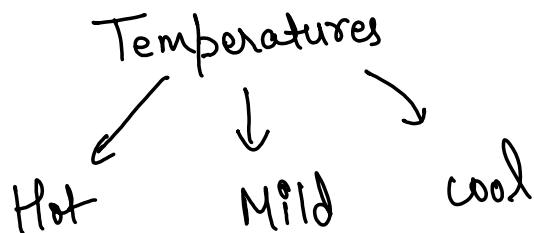
$$P(\text{Sunny}/\text{No}) = \frac{3}{5}$$

$$P(\text{Rainy}/\text{Yes}) = \frac{3}{9}$$

$$P(\text{Rainy}/\text{No}) = \frac{2}{5}$$

$$P(\text{Overcast}/\text{Yes}) = \frac{4}{9}$$

$$P(\text{Overcast}/\text{No}) = 0$$



$$P(\text{Hot}/\text{Yes}) = \frac{2}{9}$$

$$P(\text{Hot}/\text{No}) = \frac{2}{5}$$

9

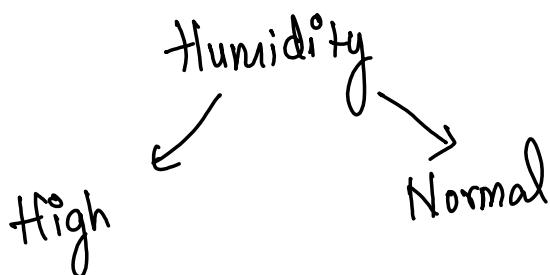
5

$$P(\text{Mild} | \text{Yes}) = \frac{4}{9}$$

$$P(\text{Mild} | \text{No}) = \frac{2}{5}$$

$$P(\text{cool} | \text{Yes}) = \frac{3}{9}$$

$$P(\text{cool} | \text{No}) = \frac{1}{5}$$

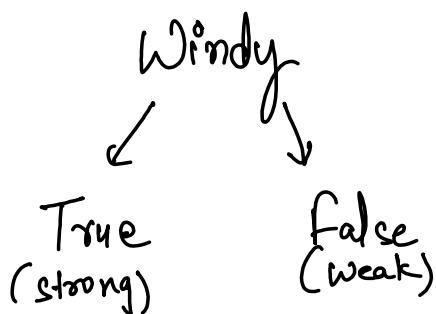


$$P(\text{High} | \text{Yes}) = \frac{3}{9}$$

$$P(\text{High} | \text{No}) = \frac{4}{5}$$

$$P(\text{Normal} | \text{Yes}) = \frac{6}{9}$$

$$P(\text{Normal} | \text{No}) = \frac{1}{5}$$



$$P(\text{True} | \text{Yes}) = \frac{3}{9}$$

$$P(\text{True} | \text{No}) = \frac{3}{5}$$

$$P(\text{False} | \text{Yes}) = \frac{6}{9}$$

$$P(\text{False} | \text{No}) = \frac{2}{5}$$

Q Outlook is sunny, temp is cool, humidity is high, wind is strong(true). Will g play?

$$\text{Sol: } P(\text{Yes} \mid \text{sunny, cool, high, true}) = P(\text{sunny} \mid \text{Yes}) \times P(\text{cool} \mid \text{Yes}) \times \\ P(\text{high} \mid \text{Yes}) \times P(\text{true} \mid \text{Yes}) \times P(\text{Yes})$$

$$= \frac{9}{14} \times \frac{2}{27} \times \frac{3}{9} \times \frac{3}{9} \times \frac{2}{9} = \frac{2}{14 \times 27} = \frac{1}{189} = 0.0052$$

$$P(\text{No} \mid \text{sunny, cool, high, true}) = P(\text{sunny} \mid \text{No}) \times P(\text{cool} \mid \text{No}) \times P(\text{high} \mid \text{No}) \times \\ P(\text{true} \mid \text{No}) \times P(\text{No})$$

$$= \frac{7}{14} \times \frac{3}{8} \times \frac{1}{5} \times \frac{4}{5}^2 \times \frac{3}{5} \\ = \frac{18}{125 \times 7} = \frac{18}{875} = 0.0205$$

$$P(\text{Yes} \mid \text{sunny, cool, high, true}) < P(\text{No} \mid \text{sunny, cool, high, true}) \\ 0.0052 < 0.0205$$

Prediction: g will not play tennis  $\Rightarrow$  No

Q outlook is rainy, temperature is hot, humidity is normal & wind is false. Will g play?

Sol.  $P(\text{Yes}) = \frac{9}{14}$   $P(\text{No}) = \frac{5}{14}$

10:30 am

$$P(\text{Yes} \mid \text{rainy, hot, normal, false}) = 0.0211$$

$$P(\text{No} \mid \text{rainy, hot, normal, false}) = 0.0045$$

∴ will Play!  $\Rightarrow$  Yes

Scenario ①: Problem of zero probability

Q Will g play, if outlook is sunny, temp is cool, humidity is normal, clothing is casual?  $P(C) = 0$

Sol.  $P(\text{Yes} \mid \text{sunny, cool, normal, casual}) = 0$

$$P(\text{No} \mid \text{sunny, cool, normal, casual}) = 0$$

\* Laplace Smoothing

$$P(\text{Yes} \mid C) = \frac{O + \alpha}{n + \alpha k}$$

$n = \# \text{ datapoints of class}$

$\alpha$  # values feature can take

$$P(Yes|C) = \frac{P(C/Yes) P(Yes)}{P(C)} = \frac{0 + \alpha}{n + \alpha k}$$

Range of  $\lambda = (1, \infty)$

Let's  $x = 1$ ,  $n = 1000$ ,  $k = 2$

$$\textcircled{1} \quad P = \frac{0 + \alpha}{n + \alpha k} = \frac{1}{1000 + 1 \times 2} = \frac{1}{1000} = \text{nearly 0}$$

overfitting

Lets  $n = 1000$ ,  $k = 2$ ,  $\alpha = 1000$

$$\text{II) } P = \frac{0 + \alpha}{n + \alpha k} = \frac{1000}{1000 + 2 \times 1000} = \frac{1000}{3000} = \frac{1}{3} = 0.33 \quad (\text{Right fit})$$

$$\text{III} \quad P = \frac{0 + 10000}{1000 + 2 \times 10000} = \frac{10000}{21000} \approx 0.5 \text{ (under fitting)}$$

Hyperparameters :  $\lambda$

Very small

very high

Naive Bayes  $\Rightarrow \propto$

overfitting

underfitting

KNN  $\Rightarrow K$

overfitting

underfitting

Scenario II: I have a lot of dimensions in my dataset



probabilities will be very low



Take log of probabilities  
(log probabilities)

(var-smoothing)

Scenario III: Outliers  $\Rightarrow \propto$  (Laplace smoothing will take care of it)

Thumb Rule:

Gaussian NB

Numerical features  
in g/p features

Naive Bayes

Bernoulli NB

Boolean features  
in g/p  
features

Multinomial NB

More than 2.  
categorical  
values in  
1..n levels

<sup>given</sup>  
in g/p features

1^n - 11  
features

values in  
g/p features

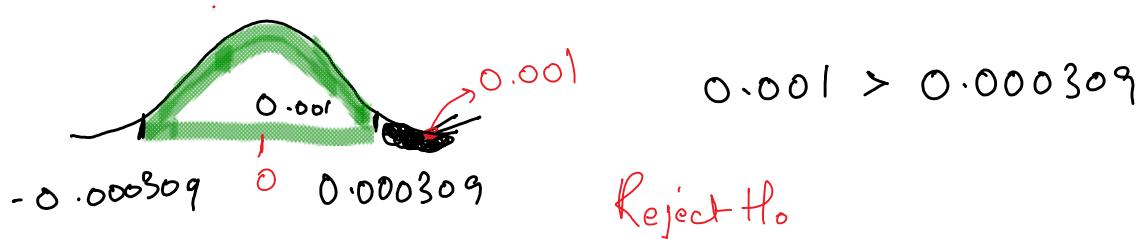
\* don't look at O/P variable

Assignment Sol" → HT

"status quo"      "claims"       $\alpha = 0.05$       CI = 0.95  
 $H_0 \Rightarrow \mu = 0$ ,  $H_A \Rightarrow \mu \neq 0$       ↗ 1.96  
 $\bar{x} = 0.1\%$ ,  $\sigma = 0.25\%$   
 $-1.96$

$$UL = 0 + 1.96 \times \frac{0.0025}{\sqrt{20}} = 0.000309$$

$$LL = 0 - 1.96 \times \frac{0.0025}{\sqrt{20}} = -0.000309$$



"Logistic Regression" → binary classification  
 ↘ Geometric    ↘ Probabilistic    ↘ Loss minimization

⇒ Assumption: data is linearly separable

$$f(x) \leftarrow y = \vec{w}^T \vec{x} + c \rightarrow w_0$$

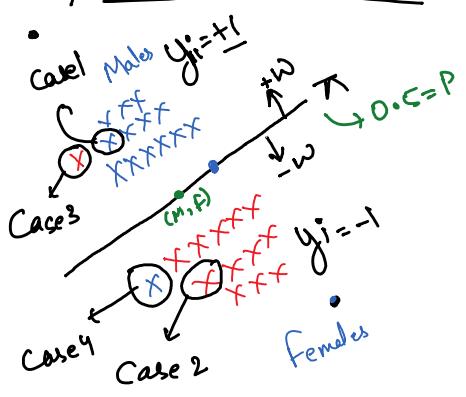
$$f(x) = w_0 x + w_0 \rightarrow \text{eqn of line}$$

$$f(x) = w_0 + w_1 x_1 + w_2 x_2 + \dots + w_n x_n \rightarrow \text{eqn of hyperplane}$$

$$\begin{aligned} f(x) &= w_0 + w_1 x_1 + w_2 x_2 + \dots + w_n x_n \\ &= w_0 + w_1 x_1 + w_2 x_2 + \dots + w_n x_n \\ &= [w_0 \ w_1 \ w_2 \ \dots \ w_n] \begin{bmatrix} x_0 \\ x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} \end{aligned}$$

$f(x) = w^T x \rightarrow \text{eqn of hyperplane passing through origin}$

### Geometric Interpretation



$$y_i = [+1, -1]$$

①  $w^T x > 0 \Rightarrow$  for +ve class

②  $w^T x < 0 \Rightarrow$  for -ve class

$\Rightarrow$  let's multiply  $y_i$  with  $w^T x_i$

$$y_i w^T x_i$$

Case 1:  $y_i = +1, w^T x_i > 0$

$$y_i w^T x_i > 0$$

Case 2:  $y_i = -1, w^T x_i < 0$

$$y_i w^T x_i > 0$$

Correct classification

Case 3:  $y_i = +1, w^T x_i < 0$

$$y_i w^T x_i < 0$$

Case 4:  $y_i = -1, w^T x_i > 0$

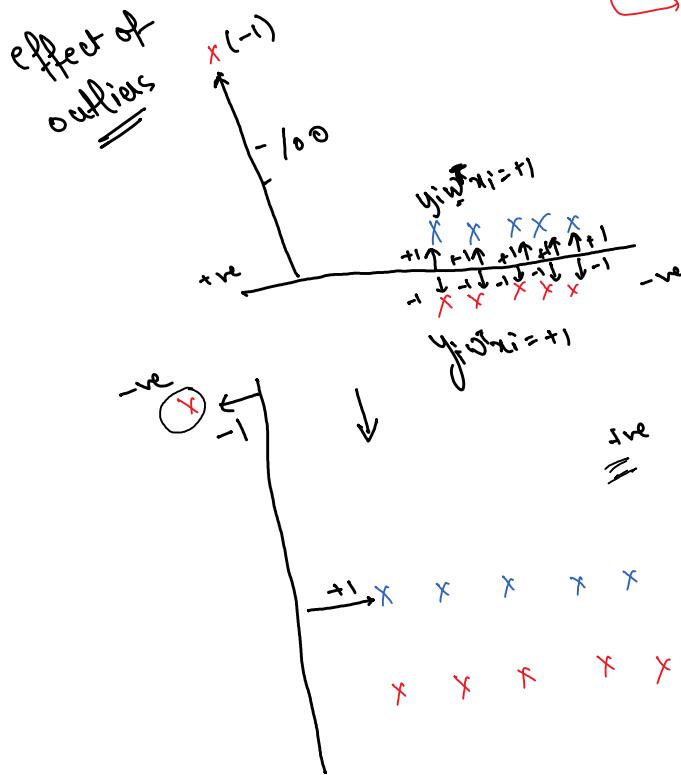
$$y_i w^T x_i < 0$$

incorrect classifications

if  $\begin{cases} y_i w^T x_i > 0 & : \text{correct classifications} \\ y_i w^T x_i < 0 & : \text{incorrect classification} \end{cases}$   $\Rightarrow$  Mathematical Objective function

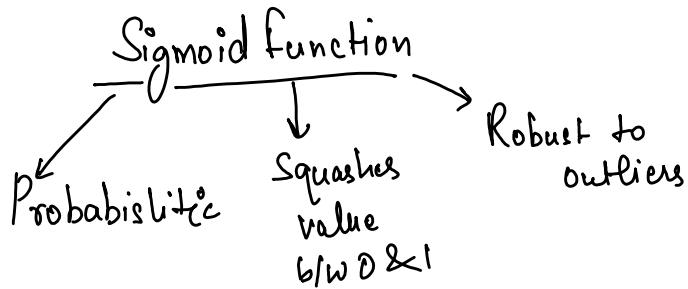
Mathematical Obj.  $F = \underset{w^T}{\arg \max} (y_i w^T x_i) = -90$

$\hookrightarrow$  Sensitive to outlier



$$\text{Mof} = 1 + 1 + 1 + 1 + 1 + 1 + 1 + 1 + 1 - 100 \\ = -90$$

- a lot of misclassifications
- Not a good model



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

$$\sigma(y_i w^T x_i) = \frac{1}{1 + e^{-y_i w^T x_i}}$$

$$MOF = \operatorname{argmax} [\sigma(y_i w^T x_i)] \Rightarrow \operatorname{argmax} \left( \frac{1}{1 + e^{-y_i w^T x_i}} \right)$$

applying log

$$\Rightarrow \operatorname{argmax} \left[ \log \left( \frac{1}{1 + e^{-y_i w^T x_i}} \right) \right] \quad \log \left( \frac{1}{a} \right) = -\log a$$

$$\Rightarrow \operatorname{argmax} [-\log(1 + e^{-y_i w^T x_i})]$$

\* loss function  $\Rightarrow \operatorname{argmin} [\log(1 + e^{-y_i w^T x_i})] \Rightarrow$  logistic loss

Defining logistic loss is equivalent to MOF

$$\begin{aligned} &= \operatorname{argmin} (\log^+ + \log^- y_i w^T x_i) \\ &= \operatorname{argmin} (-y_i w^T x_i \log^+) \\ &= \operatorname{argmax} (y_i w^T x_i) \end{aligned}$$

"Squashes Value"  
between 0 & 1

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

$x = -\infty \rightarrow 0$

$$\frac{1}{1 + e^{+\infty}} \rightarrow 0$$

$x = +\infty \rightarrow 1$

$$\frac{1}{1 + e^{-\infty}} \rightarrow 1$$

$$\frac{1}{1+\infty}$$

↓

$$\frac{1}{\infty} = 0$$

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

↓

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

Probabilistic

$$P = \frac{1}{1+e^{-Y}}$$

$$y_i = [1, 0]$$

$$P(1+e^{-Y}) = 1$$

$$P + Pe^{-Y} = 1$$

$$Pe^{-Y} = 1 - P$$

$$e^{-Y} = \frac{1-P}{P}$$

$$\frac{1}{e^Y} = \frac{1-P}{P}$$

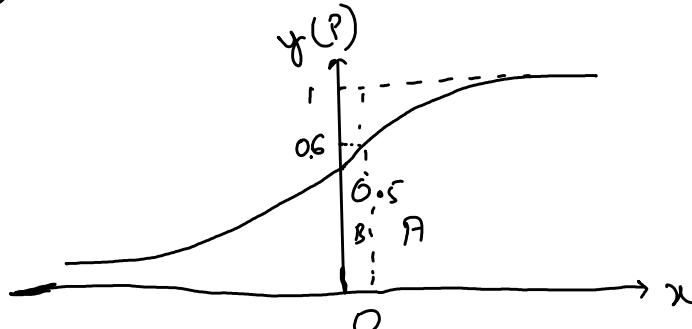
$$e^Y = \left( \frac{P}{1-P} \right) \rightarrow \text{odd's ratio}$$

Take  $\ln$  on both sides

$$y_0 = \ln \left( \frac{P}{1-P} \right) \rightarrow \text{logit function}$$

$y_i = 1, 0$

Logloss  $\Rightarrow y_i P(y_i) + (1-y_i) P(1-y_i) \rightarrow$  loss function for Log Reg  
in Probabilistic Interpretation

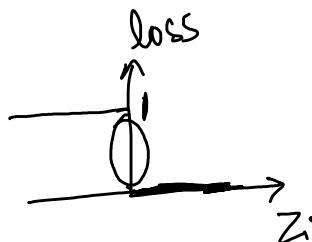
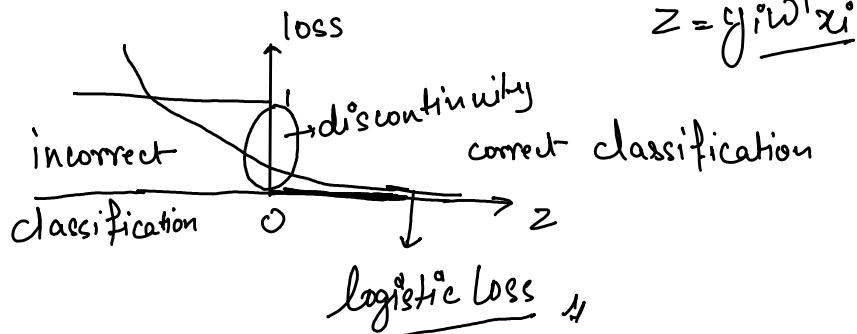


\* logit function  
you can use LogReg  
for Regression



## Loss Minimization

### 0-1 loss



Assignment:

find equivalence b/w logistic loss & log loss?

## Overfitting & Underfitting

$$w^* = \arg \min_w \sum_{i=1}^n \log(1 + e^{-y_i w^T x_i})$$

lets  $\rightarrow y_i w^T x_i = z_i$

$$= \arg \min_w \sum_{i=1}^n \log(1 + e^{-z_i})$$

$$y_i w^T x_i = z_i$$

## Regularizer

$$w^* = \arg \min_w \left[ \sum_{i=1}^n \log(1 + e^{-z_i}) + (\lambda \frac{w^T w}{\|w\|^2}) \right]$$

$w^T w = w \cdot w = \|w\|^2$   
L2 Norm  
large value

hyperparameter  $\lambda$   
 $\lambda = 0 \rightarrow$  overfitting  
 $\lambda = \text{large} \rightarrow$  underfitting

The above expression is for Ridge Regularizer (feature preservation)

→ LASSO (Abb) → \_\_\_\_\_

(feature selection)

$$w^* = \underset{w}{\operatorname{argmin}} \left[ \log(1+e^{-z_i}) + \lambda \|w\|_1 \right]$$

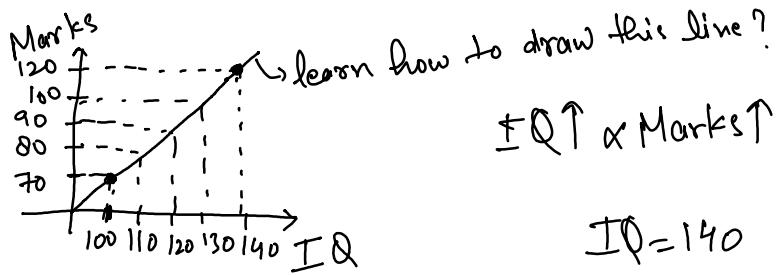
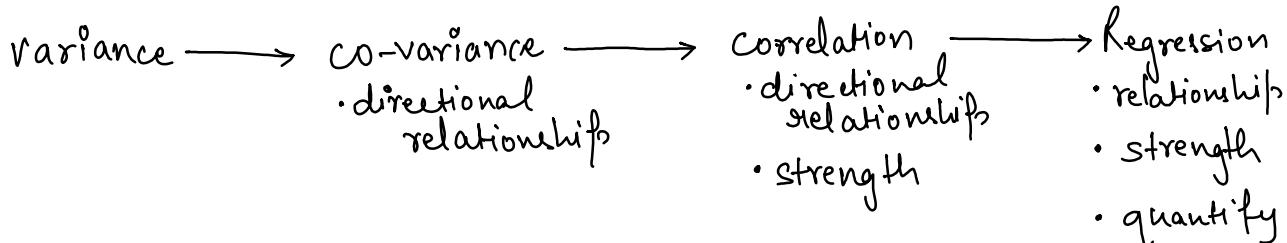
$\hookrightarrow L_1 \text{ Norm}$

LASSO is used for extracting imp. features

$$\Rightarrow \text{ElasticNet} \Rightarrow w^* = \underset{w}{\operatorname{argmin}} (\log(1+e^{-z_i}) + \lambda_1 \|w\|_1 + \lambda_2 \|w^T w\|)$$

- When the dimensions are greater than the no. of samples taken.
- Multicollinearity is handled in efficient way.

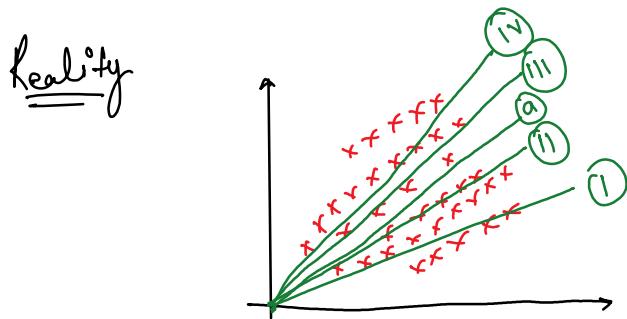
# Linear Regression



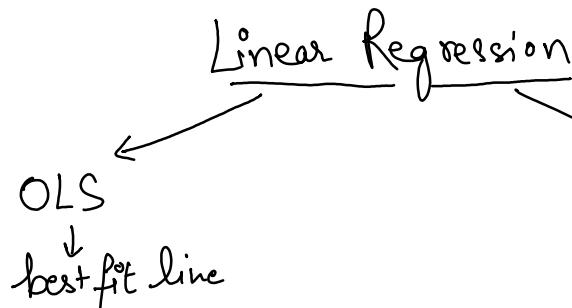
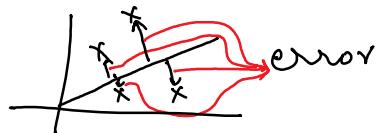
$$IQ = 140 \Rightarrow \text{Marks} = 120 \quad (\text{Prediction})$$

$$\text{Marks} = w \times IQ + b$$

O/P      ↓ Slope      g/p

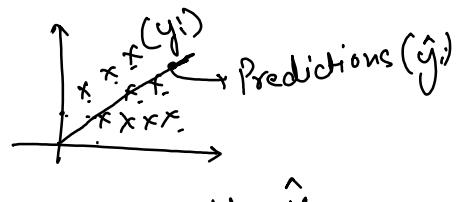


choose a line that makes min. errors.



How do you create a line?

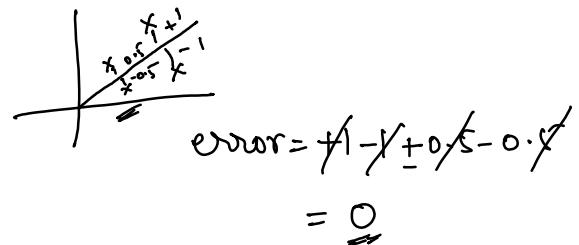
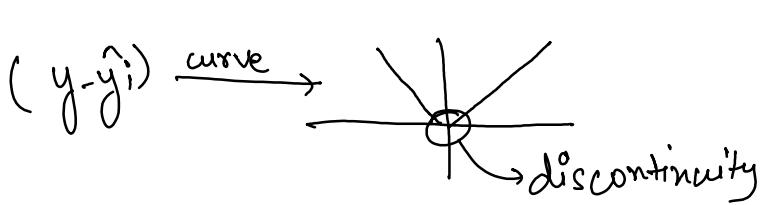
$$y = mx + b$$



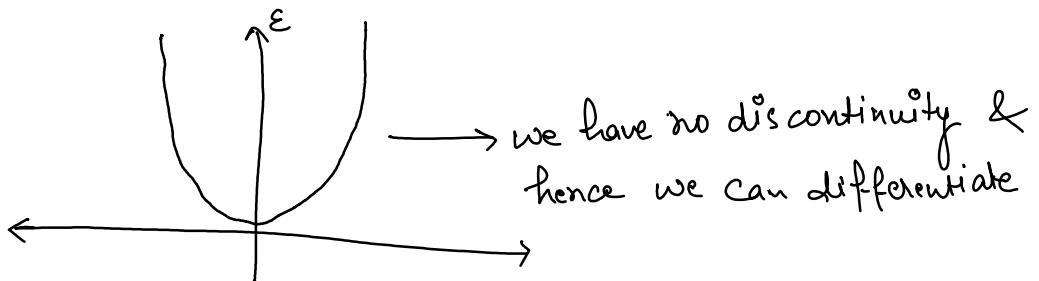
$$y = mx + b$$

slope      intercept

error =  $y_i - \hat{y}_i$



To avoid above situation, we  $(y_i - \hat{y}_i)^2$



$$\epsilon(m, b) = (y_i - \hat{y}_i)^2 = 0$$

$$\epsilon(m, b) = (y_i - (mx_i + b))^2 = 0$$

$$\epsilon(m, b) = \sum_{i=1}^n [y_i - (mx_i + b)]^2 = 0$$

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

\* Properties of derivatives:  
 $\frac{\partial x^n}{\partial x} = nx^{n-1}$

$$\frac{\partial (ax)}{\partial x} = a$$

$$\Rightarrow \sum_{i=1}^n 2[y_i - (mx_i + b)] \left( \frac{\partial y_i}{\partial b} - \cancel{\frac{\partial mx_i}{\partial b}} - \cancel{\frac{\partial b}{\partial b}} \right) = 0$$

$$\frac{d x}{d x} = 1$$

$$\Rightarrow \sum_{i=1}^n -2 [y_i - mx_i - b] = 0$$

$$\Rightarrow \sum_{i=1}^n [y_i - mx_i - b] = 0$$

$$\sum_{i=1}^n y_i - \sum_{i=1}^n mx_i - \sum_{i=1}^n b = 0$$

divide by  $n$  on both sides

$$\left( \frac{\sum_{i=1}^n y_i}{n} \right) - \frac{\sum_{i=1}^n mx_i}{n} - \frac{\sum_{i=1}^n b}{n} = \frac{0}{n}$$

$$\bar{y}_i - m\bar{x}_i - b = 0$$

$$b = \bar{y}_i - m\bar{x}_i \rightarrow \text{value for intercept}$$

$$\frac{\partial E}{\partial m} = \frac{\partial \sum_{i=1}^n [y_i - mx_i - b]^2}{\partial m}$$

$$\Rightarrow \frac{\partial [y_i - mx_i - (\bar{y}_i - m\bar{x}_i)]^2}{\partial m} = 0$$

$$\Rightarrow 2 [y_i - mx_i - \bar{y}_i + m\bar{x}_i] [0 - x_i - 0 + \bar{x}_i] = 0$$

Assignment

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

~~Assignment~~

slope  $\rightarrow$

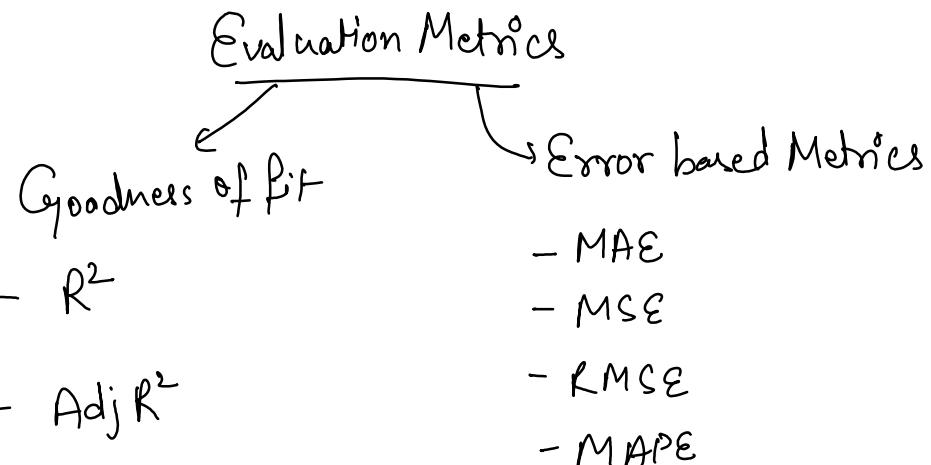
$$m = \frac{\sum (y_i - \bar{y}_i)(\bar{x}_i - x_i)}{\sum (x_i - \bar{x}_i)^2}$$

We got  $m$  &  $b$ ,

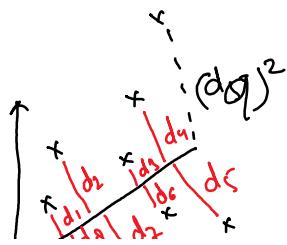
$$b = \bar{y}_i - m \bar{x}_i \quad , \quad m = \frac{\sum (y_i - \bar{y}_i)(\bar{x}_i - x_i)}{\sum (x_i - \bar{x}_i)^2}$$

The above method is OLS, as we have drawn best fit line directly.

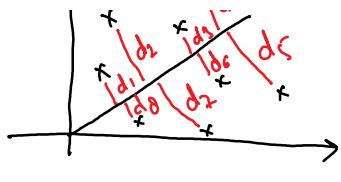
for large datasets do not use it!



MAE: Mean Absolute Error  $\Rightarrow \frac{\sum_{i=1}^n |y_i - \hat{y}_i|}{n}$



$$MAE = |d_1| + |d_2| + |d_3| + |d_4| + |d_5| + |d_6| + |d_7| + |d_8|$$



$$MAE = \frac{|d_1| + |d_2| + |d_3| + |d_4| + |d_5| + |d_6| + |d_7| + |d_8|}{8}$$

$$MAE = \frac{\sum_{i=1}^n |d_i|}{n}$$

Advantages: → same scale as that of data  
 → less sensitive to outliers

2) MSE  $\Rightarrow$  Mean Squared Error

$$\Rightarrow MSE = \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{n} \rightarrow \text{loss function for LR}$$

- Not easily interpretable
- sensitive to outliers

3) RMSE: Root Mean Squared Error

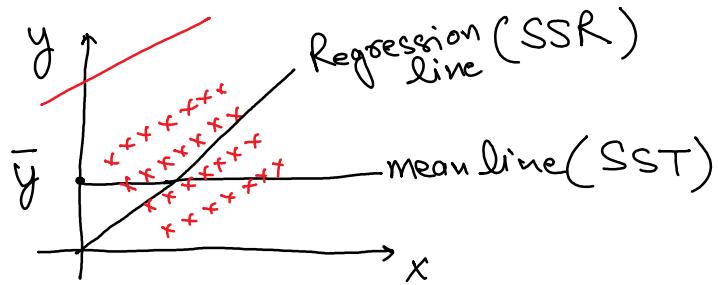
$$RMSE = \sqrt{MSE} = \sqrt{n \cdot \text{sqrt}(MSE)} = \sqrt{\frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{n}}$$

- easily interpretable
- less sensitive to outliers

• MAPE  $\Rightarrow$  Read it!

## Goodness of fit

$\rightarrow R^2$  (r<sub>2</sub>-score)  $\Rightarrow$  Coeff of determination  $\Rightarrow$  tell us how well the model fits the data



formula:

$$R^2 = 1 - \frac{SSR}{SST}$$

Case 1:  $SSR=0, SST=SST$

$$R^2 = 1 - \frac{0}{SST} = 1$$

Case 2:  $SSR=SST$

$$R^2 = 1 - \frac{SSR}{SST} = 1-1=0$$

Case 3:  $SSR > SST$

$\frac{SSR}{SST} = \text{greater than } 1.$

$$R^2 = 1 - \frac{SSR}{SST} (> 1) = -ve$$

Problem with  $R^2 \Rightarrow R^2 \uparrow$  as # columns  $\uparrow$

$$\text{Adjusted } R^2 \Rightarrow \text{Adj. } R^2 = 1 - \left[ \frac{(1-R^2)n(n-1)}{(n-1-p)} \right]$$

where,  
 $n = \# \text{ datapoints}$   
 $R^2 = R^2(\text{r}^2 - \text{score})$   
 $P = \# \text{ columns}$

↓ penalize  
 ↑ increase in columns

if  $(n-1-p)$  decrease  $> (1-R^2)$  decrease

$$\text{Adj. } R^2 \downarrow$$

if  $(1-R^2)$  decrease  $> (n-1-p)$  decrease

$$\text{Adj. } R^2 \uparrow$$

$R^2 \rightarrow 0.7 \& \text{ above}$

$\text{Adj } R^2 \rightarrow 0.7 \& \text{ above}$

Multicollinearity  $\Rightarrow$  Before Modelling

↳ One column is highly correlated with other columns

$$w \Rightarrow \{ \underbrace{f_1, f_2, f_3}_{\downarrow = 1, 2, 3} \}$$

$$\tilde{w} \Rightarrow \{ \underbrace{0, 3.5, 3}_{\cancel{f_2}} \}$$

automatically/arbitrarily changed

$$y = f_1 + 2f_2 + 3f_3$$

$$y = f_1 + \cancel{\frac{2}{1.5} f_1} + 3f_3$$

$$f_1 + 2f_2 + 5f_3$$

$$1.5f_1 + 2f_2 + 2f_3 \xrightarrow{*} 2.5f_1 + 2f_2 + nf_3$$

( $f_2$ )

$$\underline{1.5f_2 + 2f_2 + 3f_3} \xrightarrow{*} 3.5f_2 + 3f_3 + 0f_1$$

output gets disturbed!



Marks IQ  
1 0.7  
0.7 1

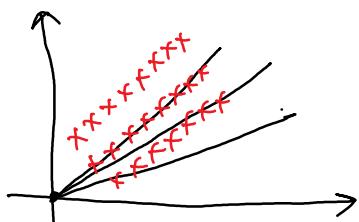
0.70

How to detect it?  $\rightarrow$  correlation matrix  
 $\hookrightarrow$  drop one of the columns.

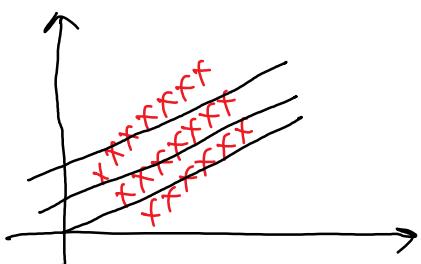
$\rightarrow VIF \Rightarrow$  variance inflation factor  $= \frac{1}{1 - R^2}$   
 $[1, \infty]$

VIF > 5

Gradient Descent



$y = mx + b$  / you can rotate this line with the help of slope.



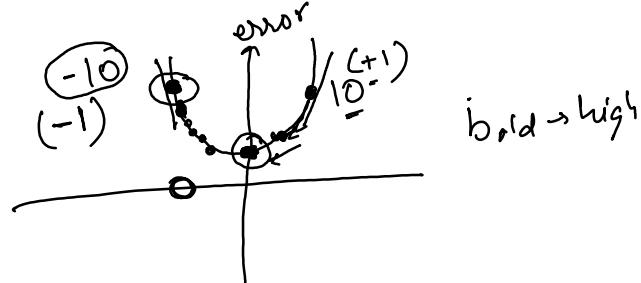
$y = mx + b$  / by changing intercept the line can be moved up or down

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

Steps:  $m = \text{constant}$ ,  $b = \text{variable}$   
 1) take any random value of  $b$

$$2) \frac{\partial L}{\partial b} = \frac{d}{db} (y_i - mx_i - b)^2 = -2(y_i - mx_i - b)$$

$$3) b_{\text{next}} = b_{\text{old}} - \eta \underbrace{\text{slope}}_{(+ve)}$$



$$b_{\text{next}} = b_{\text{old}} + \eta \text{slope}$$

$\eta$  hyperparameter  $\downarrow$  learning rate  $(\eta) \Rightarrow [0.001, 1]$   
 or Step size

effects of learning rate  $\Rightarrow$  greater the learning rate, the faster the gradient descent, but oscillate around min.

$\Rightarrow$  smaller learning rate, the slower the gradient descent, but reaches minima.

Steps:

① take any random value of  $m$  &  $b$

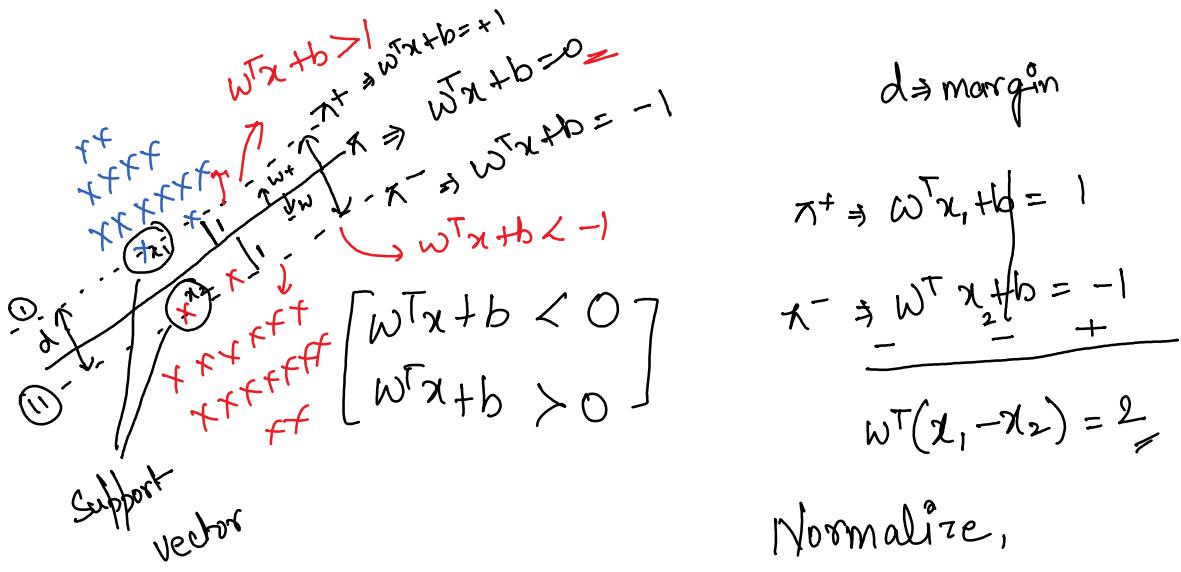
$$\text{② } \frac{\partial L}{\partial b} = -2(y_i - mx_i - b), \quad \frac{\partial L}{\partial m} = -2(y_i - mx_i - b)x_i$$

$$\text{③ } m_{\text{next}} = m_{\text{old}} - \eta \underline{(\frac{\partial L}{\partial m})}$$

(111)

$$b_{\text{next}} = b_{\text{old}} - \eta \left( \frac{\partial L}{\partial b} \right) \quad , \quad m_{\text{next}} = m_{\text{old}} - \eta \left( \frac{\partial L}{\partial m} \right)$$

# Support Vector Machine



Normalize,

$$\frac{w^T(x_1 - x_2)}{\|w\|} = \frac{2}{\|w\|}$$

$$x_1 - x_2 = \frac{2}{\|w\|} = d$$

$$\text{Mof} \Rightarrow f(x) = \arg \max \frac{2}{\|w\|}$$

for each datapoint in negative zone ,

$$w^T x + b < 0 - @$$

for each datapoint in positive zone ,

$$\omega^T x + b > 0 \rightarrow \textcircled{b}$$

Simplifying:  $y_i(\omega^T x_i + b) \Rightarrow$  used for prediction.

Case 1:  $y_i = +ve, \omega^T x_i + b > 0$       Case 2:  $y_i = -ve, \omega^T x_i + b < 0$

$y_i(\omega^T x_i + b) > 0$   $\downarrow$        $y_i(\omega^T x_i + b) < 0$   $\downarrow$

Correct classifications

Case 3:  $y_i = -ve, \omega^T x_i + b > 0$       Case 4:  $y_i = +ve, \omega^T x_i + b < 0$

$y_i(\omega^T x_i + b) < 0$   $\swarrow$        $y_i(\omega^T x_i + b) < 0$   $\swarrow$

Incorrect classification

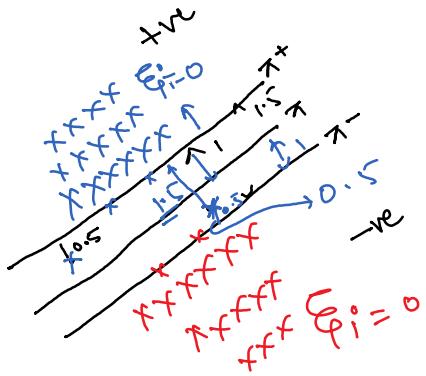
for correct classification,  $y_i(\omega^T x_i + b) \geq 1$   
(w.r.t SVM)

$$MDF \Rightarrow \arg \max \frac{2}{\|\omega\|}$$

$$MDF \Rightarrow f(x) = \arg \max \frac{2}{\|\omega\|} \text{ such that } y_i(\omega^T x_i + b) \geq 1$$

↳ Hard margin SVM

### Soft Margin SVM



$$y_i(w^T x_i + b) = 0.5$$

$$y_i(w^T x_i + b) = -0.5$$

$$\hookrightarrow y_i(w^T x_i + b) = 1 - \frac{1.5}{\epsilon_i}$$

$\epsilon_i$ : To measure how far a datapoint is in opposite direction from the right plane

$$\text{MoF} \Rightarrow f(x) = \underset{w, b}{\operatorname{argmax}} \frac{2}{\|w\|} + C \underbrace{\sum_{i=1}^n \epsilon_i}_{\text{loss}}$$

hyperparameter

$$\text{MoF} \Rightarrow f(x) = \underset{w, b}{\operatorname{argmin}} \frac{\|w\|}{2} + C \underbrace{\sum_{i=1}^n \epsilon_i}_{\text{loss}}$$

regularizer

$C \uparrow \rightarrow$  more focus on  $\epsilon_i \rightarrow$  less error  $\rightarrow$  overfitting  
 $C \downarrow \rightarrow$  less " " "  $\rightarrow$  more "  $\rightarrow$  underfitting

$\lambda \uparrow \rightarrow$  more errors  $\rightarrow$  underfitting

$\lambda \downarrow \rightarrow$  less error  $\rightarrow$  overfitting

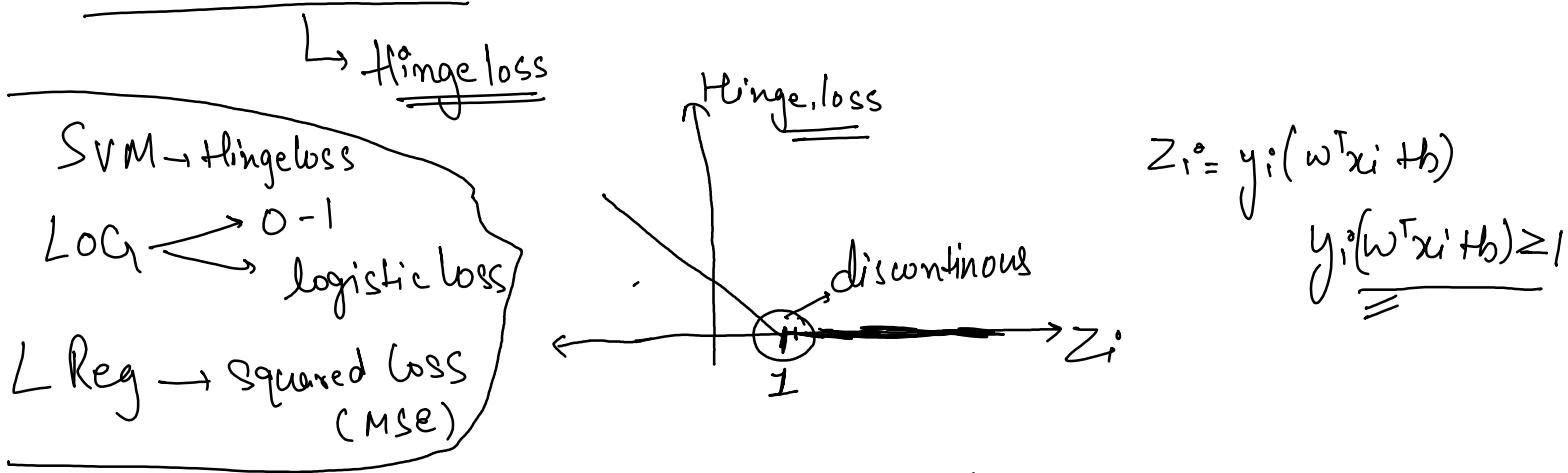
Don't ever use this

$$f(x) = \sum_{i=1}^n \epsilon_i + \frac{\lambda \|w\|}{2}$$

$C \propto \frac{1}{\lambda}$

Hyperparameter.

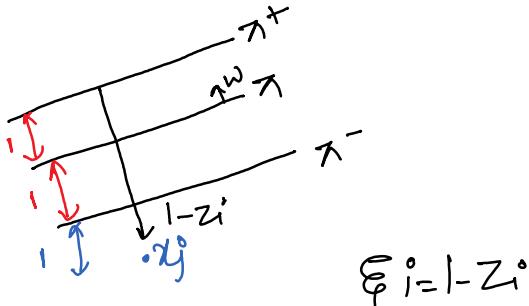
### Loss Minimization:



$x_i^* \rightarrow \epsilon_i = 0$       Hinge loss =  $\max(0, 1 - z_i^*)$

$z_i^* = y_i(w^\top x_i + b)$

$z_i^* \geq 1$



$\epsilon_i^* = 1 - z_i^*$

$y_i(w^\top x_i + b) = -2$

$$\epsilon_i = 3$$

$$z_i^{\circ} = -2$$

$$\epsilon_i^{\circ} = 1 - (-2) = 3$$

① for correct classification,  $z_i \geq 1$ ,  $\epsilon_i^{\circ} = 2$  (assume)

$$\text{Hinge loss} = \max(0, 1 - z_i^{\circ}) \Rightarrow \max(0, 1 - 2) \\ \Rightarrow \max(0, -1) = 0$$

② for incorrect classification,  $\epsilon_i^{\circ} = 1 - z_i^{\circ}$ ,  $\epsilon_i^{\circ} > 1$

$$\text{Hinge loss} = \max(0, 1 - z_i^{\circ}) = \max(0, \epsilon_i^{\circ}) = \max(0, 3) \\ = 3$$

Dual form of SVM:

$$\text{Primal: } \underset{w, b}{\operatorname{argmin}} \frac{\|w\|}{2} + C \sum_{i=1}^n \epsilon_i^{\circ}$$

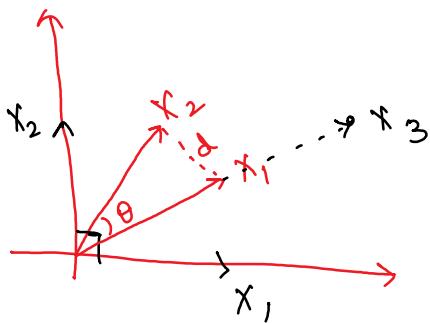
$$\text{Dual form: } \underset{\alpha_i^{\circ}}{\max} \sum_{i=1}^n \alpha_i^{\circ} - \frac{1}{2} \sum_{i=1}^n \sum_{j=1}^n \alpha_i^{\circ} \alpha_j^{\circ} y_i y_j (x_i^T x_j) \quad \text{Labeled "similarity"}$$

$\alpha_i^{\circ} \Rightarrow$  denote support vectors       $\alpha_i^{\circ} > 0$  (for support vectors)

$(x_i^T x_j^{\circ}) \Rightarrow$  how similar two points are !

Cosine similarity:

$$\underline{(\cos\theta)}$$

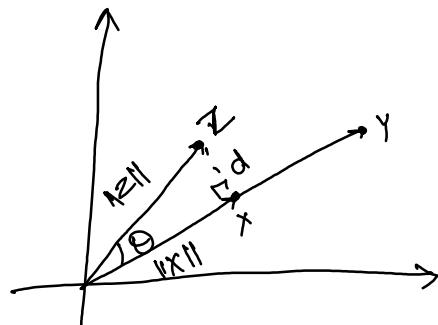


$$\cos\theta = \frac{\mathbf{x}_1 \cdot \mathbf{x}_2}{\|\mathbf{x}_1\| \|\mathbf{x}_2\|}$$

$$\theta = 90^\circ, \cos\theta = 0, \text{ cosine distance } \uparrow$$



$$\theta = 0^\circ, \cos\theta = 1, \text{ cosine distance } = 0 \downarrow$$



(x, z)

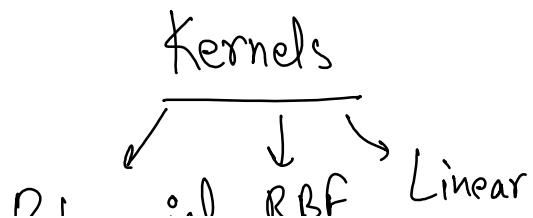
$$\theta = 30^\circ, \text{ cosine similarity} = \frac{\sqrt{3}}{2}, \text{ cosine distance } \uparrow$$

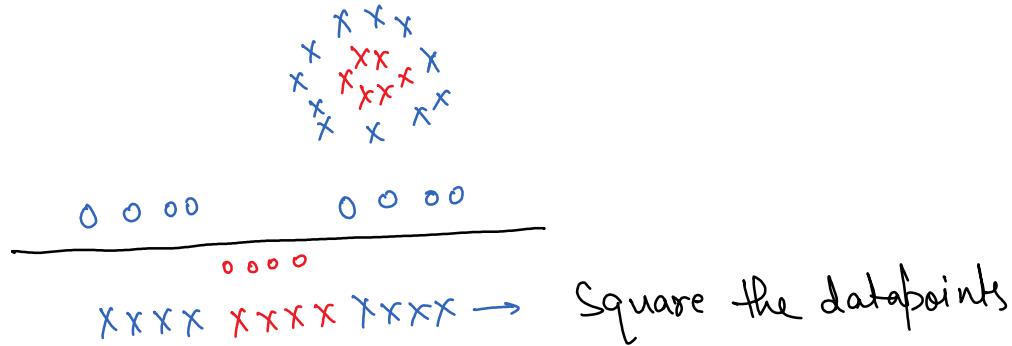
(x, y)

$$\theta = 0^\circ, \text{ similarity} = \cos\theta = 1$$

Angle b/w  
two vectors  $\cos\theta = \frac{\vec{x}_1 \cdot \vec{x}_2}{\|\mathbf{x}_1\| \|\mathbf{x}_2\|}$ , lets say  $\|\mathbf{x}_1\| = \|\mathbf{x}_2\| = 1$

$$\cos\theta = \vec{x}_1 \cdot \vec{x}_2 \xrightarrow{\leftarrow A} \mathbf{x}_1^T \cdot \mathbf{x}_2$$





"Kernel trick": applying kernels to transform datapoints to make them linearly separable.

$$\text{Polynomial Kernel: } K(x_1, x_2) = (x_1^T x_2 + c)^d$$

Quadratic function,  $d=2$

$$ax^2 + bx + c$$

$$x_1 = \begin{bmatrix} x_{11} & x_{12} \end{bmatrix}$$

$$x_2 = \begin{bmatrix} x_{21} & x_{22} \end{bmatrix} \rightarrow 2d$$

$$(1 + x_1^T x_2)^2$$

$$\Rightarrow (1 + \begin{bmatrix} x_{11} & x_{12} \end{bmatrix} \begin{bmatrix} x_{21} \\ x_{22} \end{bmatrix})^2$$

$$\Rightarrow [1 + (x_{11}x_{21} + x_{12}x_{22})]^2$$

$$\Rightarrow [1 + x_{11}x_{21} + x_{12}x_{22}]^2 \Rightarrow (a+b+c)^2 = a^2 + b^2 + c^2 + 2ab + 2bc + 2ca$$

$$\Rightarrow [1^2 + x_{11}^2 x_{21}^2 + x_{12}^2 x_{22}^2 + 2x_{11}x_{21}x_{12}x_{22} + 2x_{11}x_{21} + 2x_{12}x_{22}]$$

$x_{11}^2 \quad x_{21}^2 \quad x_{12}^2 \quad x_{22}^2 \quad \sqrt{2}x_{11}x_{12} \quad \sqrt{2}x_{21}x_{22} \quad \sqrt{2}x_{11}, \sqrt{2}x_{21}, \sqrt{2}x_{12}, \sqrt{2}x_{22}$

2d  $x_1 \rightarrow x_1' (6d)$        $x_2 \xrightarrow{(2d)} x_2' \quad (\text{co-ordinates})$

$(1, x_{11}^2, x_{12}^2, \sqrt{2}x_{11}, \sqrt{2}x_{12}, \sqrt{2}x_{11}x_{12})$        $(1, x_{21}^2, x_{22}^2, \sqrt{2}x_{21}, \sqrt{2}x_{22}, \sqrt{2}x_{21}x_{22})$

6d      6d

Mercer's Theorem: Kernel converts the d-dim dataset into (Kernel trick) d' dim dataset such  $d' > d$ .

RBF (Radial Basis function)

$$k(x_1, x_2) = e^{\frac{-||x_1 - x_2||^2}{2\sigma^2}} = e^{-\frac{d^2}{2\sigma^2}}$$

Case 1:  $||x_1 - x_2|| = d \rightarrow \text{distance}$

$$\downarrow k = \frac{1}{12\sigma^2} \quad | \quad d^2 \rightarrow d^{2\sigma^2} \rightarrow e^{-\frac{d^2}{2\sigma^2}}$$

$$\downarrow K = \frac{1}{e^{-\frac{d^2}{2\sigma^2}}} \quad d^2 \rightarrow d^2 \rightarrow e^{-\frac{d^2}{2\sigma^2}}$$

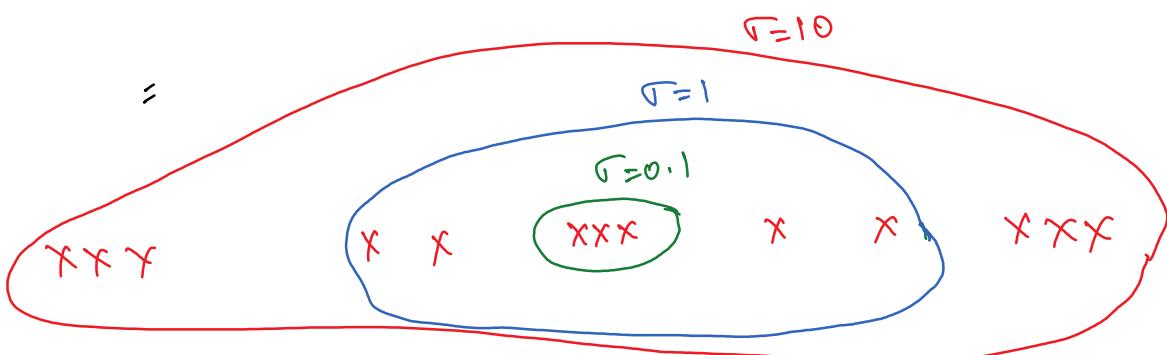
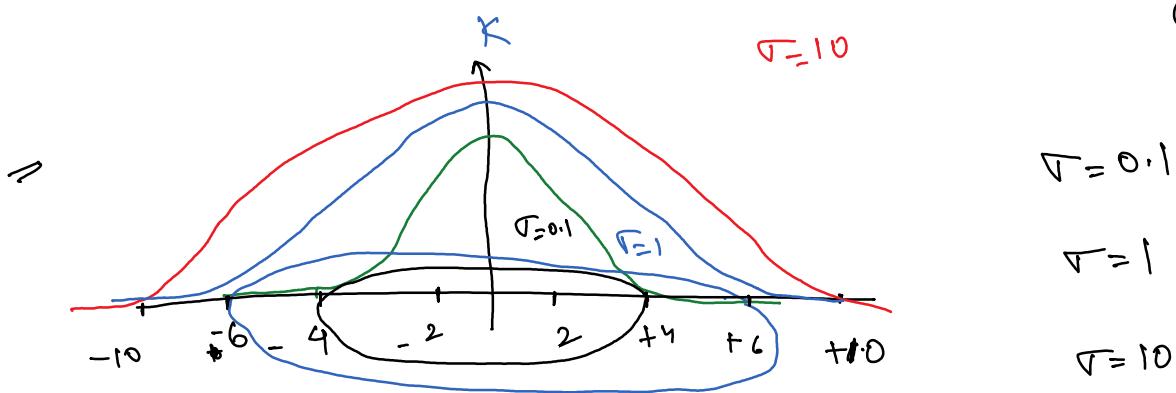
$$d=2 \rightarrow d^2 \rightarrow 4 \Rightarrow \frac{1}{e^{4\sigma^2}} \Rightarrow \frac{1}{e^4} \downarrow = k \downarrow$$

Case 2°  $\uparrow K = \frac{1}{e^{-\frac{d^2}{2\sigma^2}}} \uparrow$

$\uparrow \sigma \rightarrow \sigma^2 \rightarrow \frac{d^2}{2\sigma^2} \downarrow$

$\sigma = 2 \quad \sigma = 3$   
 $e^{-\frac{1}{2\sigma^2}} \quad e^{-\frac{1}{2\sigma^2}}$   
 $e^{-\frac{1}{8}} \approx e^{-\frac{1}{18}}$

$\frac{1}{e^{d^2/2\sigma^2}} \uparrow$



Sklearn documentation  $\Rightarrow \gamma$

$$\rho^{-\frac{d^2}{2\sigma^2}} \Rightarrow \rho^{-d^2\gamma}$$

Sklearn documentation  $\Rightarrow$   $\gamma$

$$e^{-\frac{d^2}{2\sigma^2}} \Rightarrow e^{-d^2\gamma}$$

$$\gamma = \frac{1}{2\sigma^2}$$

\*  $\gamma \propto$  similarity  $\propto \frac{1}{r}$

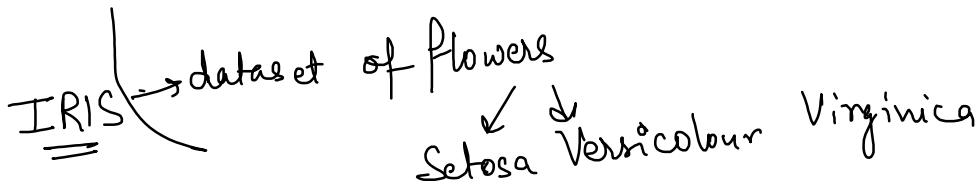
$$* K \propto \frac{1}{d}$$

$$\Rightarrow \gamma \uparrow \rightarrow r^2 \uparrow \rightarrow \frac{1}{2\sigma^2} \rightarrow \frac{1}{2\sigma^2} \left. \right\} \rightarrow r \downarrow \rightarrow \text{sim} \uparrow$$

\* Use RBF, if you have no idea which kernel to use!

## DECISION TREES AND RF

Saturday, October 21, 2023 7:56 AM



$\Rightarrow$  G/P  $\Rightarrow$  Sepal length, petal length, sepal width, petal width  
 $(SL)$                    $(PL)$                    $(SW)$                    $(PW)$

If else block:

if  $PL < a$ :

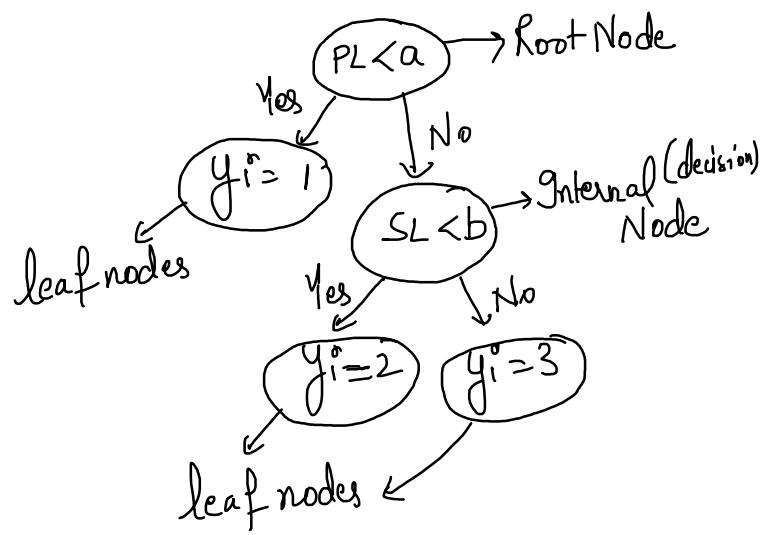
$$y_i^o = S \quad (1)$$

else: if  $SL < b$ :

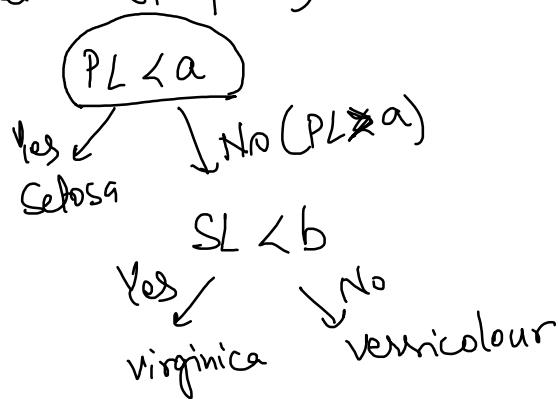
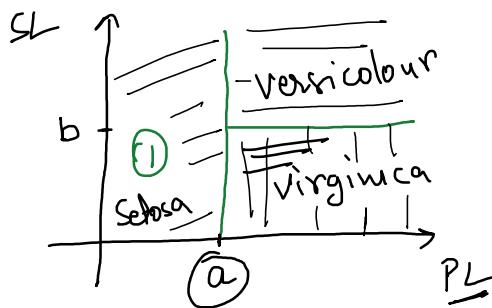
$$y_i^o = \text{Vernicolour} \quad (2)$$

else:

$$y_i^o = \text{Virginica} \quad (3)$$

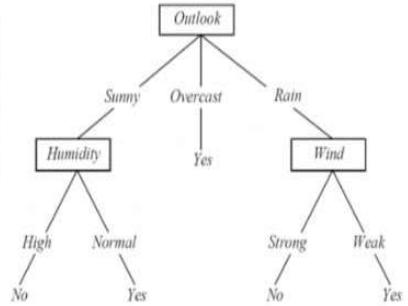


Recursive Partitioning: (axis parallel hyperplane)



DT  
 ↗ Entropy  
 ↗ Gini Impurity  
 ↗ Information Gain

| Outlook  | Temperature | Humidity | Windy | PlayTennis |
|----------|-------------|----------|-------|------------|
| Sunny    | Hot         | High     | False | No         |
| Sunny    | Hot         | High     | True  | No         |
| Overcast | Hot         | High     | False | Yes        |
| Rainy    | Mild        | High     | False | Yes        |
| Rainy    | Cool        | Normal   | False | Yes        |
| Rainy    | Cool        | Normal   | True  | No         |
| Overcast | Cool        | Normal   | True  | Yes        |
| Sunny    | Mild        | High     | False | No         |
| Sunny    | Cool        | Normal   | False | Yes        |
| Rainy    | Mild        | Normal   | False | Yes        |
| Sunny    | Mild        | Normal   | True  | Yes        |
| Overcast | Mild        | High     | True  | Yes        |
| Overcast | Hot         | Normal   | False | Yes        |
| Rainy    | Mild        | High     | True  | No         |



Entropy → Randomness

$$\hookrightarrow H_D(Y) = - \sum_{i=1}^n p_i \lg(p_i)$$

$\hookrightarrow \lg = \log_2$

Parent's Entropy

$$Y = 9$$

$$N = 5$$

$$P(Y) = \frac{9}{14} \quad P(N) = \frac{5}{14}$$

$$H_D(Y) = -P(Y) \lg(P(Y)) - P(N) \lg(P(N))$$

$$= -\frac{9}{14} \times \lg\left(\frac{9}{14}\right) - \frac{5}{14} \times \lg\left(\frac{5}{14}\right)$$

$$= 0.94$$

entropy of each column:

$$H_D(Y) = 0.94$$

|         |  |
|---------|--|
| outlook | 5 (2Y, 3N) → $-\frac{2}{5} \lg\left(\frac{2}{5}\right) - \frac{3}{5} \lg\left(\frac{3}{5}\right) = 0.97$<br>4 (4Y, 0N) → $-\frac{4}{4} \lg\left(\frac{4}{4}\right) - 0 \lg(0) = 0$<br>5 (3Y, 2N) → $-\frac{3}{5} \lg\left(\frac{3}{5}\right) - \frac{2}{5} \lg\left(\frac{2}{5}\right) = 0.97$ |
|---------|--|

$$\text{Weighted entropy} = \frac{5}{14} \times 0.97 + \frac{4}{14} \times 0 + \frac{5}{14} \times 0.97$$

[ $H_D(Y, \text{outlook})$ ]

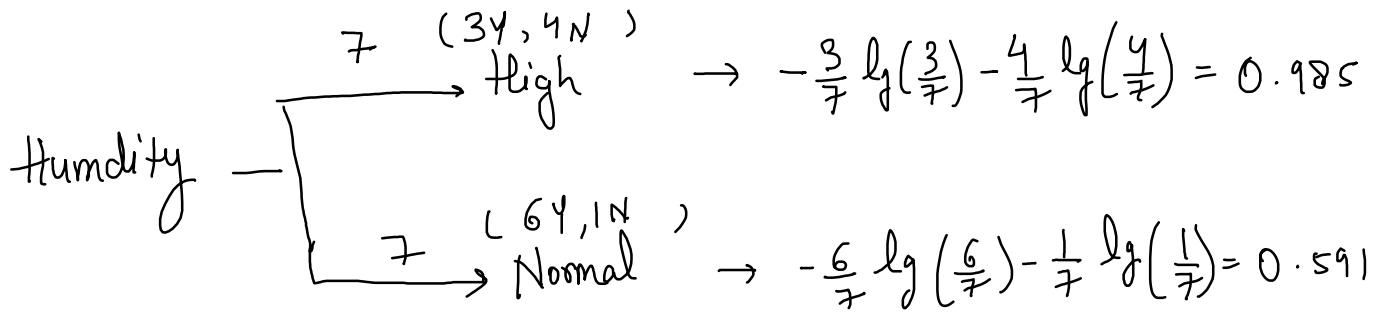
$$= 0.97 \left( \frac{5}{14} + \frac{5}{14} \right) = 0.97 \times \frac{5}{7} = 0.69$$

|             |   |
|-------------|---|
| Temperature | 4 (2Y, 2N) → $-\frac{2}{4} \lg\left(\frac{2}{4}\right) - \frac{2}{4} \lg\left(\frac{2}{4}\right) = 1$<br>6 (4Y, 2N) → $-\frac{4}{6} \lg\left(\frac{4}{6}\right) - \frac{2}{6} \lg\left(\frac{2}{6}\right) = 0.918$<br>4 (3Y, 1N) → $-\frac{3}{4} \lg\left(\frac{3}{4}\right) - \frac{1}{4} \lg\left(\frac{1}{4}\right) = 0.811$ |
|-------------|---|

$$H_D(Y, \text{temperature}) = \frac{4}{14} \times 1 + \frac{6}{14} \times 0.918 + \frac{4}{14} \times 0.811$$

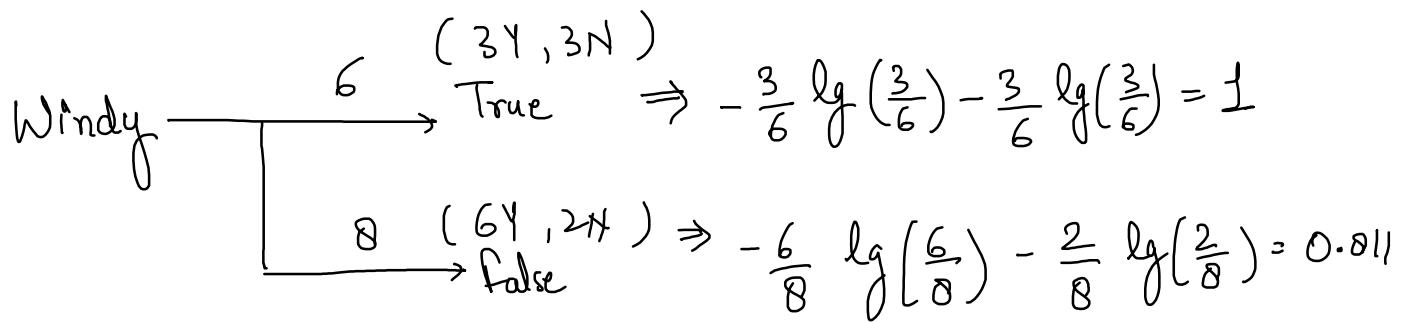
$$\Rightarrow 0.91$$

$$7 (3Y, 4N) \rightarrow -3 \lg(3) - 4 \lg(4) = \text{unclear}$$



$$H_D(Y, \text{humidity}) = \frac{7}{14} \times 0.985 + \frac{7}{14} \times 0.591$$

$$= \frac{1}{2}(0.985 + 0.591) = 0.788$$



$$H_D(Y, \text{windy}) = \frac{6}{14} \times 1 + \frac{8}{14} \times 0.811 = 0.89$$

2 ways to choose the column for split!

1) choose column with lowest weighted entropy -

outlook, temperature, humidity, windy

$\leftarrow (0.69)$     0.91,    0.78, 0.89

0.69,      0.91,      0.78, 0.89  
 best split ←

2) Information Gain  $\Rightarrow IG(Y) = H_D(Y) - \text{weighted entropy of features}$

$$\text{outlook} \Rightarrow IG_O(Y) = 0.94 - 0.69 = 0.25$$

$$\text{temperature} \Rightarrow IG_T(Y) = 0.94 - 0.91 = 0.03$$

$$\text{humidity} \Rightarrow IG_H(Y) = 0.94 - 0.78 = 0.16$$

$$\text{windy} \Rightarrow IG_W(Y) = 0.94 - 0.89 = 0.05$$

Since, outlook has highest Information Gain, we will be choosing outlook for first split.

### \* Properties of Entropy:

$$H_D(Y) = -P(y_+) \lg P(y_+) - P(y_-) \lg P(y_-)$$

$$\underline{\text{Case 1:}} \quad P(y_+) = 0.99 \quad P(y_-) = 0.01$$

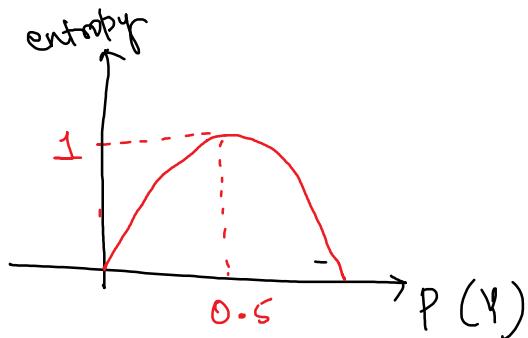
$$H_D(Y) = -0.99 \lg(0.99) - 0.01 \lg(0.01) = 0.08$$

$$\underline{\text{Case 2}}: \quad P(y_+) = 0.5 \quad P(y_-) = 0.5$$

$$H_D(Y) = -0.5 \lg(0.5) - 0.5 \lg(0.5) = 1$$

$$\underline{\text{Case 3}}: \quad P(y_+) = 0 \quad P(y_-) = 1$$

$$H_D(Y) = -0 \lg(0) - 1 \lg(1) = 0$$



Gini Impurity

$$I_G(Y) = 1 - \sum_{i=1}^n (p_i)^2$$

for  $n$ -class,

$$I_G(Y) = 1 - [p(y_1)^2 + p(y_2)^2 + p(y_3)^2 + \dots + p(y_n)^2]$$

for binary,

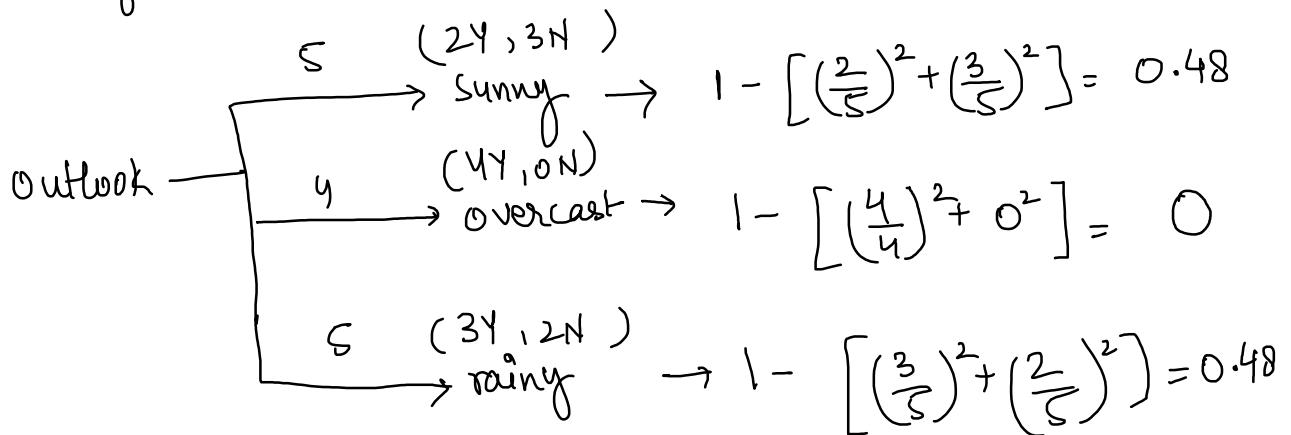
$$I_G(Y) = 1 - [p(y_1)^2 + p(y_2)^2]$$

Parent Gini Impurity :  $P(Y) = 9/14$        $P(N) = 5/14$

$$I_G(Y) = 1 - \left[ P(Y)^2 + P(N)^2 \right] = 1 - \left[ \left( \frac{9}{14} \right)^2 + \left( \frac{5}{14} \right)^2 \right]$$

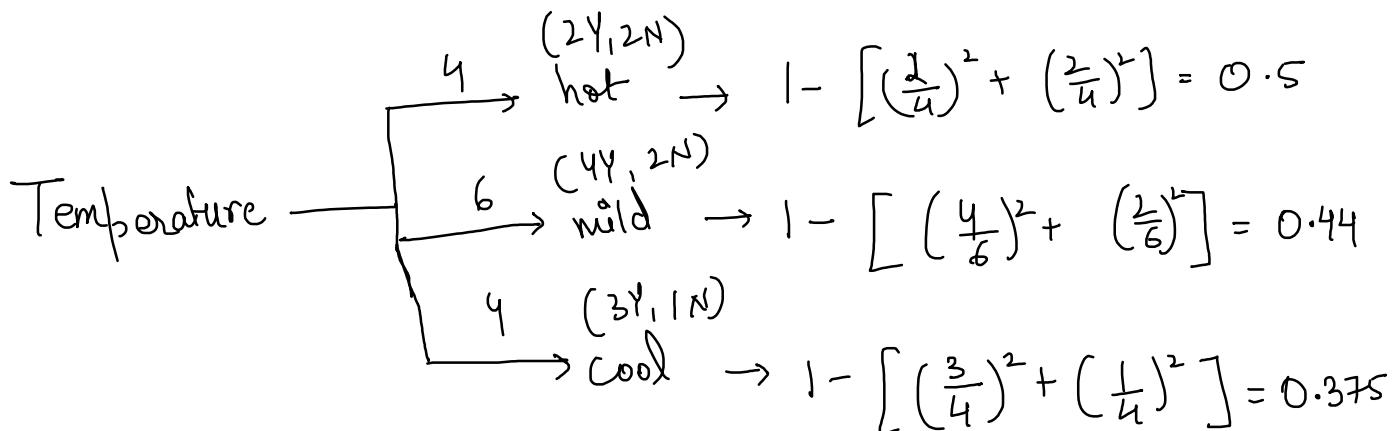
$$I_G(Y) = 0.459$$

Gini Impurity for features:

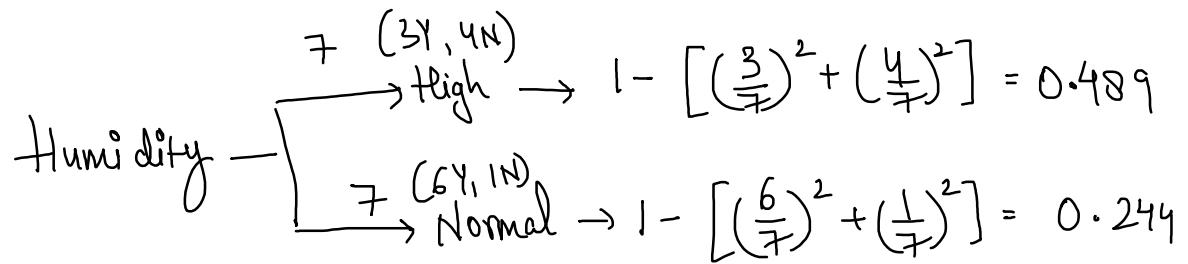


Weighted gini impurity,  $I_G(O) = \frac{5}{14} \times 0.48 + \frac{4}{14} \times 0 + \frac{5}{14} \times 0.48$

$$= 0.342$$

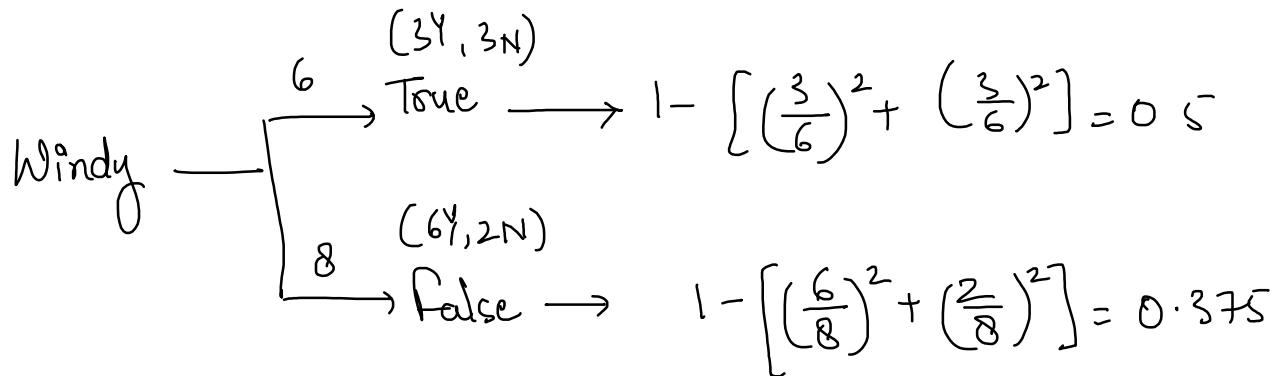


$$I_G(T) = \frac{4}{14} \times 0.5 + \frac{6}{14} \times 0.44 + \frac{4}{14} \times 0.375 = 0.44$$



$$I_G(H) = \frac{7}{14} \times 0.489 + \frac{7}{14} \times 0.244$$

$$= 0.367$$

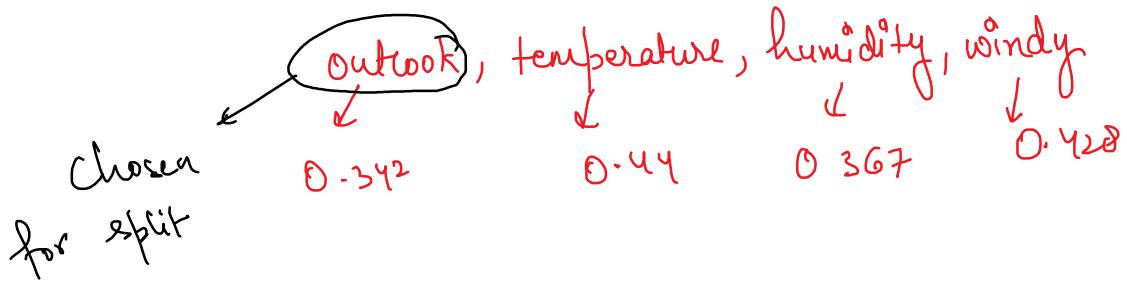


$$I_G(W) = \frac{6}{14} \times 0.5 + \frac{8}{14} \times 0.375$$

$$= 0.428$$

Two ways to choose column for split:

1) Choose column with lowest Gini Impurity



2) Information Gain =  $I_G(Y)$  - weighted Gini Impurity

$$\text{outlook, } I_G(Y) = 0.459 - 0.342 = 0.117$$

$$\text{temperature, } I_G(Y) = 0.459 - 0.44 = 0.015$$

$$\text{humidity, } I_G(Y) = 0.459 - 0.367 = 0.092$$

$$\text{windy, } I_G(Y) = 0.459 - 0.428 = 0.031$$

outlook is chosen, as it has highest information gain

Properties of Gini Impurity

$$\underline{\text{Case 1: }} P(Y+) = 0.5$$

$$I_G(Y) = 1 - [0.5^2 + 0.5^2] = 0.5$$

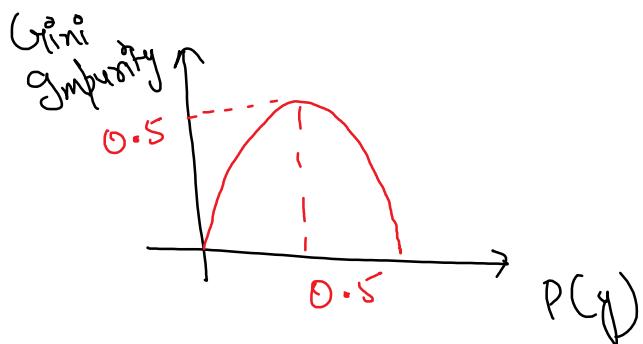
$$P(Y+) = 0.5$$

Case 2:

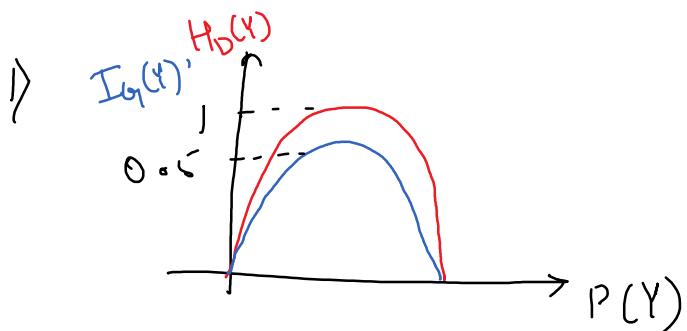
$$P(Y+) = 1$$

$$P(Y-) = 0$$

$$IG(Y) = 1 - [1^2 + 0^2] = 0$$



Comparison b/w Gini Impurity & Entropy:



2) Computation cost :

entropy is harder to calculate, higher computational cost  
 Gini Impurity " easier to calculate, lower computational cost

for large datasets  $\rightarrow$  use Gini Impurity

## Hyperparameters:

1 → "criterion" → gini, impurity, entropy

2 → max\_depth →

max\_depth ↑ → overfitting chances ↑

max\_depth ↓ → underfitting chances ↑

Thursday, December 28, 2023 8:30 AM



## BAGGING AND BOOSTING

Sunday, October 22, 2023 8:32 AM

### Ensembles

group of musicians

In m/c learning

group of models

### Ensemble

#### Bagging

Random Forest

#### Boosting

Gradient  
Boosting

Adaboost

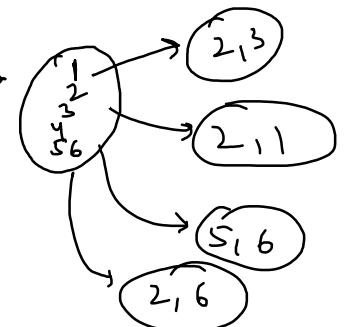
↳ image classification

↳ internet applications

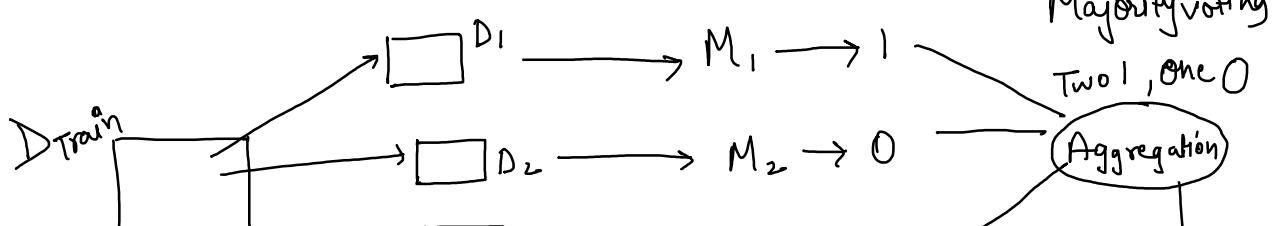
### BAGGING

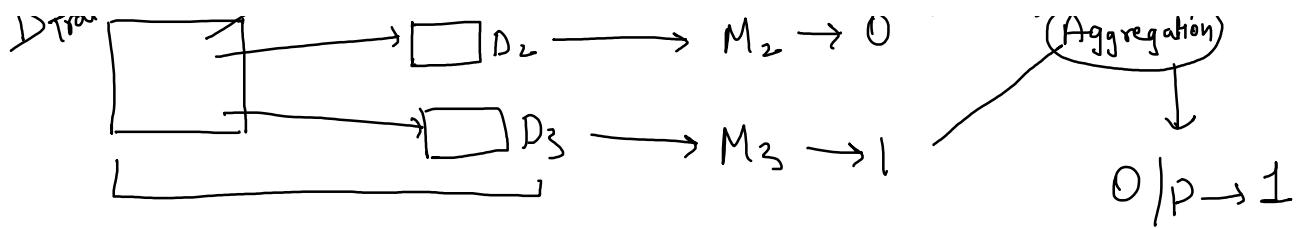
(statistical) Bootstrapping  
term

Aggregation  
Sampling with replacement



### Bagging Flow



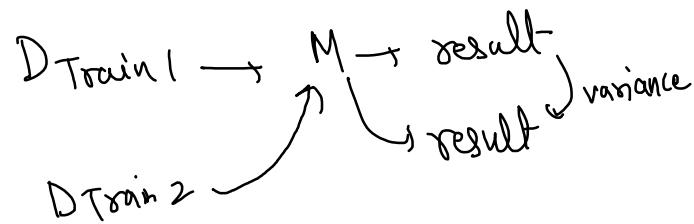


Regression: Mean/Median instead of voting

When to use Bagging?

- Bagging is used when you have low bias & high variance  
 ↴  
Overfitting

bias → underfitting  
 variance → overfitting      } explore!



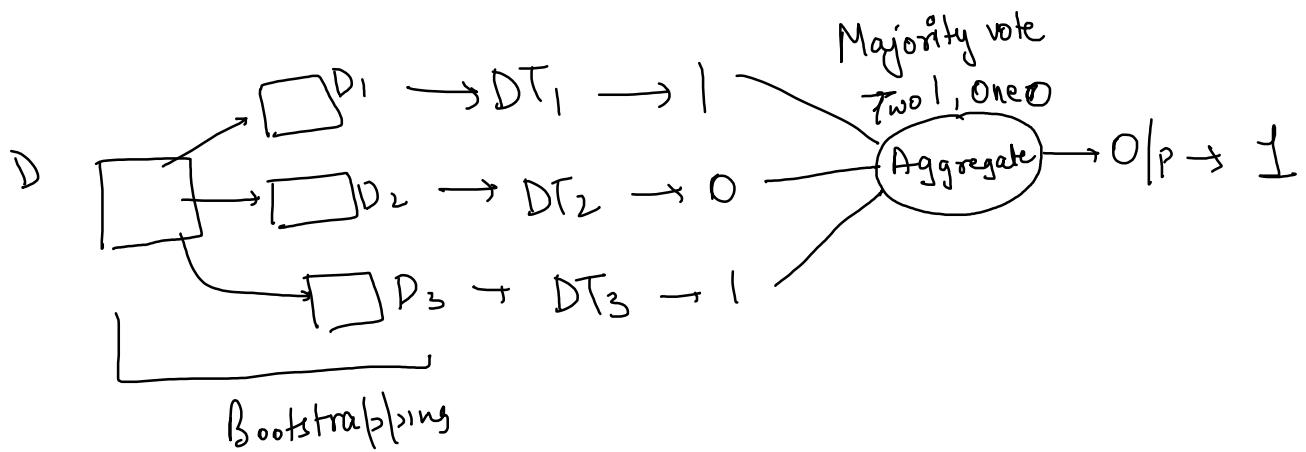
Random Forest → (group of trees) → Decision Trees

Q How do you create a situation of low bias & high variance with the help of DT?  
 ↓  
 overfitting

Sol: Create a DT of reasonable depth

DT → max-depth ↑ → height of tree ↑ → overfitting ↑

Flow



\* Models should be very different from each other

FF = low bias & high variance + row sampling + Column Sampling  
 ↓ → max-features  
 DT + Aggregation

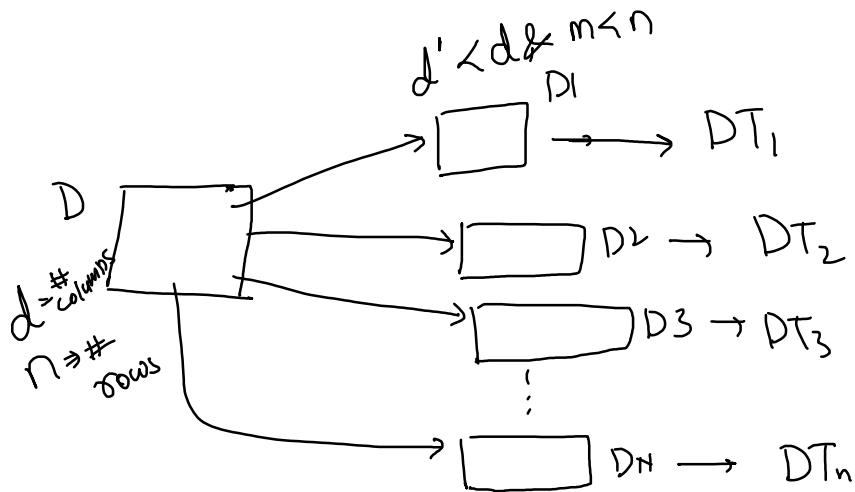
|   | GPA | IQ  | Extra-Cur. | Social | Placed |
|---|-----|-----|------------|--------|--------|
| • | 7   | 110 | 10         | 9      | 1      |
|   | 8   | 112 | 9          | 8      | 0      |
|   | 9   | 120 | 8          | 7      | 0      |
|   | 10  | 135 | 7          | 6      | 1      |

D<sub>1</sub>

D<sub>2</sub>

O/p

| <u>D<sub>1</sub></u> | <u>D<sub>2</sub></u> | <u>D<sub>3</sub></u> | <u>D<sub>4</sub></u> |
|----------------------|----------------------|----------------------|----------------------|
| C_GPA                | Extra-Curricular     | O/P                  | IQ                   |
| 10                   | 8                    | 0                    | 135                  |
| 9                    | 7                    | 1                    | 120                  |
| 7                    | 10                   | 1                    | 110                  |
|                      |                      |                      | 9 1                  |



$DT_1, DT_2, DT_3, \dots, DT_n$  will be different because we are feeding

different samples to them  
(as i/p)

i/p  $\Rightarrow$  input-

o/p  $\Rightarrow$  output

B/w  $\Rightarrow$  between

#  $\Rightarrow$  number of

### Hyperparameters:

$\Rightarrow$  # models  $\Rightarrow$  n\_estimator  $\Rightarrow [50 - 2000]$

$\Rightarrow$  Row sampling rate  $\Rightarrow \frac{m}{n}$

$\uparrow \propto$  variance  $\uparrow$

$\Rightarrow$  Low sampling rate  $\Rightarrow \frac{m}{n}$



$\Rightarrow$   $C / V_i$   $\rightarrow$   $V_i$

$\Rightarrow$   $\times$  variance ↑  
(Overfitting)

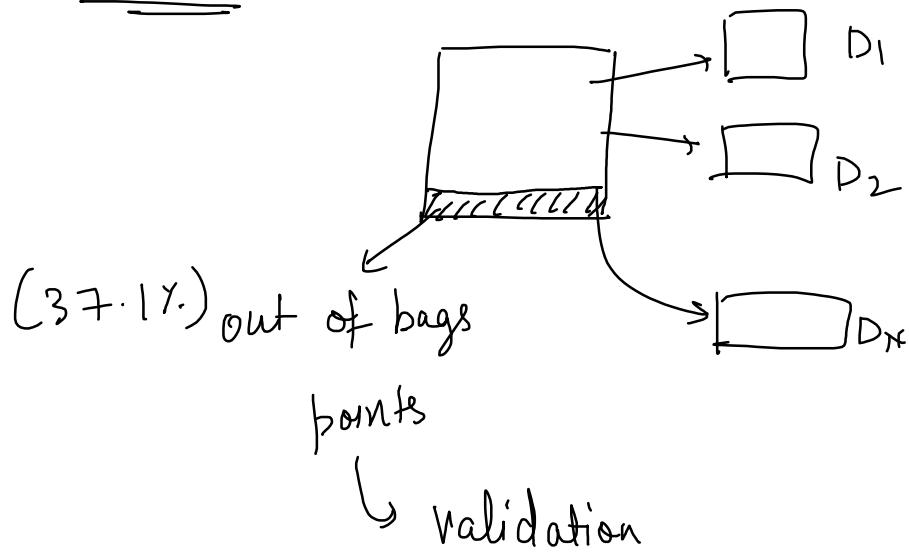
$\Rightarrow$  Column-sampling rate  $\Rightarrow \frac{d'}{d}$

⇒ max-features

$\Rightarrow$  max-depth

$$\Rightarrow n_{\text{jobs}} = -1$$

Oob Score



model = RandomforestClassifier(oob\_score=True)

`oob_score` → high, RF model is good  
→ low, RF . . . P 1

↳ low, RF "bad"

### Advantages:

→ Feature importance :

100 columns  
↓

20 to 30 columns

↳ 80% of info

### Disadvantages:

1 → Black box

2 → no loss function

Boosting  
= Sequential

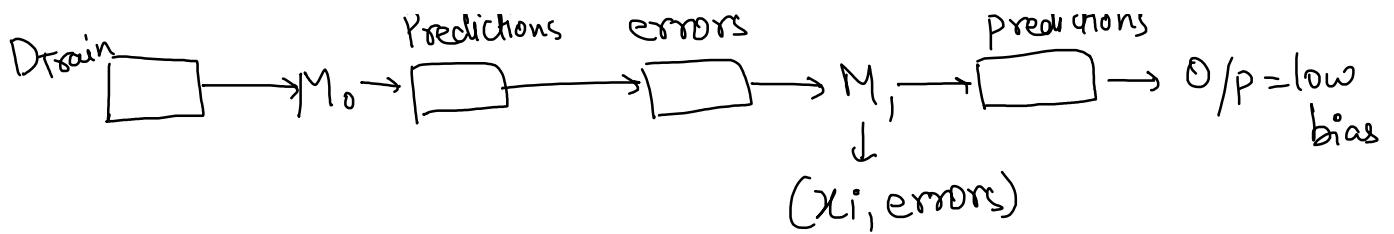
1) Bagging  $\Rightarrow$  low bias & high variance

2) Boosting  $\Rightarrow$  high bias & low variance  $\Rightarrow$  DT  $\Rightarrow$  Decision Stump  
 $\downarrow$

DT of depth = 1

### Flowchart:





Steps

0)  $D_{\text{Train}} = \{x_i^i, y_i^i\}_{i=1}^n \rightarrow M_0 \rightarrow \text{predictions} \rightarrow \text{errors}$

$\hookrightarrow [h_0(x)] \quad \downarrow$   
 mathematical fn  $y_i - \hat{y}_i$   
 $\Downarrow$

$e_i^0 = y_i^i - h_0(x) = \text{errors}$

1)  $M_1 \rightarrow \{x_i^i, e_i^0\}_{i=1}^n \rightarrow \text{errors } [e_i^0 = y_i^i - h_0(x)]$

$\underbrace{h_1(x)}_{\text{predictions}}$

Model at end of first stage:

$\underbrace{f_1(x)}_{\text{predictions}} = x_0 h_0(x) + x_1 h_1(x)$

Boosting  $\Rightarrow$  high bias + additive combine

2)  $M_2 \rightarrow \{x_i, e_i^0\}$   $e_i^0 = y_i^i - f_1(x)$

$\underbrace{h_2(x)}_{\text{predictions}}$

result at end of 2nd stage

$\underbrace{f_2(x)}_{\text{predictions}} = x_0 h_0(x) + x_1 h_1(x) + x_2 h_2(x)$

for  $k^{\text{th}}$  stage

$$f_k(x) = \sum_{i=0}^k \alpha_i h_i(x)$$

$\Rightarrow k = \# \text{ models}$

~~repeat~~

Residuals & loss functions:

$$L[y_i, f_k(x)] = [y_i - f_k(x)]^2$$

$\downarrow$   
 $\hat{y}_i$

$$\frac{\partial L(y_i, f_k(x))}{\partial f_k(x)} = -2[y_i - f_k(x)]$$

$$\leftarrow \frac{-\partial L(y_i, f_k(x))}{\partial f_k(x)} = y_i - f_k(x) = \text{error}$$

negative  
gradient  
↓  
pseudo-residuals

$$(x_i, e_i)^T$$

$$(x_i, \frac{-\partial L(y_i, f_k(x))}{\partial f_k(x)})$$

LgR, LR, SVM

## Gradient Boosting

$\mathcal{G}_P \Rightarrow \{(x_i^*, y_i^*)\}_{i=1}^n + \text{differentiable loss function } L(y_i, f_k(x_i))$

$$0) f_0 = \underset{\gamma}{\operatorname{argmin}} \sum_{i=1}^n L(y_i, \gamma) \quad \rightarrow \gamma = \bar{y}^*$$

▷ for  $m=1$  to  $M$

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

residuals for 1<sup>st</sup> stage  $m=1$

$$r_1 = - \frac{\partial L(y_i, f_0(x_i))}{\partial f_0(x_i)}$$

2)  $f_m(x)$  that can fit on pseudo-residuals, train with  $\{x_i, r_m\}$

$$3) \gamma_m = \underset{\gamma}{\operatorname{argmin}} L \left[ y_i, \underbrace{f_{m-1}(x_i)}_{\text{variable}} + \underbrace{r_m f_m(x_i)}_T \right]$$

$$3) \gamma_m = \underset{\gamma}{\operatorname{argmin}} \ L \left[ y_i, \underbrace{f_{m-1}(x_i^*)}_{\text{constant}} + \underbrace{\gamma f_m(x_i)}_{\text{constant}} \right]$$

$$4) f_m = f_{m-1}(x) + \gamma_m \underline{h_m(x)}$$

New prediction = old prediction + models (combined additively)

hyperparameter:  $m \Rightarrow \# \text{models}$

$m \uparrow \propto \text{bias} \downarrow \propto \text{variance} \uparrow$

$$\underline{\text{Shrinkage}}: f_m(x) = f_{m-1}(x) + \gamma f_m(x)$$

$\gamma$  learning rate  
 $0 < \gamma < 1$

$\gamma$  reduces  $\gamma_m$  which in turns reduces overfitting

$$m = DT$$

GBDT  $\rightarrow$  drawback  $\rightarrow$  very slow  
 $\downarrow$   
 optimization (Taylor series)

Optimized (Taylor series)



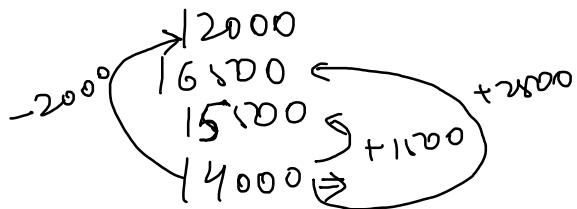
Xg Boost

↳ pip install xgboost

example:

$$y_i^o$$

$$\hat{y}_i = r$$



$$\bar{\hat{y}}_i^o = 14500$$

$$\frac{\partial L}{\partial r} = - \sum_{i=0}^n (y_i^o - r_i)$$

$$L = \frac{1}{2} (12000 - r_1)^2 + \frac{1}{2} (14000 - r_1)^2 + \frac{1}{2} (15500 - r_1)^2 + \frac{1}{2} (16500 - r_1)^2$$

$$0 - \frac{\partial L}{\partial r} = -\frac{1}{2} (12000 - r_1) + \left(-\frac{1}{2}\right) (14000 - r_1) + \left(-\frac{1}{2}\right) (15500 - r_1) + \left(-\frac{1}{2}\right) (16500 - r_1)$$

$$0 = r_1 - 12000 + r_1 - 14000 + r_1 - 15500 + r_1 - 16500$$

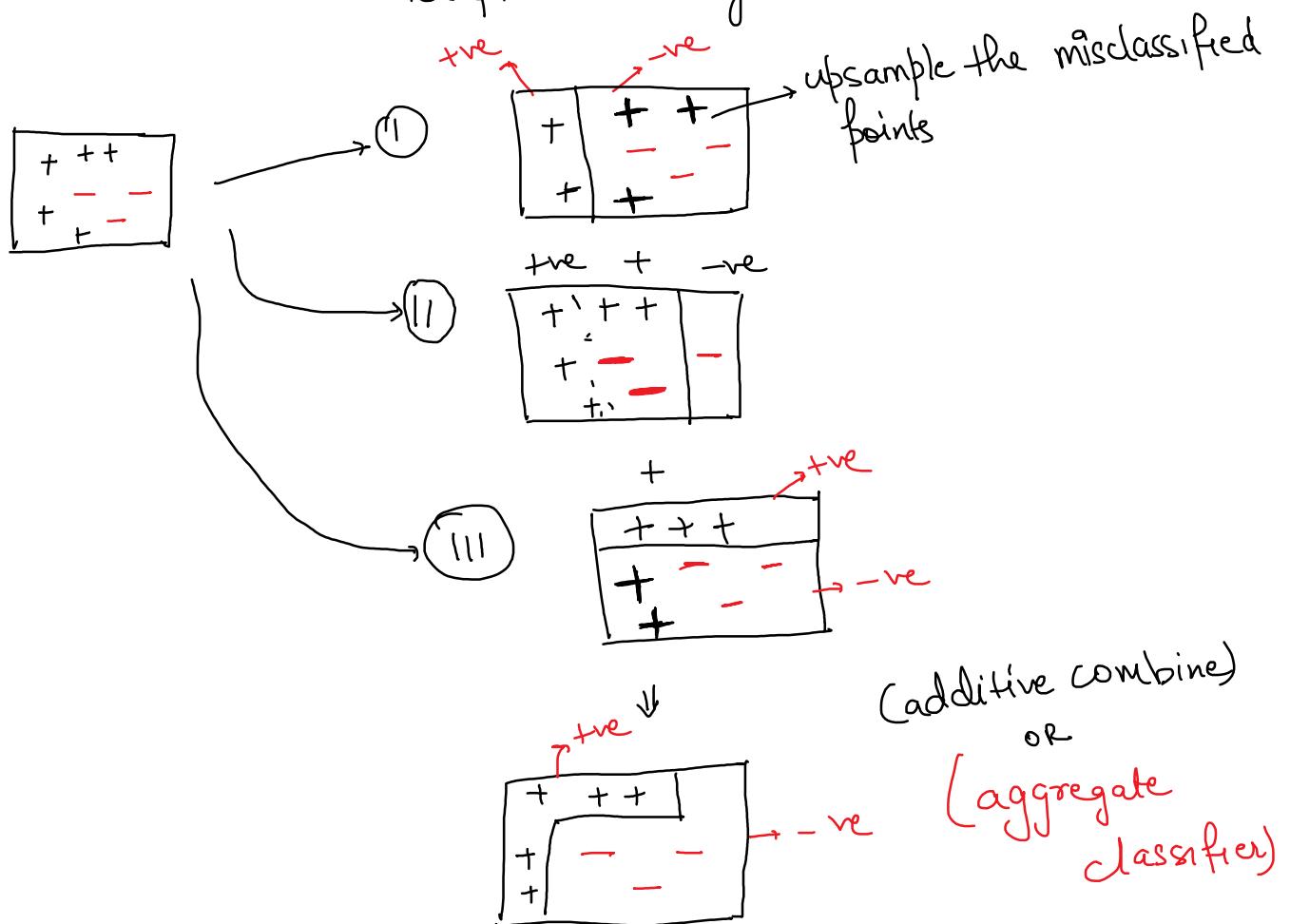
$$4r_1 - 58000 = 0$$

$$4\bar{Y}_i - 58000 = 0$$

$$\bar{Y}_i = \frac{58000}{4} = 14500 = \bar{Y}_i$$

ADABOOST

Adaptive boosting



$$C = r_1 C_1 + r_2 C_2 + r_3 C_3$$

$$x_1 \quad x_2 \quad y \quad \hat{y} \quad \text{weight} = y_n \\ 1 \quad 2 \quad . \quad . \quad 1 \quad 1 \quad 0.2$$

$$\chi = \text{error rate} \\ \dots \quad \wedge \quad \dots \quad \dots$$

|    | $x_1$ | $x_2$ | $y$ | $\hat{y}$ | weight = $y_n$ |
|----|-------|-------|-----|-----------|----------------|
| 1) | 3     | 9     | 1   | 1         | $y_n = 0.2$    |
| 2) | 2     | 4     | 0   | 1*        | $y_n = 0.2$    |
| 3) | 1     | 5     | 1   | 0*        | $y_n = 0.2$    |
| 4) | 9     | 6     | 0   | 0         | $y_n = 0.2$    |
| 5) | 5     | 7     | 0   | 0         | $y_n = 0.2$    |

$\alpha = \text{error rate}$

$$\alpha = \frac{1}{2} \ln \left( \frac{1 - \text{error}}{\text{error}} \right)$$

error  $\Rightarrow$  algebraic sum of weights of misclassified rows/points

$$\text{error} = 0.4$$

$$\alpha = \frac{1}{2} \ln \left( \frac{1 - 0.4}{0.4} \right) = \frac{1}{2} \ln \left( \frac{0.6}{0.4} \right) = 0.20$$

$$\text{new weights for correctly classified points} = e^{-\alpha} \times \text{old weights} = e^{-0.2} \times 0.2 = 0.16$$

$$\text{new weights for incorrectly classified points} = e^{\alpha} \times \text{old weight} = e^{0.2} \times 0.2 = 0.24$$

|   | $x_1$ | $x_2$ | $y$ | $\hat{y}$ | weights  | new weights | Normalize the weight |
|---|-------|-------|-----|-----------|----------|-------------|----------------------|
| 3 | 9     | 1     | 1   | 1         | 0.2      | 0.16        | $0.16/0.96$ $y_6$    |
| 2 | 4     | 0     | 1*  | 1*        | 0.2      | 0.24        | $0.24/0.96$ $y_4$    |
| 1 | 5     | 1     | 0*  | 0*        | 0.2      | 0.24        | $0.24/0.96$ $y_4$    |
| 9 | 6     | 0     | 0   | 0         | 0.2      | 0.16        | $0.16/0.96$ $y_6$    |
| 5 | 7     | 0     | 0   | 0         | 0.2      | 0.16        | $0.16/0.96$ $y_6$    |
|   |       |       |     |           | <u>1</u> | <u>0.96</u> | <u>1</u>             |



| $x_1$ | $x_2$ | $y$ | $\hat{y}$ | weights | New Weights | Normalized Weight Range |
|-------|-------|-----|-----------|---------|-------------|-------------------------|
|-------|-------|-----|-----------|---------|-------------|-------------------------|

|   |   |   |   |   |          |             |               |                                    |
|---|---|---|---|---|----------|-------------|---------------|------------------------------------|
| ① | 3 | 9 | 1 | 1 | 0.2      | 0.16        | $y_6 = 0.167$ | 0 - 0.167 $\xrightarrow{+0.7}$     |
| ② | 2 | 4 | 0 | 1 | 0.2      | 0.24        | $y_4 = 0.25$  | 0.167 - 0.417 $\xrightarrow{+0.1}$ |
| ③ | 1 | 5 | 1 | 0 | 0.2      | 0.24        | $y_4 = 0.25$  | 0.417 - 0.667                      |
| ④ | 9 | 6 | 0 | 0 | 0.2      | 0.16        | $y_6 = 0.167$ | 0.667 - 0.834                      |
| ⑤ | 5 | 7 | 0 | 0 | 0.2      | <u>0.16</u> | $y_6 = 0.167$ | 0.834 - 1                          |
|   |   |   |   |   | <u>1</u> | <u>0.96</u> | <u>1</u>      |                                    |

Randomly choose 5 numbers b/w 0 & 1

$$[0.1, 0.3, 0.4, 0.5, 0.7]$$

① ② ③ ③ ④

The above is upsample.

## Curse of Dimensionality

binary classification  $(0, 1) \Rightarrow f_1, f_2, f_3$

$$\Rightarrow 2^3 = 8$$

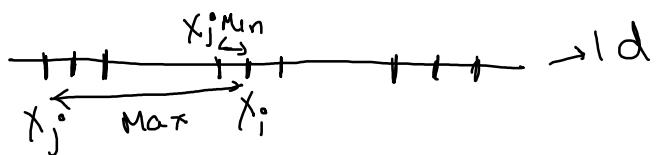
for 10 features  $= 2^{10}$

for 100 features  $= 2^{100}$

for n features  $= 2^n$

$\Rightarrow$  Hughes Phenomenon  $\Rightarrow$  Whenever the # dimensions  $\uparrow$ , the model performance  $\downarrow$

$\Rightarrow$  Distance functions  $\Rightarrow$  distance functions loses meaning in higher dimensions (especially euclidean)



$$\text{dist\_min} = \min d[x_i^*, x_j]$$

$$\text{dist\_max} = \max d[x_i^*, x_j]$$

$$\frac{\text{dist\_max} - \text{dist\_min}}{\text{dist\_min}} > 0 \quad \text{for } 1d, 2d, 3d$$

if all points are equidistant, as you have higher dimension

$$\text{dist\_max} = \text{dist\_min}$$

$$\text{dist\_max} - \text{dist\_min} = 0$$

In NLP, in order to avoid this issue, you use similarity

- ↳ hamming distance

3) As  $d \uparrow$ , chances of overfitting  $\uparrow$

### Feature Extraction

- ↳ Transformation of features

- ↳ new features

### PCA: Principal Component Analysis

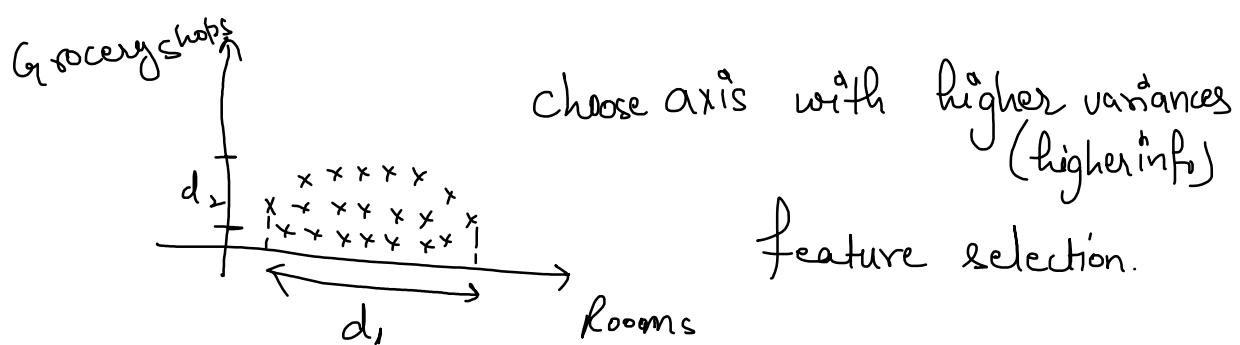
- ↳ reduces the dimensions to the best possible lowest dimensions to capture the essence of data

### Benefits:

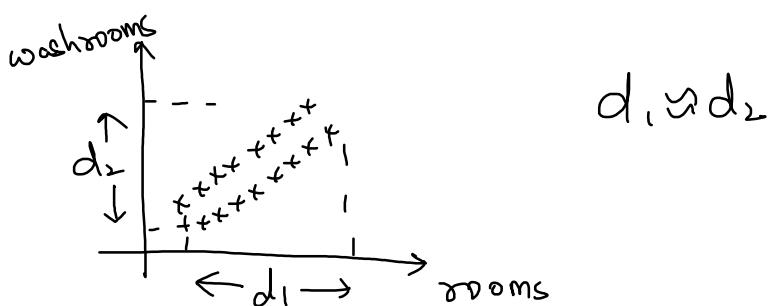
- Faster Execution
- visualization (except PCA)

Basic Intuition :

| Rooms | Grocery-shops | Price of flat |
|-------|---------------|---------------|
| 3     | 2             | 60            |
| 4     | 0             | 130           |
| 2     | 6             | 170           |
| 5     | 7             | 90            |



Rooms      washrooms      Price



feature extractions ↴

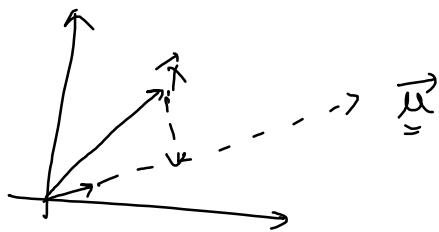
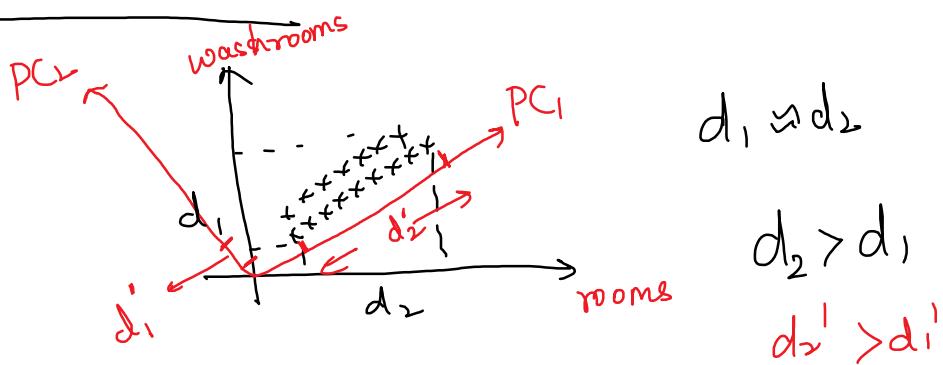
Transformed

Rooms & washrooms  $\Rightarrow$  size of flats

Rooms      washrooms  $\Rightarrow$  2d       $\xrightarrow{FE}$       size  $\rightarrow$  1d

feature extraction  $\Rightarrow$  creates new features from old features & choose a subset of features with higher importances

## Geometric Intuition of PCA:



Projection of  $x$  on  $u \Rightarrow \frac{\vec{u} \cdot \vec{x}}{\|\vec{u}\|}, \|\vec{u}\| = 1$

$$\Rightarrow \vec{u} \cdot \vec{x} \xrightarrow{\text{LA}} (\vec{u}^\top \cdot \vec{x})$$

The unit vector with higher variance is chosen as the right axis.

$$\text{MoF} \Rightarrow \sum_{i=1}^n \left( \vec{x}_i - \vec{\mu} \right)^2 \xrightarrow{\substack{\text{original dataset} \\ \Rightarrow}} \sum_{i=1}^n \left( \vec{u}^\top \cdot \vec{x}_i - \vec{u}^\top \cdot \vec{\mu} \right)^2 \xrightarrow{\substack{\text{mean} \\ \downarrow \\ n}} \text{Rayleigh Quotient (1950s)}$$

2)

Covariances  $\Rightarrow$  tells us about the relationship b/w matrix features

$$\begin{matrix} & X_1 & X_2 \\ X_1 & \text{Var}(X_1) & \text{Cov}(X_1, X_2) \\ X_2 & \text{Cov}(X_1, X_2) & \text{Var}(X_2) \end{matrix}$$

Covariance matrix

$$A\vec{x} = \lambda\vec{x}$$



Eigen decomposition  $\rightarrow$  eigen value & eigen vector

matrix

"largest eigen vector of covariance always points in the direction of largest variance."

Steps

1  $\xrightarrow{\substack{m=0 \\ \text{Mean Centering}}}$  not a mandatory step  
but algo works well

Standardization

2 Find covariance Matrix:

$$f_1 \quad f_2 \quad f_3$$

$$f_1 \quad \left[ \begin{matrix} \text{Var}(f_1) & \text{Cov}(f_1, f_2) & \text{Cov}(f_1, f_3) \end{matrix} \right]$$

$$\begin{matrix} f_1 \\ f_2 \\ f_3 \end{matrix} \begin{bmatrix} \text{Var}(f_1) & \text{Cov}(f_1, f_2) & \text{Cov}(f_1, f_3) \\ \text{Cov}(f_1, f_2) & \text{Var}(f_2) & \text{Cov}(f_2, f_3) \\ \text{Cov}(f_1, f_3) & \text{Cov}(f_2, f_3) & \text{Var}(f_3) \end{bmatrix} \begin{matrix} \text{Cov}(f_1, f_2) \\ " \\ \text{Cov}(f_2, f_1) \end{matrix}$$

3) → Eigen decomposition of covariance matrix:

$$\begin{matrix} f_1 & f_2 & f_3 \\ \downarrow & \downarrow & \downarrow \\ \text{eigenvalue} \rightarrow \lambda_1 & \lambda_2 & \lambda_3 \\ \downarrow & \downarrow & \downarrow \\ PC_1 & PC_2 & PC_3 \end{matrix} \begin{matrix} \text{Info} \\ PC_1 > PC_2 > PC_3 \end{matrix} \begin{matrix} \text{comparison} \end{matrix}$$

If you choose  $\lambda_1$ : it will be 1d

" " "  $\lambda_1 \& \lambda_2$ : " " " 2d

How to transform points for 3d to 1d?

Lets dataset  $\rightarrow$  1000 rows 3 columns

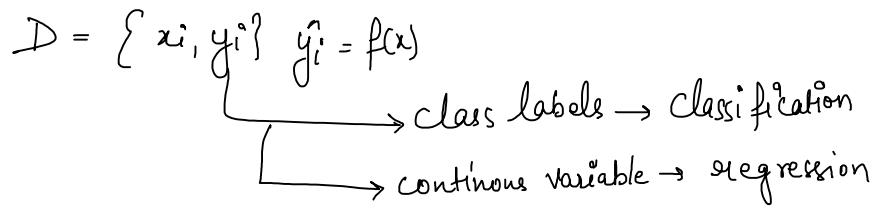
Shape of unit vector =  $[1 \times 3] \underbrace{[ \text{rows, columns} ]}_{\text{Transpose}}$

$$X \cdot U^T = \left[ \quad \right]_{1000 \times 3} \left[ \quad \right]_{3 \times 1}$$

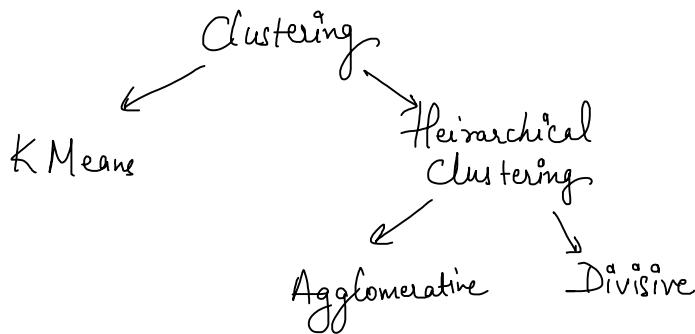
$$= \left[ \quad \right]_{1 \times 1}$$

- Transformation of 2d to 1d? assignment

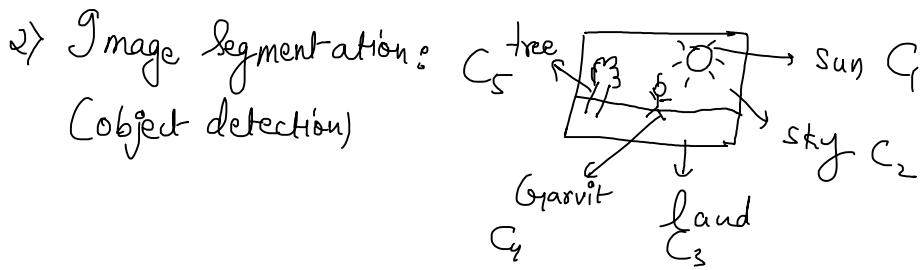
## Clustering



$D = \{x_i\} \rightarrow$  No class label → unsupervised learning



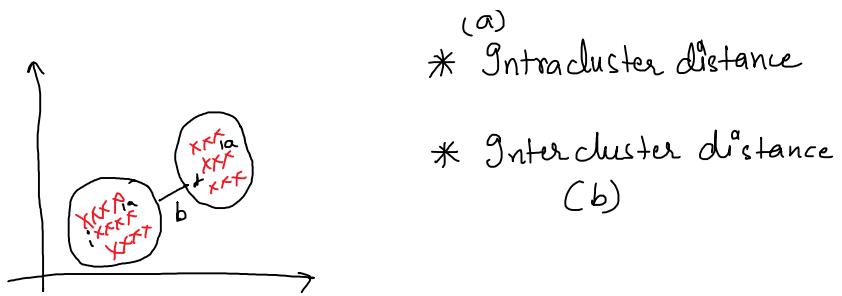
Applications: 1) commerce: group customers on the basis of location, gender, income level, product history etc



3) Review Analysis:



Metrics

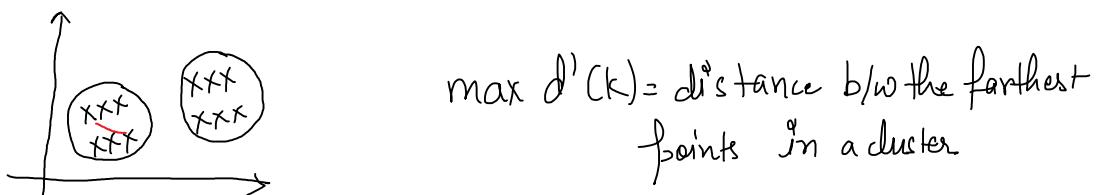
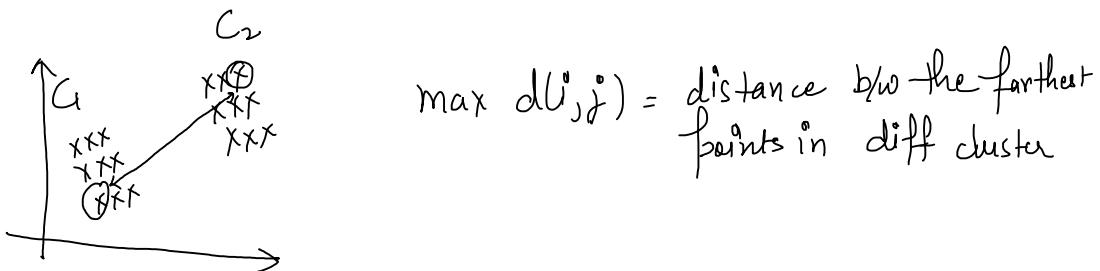


characteristics of good cluster:

- Small intracluster distance
- large intercluster distance

$$\Rightarrow \text{Dunn's Index} \uparrow \Rightarrow \frac{\max d(i, j)}{\max d(k)} \rightarrow \begin{array}{l} \text{maximal intercluster} \\ \text{distance} \end{array}$$

(0,  $\infty$ )                             $\max d(k) \rightarrow \text{maximal intracluster distance}$



$$\rightarrow \text{Silhouette's Score} \Rightarrow \frac{b - a}{\max(b, a)} \rightarrow \text{sklearn.metrics}$$

$a \in [-1, +1]$

$b \Rightarrow \text{average intercluster distance}$

$\Rightarrow$  average intrachuster distance

Case 1:  $a = \min \Rightarrow 0$ ,  $b = \max = b$

$$\text{Silhouette's Score} = \frac{b - a}{\max(b, a)} = \frac{b}{b} = 1$$

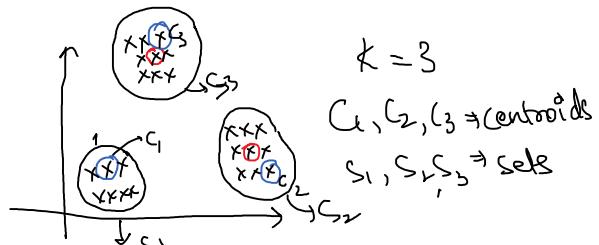
Case 2:  $b < a$ ,  $b=0$ ,  $a=a$

$$\text{Silhouette's Score} = \frac{b - a}{\max(b, a)} = \frac{-a}{a} = -1$$

Case 3:  $b=a$

$$\text{Silhouette's Score} = \frac{b - a}{\max(b, a)} = \frac{a - a}{a} = 0$$

# clusters  $\leftarrow$  K-Means  $\rightarrow$  average



$\rightarrow$  define the centroids randomly (means)

$\rightarrow$  Create clusters  $\Rightarrow$  distance of datapoint from the centroids

$$\text{Distance} = |x_i - c_i|$$

$$S_1 \cap S_2 = \emptyset$$

$$S_2 \cap S_3 = \emptyset$$

$$S_1 \cap S_3 = \emptyset$$

$\rightarrow$  after the creation of sets, calculate centroids again,

Mathematical Objective function  $C^* = \underset{C_1, C_2, \dots, C_k}{\operatorname{Argmin}} \sum_{i=1}^n \sum_{x \in S_i} \|x - C_i\|^2$

s.t.:  $x \in S_i$   
 $S_i \cap S_j = \emptyset$

↳ intracluster  
distance

Complexity theory: NP hard  $\Rightarrow$  Very complex problem

PHD  $\leftarrow \downarrow$   
Approximate Algorithm

### Lloyd's Algorithm

- 1) Randomly choose  $K$  datapoints from dataset & call them Centroids.
- 2) Assignment: For each point, select the nearest centroid with the help of distance & add the point to the corresponding cluster

- 3) Updation: Recalculate centroids

$$C_j^* = \frac{1}{|S_j|} \sum_{i=1}^n x_i \quad (x_i \in S_j)$$

- 4) Repeat step 2 & 3 till convergence

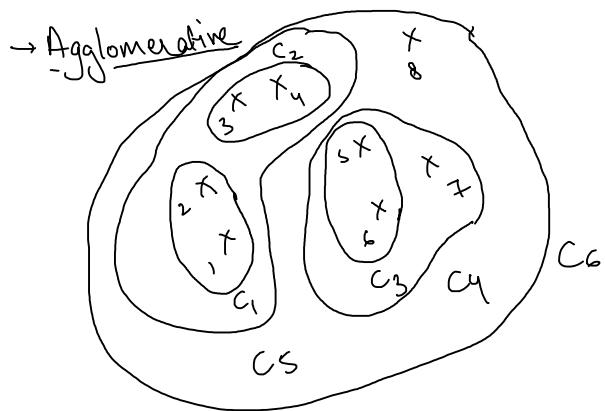
KMeans++ = [ $\text{init} = \text{'Kmeans++'}$ ]

| <u>data points</u> | <u>distance</u> |
|--------------------|-----------------|
| $x_1$              | $d_1$           |
| $x_2$              | $d_2$           |
| $x_3$              | $d_3$           |
| $\vdots$           | $\vdots$        |
| $x_n$              | $d_n$           |

the larger the distance,  
the greater the probability  
of it being picked out  
as centroid

→ KMeans is affected by outliers

### Hierarchical Clustering



→ each point is a cluster

8 clusters

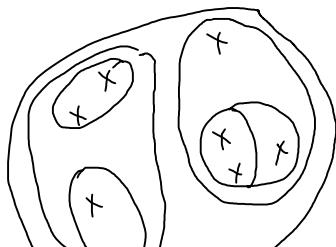
↓  
4 clusters

↓  
2 clusters

↓  
1 clusters

How to do grouping/clustering? →  $\begin{cases} \text{distance} \\ \text{similarity} \end{cases}$

### Divisive

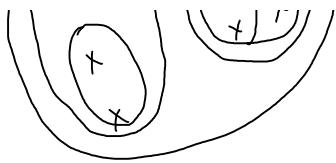


### Divisive

1 cluster

↓  
2 clusters

↓  
2 - 1...1...1...1



↓  
3 clusters  
↓

2 clusters

Dendogram: A tree like structure that records merges & splits.

## Ensembles

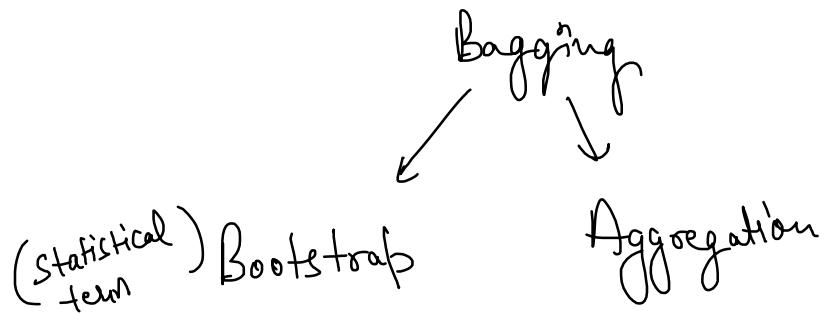
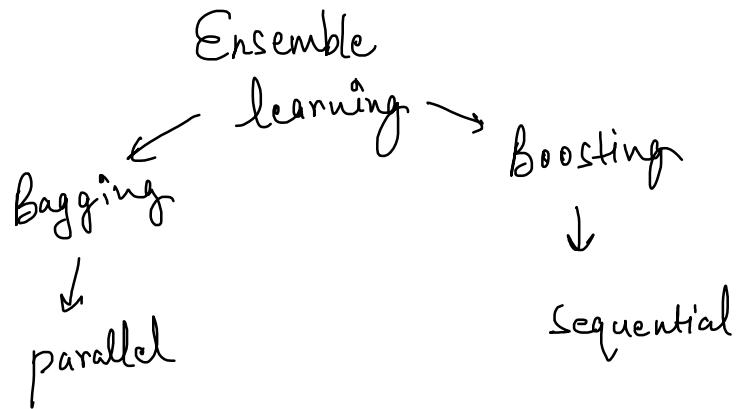
↳ groups of notes/musicians



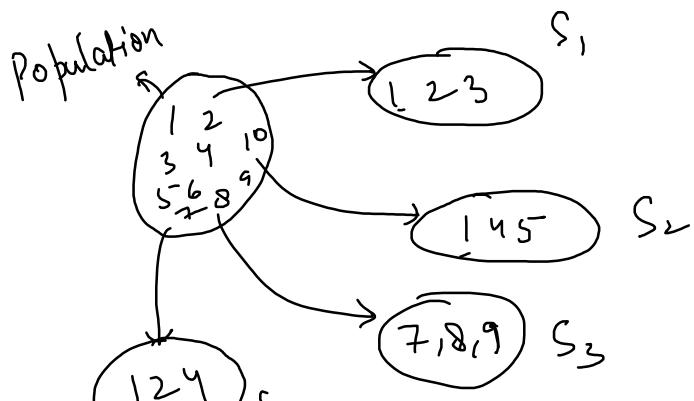
in m/c learning

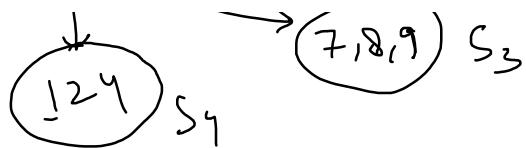


group of models

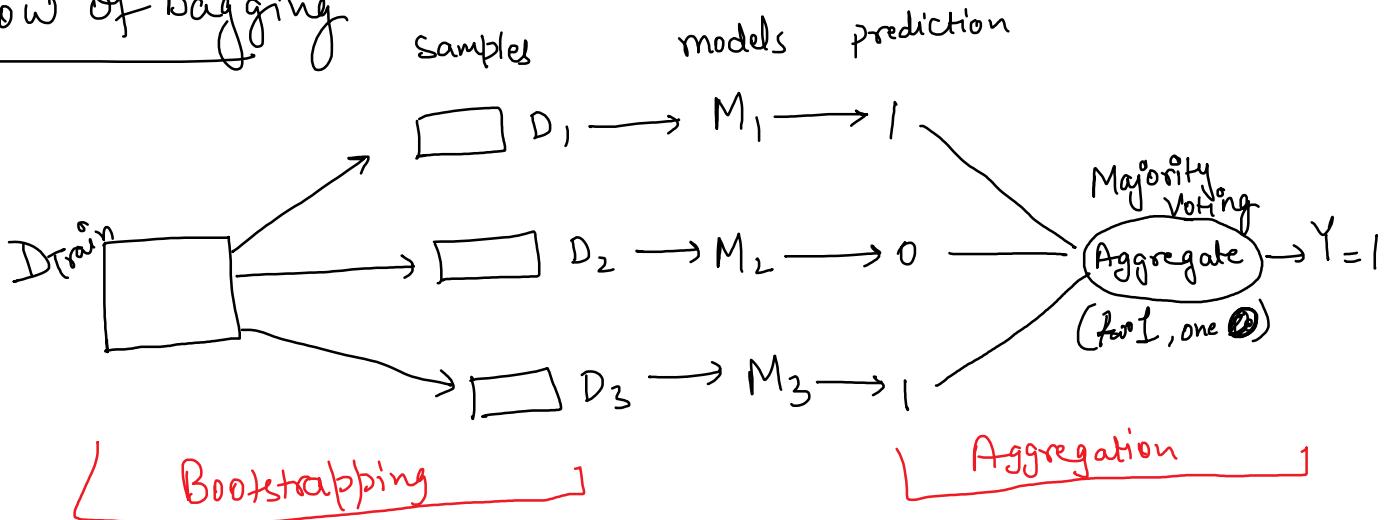


Bootstrapping → Sampling with replacement

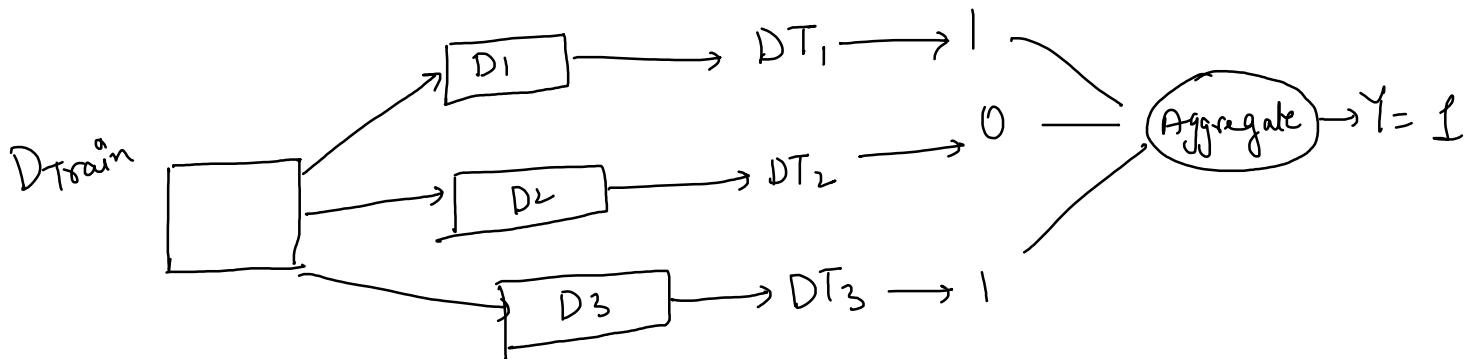




## Flow of Bagging



## Random Forest



→ In Random Forest, DT that have high variance & low bias  
 $\downarrow$   
 $DT \rightarrow \text{max-depth } \uparrow$

→ Model should be different

|    | CGPA | IQ  | Extra | Social | Placed |                            |
|----|------|-----|-------|--------|--------|----------------------------|
| 1) | 9    | 120 | 6     | 8      | 1      | $DT_1, DT_2$ <del>DT</del> |
| 2) | 7    | 110 | 7     | 5      | 0      |                            |
| 3) | 8    | 125 | 8     | 7      | 1      |                            |
| 4) | 6    | 105 | 9     | 9      | 0      |                            |

D1

| CGPA | Social | Placed |
|------|--------|--------|
| 8    | 7      | 1      |
| 6    | 9      | 0      |



DT1

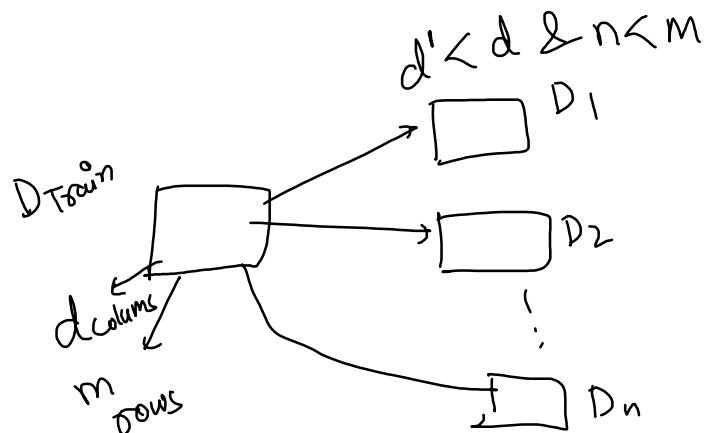
D2

| Extra | IQ  | Placed |
|-------|-----|--------|
| 8     | 125 | 1      |
| 9     | 105 | 0      |



DT2

$$RF = \underbrace{(\text{low bias} + \text{high variance})}_{\text{DT (base)}} + (\text{Row sampling}) + (\text{Column sampling}) + \text{Aggregation}$$



Oob score :

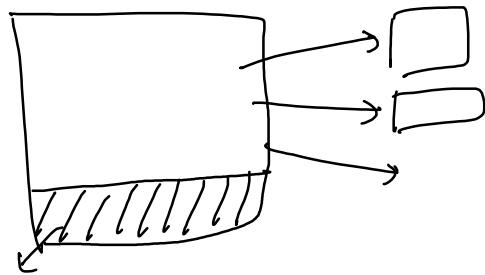
37.1%

Out of bag

samples ↓ can be used for validation

oob score ↑

oob score ≈ accuracy



→ Black box

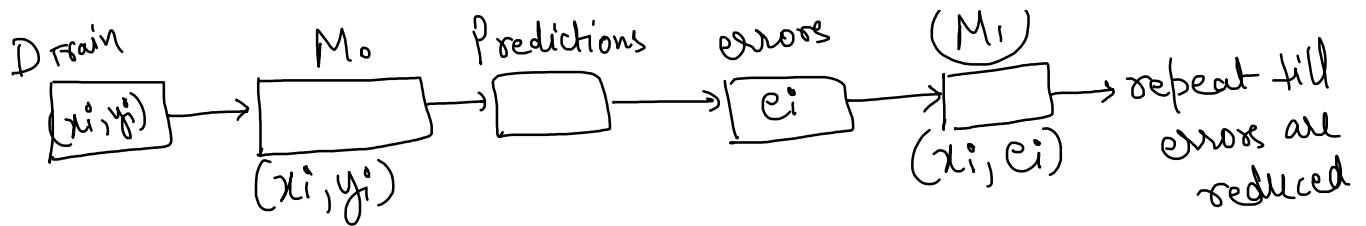
→ No mathematical objective function / loss function

Boosting  
↓  
sequential

Base Model → high bias & low variance

depth=1  
→ decision stump

$D_{train}$        $M_0$       Predictions      errors       $M_1$       repeat till



Steps

0)  $D_{\text{Train}} = \{x_i^i, y_i^i\}_{i=1}^n \rightarrow M_0 \left( \text{fits } h_0(x) \text{ on } D_{\text{Train}} \right)$

$\downarrow$   
predictions

$\downarrow$   
errors  $e_i^i = y_i^i - \hat{y}_i^i = y_i^i - h_0(x)$

$$h_0(x) = \underline{y} = \underbrace{m_1 x_1 + m_2 x_2}_{h_0(x)}$$

1)  $M_1 \rightarrow (x_i^i, e_i^i)$

$\underbrace{h_1(x)}_{\text{Model at end stage}}$

$e_i^i \leftarrow \underset{\substack{\text{prediction} \\ \text{at end of} \\ \text{stage } 1}}{\text{prediction}} \leftarrow f_1(x) = \gamma_0 h_0(x) + \gamma_1 h_1(x)$

$\underbrace{\gamma_0 h_0(x) + \gamma_1 h_1(x)}_{\text{weighted sum of functions}}$

2)  $(x_i^i, e_i^i) \rightarrow M_2$

$\underbrace{h_2(x)}_{\text{Model at end stage}}$

$$f_2(x) = \overbrace{\gamma_0 h_0(x) + \gamma_1 h_1(x) + \gamma_2 h_2(x)}^{f_1(x)}$$

$$f_2(x) = \underline{f_1(x)} + \gamma_2 h_2(x)$$

for  $\underline{k^{\text{th}}}$  stage:  $f_k(x) = \sum_{i=0}^k \underline{\gamma_i h_i(x)}$

for  $\underline{k^{\text{th}}}$  stage:  $f_k(x) = \sum_{i=0}^k \underline{r_i h_i(x)}$

$K = \# \text{models}$

$$\begin{aligned}
 &= r_0 f_0(x) + \sum_{i=1}^k r_i h_i(x) \\
 &= r_0 f_0(x) + r_1 f_1(x) + \sum_{i=2}^k r_i f_i(x) \\
 &= f_0(x) + \sum_{i=1}^k r_i f_i(x)
 \end{aligned}$$

## Residuals & loss functions

$$L[y_i, f(x)] = \underset{\text{prediction}}{[y_i - f_k(x)]^2} = [y_i - z_i]^2$$

$$\frac{\partial L}{\partial z_i} = -\odot [y_i - z_i]$$

$\frac{\partial}{\partial} \rightarrow \text{gradient}$

$$\frac{\partial L}{\partial z_i} = -\text{error} \quad \Rightarrow \text{error} = -\frac{\partial L}{\partial z_i}$$

errors  $\Rightarrow$  negative gradient / pseudo-residual

$$\underline{\underline{G_B}}$$

$\mathcal{G}/P \rightarrow (x_i, y_i) \Rightarrow$  differentiable loss fn  $L(y_i, f(x))$

(1)  $\Gamma = \text{argmin} \sum_{i=1}^n L(y_i, r)$   $r \in \dots$

$$0) \quad f_0 = \underset{r}{\operatorname{argmin}} \sum_{i=1}^n L(y_i, r) \rightarrow r = \bar{y}_i$$

1) for  $m=1$  to  $M$

$$\gamma_m = - \left[ \frac{\partial L(y, f_{m-1}(x_i))}{\partial F_{m-1}(x_i)} \right]$$

for  $m=1$ ,

$$\gamma_1 = - \left[ \frac{\partial L(y, f_0(x_i))}{\partial F_0(x_i)} \right]$$

$$2) \quad \gamma_m = \underset{r}{\operatorname{argmin}} \quad L \left[ y_i, f_{m-1}(x_i) + \gamma_m h_m(x_i) \right]$$

for  $m=1$  predictions for  $i$  model

$$L \left[ y_i, \underbrace{f_0(x_i) + \gamma_1 h_1(x_i)} \right]$$

$$4) \quad f_m(x_i) = f_{m-1}(x_i) + \gamma_m h_m(x) \quad \gamma = \text{weight of models}$$

New Pred = Old Pred + additive combinations of models

Boosting models are very easy to overfit.

### Shrinkage

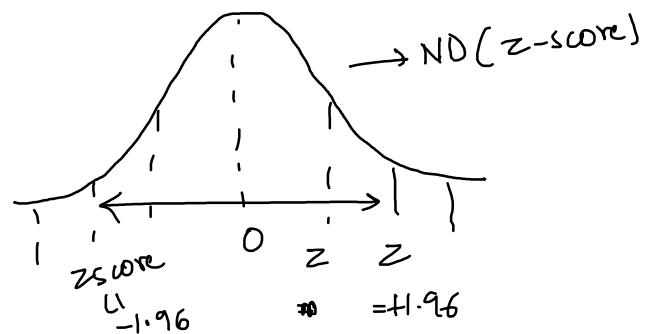
$$f_m(x) = f_{m-1}(x) + \underbrace{\gamma_m h_m(x)}_{\text{learning rate}} \quad \text{to avoid overfitting}$$
$$[0 - 1]$$

XGBoost → Optimized version of GB  
↓ Series  
Taylor Expansion

## Z-score & calculate probability values

$$Z = \frac{x - \mu}{\sigma}$$

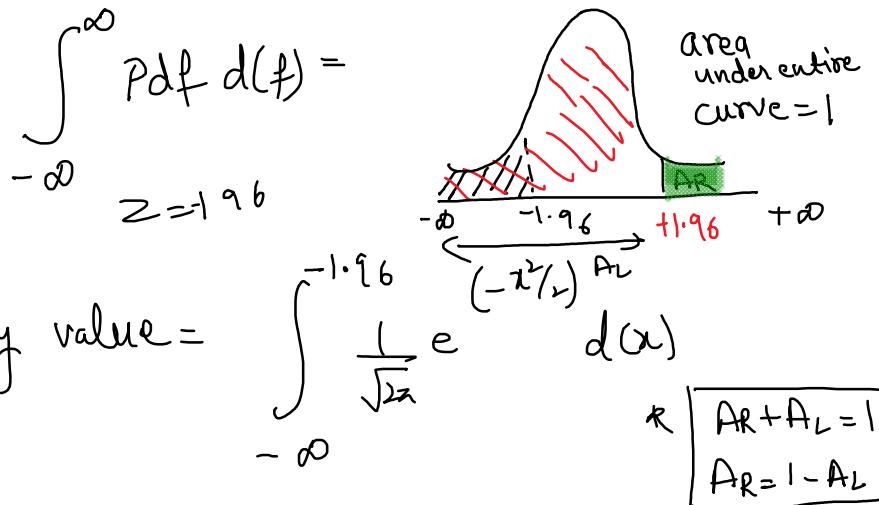
area under curve = probability



Q Z-score, can we connect with probability or can we get probability value for z-score?

Sol.

calculate probability value =

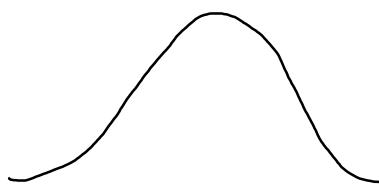


$$\begin{aligned} AR + AL &= 1 \\ AR &= 1 - AL \end{aligned}$$

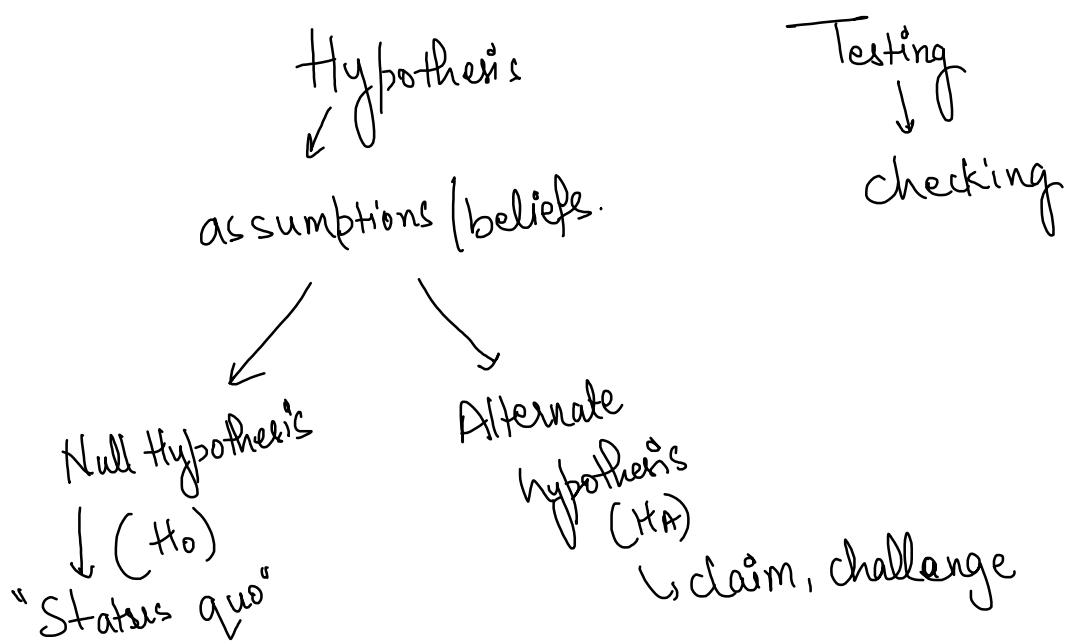
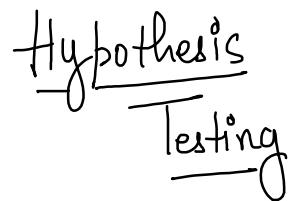
instead of integration, we will use z-table

\* In z-table, area is always calculated from extreme left to the desired z-value!

$$\underline{\text{Z-score}} \Rightarrow \underline{\frac{x - \mu}{\left(\frac{\sigma}{\sqrt{n}}\right)}}$$



the reason you are dividing  $\sigma$  by  $\sqrt{n}$  is that you are paying a penalty for using sample instead of pop



Police claims that a person <sup>is</sup> a criminal?

$H_0$ : Person is innocent.

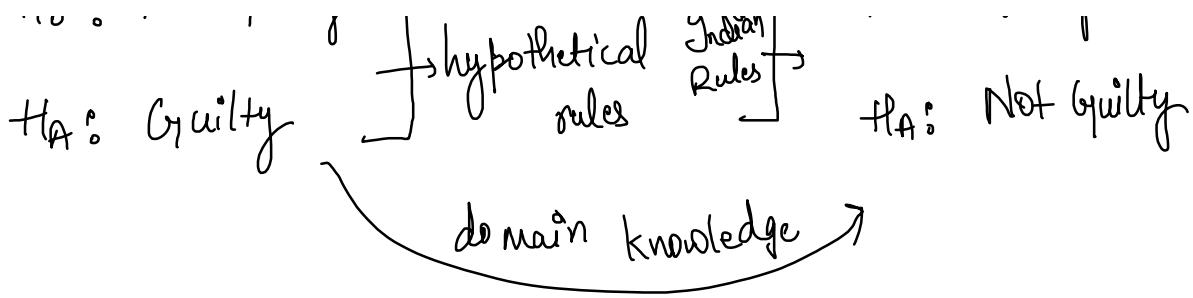
Def.: Person is a criminal

Q Bride claims that groom has taken dowry?  
Plaintiff

$H_0$  : Not Guilty      }  
 $H_1$  : Guilty              } hypothetical  
                              after

H<sub>0</sub> : Guilty

DP : Not guilty



Q India is going to win the World Cup.

$H_0$ : Any other team can win.

$H_A$ : India is winning world.

Q I claim, that avg salary of an electrical engg changed from \$50,000

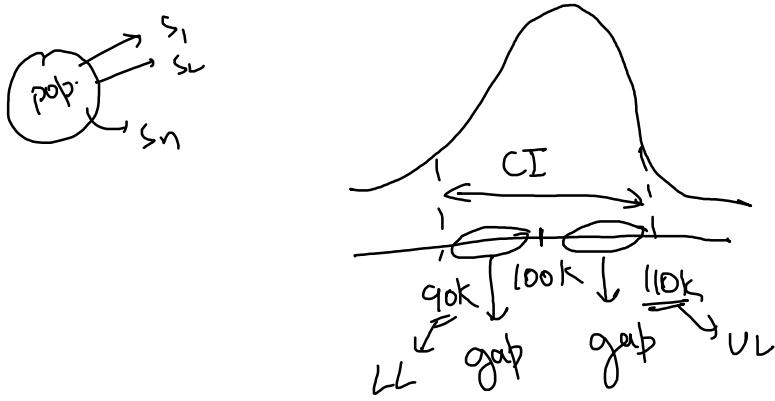
Sol.  $H_0: \mu = \$50,000$

$H_A: \mu \neq \$50,000$

Build the criteria to test hypothesis:

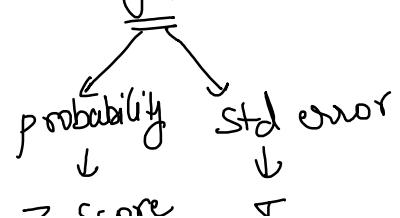
→ Acceptance Region Method

Q Data Scientists earn \$100,000 salary on avg?



$$qok = 100k - gap$$

$$100k = 100k + \underline{gap}$$



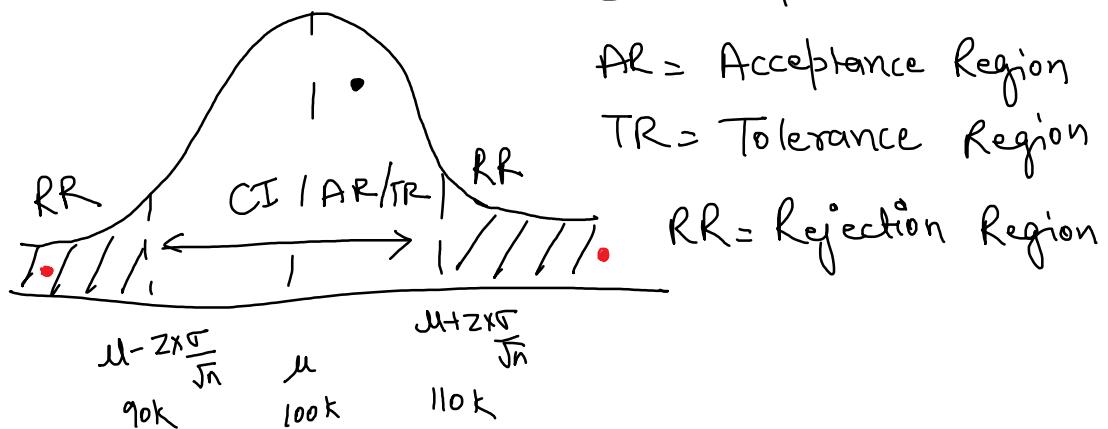
$\mu$  gap  $\sigma$  gap  $\cdot \bar{X}$

Power  $\downarrow$   $\rightarrow$   $\downarrow$   
 $Z$ -score  $\frac{\bar{X} - \mu}{\sigma/\sqrt{n}}$

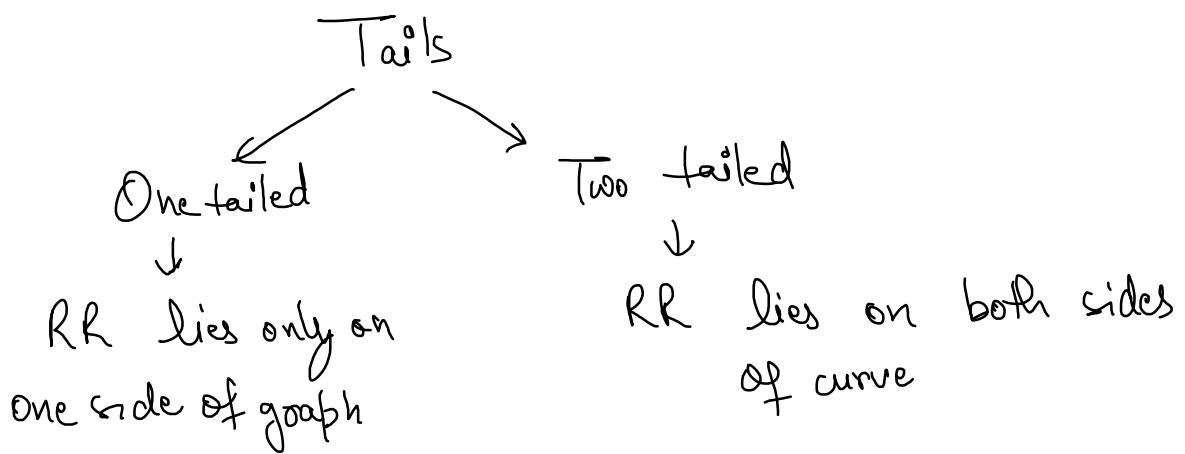
$$UL = \mu + Z \times \frac{\sigma}{\sqrt{n}}$$

Margin of error

$$LL = \mu - Z \times \frac{\sigma}{\sqrt{n}}$$



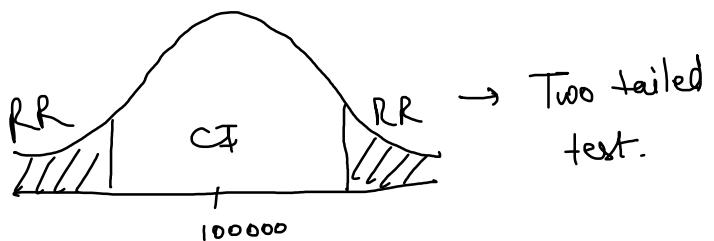
$$RR + AR + TR = 1 \Rightarrow \text{Acceptance Region} + \text{Rejection Region} = 1$$



$\Leftrightarrow$

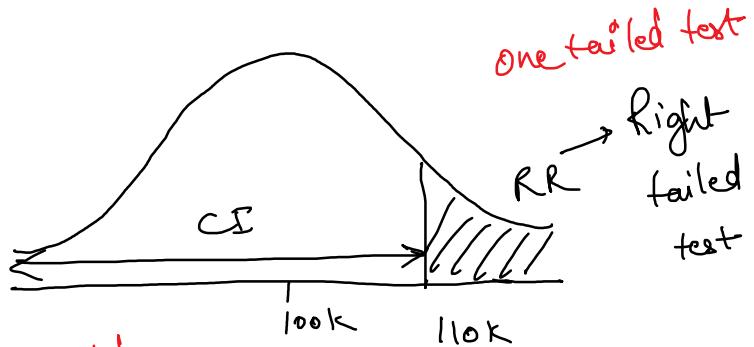
$$\mu = \$100,000$$

$$\mu \neq \$100,000$$



$H_0: \mu \leq \$100,000$

$H_1: \mu > \$100,000$



$H_0: \mu \geq \$100,000$

$H_1: \mu < \$100,000$

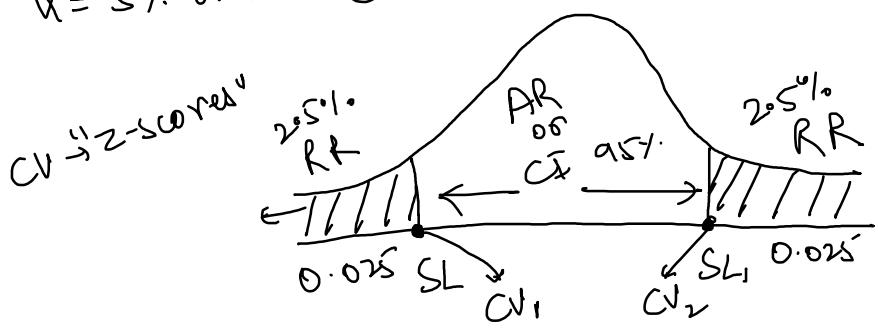


## 2) Critical Value Method

$\alpha = 5\% \text{ or } 0.05 \text{ (default)}$

→ where to build Rejection Region)

↓  
significance level ( $\alpha$ )



$$CV_1 = -1.96$$

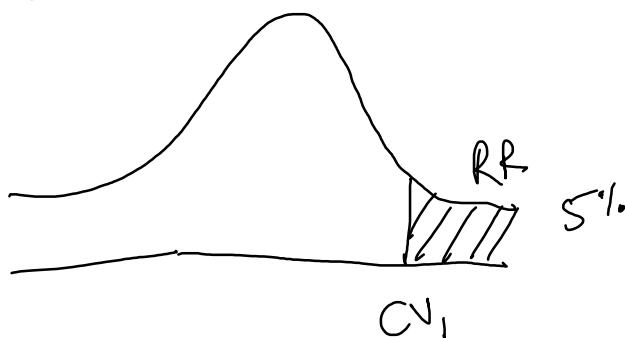
$$CV_2 = +1.96$$

$$\ast \quad CI + SL = 1$$

$$CI = 1 - SL \quad | \quad SL = 1 - CI$$

in this scenario,  $CI = 0.95$   
 $SL = 1 - 0.95 = 0.05$

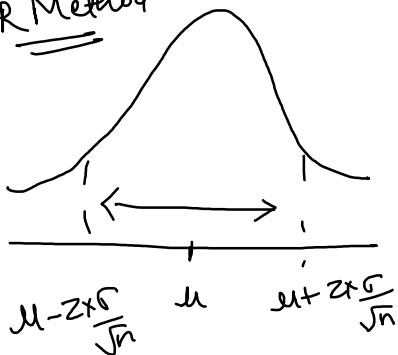
$\alpha = 0.05 \text{ or } 5\%$



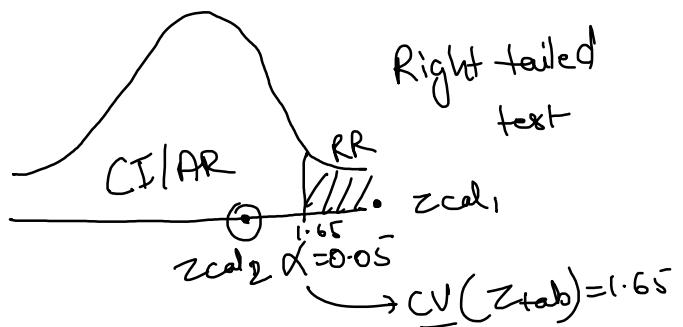
$$CV_1 = 1.65$$

Steps to perform AR & CV Method for Testing:

1) AR Method



2) Critical Value Method



$$Z_{cal} = \frac{x - \mu}{\frac{\sigma}{\sqrt{n}}}$$

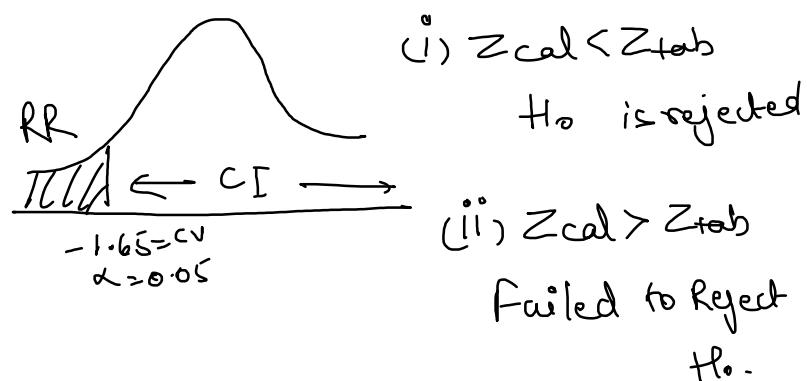
(i)  $Z_{cal} > Z_{tab}$

$H_0$  is Rejected

(ii)  $Z_{cal} < Z_{tab}$

Failed to Reject  $H_0$

Left tailed test



\* P-value Method : The probability of your null hypothesis to be true.

③ \* P-value Method : The probability of your null hypothesis to be true.

compare your pvalue with  $\alpha$ .

probability  
↓ area under curve.

if pvalue <  $\alpha$

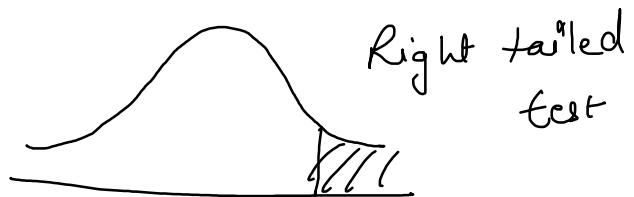
Reject  $H_0$

Q A principal of school claims that students have above average IQ. A random sample of 30 students is taken with a mean of 112.5. The mean & std dev of population is 100 & 15. Test your hypothesis.

Sol. ①  $H_0 : \mu \leq 100$

② Need to check whether test is one tailed or two tailed

$H_A : \mu > 100$



③  $\mu = 100, \sigma = 15, X = 112.5$

AR/TR

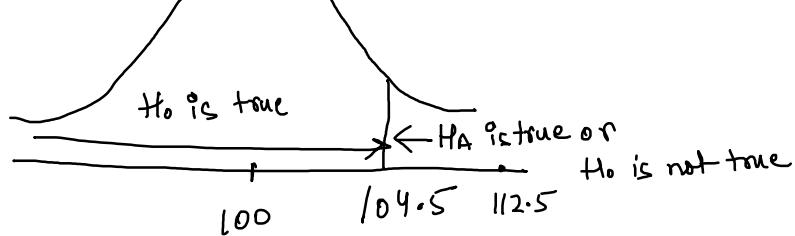
$$UL = \mu + Z \times \frac{\sigma}{\sqrt{n}} = 100 + 1.65 \times \frac{15}{\sqrt{30}} = 104.5$$

z-value  
1.65



$$LL = \mu - Z \times \frac{\sigma}{\sqrt{n}} = 100 - 1.65 \times \frac{15}{\sqrt{30}} = 95.5$$

$X = 112.5$



lets compare  $112.5$  with  $104.5$

$$112.5 > 104.5$$

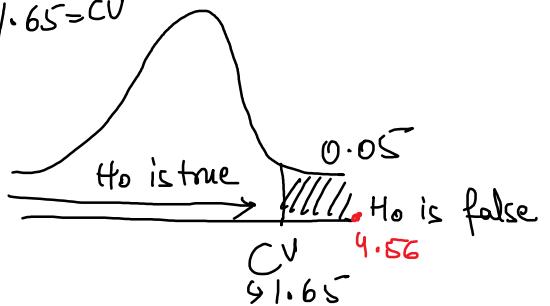
hence, Reject  $H_0$

## 2) Critical Value Method

$\alpha = 0.05$ , test is right tailed

$$Z_{\text{cal}} = \frac{x - \mu}{\sigma / \sqrt{n}}$$

$$, Z_{\text{tab}} (\alpha = 0.05) = 1.65 = CV$$



$$Z_{\text{cal}} = \frac{112.5 - 100}{15 / \sqrt{30}} = \frac{12.5}{15 / \sqrt{30}} = 4.56$$

$$\therefore Z_{\text{cal}} > Z_{\text{tab}}$$

$\therefore$  Reject  $H_0$

## 3) Pvalue Method:

$$\alpha = 0.05, Z_{\text{cal}} = \frac{112.5 - 100}{15 / \sqrt{30}} = 4.56$$



$$P(Z_{\text{cal}} = 4.56) = 1 - F_L = 0.0000034$$

compare pvalue with significance level

pvalue  $\leftarrow$   $\alpha$   
(less than)

$$(0.000034) < 0.05$$

Hence, reject  $H_0$

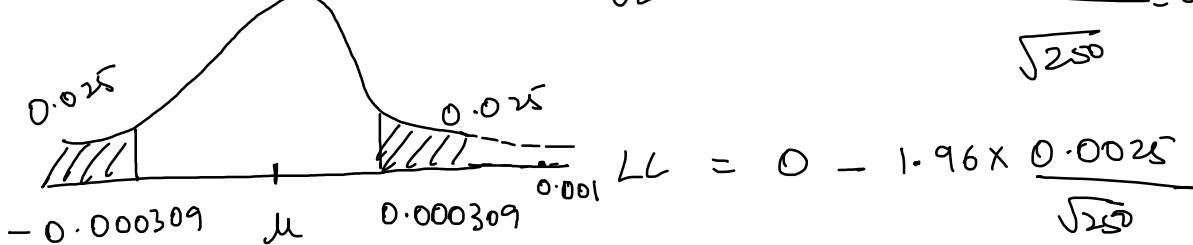
$$\mu = 0, \bar{x} = 0.1\%, \sigma = 0.25\%$$

Q A researcher has agreed upon a data of daily return of portfolio of call option over a recent 250 days period.  
The mean of daily return is  $0.1\%$  & std dev is  $0.25\%$ .  
The researcher claims the mean daily portfolio is not 0.  
Construct CI at 95% & test the belief.

Sol:  $H_0 \Rightarrow \mu = 0$  test is two tailed.  $CI + \frac{\alpha}{2} = 1$

$$H_A \Rightarrow \mu \neq 0 \quad CI = 0.95, \alpha = 1 - 0.95 = 0.05$$

AR  $\bar{x} = 0.1\%, \mu = 0, \sigma = 0.0025$   $\text{TS} \rightarrow$   $UL = 0 + 1.96 \times \frac{0.0025}{\sqrt{250}} = 0.000309$



$$LL = 0 - 1.96 \times \frac{0.0025}{\sqrt{250}} = -0.000309$$

$$-0.000309 < 0.1\% > 0.000309$$

$$0.0017$$

Reject  $H_0$ .

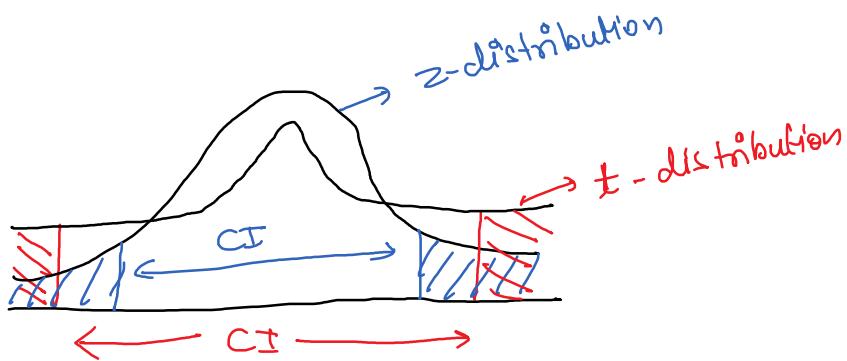
CV Method & Pvalue is your assignment!

T-distribution

low confidence

samples are very less  
( $n < 30$ )

$$t = \frac{x - \mu}{\frac{s}{\sqrt{n}}}$$



Degrees of freedom: logically independent values

$$\begin{array}{c} \text{5 values} \\ \left\{ \begin{array}{l} +3 \\ -3 \\ +2 \\ -2 \\ \text{avg} = 5 \\ +1 \\ -1 \\ 0 \\ +0 \\ x \\ x = 7 \end{array} \right. \end{array}$$

g have 4 logically  
independent values!

for  $n$  values, degrees of freedom =  $(n-1)$

↳ logically independent  
values

$t_{\text{cal}}$ ,  $t_{\text{tab}}$ ,  $\chi$ , df

$t_{\text{tab}}$  compare it with  $t_{\text{cal}}$

Q A company manufactures car batteries with average life span of 2 years or more. An engineer believes this value to be less. Using 10 samples, he measured the life span & found it to be 1.8 years with a std dev of 0.15.

At 99% CI, is there enough evidence to reject  $H_0$ .

Sol.

$$H_0 : \mu \geq 2$$

$$H_A : \mu < 2$$



$$CI = 99\% = 0.99$$

$$\alpha = 1 - 0.99 \\ = 0.01$$

$$n = 10, \quad df = 10 - 1 = 9 \quad t_{\text{tab}} = 2.821 \\ (0.01)$$

$$t_{\text{cal}} = \frac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}} = \frac{1.8 - 2}{\frac{0.15}{\sqrt{10}}} = -4.25$$

$$t_{\text{cal}} < t_{\text{tab}}$$

Reject  $H_0$

## Errors in Hypothesis



- $H_0$  is true but  $H_A$  is accepted
- $H_A$  is true but  $H_0$  is accepted

|                     |   | Actual<br>$H_0$ is true | $H_0$ is false ( $H_A$ is true) |
|---------------------|---|-------------------------|---------------------------------|
| Predicted<br>Values | ( $H_0$ accepted)                                 |                         |                                 |
|                     | $H_0$ rejected                                    | Type I error<br>FP      | ✓                               |
|                     | Failed to<br>reject $H_0$<br>( $H_A$ is rejected) | ✓                       | FN<br>Type II error             |

Type I error = significance level

Type II error =  $\beta$

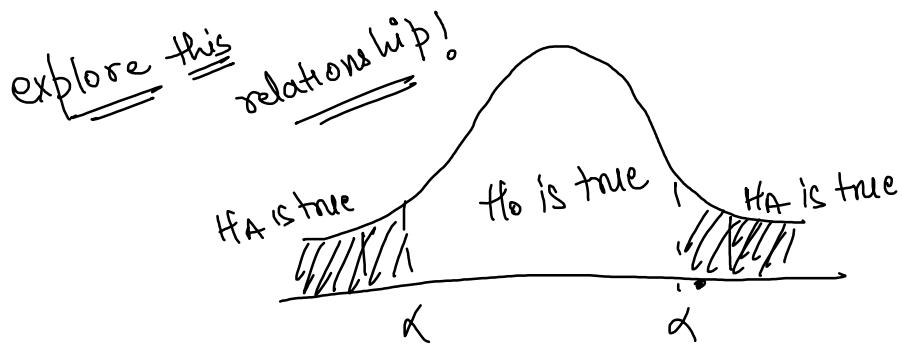
↳ Power of Test

↓  
Ability of the test to make right decisions.

Power of test  $\propto$  no. of sample / sample size

$$\boxed{\beta = 1 - \text{Power}}$$

\* Relationship b/w Type I & Type II error



$$\downarrow \text{Type I} \propto \frac{1}{\text{Type II} \uparrow}$$

### CHI SQUARE Test

- ↳ Non-Parametric Test  
(no assumptions taken)
- ↳ char-char situations  
(categories)

Degrees of freedom =  $(r-1)(c-1)$

rows                          columns

Q Whether there is a relationship b/w gender and result?

|        |   | Result |        |
|--------|---|--------|--------|
|        |   | (Pass) | (Fail) |
| Gender | M | 60     | 40     |
|        | F | 24     | 32     |

Sol.

$H_0$ : There is no relationship b/w gender & result  
 $H_A$ : There is relationship b/w gender & result

| Gender | Result | Pass | Fail | Total |
|--------|--------|------|------|-------|
| M      |        | 60   | 40   | 100   |
| F      |        | 24   | 32   | 56    |
|        | Total  | 84   | 72   | 156   |

Total males = 100, Total females = 56, total pass = 84, total fail = 72

### Expected value

$$\text{expected value} = \frac{\text{Total males} \times \text{total pass}}{\text{total no. of people}} = EV1$$

(total males who passed)

$$\text{expected value} = \frac{\text{total females} \times \text{total pass}}{\text{total no. of people}} = EV2$$

(total female passed)

$$\text{expected value} = \frac{\text{total males} \times \text{total failed}}{\text{total no. of people}} = EV3$$

(total males failed)

$$\frac{\text{expected value}}{\text{(total female failed)}} = \frac{\text{total females} \times \text{total failed}}{\text{total no of people}} = EV4$$

| Result | Pass       | Fail              | Total |
|--------|------------|-------------------|-------|
| Gender |            |                   |       |
| M      | 53.8 (EV1) | 46.1 (EV3) = 100. |       |
| F      | 30.1 (EV2) | 25.8 (EV4) = 56   |       |
| Total  | <u>84</u>  | <u>72</u>         |       |

$$EV1 = \frac{100 \times 84}{156} = 53.8$$

$$EV2 = \frac{56 \times 84}{156} = 30.1$$

$$EV3 = \frac{100 \times 72}{156} = 46.1$$

$$EV4 = \frac{56 \times 72}{156} = 25.8$$

Calculate  $\chi^2$ :

$$\chi^2 = \frac{(Actual - Expected)^2}{Expected}$$

$$\textcircled{I} \quad \frac{(60 - 53.8)^2}{53.8} = 0.71$$

$$\textcircled{II} \quad \frac{(40 - 46.1)^2}{46.1} = 0.81$$

$$\textcircled{III} \quad \frac{(24 - 30.1)^2}{30.1} = 1.23$$

$$\textcircled{IV} \quad \frac{(32 - 25.8)^2}{25.8} = 1.48$$

$$\chi^2_{\text{cal}} = 0.71 + 1.23 + 0.81 + 1.48 \\ = 4.23$$

Tabulate  $\chi^2$ :  $\alpha = 0.05$ ,  $df = (r-1)(c-1)$   
 $= (2-1)(2-1)$   
 $= 1 \times 1 = 1$

$$\chi^2_{\text{tab}} = 3.841$$

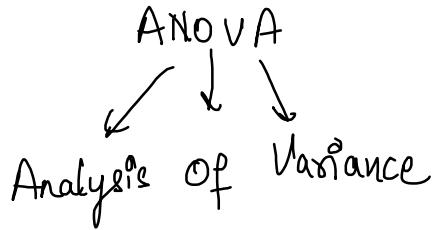
Comparing  $\chi^2_{\text{cal}}$  and  $\chi^2_{\text{tab}}$ :

$$\chi^2_{\text{cal}} > \chi^2_{\text{tab}}$$

Reject  $H_0$

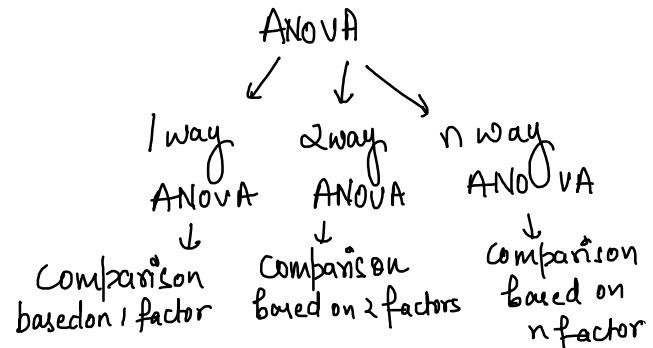
# ANOVA

Wednesday, November 29, 2023 8:08 AM



→ extension of z-test/t-test

→ variances (std dev) are analyzed



$\rightarrow$  Fischer  
F-statistic

① → It is used to compare variances b/w groups

$$\textcircled{II} \rightarrow F = \frac{\text{variance of 1st group}}{\text{variance of 2nd group}} = \frac{SD_1^2}{SD_2^2} \Rightarrow SD_1 > SD_2$$

ANOVA:

One Way ANOVA

Q1: To assess the significance of possible variation in performance in a certain test between the convent schools of a city, a common test was given to a number of students taken at random from the 5<sup>th</sup> class of the 3 schools concerned. The result is given as follows:

| A  | B  | C  |
|----|----|----|
| 9  | 13 | 14 |
| 11 | 12 | 13 |
| 13 | 10 | 17 |
| 9  | 15 | 7  |
| 8  | 5  | 9  |

Make the Analysis of Variance of the given data. (Null Hypo: No Significance Variation in the schools).

Solution:

Null Hypothesis = No variation between schools  
Alt. Hypothesis = There is variation between schools

$$n = 15$$

$$C = 3$$

$$\bar{X}_A = \frac{50}{5} = 10$$

$$\bar{X} = \frac{10+11+12}{3}$$

$$\bar{X}_B = \frac{55}{5} = 11$$

$$\bar{X} = \frac{33}{3} = 11$$

$$\bar{X}_C = \frac{60}{5} = 12$$

| Source of Variation | Sum of Square | Degrees of freedom | Mean Square                   | F                            |
|---------------------|---------------|--------------------|-------------------------------|------------------------------|
| Between the Sample  | SSC = 10      | (c-1) = 2          | MSC = SSC/df1 = 5             | F = MSC/MSE                  |
| Within the sample   | SSE = 138     | (n-c) = 12         | MSE = SSE/df2 = 138/12 = 11.5 | $F = \frac{5}{11.5} = 0.435$ |

$$\begin{array}{ccccccc}
 \underline{\underline{SSC}} & \bar{x}_A - \bar{x} & (x_A - \bar{x})^2 & (\bar{x}_B - \bar{x}) & (x_B - \bar{x})^2 & (\bar{x}_C - \bar{x}) & (\bar{x}_C - \bar{x})^2 \\
 & |0-1| = -1 & 1 & |1-1|=0 & 0 & |2-1|=1 & 1 \\
 & |0-1| = -1 & 1 & |1-1|=0 & 0 & |2-1|=1 & 1 \\
 & |0-1| = -1 & 1 & |1-1|=0 & 0 & |2-1|=1 & 1 \\
 & |0-1| = -1 & 1 & |1-1|=0 & 0 & |2-1|=1 & 1 \\
 & |0-1| = -1 & 1 & |1-1|=0 & 0 & |2-1|=1 & 1 \\
 & & \underline{\underline{5}} & & \underline{\underline{0}} & & \underline{\underline{5}}
 \end{array}$$

$$SSC = 5 + 0 + 5 = 10$$

$$\text{Degrees of freedom} = C-1 = 3-1 = 2$$

$$MSE = \frac{\underline{\underline{SSC}}}{\text{Degrees of freedom}} = \frac{10}{2} = 5$$

Within Sample:

$$\begin{array}{ccccccc}
 \underline{\underline{SSE}} & A - \bar{x}_A & (A - \bar{x}_A)^2 & (B - \bar{x}_B) & (B - \bar{x}_B)^2 & (C - \bar{x}_C) & (C - \bar{x}_C)^2 \\
 & 9 - 10 = -1 & 1 & 13 - 11 = 2 & 4 & 14 - 12 = 2 & 4 \\
 & 11 - 10 = 1 & 1 & 12 - 11 = 1 & 1 & 13 - 12 = 1 & 1 \\
 & 13 - 10 = 3 & 9 & 10 - 11 = -1 & 1 & 17 - 12 = 5 & 25 \\
 & 9 - 10 = -1 & 1 & 15 - 11 = 4 & 16 & 7 - 12 = -5 & 25 \\
 & 8 - 10 = -2 & 4 & 5 - 11 = 6 & \underline{\underline{36}} & 9 - 12 = -3 & 9 \\
 & & \underline{\underline{16}} & & \underline{\underline{58}} & & \underline{\underline{64}}
 \end{array}$$

$$SSE = 16 + 58 + 64 = 138 \quad \left. \right\} MSE = 138/12 = 11.5$$

$$f_{cal} = 0.435 ; \quad f_{tab} \rightarrow \begin{cases} \stackrel{df_1}{j_1} = 2 \text{ (smaller value)} \\ \stackrel{df_2}{j_2} = 12 \text{ (higher value)} \end{cases}$$

$$f_{tab} = 3.89$$

Compare  $f_{cal}$  with  $f_{tab}$ :

$$f_{cal} < f_{tab}$$

$$0.435 < 3.89$$

Failed to Reject  $H_0$

## 2-way ANOVA

The following data represents the number of Units of Tablet production (in thousands) per day by five different technicians by using 4 different machines.

- Tell whether the mean productivity of the different machines are same?
- Test whether the 5 technicians differ w.r.t. the mean productivity?

| Machines →           | A  | B  | C       | D       |
|----------------------|----|----|---------|---------|
| .....<br>Technicians |    |    |         |         |
| P ↓                  | 54 | 48 | 57 - 50 | 46 - 50 |
| Q                    | 56 | 50 | 62 - 50 | 53 - 50 |
| R                    | 44 | 46 | 54 - 50 | 42 - 50 |
| S                    | 53 | 48 | 56 - 50 | 44 - 50 |
| T                    | 48 | 52 | 59 - 50 | 48 - 50 |

$$MSC = \frac{338.8}{3} = 112.93$$

$$MSR = \frac{158}{4} = 39.5$$

$$MSE = \frac{67.2}{12} = 5.6$$

| Source of Variance  | Sum of Squares | Degree of Freedom      | Mean sum of squares    | F                 |
|---------------------|----------------|------------------------|------------------------|-------------------|
| Between the columns | $SSC = 338.8$  | $df = c-1 = 3$         | $MSC = SSC/(c-1)$      | $MSC/MSE = 20.16$ |
| Between the rows    | $SSR = 158$    | $df = r-1 = 4$         | $MSR = SSR/(r-1)$      | $MSR/MSE = 7.653$ |
| Residual Errors     | $SSE = 67.2$   | $df = (c-1)(r-1) = 12$ | $MSE = SSE/(c-1)(r-1)$ |                   |
| Total Sum of Square | $SST = 564$    | $df = n-1$             |                        |                   |

Sol ① → Calculate Grand total

$$\text{Mid value} \Rightarrow \frac{42+62}{2} = 52 \approx 50$$

A

B

C

D

Total



|       | A              | B               | C               | D                | Total                                |
|-------|----------------|-----------------|-----------------|------------------|--------------------------------------|
| P     | $54 - 50 = 4$  | $48 - 50 = -2$  | 7               | -4               | $\sum S$                             |
| Q     | $56 - 50 = 6$  | $60 - 50 = 0$   | 12              | 3                | 21                                   |
| R     | $44 - 50 = -6$ | $46 - 50 = -4$  | 4               | -8               | -14                                  |
| S     | $53 - 50 = 3$  | $48 - 50 = -2$  | 6               | -6               | 1                                    |
| T     | $48 - 50 = -2$ | $52 - 50 = 2$   | 9               | -2               | 7                                    |
| Total | $\overline{5}$ | $\overline{-6}$ | $\overline{38}$ | $\overline{-17}$ | $\frac{\text{Grand Total}}{20} = 20$ |

Correction factors =  $\frac{T^2}{N} = \frac{(\text{Grand Total})^2}{N} = \frac{20 \times 20}{20} = 20$

$$\underline{\underline{SSC}} = \left( \frac{\sum A}{n_A} \right)^2 + \left( \frac{\sum B}{n_B} \right)^2 + \left( \frac{\sum C}{n_C} \right)^2 + \left( \frac{\sum D}{n_D} \right)^2 - \frac{T^2}{N}$$

$$= \frac{(5)^2}{5} + \frac{(-6)^2}{5} + \frac{38^2}{5} + \frac{(-17)^2}{5} - \frac{20^2}{20}$$

$$= 338.8$$

$$\underline{\underline{SSR}} \Rightarrow \left( \frac{\sum P}{n_P} \right)^2 + \left( \frac{\sum Q}{n_Q} \right)^2 + \left( \frac{\sum R}{n_R} \right)^2 + \left( \frac{\sum S}{n_S} \right)^2 + \left( \frac{\sum T}{n_T} \right)^2 - \frac{T^2}{N}$$

$$= \frac{5^2}{4} + \frac{21^2}{4} + \frac{(-14)^2}{4} + \frac{1^2}{4} + \frac{7^2}{4} - \frac{20^2}{20}$$

$$= 158$$

SST = Sum of Squares of all observation residuals -  $\frac{T^2}{N}$

$$= 4^2 + 6^2 + (-6)^2 + 3^2 + (-2)^2 + \dots - + (-6)^2 + (-2)^2 - \frac{20^2}{20}$$

$$= 564$$

$$\underline{\underline{SSE}} = SST - (SSC + SSR) = 564 - (338.8 + 158)$$

$$= 67.2$$

Test a)  $F_{cal} = 20.16$

b)  $F_{cal} = 7.05$

$$F_{tab} (\alpha = 0.05) = 3.49$$

$$F_{tab} (\alpha = 0.05) = 3.26$$

$$v_1 = 3$$

$$v_1 = 4$$

$$v_2 = 12$$

$$v_2 = 12$$

$$F_{tab} < F_{cal}$$

$$F_{tab} < F_{cal}$$

Yes, there is significant  
in .

Yes, there is significant  
in .

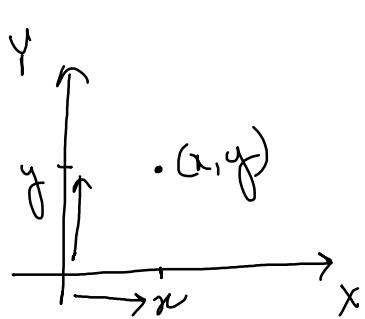
Yes, there is significant  
variation between the columns

Rejected H<sub>0</sub>

Yes, there is significant  
variation b/w the rows

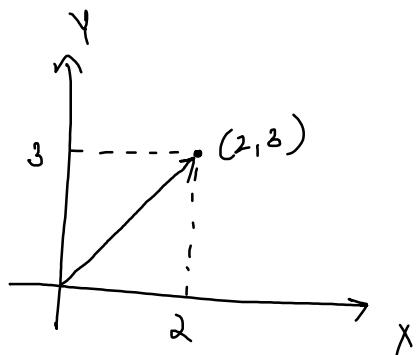
Rejected H<sub>0</sub>

# Linear Algebra



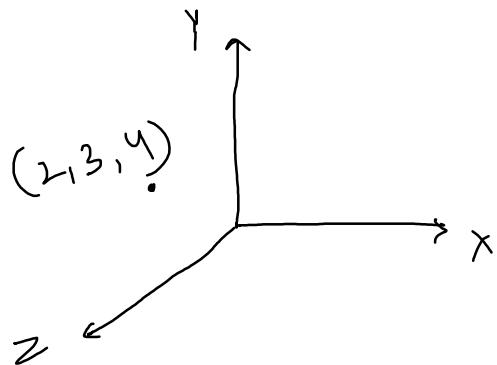
Quantities

- Scalar  $\downarrow$  Magnitude
- Vector  $\downarrow$  Magnitude & direction



$\begin{array}{c} + \\ \overrightarrow{AB} \\ - \end{array}$  1 km  
 $\begin{array}{c} \uparrow \\ (+, -) \end{array}$   
 = distance =  $1+1 = 2 \text{ km}$   
 = displacement =  $1-1 = 0 \text{ km}$

$$\text{Vectors} = [2 \ 3] \Rightarrow 2d$$



$$\text{Vector} = [2 \ 3 \ 4] \Rightarrow 3d$$

↓

$$\text{Vector} = [2 \ 3 \ 4 \ 1 \ 5 \ 6] \Rightarrow 6 \text{ dimensions}$$

↓

vector in  $n$ d =  $[2 \ 3 \ 4 \ \dots \ n] \Rightarrow n$  dimensions.

MATRIX  $\Rightarrow$  it is a table of numbers.

cols

Rows  $\rightarrow$

$$\begin{bmatrix} 1 & 2 & 3 & 4 \\ 5 & 6 & 7 & 8 \\ 9 & 10 & 11 & 12 \\ 13 & 14 & 15 & 16 \end{bmatrix} \quad 4 \times 4$$



Addition

$$\begin{bmatrix} a & b \\ c & d \end{bmatrix} + \begin{bmatrix} e & f \\ g & h \end{bmatrix} \Rightarrow \begin{bmatrix} a+e & b+f \\ c+g & d+h \end{bmatrix}$$

$\Leftrightarrow$

$$\begin{bmatrix} 1 & 2 & 3 \\ 5 & 6 & 7 \end{bmatrix} + \begin{bmatrix} 2 & 4 & 8 \\ 9 & 10 & 11 \end{bmatrix} = \begin{bmatrix} 3 & 6 & 11 \\ 14 & 16 & 18 \end{bmatrix}$$

Multiplication

$\begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix} \begin{bmatrix} 2 \\ 1 \end{bmatrix} \Rightarrow [1 \times 2 + 2 \times 1]_{1 \times 1} \Rightarrow [4]$

$$\begin{matrix} 1 \times 2 \\ | & | \\ 1 & 1 \end{matrix} \xrightarrow{\quad} 1 \times 1$$

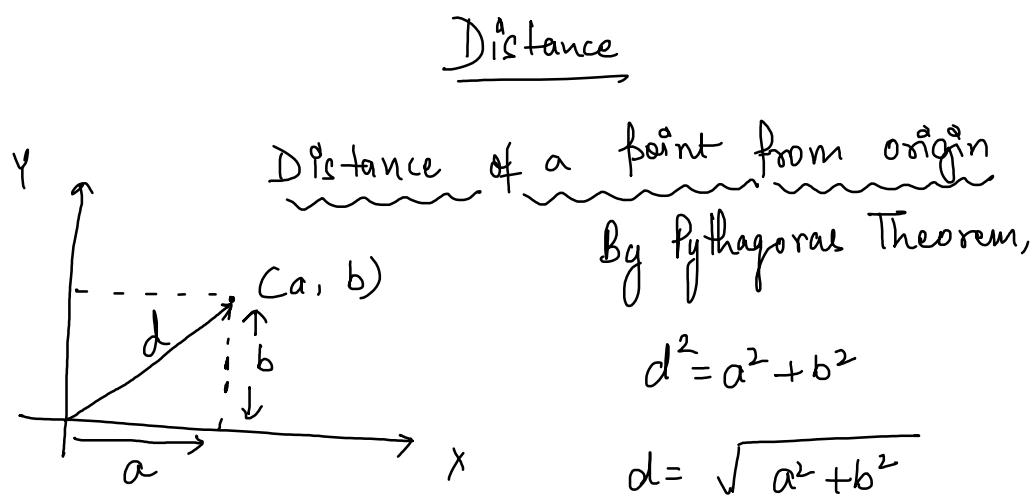
$$\begin{bmatrix} a & b \\ c & d \end{bmatrix}_{2 \times 2} \times \begin{bmatrix} e & f \\ g & h \end{bmatrix}_{2 \times 2} \Rightarrow \begin{bmatrix} ae+gb & af+bh \\ ce+dg & cf+dh \end{bmatrix}_{2 \times 2}$$

In order to perform matrix multiplication,

no. of columns in first matrix = no. of rows in Second Matrix

a)  $a_{m \times n} \times b_{p \times q} \Rightarrow$  No

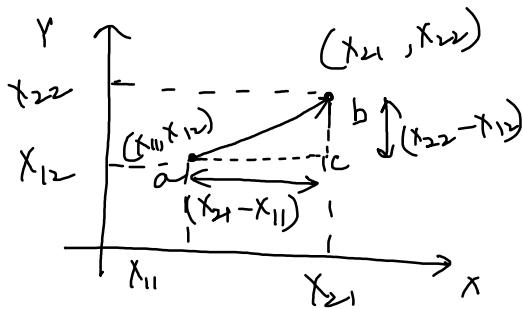
b)  $a_{m \times n} \times b_{n \times q} \Rightarrow C_{m \times q}$



lets extend this idea to n dimension

distance,  $d = \sqrt{a^2 + b^2 + c^2 + d^2 + \dots + n^2}$

Distance b/w two points :



$$a = [x_{11} \ x_{12}]$$

$$b = [x_{21} \ x_{22}]$$

By Pythagoras Theorem,

$$d^2 = (x_{21} - x_{11})^2 + (x_{22} - x_{12})^2$$

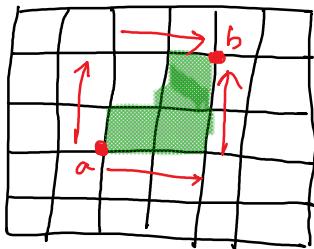
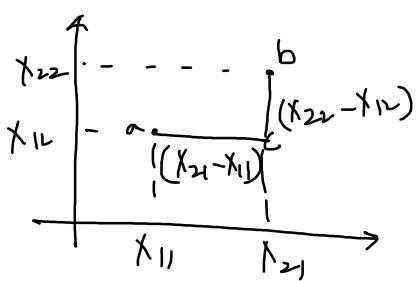
$$d = \sqrt{(x_{21} - x_{11})^2 + (x_{22} - x_{12})^2} = \sqrt{(x_{11} - x_{21})^2 + (x_{12} - x_{22})^2}$$

~~REMARK~~

Euclidean distance =  $\left[ \sum_{i=1}^n |x_{1i} - x_{2i}|^2 \right]^{1/2}$

→ L2 Norm

Manhattan Distance



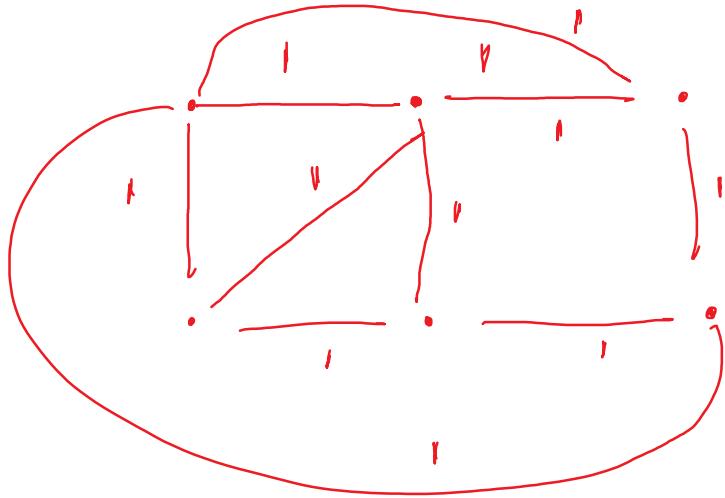
$$d = |x_{21} - x_{11}| + |x_{22} - x_{12}|$$

$$d = |x_{11} - x_{21}| + |x_{12} - x_{22}|$$

~~REMARK~~

~~REDEF~~

$$d = \sum_{i=1}^n |x_{1i} - x_{2i}| \rightarrow L_1 \text{ Norm}$$



when you have high dimensional data, use Manhattan distance.

### Minkowski distance

$$d = \left[ \sum_{i=1}^n |x_{1i} - x_{2i}|^p \right]^{\frac{1}{p}} \rightarrow L_p \text{ Norm}$$

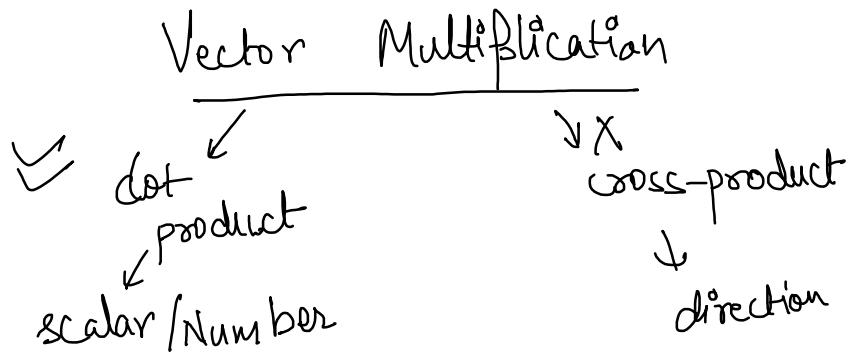
$p = 1, 2, 3, \dots$

lets put  $p=1$ ,

$$d = \sum_{i=1}^n |x_{1i} - x_{2i}| \rightarrow \text{Manhattan distance}$$

lets put  $p=2$ ,  $d = \left[ \sum_{i=1}^n |x_{1i} - x_{2i}|^2 \right]^{\frac{1}{2}} \rightarrow \text{Euclidean distance}$

lets put  $p=2$ ,  $d = \left[ \sum_{i=1}^n |x_i^o - x_2^o| \right] \Rightarrow$  Euclidean distance



dot product in linear algebra,

$$a = [a_1, a_2, a_3, \dots, a_n]_{1 \times n} = a^T$$

$$b = [b_1, b_2, b_3, \dots, b_n]_{1 \times n}$$

$$a \cdot b = [a_1 b_1 + a_2 b_2 + a_3 b_3 + \dots + a_n b_n]_{1 \times 1}$$

Vectors Representation

default  $\leftarrow$  Column vector

$$\begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_n \end{bmatrix}$$

Row vector

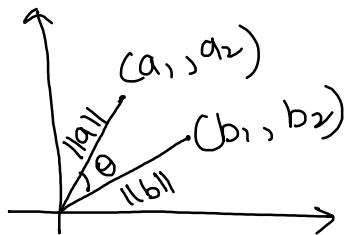
$$[a_1, \dots, a_n]$$

$$\vec{a} = [a_1, a_2, a_3, \dots, a_n]_{1 \times n} \quad b = \begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_n \end{bmatrix}$$

$$\vec{a} = [a_1 \ a_2 \ a_3 \ \dots \ a_n]_{n \times 1} \quad b = \begin{bmatrix} b_1 \\ b_2 \\ b_3 \\ \vdots \\ b_n \end{bmatrix}_{n \times 1}$$

$$\vec{a}^T \cdot b = [a_1 b_1 + a_2 b_2 + a_3 b_3 + \dots + a_n b_n]_{1 \times 1}$$

Angle b/w vectors



(Geometric)  
dot product

$$a \cdot b = ||a|| \cdot ||b|| \cos \theta$$

$$\vec{a}^T \cdot b = [a_1 b_1 + a_2 b_2] = \frac{a \cdot b}{\text{(linear algebra way)}}$$

$$||a|| \cdot ||b|| \cos \theta = [a_1 b_1 + a_2 b_2]$$

$$\cos \theta = \frac{[a_1 b_1 + a_2 b_2]}{||a|| \cdot ||b||}$$

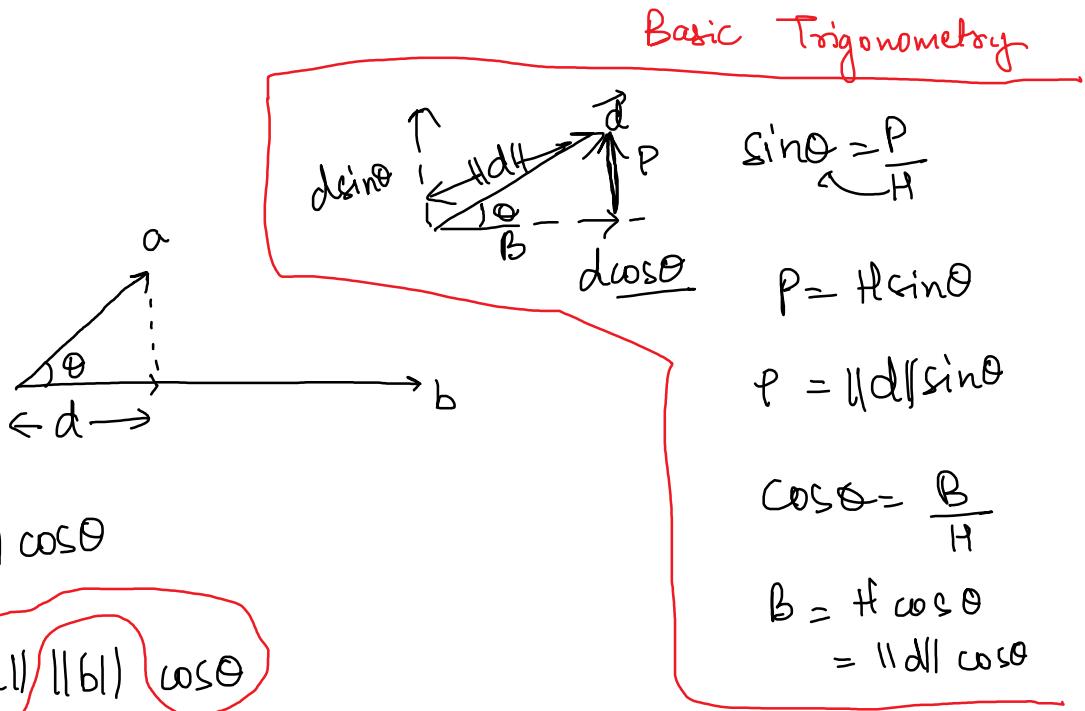
\*

$$\theta = \cos^{-1} \frac{[a_1 b_1 + a_2 b_2]}{||a|| \cdot ||b||}$$

where,  $\|a\| = \sqrt{a_1^2 + a_2^2 + \dots + a_n^2}$

$$\|b\| = \sqrt{b_1^2 + b_2^2 + \dots + b_n^2}$$

Projection



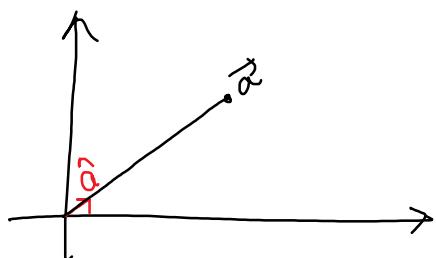
$$a \cdot b = d \|b\|$$

$$d = \frac{a \cdot b}{\|b\|}$$

Projection of vector  $a$  on  $b$

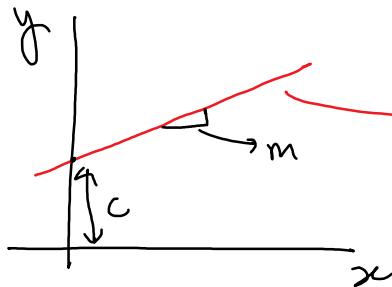
Unit Vector → vector with magnitude 1  
→ gives information about direction

$$\hat{a} = \frac{\vec{a}}{\|a\|}$$



## Lines & Planes

### Line



$c=0$  (line passing through origin)

$$y = mx + c$$

$$\text{Slope}(m) = \frac{y_2 - y_1}{x_2 - x_1} = \tan\theta = \frac{\Delta y}{\Delta x}$$

simple geometry      trigonometrical geometry      calculus

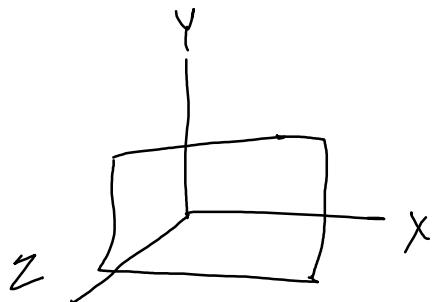
General Equation of line,

$$= [ax + by + c = 0]$$

$$by = -ax - c$$

$$y = -\frac{a}{b}x - \frac{c}{b}$$

### Plane



General Equation is :  $ax + by + cz + d = 0$   
 ↓ change coeff

$$w_1x + w_2y + w_3z + w_0 = 0$$

↓ change axis name

$$w_1x_1 + w_2x_2 + w_3x_3 + w_0 = 0$$

Above 3d  $\circ$  hyperplane  $\circ$ .  $w_0 + [w_1x_1 + w_2x_2 + \dots + w_nx_n] = 0$

$$w_0 + \mathbf{w}^\top \mathbf{x} = 0 \quad (\text{linear algebra way})$$

Let's say hyperplane is passing through origin,

$$\boxed{\mathbf{w}^\top \mathbf{x} = 0} \quad *w_0 = 0$$

Eigen value & eigen vector → vectors that do not rotate when linear transformation is applied on them

value by which  
the original vector  
gets scaled up

$$A \vec{x} = \lambda \vec{x}$$

Matrix that applies LT on vector

eigen vector

eigen value

# KNN

Thursday, December 7, 2023 8:00 AM

## K-Nearest Neighbors

↳ You are like your neighbors

KNN

$k=5$

$X_q(3B, 2R) \Rightarrow \text{Blue} \Leftarrow \text{color of } X_q$



Color of  $X_q = (3B, 4R) \Rightarrow \text{Red}$

$K \Rightarrow$  Hyperparameter  $\Rightarrow \# \downarrow$  neighbors  
number of

$K=3$

$P(R) = \frac{2}{3}$        $P(B) = \frac{1}{3}$   
Color of  $X_q = (2R, 1B) = \text{Red}$

Now,

$K=4$       Color of  $X_q = (2R, 2B) = \text{No voting}$

Q Why can't  $g$  have even number?

Sol  $K \neq 2, 4, 6, 8 \dots$

$K=4 \Rightarrow \text{color of } X_q = 2R, 2B$

$P(R) = \frac{1}{2} = 0.5$  ( $g$  don't know)

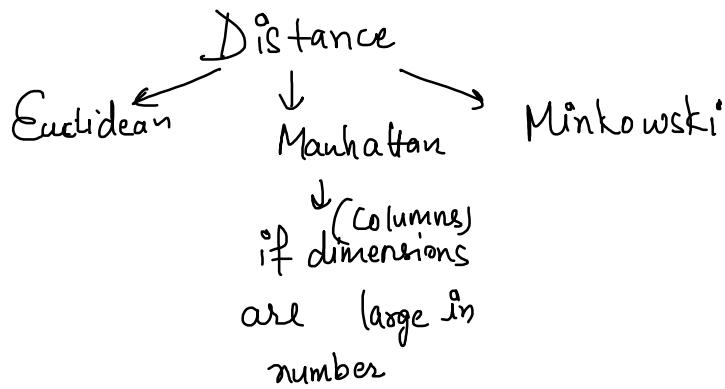
- always keep value of  $K$  as odd!

Q  $K=3, K=5, K=7, ?$  Pick wrong value of  $K$

$K=6$  → wrong value

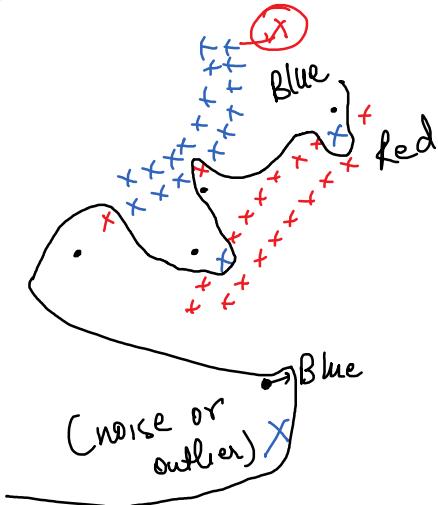
$K \Rightarrow$  Hyperparameter  $\rightarrow$  Tune your model  $\rightarrow K \uparrow / K \downarrow \rightarrow$  to get the best result

$$K = [3, 5, 7, 9, 11, \dots]$$



### Effect of $K$ :

$K=1$



- $\Rightarrow$  decision surface is not smooth
  - $\Rightarrow$  working well in training
  - $\Rightarrow$  working poorly in testing
- $\Rightarrow$  overfitting
- Training accuracy  $\Rightarrow 99\% \text{ or } 100\%$
- Test accuracy  $\Rightarrow 60\% \text{ or } 50\%$

$$K = n \quad (\# \text{ datapoints})$$

count(Blue) > count(Red)



$\Rightarrow$  working poorly in training  
 $\Rightarrow$  " " " in testing

$\rightarrow$  It is always blue.

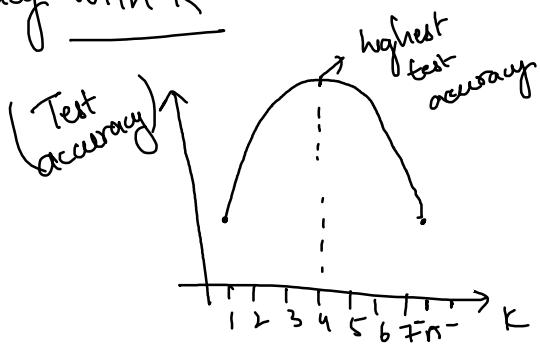
### Overfitting

- $\rightarrow$  Training accuracy is very high
- $\rightarrow$  Test accuracy is low

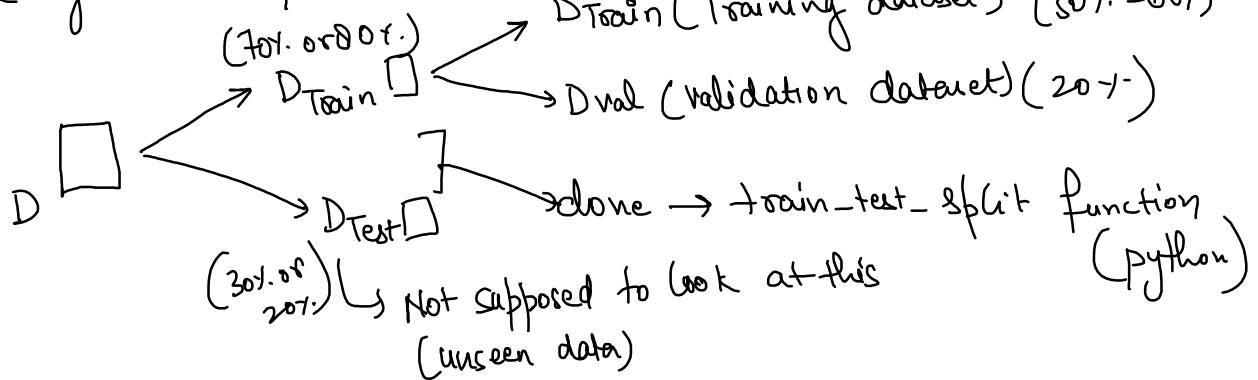
### Underfitting

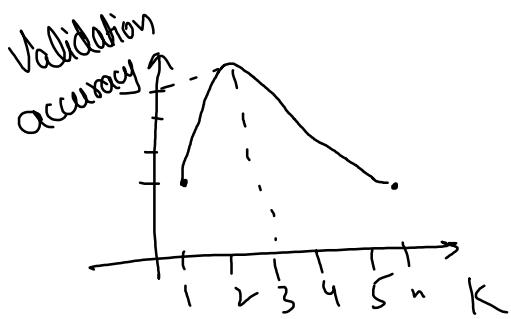
- $\rightarrow$  Training accuracy is low
- $\rightarrow$  Test accuracy is low

### Curve of accuracy with K



Choose the right value of K:





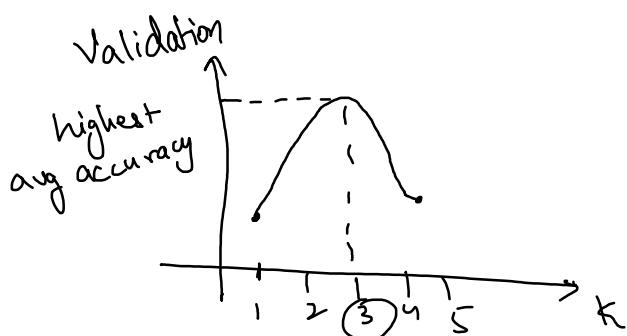
CROSS-VALIDATION → to keep test data hidden/unseen for tuning hyperparameters  
 (K-Fold)                     $K' = 4$  ,  $K = \# \text{ neighbors}$



| $K$   | Training        | Validation             |
|-------|-----------------|------------------------|
| 1     | $D_1, D_2, D_3$ | $D_4 \rightarrow a_1$  |
| 1     | $D_2, D_3, D_4$ | $D_1 \rightarrow a_2$  |
| 1     | $D_1, D_3, D_4$ | $D_2 \rightarrow a_3$  |
| 1     | $D_1, D_2, D_3$ | $D_3 \rightarrow a_4$  |
| <hr/> |                 |                        |
| 2     | $D_1, D_2, D_3$ | $D_4 \rightarrow a'_1$ |
| 2     | $D_2, D_3, D_4$ | $D_1 \rightarrow a'_2$ |
| 2     | $D_1, D_3, D_4$ | $D_2 \rightarrow a'_3$ |
| 2     | $D_1, D_2, D_3$ | $D_3 \rightarrow a'_4$ |

$a_{\text{mean}}$

$a'_{\text{mean}}$

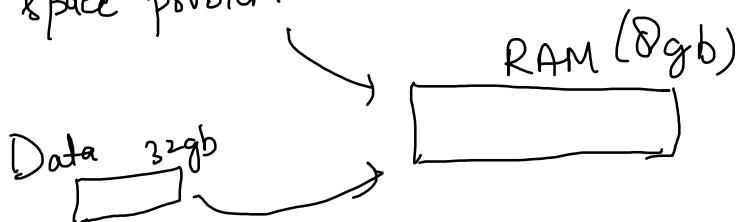


$\boxed{K=3}$

\* choose the value of  $k$  which has highest accuracy.  
 $\text{KNN}(\text{n-neighbors} = 3, \text{distance} = \text{'euclidean'}) = \text{model}$

### Disadvantages:

- 1 → lazy learner → calculations are done at time of execution
- 2 → sensitive to outlier/noise
- 3 → time complexity → takes a lot of time to execute
- 4 → space problem



### Application → Healthcare

### Evaluation Metrics (Classification)

#### CONFUSION MATRIX

|           |         | Actual  |         |
|-----------|---------|---------|---------|
|           |         | 0 (-ve) | 1 (+ve) |
| Predicted | 0 (-ve) | TN      | FN      |
|           | 1 (+ve) | FP      | TP      |

Accuracy  $\Rightarrow \frac{TP + TN}{TP + FP + TN + FN} \Rightarrow$  In some cases it is not reliable  
 $\rightarrow$  Class imbalance  
 (Blue > Red)

$$TP + FP + TN + FN \rightarrow \text{Class Imbalance} \\ (\text{Blue} > \text{Red}) \\ \rightarrow P = 0.5$$

$$\uparrow \underline{\text{Precision}} \Rightarrow \frac{TP}{TP + FP} \downarrow$$

$$\uparrow \underline{\text{Recall}} \Rightarrow \frac{TP}{TP + FN} \Rightarrow \text{Sensitivity (TPR)}$$

$$* \text{F1-Score} \Rightarrow \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

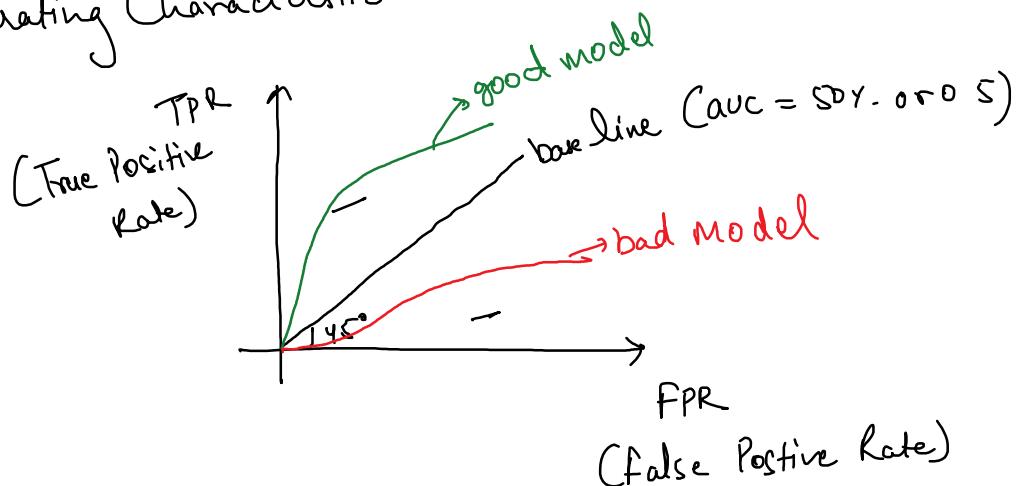
$$\underline{\text{Specificity}} \Rightarrow \frac{TN}{TN + FP}$$

$$fPR = \frac{FP}{TN + FP}$$

$$fPR = (1 - \text{specificity})$$

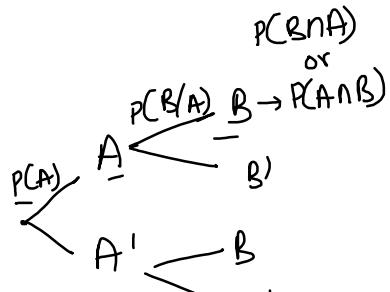
ROC-AUC curve  $\xrightarrow{\text{Area under Curve}}$

Receiver Operating Characteristic



Bayes Theorem

$$P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{P(B \cap A)}{P(B)}$$



$$P(A|B) = \frac{P(A) \times P(B|A)}{P(B)} \rightarrow \text{Law of Total Prob}$$

Posterior  
 ↑      likelihood  
 $P(A|B) = \frac{P(B|A) \times P(A)}{P(B)} \rightarrow \text{Posterior}$   
 ↓      Evidence

\* Probability & likelihood?  $\Rightarrow$  What's the difference?

Let's  $h$  = height , ( $\mu = 160 \text{ cm}, \sigma = 1$ )

Probability  $H > 170 \text{ cm}$

$P(H > 170 \text{ cm} | \mu = 160 \text{ cm}, \sigma = 1)$  = quantification of height being greater than 170 cm.

Likelihood

$P(\mu = 160 \text{ cm}, \sigma = 1 | H > 170 \text{ cm})$   $\Rightarrow$  getting best distribution for your observation

Naive Bayes  
 (ignorant)

Assumptions:

↳

### Assumptions:

- 1 → All features are independent of each other.
- 2 → All features have equal contribution in predicting the op.

$$P(C_x/x_i) = \frac{P(X_i/C_x) \times P(C_x)}{P(X_i)}$$

$C_x$  → class labels

$X_i$  → g/p variables

$$P(\text{Yes} / O, T, H, W) = ?$$

$$P(\text{No} / O, T, H, W) = ?$$

| Outlook  | Temperature | Humidity | Windy   | PlayTennis |
|----------|-------------|----------|---------|------------|
| Sunny    | Hot         | High     | False X | No         |
| Sunny    | Hot         | High     | True    | No         |
| Overcast | Hot         | High     | False ✓ | Yes        |
| Rainy    | Mild        | High     | False ✓ | Yes        |
| Rainy    | Cool        | Normal   | False ✓ | Yes        |
| Rainy    | Cool        | Normal   | True    | No         |
| Overcast | Cool        | Normal   | True    | Yes        |
| Sunny    | Mild        | High     | False X | No         |
| Sunny    | Cool        | Normal   | False ✓ | Yes        |
| Rainy    | Mild        | Normal   | False ✓ | Yes        |
| Sunny    | Mild        | Normal   | True    | Yes        |
| Overcast | Mild        | High     | True    | Yes        |
| Overcast | Hot         | Normal   | False ✓ | Yes        |
| Rainy    | Mild        | High     | True    | No         |

$$\text{if } P(\text{Yes} / O, T, H, W) > P(\text{No} / O, T, H, W)$$

then g will play  
else: No play

$$\star P(\text{Yes} / \text{outlook}) = \frac{P(\text{outlook} | \text{Yes}) P(\text{Yes})}{P(\text{outlook})}$$

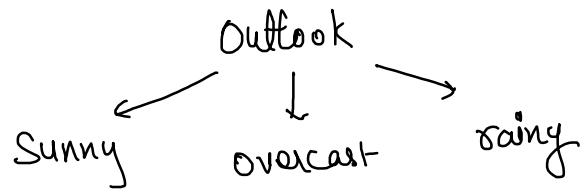
# ignore the denominator.

$$\star P(\text{No} / \text{outlook}) = \frac{P(\text{outlook} | \text{No}) P(\text{No})}{P(\text{outlook})}$$

$$P(\text{Yes}) = ? \quad | \quad P(\text{No}) = ? \quad | \quad P(\text{outlook} | \text{Yes}) = ? \quad | \quad P(\text{outlook} | \text{No}) = ?$$

Working:  $P(\text{Yes}) = \frac{9}{14}$

$$P(\text{No}) = \frac{5}{14}$$



$$P(\text{Sunny} | \text{Yes}) = \frac{2}{9}$$

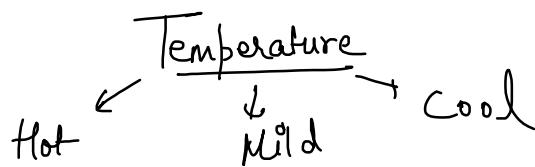
$$P(\text{Sunny} | \text{No}) = \frac{3}{5}$$

$$P(\text{Rainy} | \text{Yes}) = \frac{3}{9}$$

$$P(\text{Rainy} | \text{No}) = \frac{2}{5}$$

$$P(\text{overcast} | \text{Yes}) = \frac{4}{9}$$

$$P(\text{overcast} | \text{No}) = 0$$



$$P(\text{Hot} | \text{Yes}) = \frac{2}{9}$$

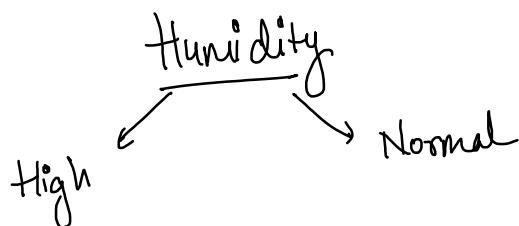
$$P(\text{Hot} | \text{No}) = \frac{2}{5}$$

$$P(\text{Mild} | \text{Yes}) = \frac{4}{9}$$

$$P(\text{Mild} | \text{No}) = \frac{2}{5}$$

$$P(\text{cool} | \text{Yes}) = \frac{3}{9}$$

$$P(\text{cool} | \text{No}) = \frac{1}{5}$$



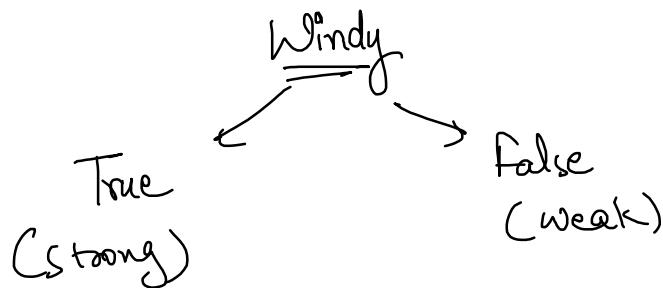
$$P(\text{High} | \text{Yes}) = \frac{3}{9}$$

$$P(\text{High} | \text{No}) = \frac{4}{5}$$

$$P(\text{High} | \text{Yes}) = \frac{1}{9}$$

$$P(\text{Normal} | \text{Yes}) = \frac{6}{9}$$

$$P(\text{Normal} | \text{No}) = \frac{1}{5}$$



$$P(\text{True} | \text{Yes}) = \frac{3}{9}$$

$$P(\text{True} | \text{No}) = \frac{3}{5}$$

$$P(\text{False} | \text{Yes}) = \frac{6}{9}$$

$$P(\text{False} | \text{No}) = \frac{2}{5}$$

Q Outlook is overcast, temp is cool, humidity is high & wind is strong. Will g play?

$$\begin{aligned}
 \text{Sol. } P(\text{Yes} | \text{overcast, cool, high, strong}) &= P(\text{overcast/Yes}) \times P(\text{cool/Yes}) \times P(\text{high/Yes}) \\
 &\quad \times P(\text{strong/Yes}) \times P(\text{Yes}) \\
 &= \frac{4}{9} \times \frac{3}{9} \times \frac{2}{9} \times \frac{5}{9} \times \frac{9}{14} = 0.010
 \end{aligned}$$

$$\begin{aligned}
 P(\text{No} | \text{overcast, cool, high, strong}) &= P(\text{overcast/No}) P(\text{cool/No}) P(\text{high/No}) \\
 &\quad P(\text{strong/No}) P(\text{No}) \\
 &= 0
 \end{aligned}$$

Will g play?  $\Rightarrow$  Yes!

Q outlook is sunny, temp is cool, humidity is high, wind is strong.  
Will g play?

$$\begin{aligned} \text{Sol } P(\text{Yes}/\text{sunny, cool, high, strong}) &= P(\text{sunny}/\text{yes}) \times P(\text{cool}/\text{Yes}) \times P(\text{high}/\text{Yes}) \times \\ &\quad \underbrace{\dots}_{P(\text{strong}/\text{Yes})} \times p(\text{Yes}) \\ &= \frac{2}{9} \times \frac{3}{7} \times \frac{3}{9} \times \frac{1}{7} = \frac{1}{7 \times 27} = \frac{1}{189} = 0.0052 \end{aligned}$$

$$P(\text{No}/\text{sunny, cool, high, strong}) = \frac{8}{14} \times \frac{3}{5} \times \frac{1}{5} \times \frac{4}{5} \times \frac{3}{5} = \frac{18}{875} = 0.02$$

$$0.02 > 0.0052 =$$

$$P(\text{No}/\text{sunny, cool, high, strong}) \quad \swarrow \quad P(\text{Yes}/\text{sunny, cool, high, strong})$$

No g will not play!

Problem of Zero Probability:

$$\begin{aligned} Q \quad P(\text{Yes}/\text{cloudy, cool, high, weak}) &= P(\text{cloudy}/\text{Yes}) \times P(\text{cool}/\text{Yes}) \times \\ &\quad P(\text{high}/\text{Yes}) \times P(\text{weak}/\text{Yes}) \times P(\text{Yes}) = 0 \end{aligned}$$

$$\begin{aligned} P(\text{No}/\text{cloudy, cool, high, weak}) &= P(\text{cloudy}/\text{No}) \times P(\text{cool}/\text{No}) \times P(\text{high}/\text{No}) \\ &\quad \times P(\text{weak}/\text{No}) \times P(\text{No}) = 0 \end{aligned}$$

## \* Laplace Smoothing

$\downarrow$   
(Var Smoothing)

$$P(Y_{\text{es}}|C) = \frac{0 + \alpha \rightarrow \text{a very small no}}{n + \alpha k}$$

↑  
# datapoints  
↓  
(no of values your feature can take)

$$P(Y_{\text{es}}|C) = P(C|Y_{\text{es}})P(Y_{\text{es}}) = \frac{0 + \alpha}{n + \alpha k}$$

Range of  $\alpha = (1, \infty)$

$\lambda \rightarrow$  hyperparameter

Effect of  $\alpha$ :  $n=1000$ ,  $k=2$

①  $\lambda = 1$   $\Rightarrow P = \frac{0+1}{1000+2 \times 1} = \frac{1}{1002} \approx 0 \rightarrow$  overfitting

↑ high variance

②  $\lambda = 10000 \Rightarrow P = \frac{0+10000}{1000+2 \times 10000} = \frac{10000}{21000} \approx 0.5 \rightarrow$  underfitting

↑ high bias

$=$

Var-smoothing  $\Rightarrow$  —

$$\frac{0}{n+\alpha k} = \frac{0}{n} = 0$$

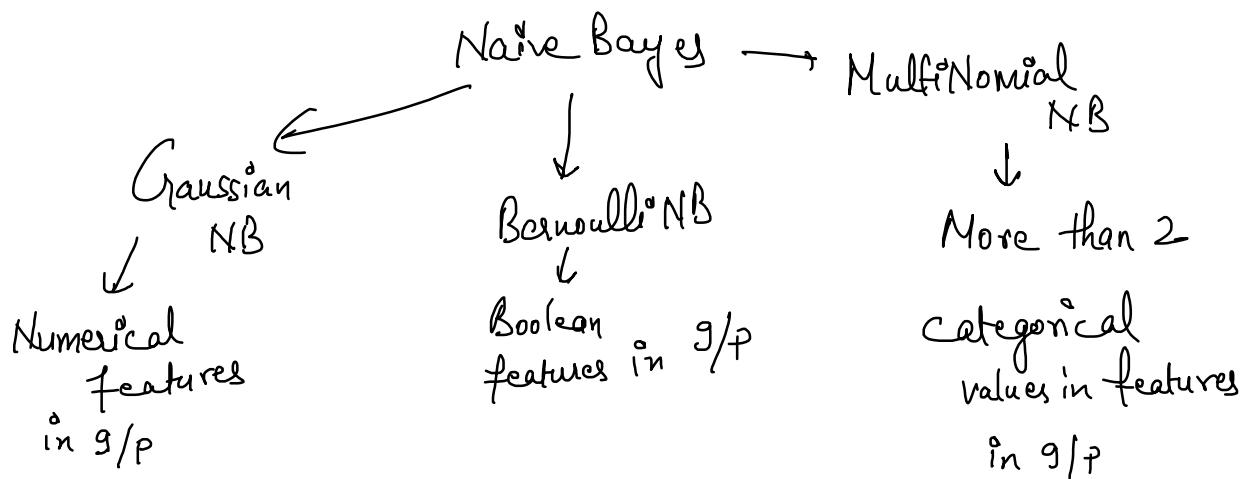
## Scenarios:

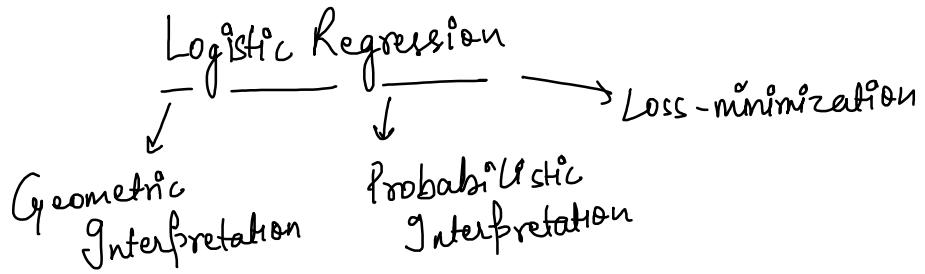
① g have a lot of dimensions:

↳ probabilities will be very low  $\rightarrow$  0

Take  $\downarrow$  log of probabilities

② X will take care of outliers.





→  $g_t$  is used for binary classification.

→ Assumption: → Data is linearly separable.

Equation of line:  $f(x) \leftarrow y = mx + c \rightarrow w_0$

$$y = w_0 + w_1 x_1$$

Equation of Plane:  $Ax + By + c = 0$

$$Ax + By + Cz + D = 0$$

$$D + Ax + By + Cz = 0$$

$$w_0 + w_1 x_1 + w_2 x_2 + w_3 x_3 = 0$$

$$w_0 + [w_1 x_1 + w_2 x_2 + w_3 x_3] = 0$$

$$w_0 + \sum_{i=1}^3 w_i x_i = 0$$

plane is passing through origin,

$$\sum_{i=1}^3 w_i x_i = 0 \quad \begin{bmatrix} w_1 & w_2 & w \end{bmatrix} \times \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$$

$$\sum_{i=1}^3 w_i x_i = 0 \rightarrow [w_1, w_2, w_3]^T \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix}_{3 \times 1}$$

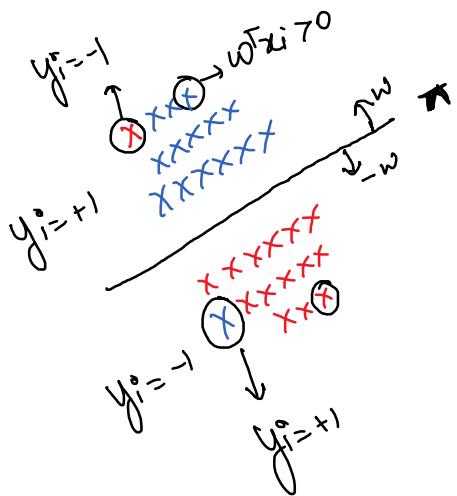
$w^o \Rightarrow$  weight vector

$x^i \Rightarrow$  datapoint vector

$$w^T x^i = 0 \rightarrow \text{for a point } i \text{ in matrix}$$

$$f(x) = w^T x$$

Geometric Interpretation  $\Rightarrow y_i = [+1, -1]$



①  $w^T x_i > 0$  for +ve class

②  $w^T x_i < 0$  for -ve class

Let's multiply  $y_i$  (class label) with  $w^T x_i$

$$\textcircled{a} \rightarrow y_i w^T x_i$$

Case 1:  $y_i = +1, w^T x_i = +ve$

$$\therefore y_i w^T x_i > 0$$

Case 2:  $y_i = -1, w^T x_i = -ve$

$$y_i w^T x_i > 0$$

Correct classification

Case 3  $y_i = -1$ ,  $w^T x_i = \text{true}$

$$\therefore y_i w^T x_i < 0$$

Case 4  $y_i = +1$ ,  $w^T x_i = \text{true}$

$$\therefore y_i w^T x_i < 0$$

incorrect classification

Case 1 & Case 2: correct classification

Case 3 & Case 4: incorrect classification.

if  $\begin{cases} y_i w^T x_i \geq 0 & ; \text{ correct classification} \\ y_i w^T x_i < 0 & ; \text{ incorrect classification} \end{cases}$

~~Mathematical~~  
Mathematical  
Objective function  $\Rightarrow \underset{w}{\operatorname{argmax}} (y_i w^T x_i) \Rightarrow y_i = (+, -)$

Diff b/w Argmax & Max?

$$f(x) = 2x ; x_1=1, x_2=2$$

Max ( $f(x)$ )



$$f(1) = 2$$

$$f(2) = 4 \rightarrow 0/P$$

Argmax ( $f(x)$ )

$$f(1) = 2$$

$$f(2) = 4$$

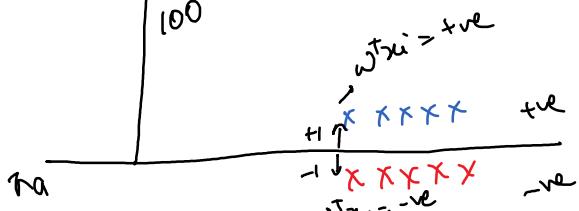
O/P

$\min \left[ \sum_{i=1}^n (y_i w^T x_i) \right]$   $\rightarrow$  it is very sensitive to outliers

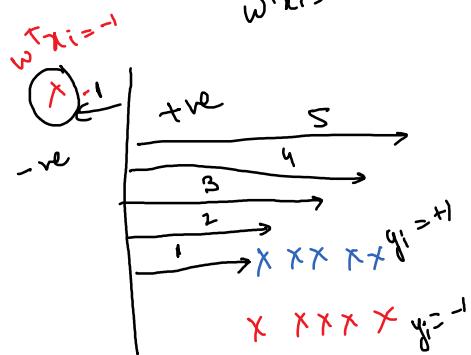
$$M_{OF} = \left[ \sum_{i=1}^n (y_i w^\top x_i) \right] \rightarrow \text{it is very sensitive to outliers}$$

outlier  

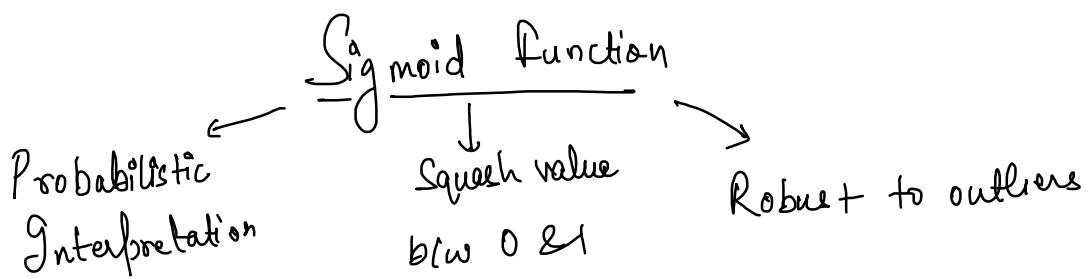

$$M_{OF} = +1+1+1+1+1 + 1+1+1+1+1 - 100$$



$M_{OF} = -90 \rightarrow \text{Model is not good}$   
 $\rightarrow 10\% \text{ of misclassification}$



$$M_{OF} = +1+2+3+4+5 - 1-2-3-4/5+1 = 4$$



$$\sigma(x) = \frac{1}{1+e^{-x}} \Rightarrow \text{expression for sigmoid function}$$

$$\sigma(y_i w^\top x_i) = \frac{1}{1+e^{-(y_i w^\top x_i)}}$$

$$\underline{M_{OF}} \Rightarrow \arg \max [ \sigma(y_i w^\top x_i) ]$$

$$y_i = \sigma(w^T x_i + b)$$

$$\Rightarrow \operatorname{argmax}_w \left[ \frac{1}{1 + e^{-y_i w^T x_i}} \right]$$

$$\Rightarrow \operatorname{argmax}_w \left[ \log \left( \frac{1}{1 + e^{-y_i w^T x_i}} \right) \right]$$

$$\log \left( \frac{1}{a} \right) = -\log a$$

$$\text{MOF} \Rightarrow \operatorname{argmax}_w \left[ \log \left( 1 + e^{-y_i w^T x_i} \right) \right] \Leftrightarrow \operatorname{argmax}_w (y_i w^T x_i)$$

Loss function  $\Rightarrow \operatorname{argmin}_w \left[ \log \left( 1 + e^{-y_i w^T x_i} \right) \right]$

Logistic Loss

Squashes bw 0 & 1

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

$\xrightarrow{-\infty} \xrightarrow{+\infty}$

$$\frac{1}{1 + e^{-(-\infty)}} \quad \frac{1}{1 + e^{-\infty}} \qquad \qquad \qquad e^{-\infty} = 0$$

$$\frac{1}{1 + e^{\infty}} \quad \frac{1}{1 + 0} \qquad \qquad \qquad \downarrow$$

$$\frac{1}{1 + \underbrace{e^{\infty}}_{\approx 7.8^\infty}} \quad \frac{1}{1} = 1$$

$$\frac{1}{1 + \infty}$$

$$\frac{1}{\infty} = 0$$

Probabilistic Way:  
 $(1, 0) \Rightarrow \text{class labels}$

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

$$P(1 + e^{-y}) = 1$$

$$P + Pe^{-y} = 1$$

$$Pe^{-y} = 1 - P$$

$$e^{-y} = \frac{1-P}{P}$$

$$\frac{1}{e^{-y}} = \left(\frac{P}{1-P}\right) \rightarrow \text{odd's ratio}$$

$$e^y = \frac{P}{1-P}$$

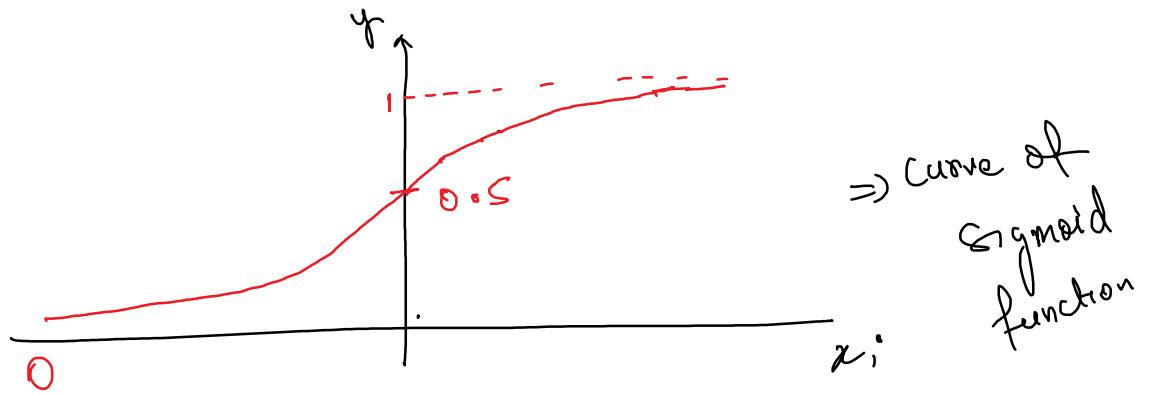
taking  $\ln$  on both sides

$$\ln e^y = \ln \left(\frac{P}{1-P}\right)$$

$y = \ln \left(\frac{P}{1-P}\right)$  logit function

$$y = \ln \left(\frac{P}{1-P}\right)$$

$y \uparrow$  --- - - - -



$$\text{Log loss} \Rightarrow - \left[ \underbrace{y_i \log p(y_i)}_{0 \text{ (class 0)}} + \underbrace{(1-y_i) \log p(1-y_i)}_{0 \text{ (class 1)}} \right]$$

Case 1:  $y_i = +ve \begin{cases} \text{geometric } = + \\ \text{probabilistic } = 1 \end{cases}$

$$\text{Geo: } \log(1 + e^{-y_i w^T x_i}) \Rightarrow \log(1 + e^{-w^T x_i}) =$$

$$\text{prob: } [y_i \log p(y_i) + (1-y_i) \log(1-y_i)] \xrightarrow{=} 0$$

$$\Rightarrow y_i \log p(y_i) \Rightarrow \log p \Rightarrow -\log \left( \frac{1}{1+e^{-w^T x_i}} \right)$$

$$\Rightarrow \log(1+e^{-w^T x_i})$$

Deriving Sigmoid from MOF:

$$\ln \left( \frac{P}{1-P} \right) \Rightarrow y_i w^T x_i$$

$$e^{\ln \left( \frac{P}{1-P} \right)} = e^{y_i w^T x_i}$$

$$\frac{P}{1-P} = e^{y_i w^T x_i}$$

$$P = (1-P) e^{y_i w^T x_i}$$

$$P = e^{y_i w^T x_i} - P e^{y_i w^T x_i}$$

$$P + P e^{y_i w^T x_i} = e^{y_i w^T x_i}$$

$$P(1 + e^{y_i w^T x_i}) = e^{y_i w^T x_i}$$

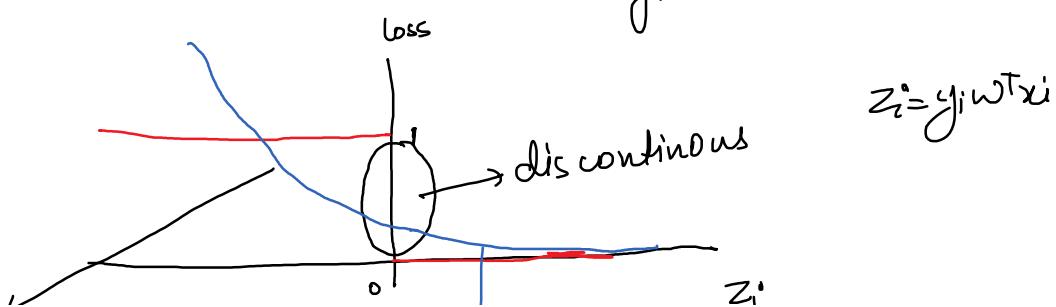
$$P = \left( \frac{e^{y_i w^T x_i}}{1 + e^{y_i w^T x_i}} \right) \quad (\div \text{ by } e^{y_i w^T x_i})$$

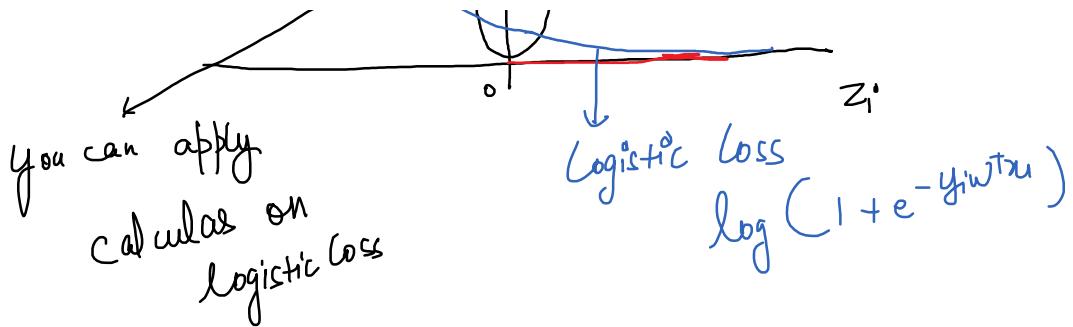
$$P = \frac{1}{\frac{1}{e^{y_i w^T x_i}} + 1} = \frac{1}{1 + \frac{1}{e^{y_i w^T x_i}}}$$

$$P = \frac{1}{1 + e^{-y_i w^T x_i}} = \sigma(y_i w^T x_i)$$

### Loss minimization (0-1 loss)

$y_i w^T x_i < 0$  i.e.





## Overfitting & Underfitting

$$w^* = \arg \min_w \sum_{i=1}^n \log(1 + e^{-y_i w^T x_i}) \quad y_i w^T x_i = z_i$$

$z_i = \infty$

$$w^* = \arg \min_w \sum_{i=1}^n \log(1 + e^{-z_i})$$

$$w^* = \arg \min_w \sum_{i=1}^n \log(1 + e^{-z_i}) \quad z_i = y_i w^T x_i$$

constant  
variable

$w^* \Rightarrow \infty$

$X_1 \quad X_2 \quad X_3 \quad Y$

## Regularization

RIDGE Regularization

$$w^* = \arg \min_w \left[ \sum_{i=1}^n \log(1 + e^{-z_i}) + \lambda [w^T w] \right]$$

$0.001 \times 10000 = 10$   
hyperparameter 10,000

LASSO Regularization

$$\omega \rightarrow 10000$$

$[0, 1, 0, 0, 0, 0, 0]$   
 $-n \quad -n \quad -n \dots -z_i \lambda, 1, 1, 1, 1, 1, 1 \rightarrow 10000$

$$w^* = \underset{w}{\operatorname{argmin}} \left[ \sum_{i=1}^n \log(1+e^{-z_i}) + \frac{\lambda}{2} \|w\|^2 \right]$$

$\omega \rightarrow 10000$

$w \downarrow 0$

LASSO creates sparsity.  $\rightarrow$  benefit  $\rightarrow$  extract out important feature

$$\text{Sparse vector} = [1, 0, 0, 0, 0, 1, 0, 0, 1, 0, 0, 1]$$

$$y = w_1 x_1 + w_2^0 x_2 + w_3^0 x_3 + w_4^0 x_4 + w_5 x_5$$

Hyperparameter ( $\lambda$ ):  $\lambda = 0$  overfitting

$\lambda = \text{very high}$  underfitting

ElasticNet: power of both regularizer.

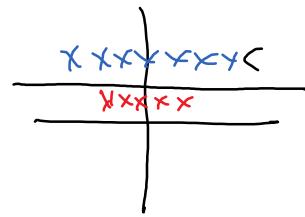
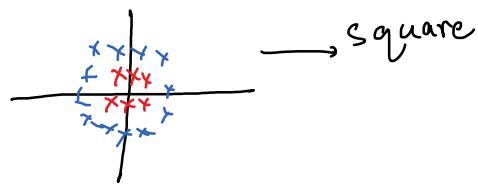
$$w^* = \underset{w}{\operatorname{argmin}} \left[ \log(1+e^{-z_i}) + \underbrace{\lambda_1 \|w\|}_\text{Lasso} + \underbrace{\lambda_2 \|w\|^2}_\text{Ridge} \right]$$

Features:  $\rightarrow$  standardized.

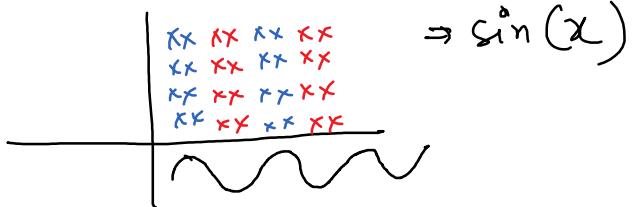
Should be linearly separable.

Ex:

(1)

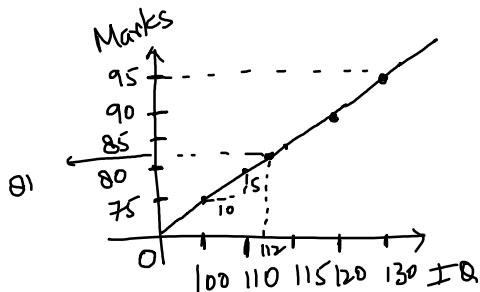


(11)



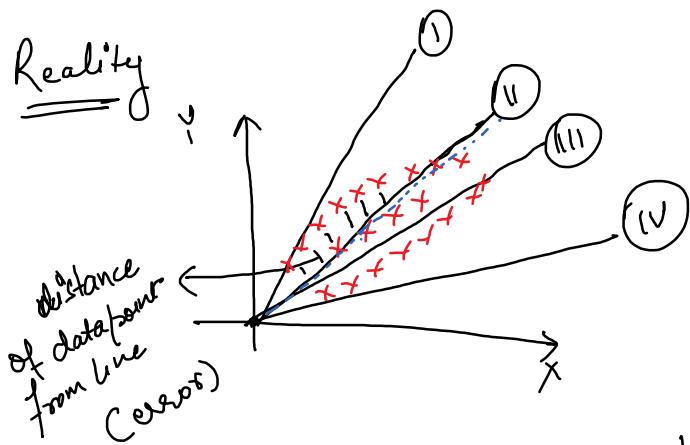
## Linear Regression

Variance → Covariance → Correlation → Regression  
 ↗ directional relationship   ↗ directional relationship   ↗ strength  
 ↗ strength   ↗ strength   ↗ quantification

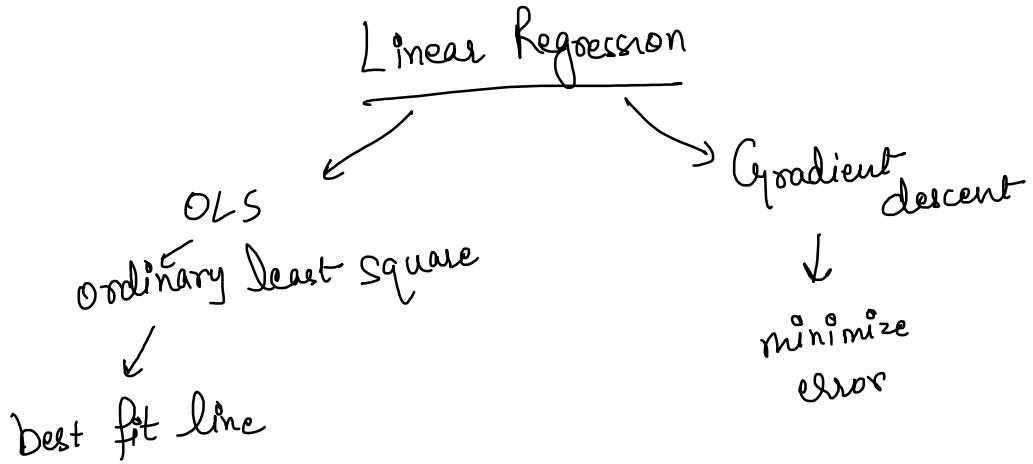


$$IQ \quad | \quad 130 = \frac{Marks}{?} \quad 95 \quad (\text{Prediction})$$

$$y = mx \\ \text{Marks} = \text{slope} \times IQ$$

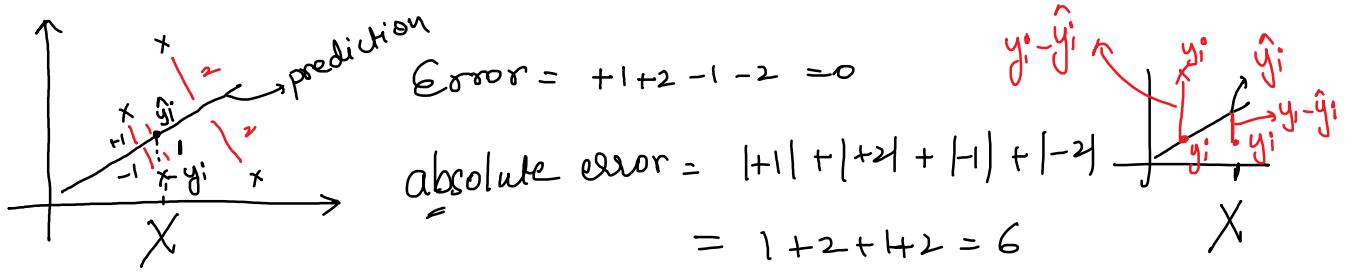


choose a line with min error!

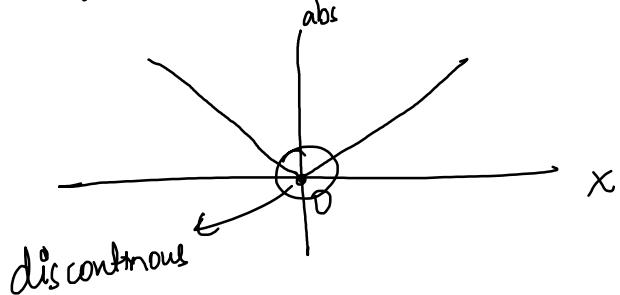


How to create a line?

$$y = w_0 + w_1 x_1 + w_2 x_2 + \dots + w_n x_n$$



$$\text{absolute error} = \sum_{i=1}^n |y_i - \hat{y}_i| = |+1| + |+2| + |-1| + |-2| = 6$$



Squared error:  $\sum (y_i - \hat{y}_i)^2 \rightarrow \text{parabola} \rightarrow \text{can apply parabola.}$

$$E(m, b) = (y_i - \hat{y}_i)^2 = 0$$

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

$$E(m, b) = (y_i - (mx_i + b))^2 = 0 \Rightarrow \sum_{i=1}^n [y_i - (mx_i + b)]^2$$

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

$$\Rightarrow \sum_{i=1}^n 2 [y_i - (mx_i + b)] \left( \frac{dy_i}{db} - \frac{d(mx_i)}{db} - \frac{db}{db} \right) = 0$$

$\frac{d y_i}{d b} = 0$   
 $\frac{d(mx_i)}{d b} = m \frac{d x_i}{d b} = m \cdot 1 = m$   
 $\frac{db}{d b} = 1$

$$\Rightarrow -2 \sum_{i=1}^n [y_i - (mx_i + b)] = 0$$

$$\rightarrow \sum_{i=1}^n [y_i - (mx_i + b)] = 0$$

$$\Rightarrow \sum_{i=1}^n [y_i - (mx_i + b)] = 0$$

$$\Rightarrow \sum_{i=1}^n [y_i - mx_i - b] = 0$$

$$\Rightarrow \sum_{i=1}^n y_i - \sum_{i=1}^n mx_i - \sum_{i=1}^n b = 0$$

Dividing LHS & RHS by  $n$ .

$$\bar{y}_i - \frac{\sum_{i=1}^n mx_i}{n} - \frac{\sum_{i=1}^n b}{n} = 0$$

(hb)

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

Value of  
Intercept

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

$$\frac{dE}{dm} = \frac{d [y_i - (mx_i + b)]^2}{dm} = 0$$

$$= \frac{d [y_i - mx_i - (\bar{y}_i - m\bar{x}_i)]^2}{dm} = 0$$

Assignment  
↓

$$\text{Value of slope} \Rightarrow m = \frac{\sum (y_i - \bar{y}_i)(\bar{x}_i - x_i)}{\sum (x_i - \bar{x}_i)^2}$$

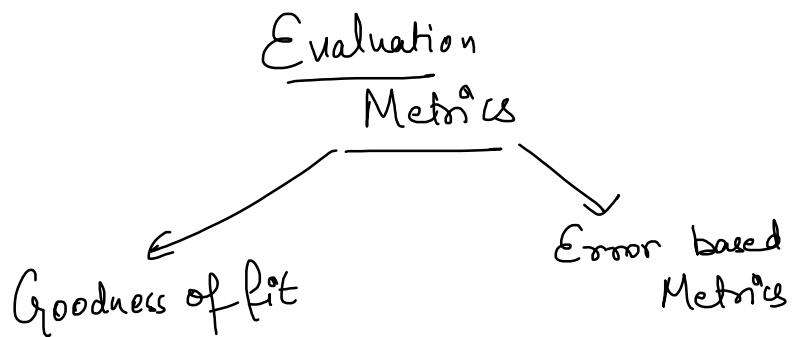
Max slope  $\Rightarrow$   $\frac{\sum (x_i - \bar{x}_i)^2}{\sum (y_i - \hat{y}_i)^2}$

We got best  $m$  &  $b$ :

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

$\Rightarrow$  With these  $m$  &  $b$ , the best line can be directly drawn.

$\Rightarrow$  Do not use for very large dataset!



MAE: Mean Absolute Error  $\Rightarrow \frac{1}{n} \sum_{i=1}^n |y_i - \hat{y}_i|$

### Advantages:

$\rightarrow$  Same scale as that of data

$\rightarrow$  less sensitive to outlier.

### Disadvantage:

$\rightarrow$  not differentiable

MSE: Mean Squared Error:  $\frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2$

### Advantages:

- differentiable
- optimized

### Disadvantages:

- sensitive to outliers.
- not in same scale as that of data

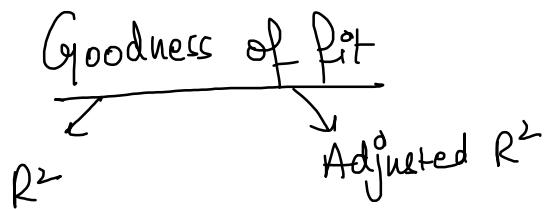
RMSE: Root mean squared error

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

- easy to interpret
- less sensitive to outlier

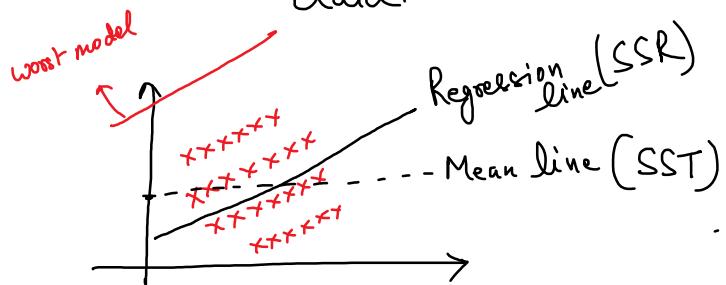
$\Rightarrow MAPE \Rightarrow$  Mean absolute percentage error:

Homework



$R^2$  (R2-score / coeff. of determination)  $\Rightarrow$  how well your model fits the data.

$$R^2 = 1 - \frac{SSR}{SST}$$



Case 1°  $SSR = 0$ ,  $SST = SST$   $\rightarrow$  best model (overfitting)

$$R^2 = 1 - \frac{0}{SST} = 1 - 0 = 1$$

Case 2°  $SSR = SST$

$$R^2 = 1 - \frac{SST}{SST} = 1 - 1 = 0 \rightarrow \text{bad model (underfitting)}$$

Case 3°  $SSR > SST \Rightarrow \frac{SSR}{SST} > 1$

$$R^2 = 1 - \left( \frac{SSR}{SST} \right) = -\text{ve} \quad (\text{blunder})$$

Problem with  $R^2$ ?

↳ as # columns  $\uparrow$ ,  $R^2$  square

Adjusted  $R^2 \Rightarrow \text{Adj } R^2 = 1 - \left[ \frac{(1-R^2)(n-1)}{(n-1-p)} \right]$

not present in  $\downarrow$   $\text{penalize the increase in columns}$

If  $(n-1-p) \text{ decrease} > (1-R^2) \text{ decrease}$   
 $\text{Adj } R^2 \downarrow$

else      Adj R<sup>2</sup> ↑

## Multicollinearity: Before Modelling

↳ One column is highly correlated with other columns

$$\begin{matrix} f_1 & f_2 & f_3 \\ w \Rightarrow \{1 & 2 & 3\} \\ \downarrow \\ w \Rightarrow \{0 & 3.5 & 3\} \end{matrix}$$

output gets disturbed!

$$\begin{array}{l} \text{O/P equation} \\ \boxed{y = 1f_1 + 2f_2 + 3f_3} \end{array}$$

$$\boxed{f_1 = 1.5f_2}$$

$$y_1 = 1.5f_2 + 2f_2 + 3f_3$$

$$y_1 = 0f_1 + 3.5f_2 + 3f_3$$

How to detect Multicollinearity?

① Correlation Matrix (0-70)

↳ drop one of the columns.

② VIF = variance inflation factor =  $\frac{1}{1 - R^2}$   
[1, ∞]

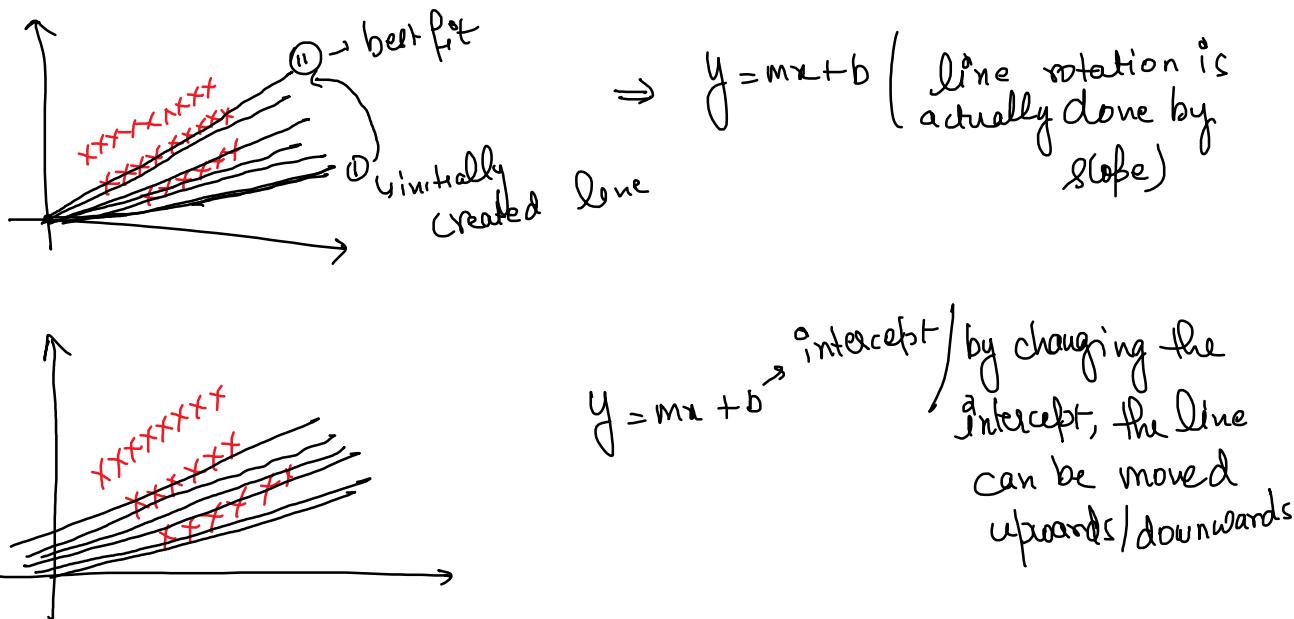
"VIF > 5"

## Gradient Descent



⇒ best fit

if ... L / no minimization



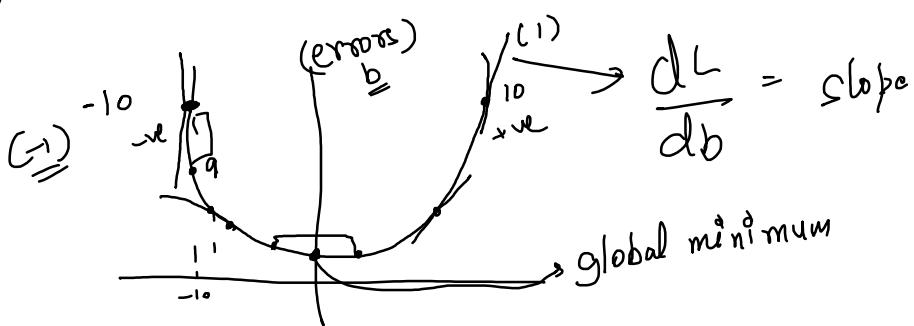
$$L = (y_i^o - \hat{y}_i)^2$$

Steps:  $m = \text{constant}$ ,  $b = \text{variable}$

1) choose any random value of  $b$

2)  $\frac{dL}{db} = \frac{d(y_i^o - mx_i - b)^2}{db} = -2 \underbrace{(y_i^o - mx_i - b)}_{\text{slope}} \frac{d}{db}$

3)  $b_{\text{next}} = b_{\text{old}} - \text{slope}$



$$\begin{aligned} b_{\text{next}} &= -10 - (-1) \\ &= -10 + 1 = 9 \end{aligned}$$

$$\begin{aligned} b_{\text{next}} &= 10 - (+1) \\ &= 9 \end{aligned}$$

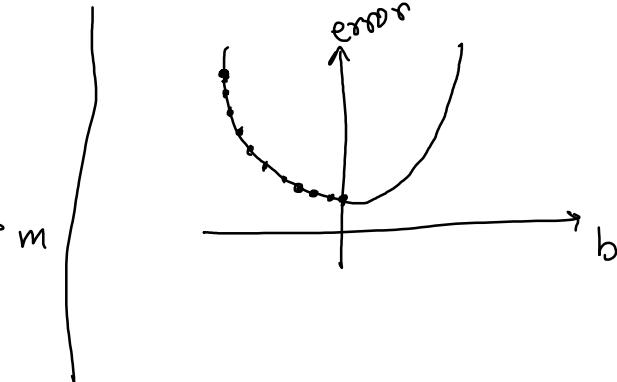
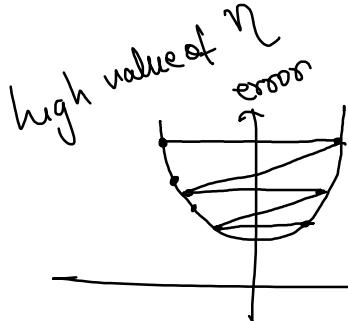
Actual steps:

1) Take random values of  $m$  &  $b$

2) Find  $\frac{\partial L}{\partial m}$  &  $\frac{\partial L}{\partial b}$

3)  $m_{\text{next}} = m_{\text{old}} - \eta \frac{\partial L}{\partial m}$

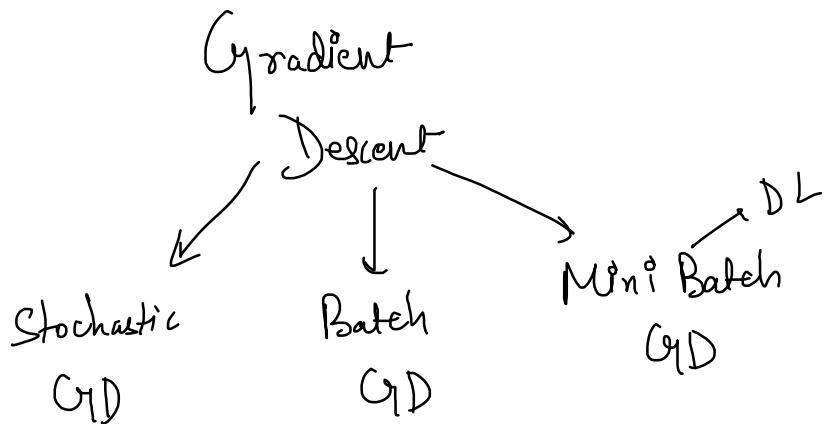
$b_{\text{next}} = b_{\text{old}} - \eta \frac{\partial L}{\partial b}$



$\eta$  = Learning rate  
or  
Stepsize

effects of  $\eta$ :  $\rightarrow$  the greater the learning rate, the faster the algorithm, but it will oscillate around min value but will never reach it!

2) the smaller the learning rate, the slower the algorithm, but it will reach min value.



Stochastic Gradient Descent → faster

→ row wise operation

$$\begin{array}{c} \text{100 rows, 100 iterations} \\ \hline \text{iterations rows/calculations} \\ 1 \rightarrow 100 \end{array} \rightarrow \begin{array}{l} 1^{\text{st}} \\ 1 \rightarrow b, m \\ 2 \rightarrow b, m \\ 3 \rightarrow b, m \\ \vdots \\ 100 \rightarrow b, m \end{array}$$

100 →  $100 \times 100 = 10,000$  calculations

$$\begin{array}{l} 2^{\text{nd}} \\ 1 \rightarrow b, m \\ 2 \rightarrow b, m \\ 3 \rightarrow b, m \\ \vdots \\ 100 \rightarrow b, m \end{array}$$

$$\begin{array}{l} 100^{\text{th}} \\ 1 \rightarrow b, m \\ 2 \rightarrow b, m \\ 3 \rightarrow b, m \\ \vdots \\ 100 \rightarrow b, m \end{array}$$

Batch Gradient descent → we in small dataset

100 rows, 100 iteration

1 iteration → 1 calculation

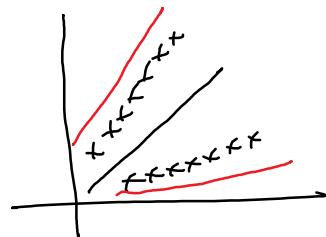
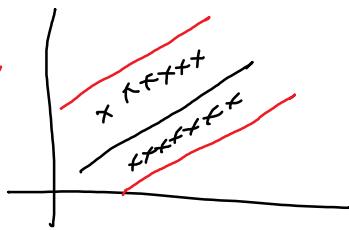
100 iteration → 100 calculations

$$\begin{array}{l} 1^{\text{st}} \text{ iteration} \rightarrow [1 - 100] \rightarrow b, m \\ 2^{\text{nd}} \text{ " } \rightarrow [1 \rightarrow 100] \rightarrow b, m \\ \vdots \\ 100^{\text{th}} \text{ iteration} \rightarrow [1 - 100] \rightarrow b, m \end{array}$$

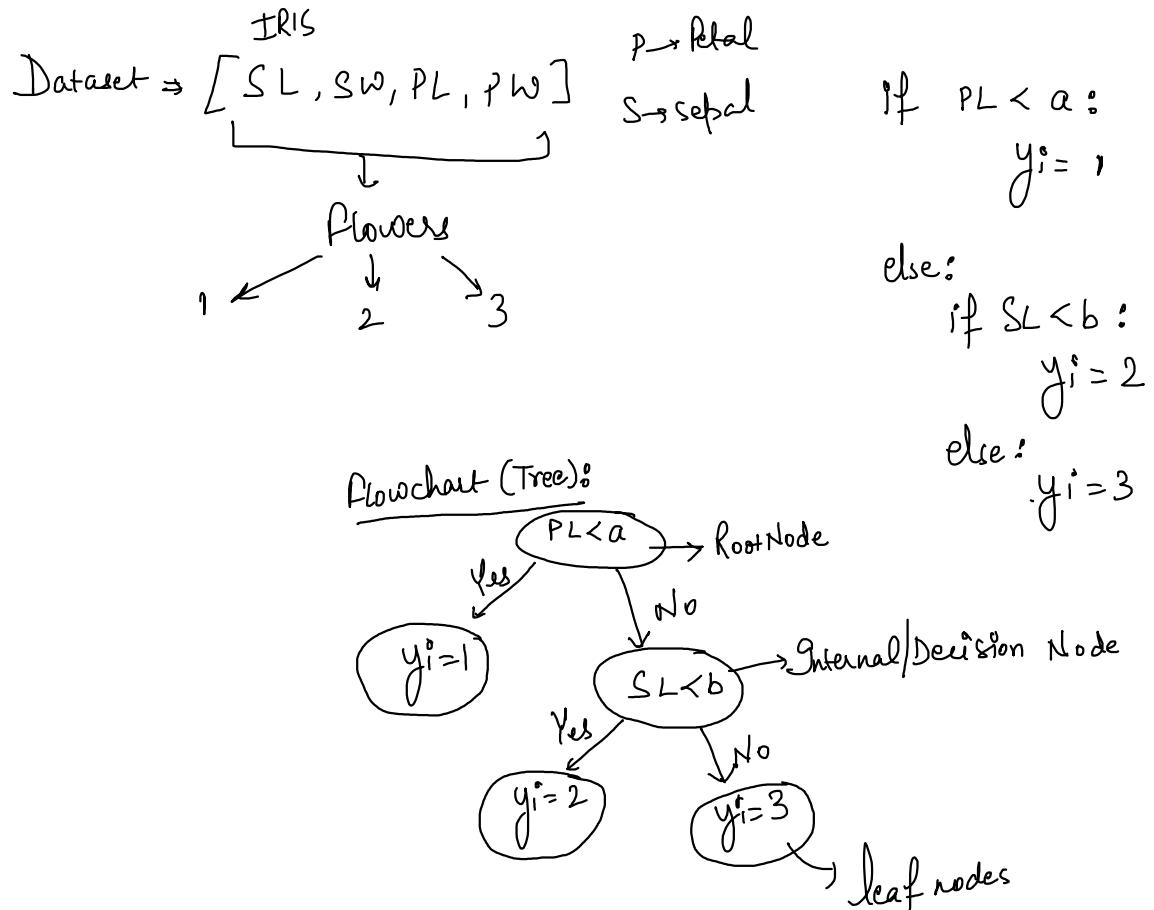
Assumptions:  $\rightarrow X$  and  $Y$  have linear relationship.  
(Linear Regression)

- $\rightarrow$  columns are independent
- $\rightarrow$  Residual (error) are normally distributed.
- $\rightarrow$  Residual (error) are homoscedastic.  $\rightarrow$  (constant variance)

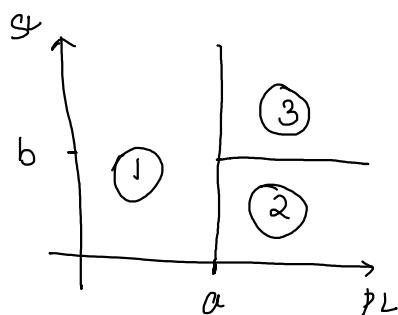
errors have  
constant  
variance  
 $\nwarrow$   
Durbin-Watson  
Test  $\approx 2$



## Decision Trees



Recursive Partitioning: (axis-parallel hyperplanes)



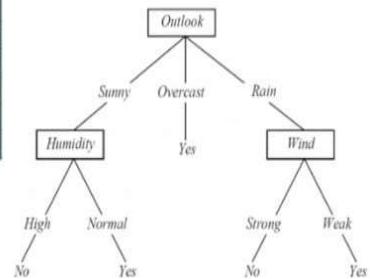
if  $PL < a$   
 $y_i^* = 1$

else  
 if  $SL < b$   
 $y_i^* = 2$

else  
 $y_i^* = 3$

DT → Entropy → Randomness  
 ↓  
 Gini Impurity  
 ↓  
 Information Gain

| Outlook  | Temperature | Humidity | Windy | PlayTennis |
|----------|-------------|----------|-------|------------|
| Sunny    | Hot         | High     | False | No         |
| Sunny    | Hot         | High     | True  | No         |
| Overcast | Hot         | High     | False | Yes        |
| Rainy    | Mild        | High     | False | Yes        |
| Rainy    | Cool        | Normal   | False | Yes        |
| Rainy    | Cool        | Normal   | True  | No         |
| Overcast | Cool        | Normal   | True  | Yes        |
| Sunny    | Mild        | High     | False | No         |
| Sunny    | Cool        | Normal   | False | Yes        |
| Rainy    | Mild        | Normal   | False | Yes        |
| Sunny    | Mild        | Normal   | True  | Yes        |
| Overcast | Mild        | High     | True  | Yes        |
| Overcast | Hot         | Normal   | False | Yes        |
| Rainy    | Mild        | High     | True  | No         |



Entropy:

$$H_D(Y) = - \sum_{i=1}^n p_i \lg(p_i)$$

$\lg \Rightarrow \log_2$

$$H_D(Y) = - P(Y_+) \lg P(Y_+) - P(Y_-) \lg P(Y_-) \Rightarrow \text{for binary classification}$$

Parents Entropy: Yes = 9      No = 5      Total = 14

$$\begin{aligned} H_D(Y) &= - P(\text{Yes}) \lg P(\text{Yes}) - P(\text{No}) \lg P(\text{No}) \\ &= - \frac{9}{14} \lg \left( \frac{9}{14} \right) - \frac{5}{14} \lg \left( \frac{5}{14} \right) = 0.94 \end{aligned}$$

entropy of each column:

$H_D(Y)$  0.94

Outlook  $\begin{cases} 5 \xrightarrow{(2Y, 3N)} \text{Sunny} \xrightarrow{-\frac{2}{5} \lg \left( \frac{2}{5} \right) - \frac{3}{5} \lg \left( \frac{3}{5} \right) = 0.97} \\ 4 \xrightarrow{(4Y, 0N)} \text{Overcast} \xrightarrow{-\frac{4}{4} \lg \left( \frac{4}{4} \right) - 0 \lg 0 = 0} \\ 5 \xrightarrow{(3Y, 2N)} \text{Rainy} \xrightarrow{-\frac{3}{5} \lg \left( \frac{3}{5} \right) - \frac{2}{5} \lg \left( \frac{2}{5} \right) = 0.97} \end{cases}$

Weighted entropy  $\Rightarrow H_D(Y_{\text{outlook}}) = \frac{5}{14} \times 0.97 + \frac{4}{14} \times 0 + \frac{5}{14} \times 0.97$

$$= \frac{1}{4} \times (0.97 + 0.97) = 0.69$$

Temperature

|   |            |  |
|---|------------|--|
| $\begin{array}{l} 4 \\ \downarrow \\ \text{Hot} \end{array}$  | $(2Y, 2N)$ | $\Rightarrow -\frac{2}{4} \lg \frac{2}{4} - \frac{2}{4} \lg \frac{2}{4} = 1$ |
| $\begin{array}{l} 6 \\ \downarrow \\ \text{Mild} \end{array}$ | $(4Y, 2N)$ | $= -\frac{4}{6} \lg \frac{4}{6} - \frac{2}{6} \lg \frac{2}{6} = 0.91$        |
| $\begin{array}{l} 4 \\ \downarrow \\ \text{Cold} \end{array}$ | $(3Y, 1N)$ | $= -\frac{3}{4} \lg \frac{3}{4} - \frac{1}{4} \lg \frac{1}{4} = 0.81$        |

$$\text{Weighted entropy of } H_D(Y, \text{Temperature}) = \frac{4}{14} \times 1 + \frac{6}{14} \times 0.91 + \frac{4}{14} \times 0.81$$

$$= 0.91$$

Humidity

|   |            |   |
|---|------------|---|
| $\begin{array}{l} 7 \\ \downarrow \\ \text{High} \end{array}$   | $(3Y, 4N)$ | $\Rightarrow -\frac{3}{7} \lg \frac{3}{7} - \frac{4}{7} \lg \frac{4}{7} = 0.98$ |
| $\begin{array}{l} 7 \\ \downarrow \\ \text{Normal} \end{array}$ | $(6Y, 1N)$ | $\Rightarrow -\frac{6}{7} \lg \frac{6}{7} - \frac{1}{7} \lg \frac{1}{7} = 0.59$ |

$$H_D(Y, \text{Humidity}) = \frac{7}{14} \times 0.98 + \frac{7}{14} \times 0.59 = 0.78$$

windy

|  |            |   |
|--|------------|---|
| $\begin{array}{l} 6 \\ \downarrow \\ \text{True} \end{array}$  | $(5Y, 3N)$ | $\Rightarrow -\frac{5}{6} \lg \frac{5}{6} - \frac{3}{6} \lg \frac{3}{6} = 1$    |
| $\begin{array}{l} 8 \\ \downarrow \\ \text{False} \end{array}$ | $(6Y, 2N)$ | $\Rightarrow -\frac{6}{8} \lg \frac{6}{8} - \frac{2}{8} \lg \frac{2}{8} = 0.81$ |

$$H_D(Y, \text{Windy}) = \frac{6}{14} \times 1 + \frac{8}{14} \times 0.81 = 0.89$$

$\Rightarrow$  Two ways to choose columns for first split:

1 → choose column with lowest weighted entropy -

| Outlook | temperature | windy | humidity |
|---------|-------------|-------|----------|
| ↓       | ↓           | ↓     | ↓        |

0.69      0.91      0.89      0.78

2 → Calculate Information Gain:

$$IG(Y) = \text{Parents entropy} - \frac{\text{column entropy}}{\text{(weighted)}}$$

$$\text{outlook} \Rightarrow IG_O(Y) = 0.94 - 0.69 = 0.25 \rightarrow \text{choose this for first split}$$

$$\text{temperature} \Rightarrow IG_T(Y) = 0.94 - 0.91 = 0.03$$

$$\text{Humidity} \Rightarrow IG_H(Y) = 0.94 - 0.78 = 0.16$$

$$\text{Windy} \Rightarrow IG_W(Y) = 0.94 - 0.89 = 0.05$$

### Properties of Entropy:

$$H_D(Y) = -P(y_+) \lg P(y_+) - P(y_-) \lg P(y_-)$$

Case 1:  $P(y_+) = 0.99, P(y_-) = 0.01$

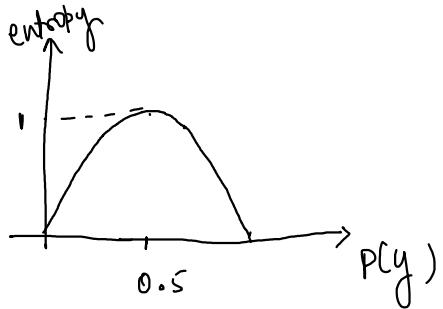
$$H_D(Y) = -0.99 \lg 0.99 - 0.01 \lg 0.01 = 0.08$$

Case 2:  $P(y_+) = 0.5, P(y_-) = 0.5$

$$H_D(Y) = -0.5 \log 0.5 - 0.5 \log 0.5 = 1$$

Case 3:  $P(y_+) = 1$        $P(y_-) = 0$

$$H_D(Y) = -1 \log 1 - 0 \log 0 = 0$$



Gini Impurity ( $I_G$ )     $I_G \neq H_D$

$$I_G(Y) = 1 - \sum_{i=1}^n (p_i)^2$$

for binary classification,

$$I_G(Y) = 1 - [P(y_+)^2 + P(y_-)^2]$$

for multiclass,

$$I_G(Y) = 1 - [P(y_1)^2 + P(y_2)^2 + P(y_3)^2 + \dots + P(y_n)^2]$$

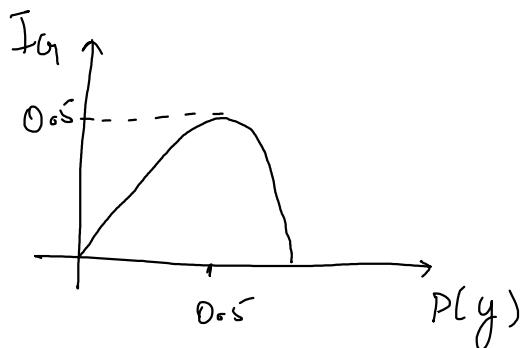
Properties of Gini Impurity:

Case 1:  $P(y_+) = 0.5$        $P(y_-) = 0.5$

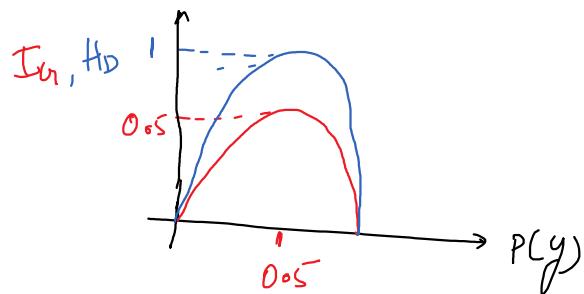
$$I_G(Y) = 1 - [P(y_+)^2 + P(y_-)^2] = 1 - [0.5^2 + 0.5^2] = 0.5$$

$$\text{Case 2: } P(y_+) = 1, \quad P(y_-) = 0$$

$$I_G(Y) = 1 - [1^2 + 0^2] = 0$$



Comparison b/w Gini Impurity & Entropy:



## ② Computational Cost:

entropy is harder to calculate, higher computational cost

$I_G$  is easier to calculate, lower computational cost

for larger datasets, use Gini Impurity!

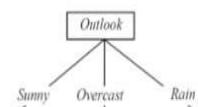
Gini Impurity (Parent)

$$I_G(Y) = 1 - \left[ P(\text{Yes})^2 + P(\text{No})^2 \right]$$

$$= 1 - \left[ \left( \frac{9}{14} \right)^2 + \left( \frac{5}{14} \right)^2 \right]$$

.....

| Outlook  | Temperature | Humidity | Windy | PlayTennis |
|----------|-------------|----------|-------|------------|
| Sunny    | Hot         | High     | False | No         |
| Sunny    | Hot         | High     | True  | No         |
| Overcast | Hot         | High     | False | Yes        |
| Rainy    | Mild        | High     | False | Yes        |
| Rainy    | Cool        | Normal   | False | Yes        |
| Rainy    | Cool        | Normal   | True  | No         |
| Overcast | Cool        | Normal   | True  | Yes        |
| Sunny    | Mild        | High     | False | No         |
| Sunny    | Cool        | Normal   | False | Yes        |
| Rainy    | Mild        | Normal   | False | Yes        |
| Sunny    | Mild        | Normal   | True  | Yes        |
| Overcast | Mild        | High     | True  | Yes        |

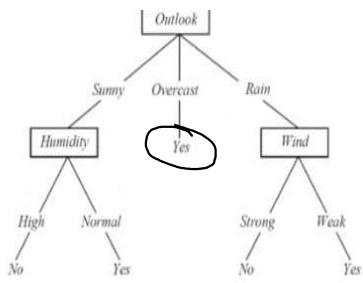


$$= 1 - \left[ \frac{10}{14} \right]$$

$$= 1 - \left[ \frac{10}{14} \right] = 0.429$$

$$= 0.429$$

|          |      |        |       |     |
|----------|------|--------|-------|-----|
| Rainy    | Mild | Normal | False | Yes |
| Sunny    | Mild | Normal | True  | Yes |
| Overcast | Mild | High   | True  | Yes |
| Overcast | Hot  | Normal | False | Yes |
| Rainy    | Mild | High   | True  | No  |



outlook

$$\begin{aligned} &\xrightarrow{5} \text{Sunny } (2Y, 3N) \Rightarrow 1 - \left[ \left( \frac{2}{5} \right)^2 + \left( \frac{3}{5} \right)^2 \right] = \frac{12}{25} = 0.48 \\ &\xrightarrow{4} \text{Overcast} \Rightarrow 1 - 1^2 = 0 \\ &\xrightarrow{5} \text{Rainy } (3Y, 2N) \Rightarrow 1 - \left[ \left( \frac{3}{5} \right)^2 + \left( \frac{2}{5} \right)^2 \right] = 0.48 \end{aligned}$$

$$\text{Weighted Gini Impurity} \Rightarrow \frac{5}{14} \times 0.48 + 0 \times \frac{4}{14} + \frac{5}{14} \times 0.48 = 0.342$$

Temperature

$$\begin{aligned} &\xrightarrow{4} \text{Hot } (2Y, 2N) \Rightarrow 1 - \left[ \left( \frac{2}{4} \right)^2 + \left( \frac{2}{4} \right)^2 \right] = 0.5 \\ &\xrightarrow{6} \text{Mild } (4Y, 2N) \Rightarrow 1 - \left[ \left( \frac{4}{6} \right)^2 + \left( \frac{2}{6} \right)^2 \right] = 0.444 \\ &\xrightarrow{4} \text{Cool } (3Y, 1N) \Rightarrow 1 - \left[ \left( \frac{3}{4} \right)^2 + \left( \frac{1}{4} \right)^2 \right] = 0.375 \end{aligned}$$

$$W.G.I \quad I_G(T) = \frac{4}{14} \times 0.5 + \frac{6}{14} \times 0.444 + \frac{4}{14} \times 0.375 = 0.44$$

Humidity

$$\begin{aligned} &\xrightarrow{7} \text{High } (3Y, 4N) \Rightarrow 1 - \left[ \left( \frac{3}{7} \right)^2 + \left( \frac{4}{7} \right)^2 \right] = 0.489 \\ &\xrightarrow{7} \text{Normal } (6Y, 1N) \Rightarrow 1 - \left[ \left( \frac{6}{7} \right)^2 + \left( \frac{1}{7} \right)^2 \right] = 0.244 \end{aligned}$$

$$I_G(H) = \frac{7}{14} \times 0.489 + \frac{7}{14} \times 0.244 = 0.367$$

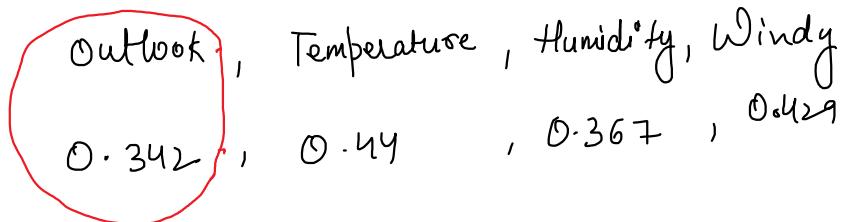
Windy

$$\begin{aligned} &\xrightarrow{6} \text{True } (3Y, 3N) \Rightarrow 0.5 \\ &\xrightarrow{8} \text{False } (6Y, 2N) \Rightarrow 1 - \left[ \left( \frac{6}{8} \right)^2 + \left( \frac{2}{8} \right)^2 \right] = 0.375 \end{aligned}$$

$$I_{G_1}(W) = \frac{6}{14} \times 0.5 + \frac{8}{14} \times 0.375 = 0.429$$

Two ways to choose:

1) Choose column with least <sup>(W.G.I)</sup> Gini Impurity for the split



2) Information Gain  $\Rightarrow$  Parent Gini Impurity - Weighted Gini Impurity of columns

$$\text{outlook} \Rightarrow IG_{G_1}(Y) \Rightarrow 0.459 - 0.342 = 0.117$$

$$\text{temperature} \Rightarrow IG_{G_T}(Y) \Rightarrow 0.459 - 0.440 = 0.019$$

$$\text{humidity} \Rightarrow IG_{G_H}(Y) \Rightarrow 0.459 - 0.367 = 0.092$$

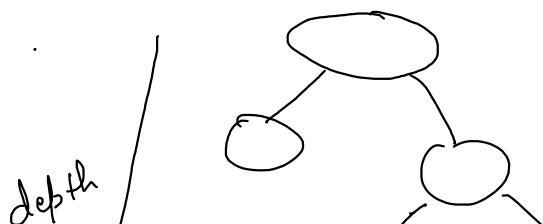
$$\text{Windy} \Rightarrow IG_{G_W}(Y) \Rightarrow 0.459 - 0.429 = 0.03$$

Choose outlook because it has maximum information Gain

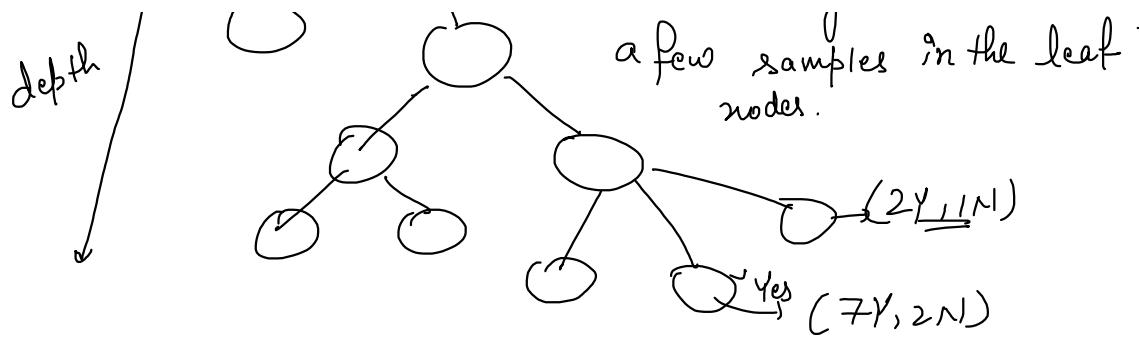
When to stop a tree?

a) Pure Node  $\rightarrow (24, 2N)$

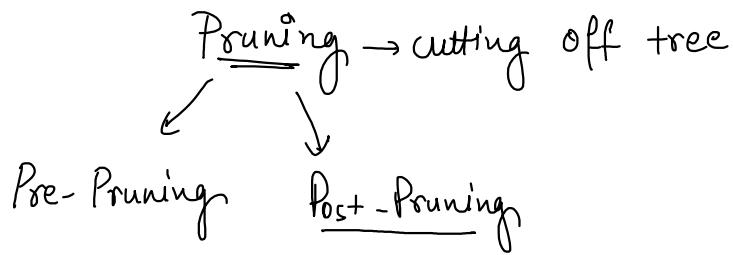
b) You can't grow a tree because you lack datapoints



we don't grow tree with  
a few samples in the leaf



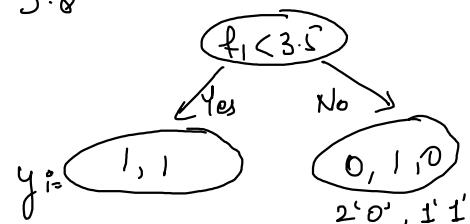
Hyperparameter  $\Rightarrow$  Max\_depth  $\Rightarrow$  height of tree



Splitting a Numerical feature

| $f_1$ | $y$ | ① → sort the values of feature |
|-------|-----|--------------------------------|
| 2.2   | 1   |                                |
| 2.6   | 1   |                                |
| 3.5   | 0   |                                |
| 3.8   | 0   | ② $f_1 < 2.2$                  |
| 4.6   | 1   | $f_1 < 2.6$                    |
| 5.3   | 0   | $f_1 < 3.5$                    |

$$\begin{array}{ll} \textcircled{2} & f_1 < 2.2 \\ & f_1 < 2.6 \\ & f_1 < 3.5 \\ & f_1 < 3.8 \end{array}$$



Feature Engineering :  $\Rightarrow$  Categorical column : PINCODE

$$\frac{\text{Pinode}}{(P_j)}$$

↓  
numerical column

$$(P_j^*)$$

$P(y_i=1) \quad P(y_i=0)$

$$P(y_i=1 / p_j) = \frac{P_j \cap y_i=1}{P_j}$$

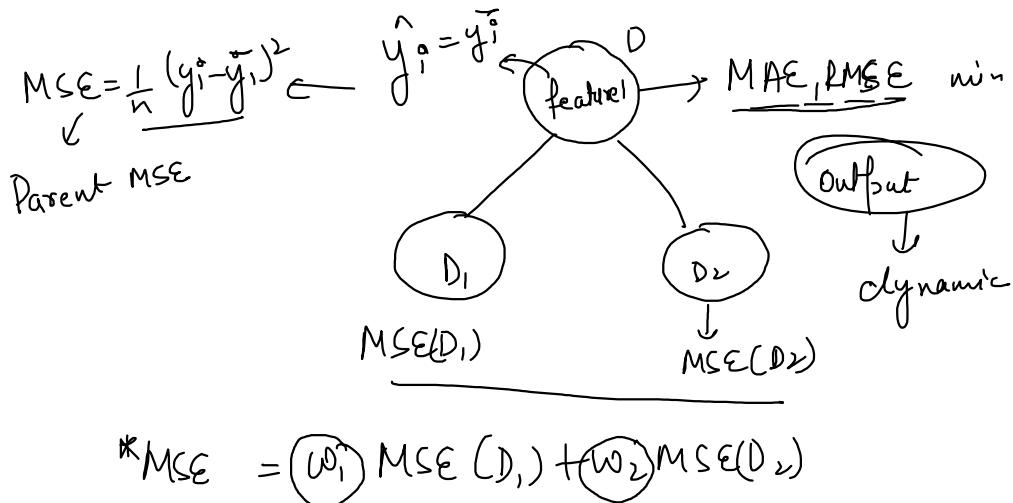
numerical column  
conditional probability

$\Rightarrow$  drop it!

## Regression

$\rightarrow$  Classification  $\Rightarrow$  Gini Impurity, Entropy

$\rightarrow$  Regression  $\Rightarrow$  MAE, MSE, RMSE



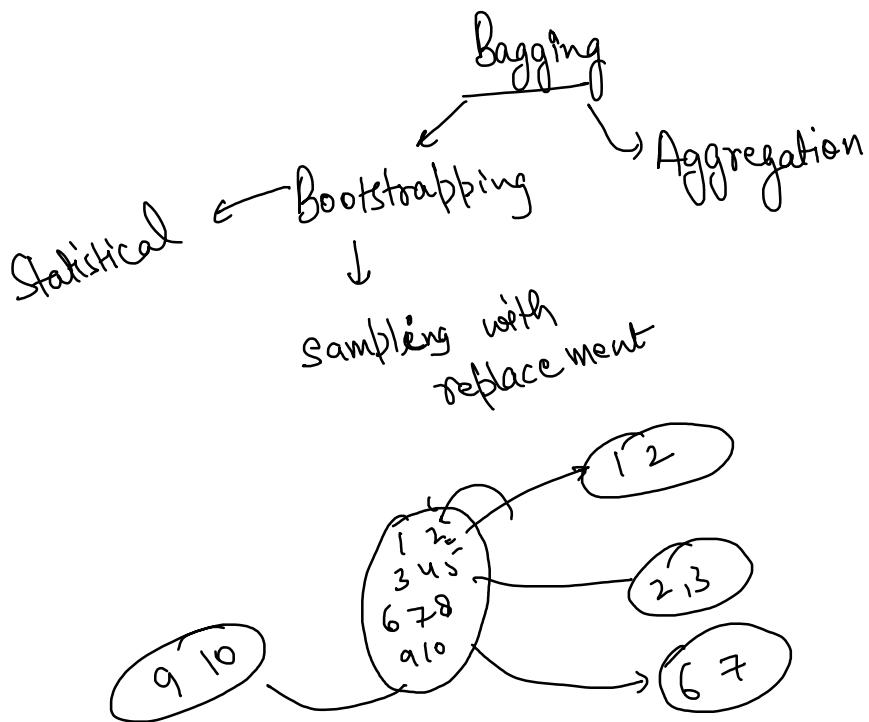
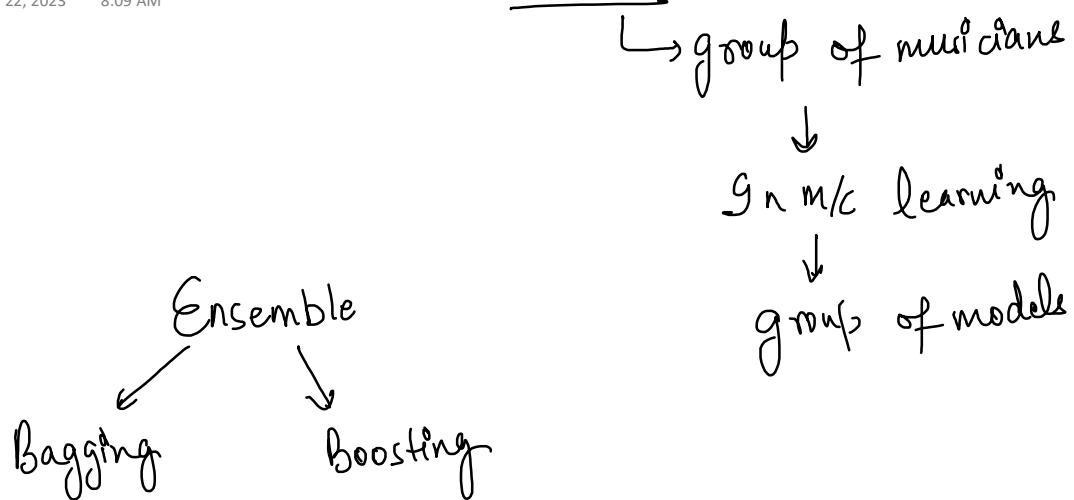
Advantages:

$\rightarrow$  fit is very interpretable

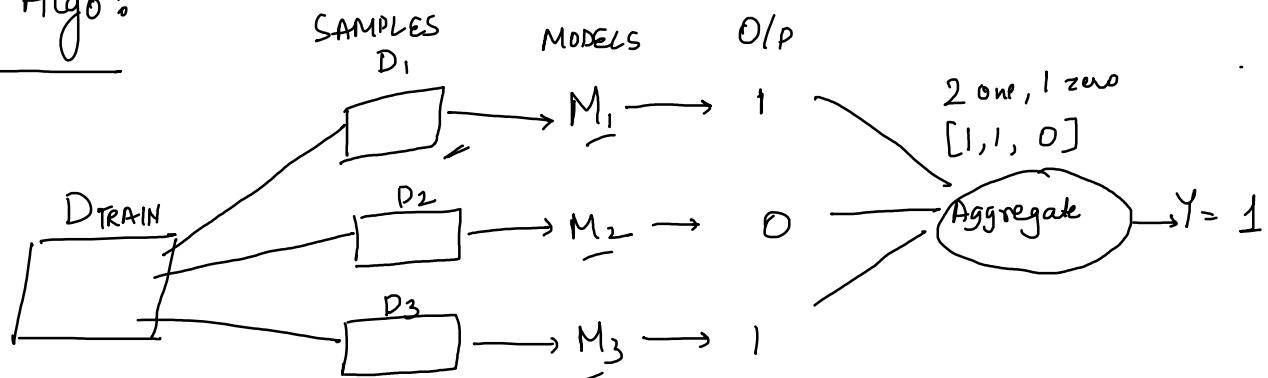
$\rightarrow$  important features

ML

## Ensembles



## Bagging Algo:



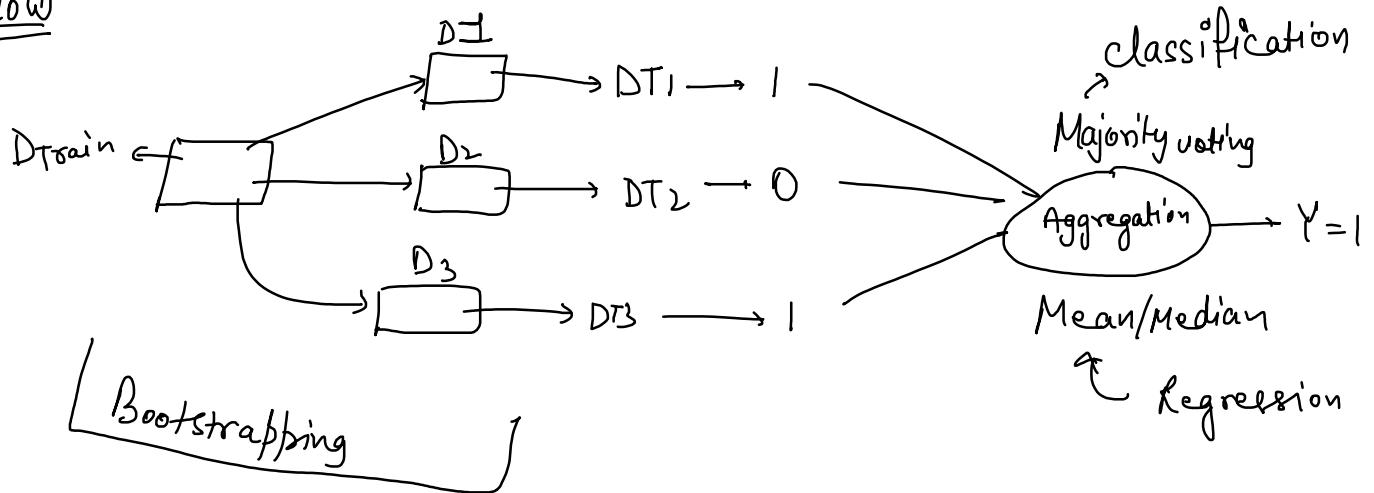
Models  $\Rightarrow$  low bias and high variance  $\rightarrow$  overfitting  
 underfitting  $\qquad\qquad\qquad$  overfitting

Random Forest  $\rightarrow$  (group of decision trees)

$\Rightarrow$  Decision Tree with good depth

DT  $\rightarrow$  "Max\_depth"  $\uparrow \rightarrow$  height of tree  $\uparrow \rightarrow$  overfitting  $\uparrow$ .

Flow



# Models should be diff from each other

| CGPA | IQ  | Extra | Social | Placed | $\rightarrow$ D_train |
|------|-----|-------|--------|--------|-----------------------|
| - 7  | 110 | 10    | 9      | 1      |                       |
| - 8  | 112 | 9     | 8      | 0      |                       |
| - 9  | 120 | 8     | 7      | 0      |                       |
| - 10 | 135 | 7     | 6      | 1      |                       |

|   |    |     |    |   |   |
|---|----|-----|----|---|---|
| - | 7  | 110 | 10 | 9 | 1 |
| - | 8  | 112 | 9  | 8 | 0 |
| - | 9  | 120 | 8  | 7 | 0 |
| - | 10 | 135 | 7  | 6 | 1 |

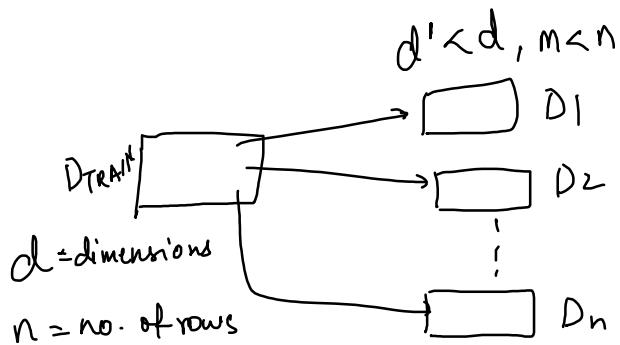
D<sub>1</sub>

D<sub>2</sub>

| CGPA | Extra | Placed | I & | SOCIAL | PLACED |
|------|-------|--------|-----|--------|--------|
| 7    | 10    | 1      | 110 | 9      | 1      |
| 8    | 9     | 0      | 112 | 8      | 0      |

In order to have different samples, you are doing row sampling  
 & column sampling.

RF  $\Rightarrow$  low bias & high variance + Row sampling + Column Sampling  
 ↓  
 max\_depth + Aggregation

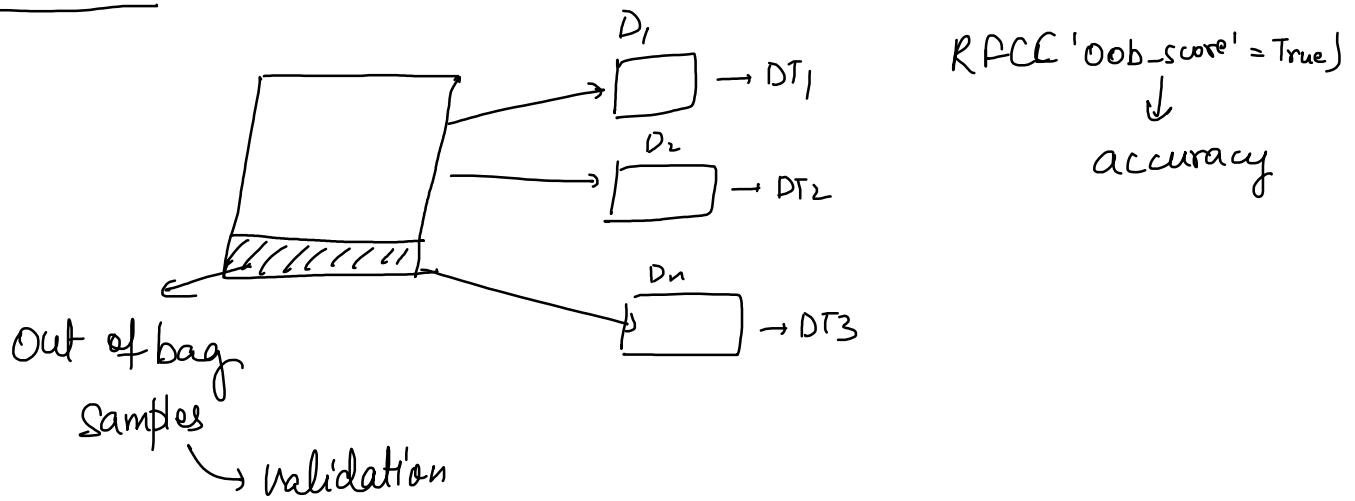


$i/p = \text{input}$        $m/c = \text{machine}$   
 $o/p = \text{output}$   
 $b/w = \text{between}$

Hyperparameters!

- $\Rightarrow \# \text{ models} \Rightarrow n_{\text{estimators}} = [50 - 2000]$   
 $\Rightarrow \text{Row sampling rate} \Rightarrow \frac{m}{n}$   
 $\Rightarrow \text{column sampling} \Rightarrow \text{max\_features} \Rightarrow [\text{'auto'}, \text{'sqrt'}, \text{'lg'}, 0.7, 0.5]$   
 $\Rightarrow \text{column sampling rate} \Rightarrow \frac{d'}{d}$   
 $\Rightarrow \text{max\_depth}$   
 $\Rightarrow n_{\text{jobs}} = -1$

### Oob score



(A) RF  $\Rightarrow [100] \Rightarrow \text{oob-score} \rightarrow$  high (good model)  
 $\rightarrow$  low (you need to improve the model)

## Advantages:

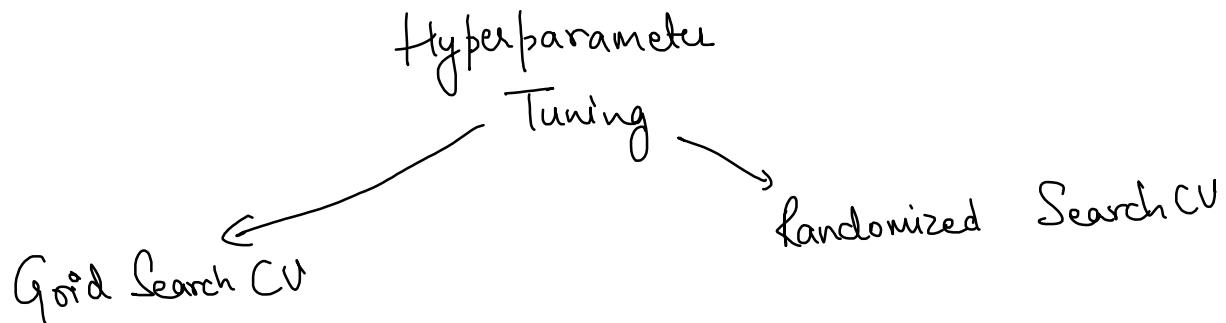
→ Features importances : 100 columns

↳ 30 to 40 columns → 80% of information

## Disadvantages:

→ Black box

→ No Mathematical function



$$\rightarrow n\_estimators \Rightarrow [100, 200, 300, 400] \Rightarrow^4$$

$$\text{max\_depth} \Rightarrow [5, 10, 15, 20] \Rightarrow^4$$

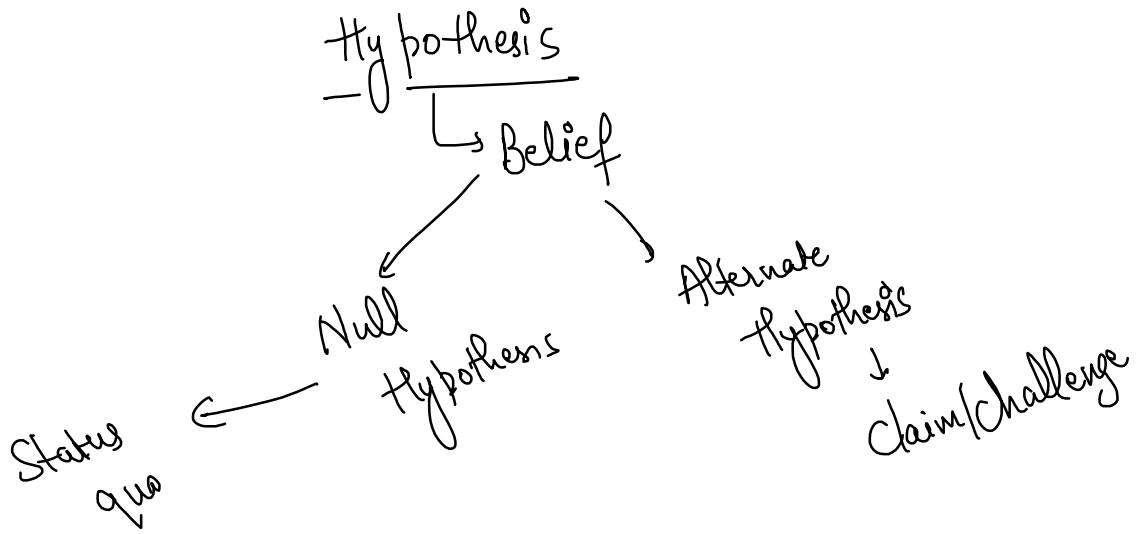
$$\text{max\_features} \Rightarrow [\text{auto}, \sqrt, \sqrt{2}] \Rightarrow^3$$

→ same HPT

→ # model = 10

best-parameter.

$$\# GSCV = 4 \times 4 \times 3 = 48$$



Q India is going to win World cup  $\rightarrow$  I claim

$H_0$ : Any team can win

$H_A$ : India will win

Q Police claims that a person is a criminal?

$H_0$ : innocent

$H_A$ : Criminal

Q Bride claims that groom has taken dowry?

$H_0$ : innocent

Court

Theory  $H_A$ : groom has taken dowry / guilty

$H_0$ : guilty

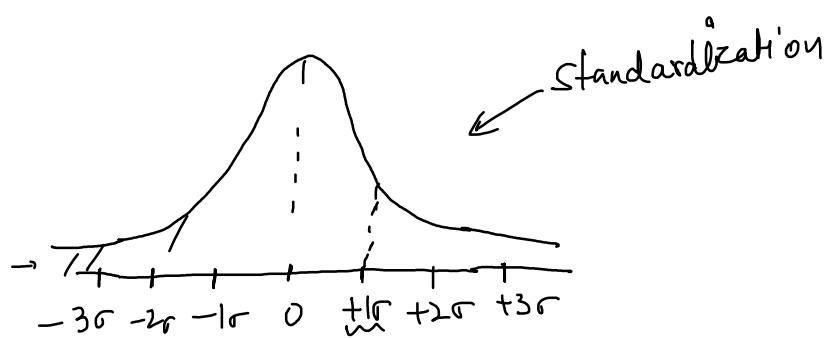
"from aim"  
Reality

$H_A$ : innocent

## Z-score & probability values

$$Z = \frac{x - \mu}{\sigma}$$

Area under curve = prob.

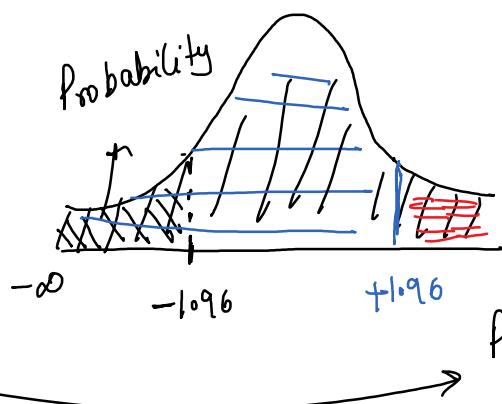


Q Can we connect z-score with prob OR can we get probability for z-score?

$$-\infty \leq Z \leq \infty$$

$$\int_{-\infty}^{\infty} \text{PDF} \Rightarrow \text{Probability}$$

Z-table



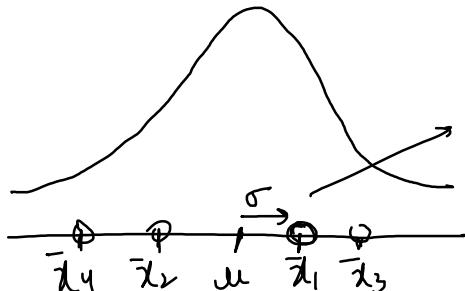
$$\text{Prob} = \int = \text{AVC}$$

$$A_R + A_L = 1 \quad | \quad A_R = 1 - A_L$$

$$\text{Prob} = \int_{-\infty}^{-1.96} \text{PDF}$$

Instead of integration, we will use z-table; In z-table, area is calculated from extreme left to the desired value.

(X - μ)



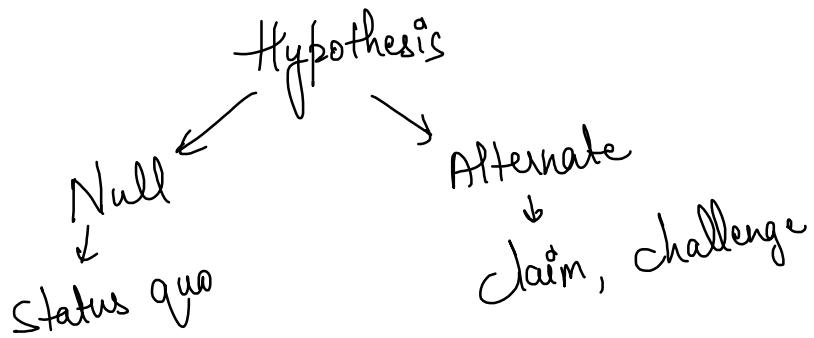
Std error  $\Rightarrow \frac{\sigma}{\sqrt{n}}$

$$Z = \frac{x - \mu}{\left(\frac{\sigma}{\sqrt{n}}\right)}$$

$$\bar{x}_1 \bar{x}_2 \underline{\mu} \bar{x}_3$$

$$\left(\frac{\sigma}{\sqrt{n}}\right)$$

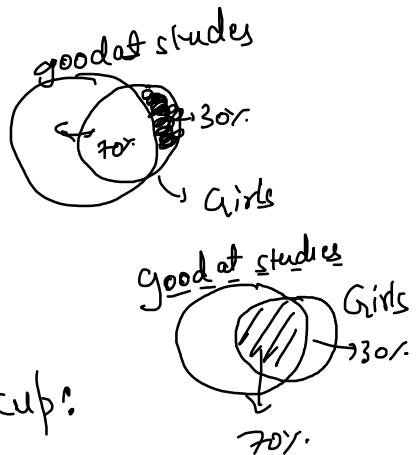
Hypothesis  
Testing



Q g claim. girls are good at studies:

$H_0$ : boys are good at studies.

$H_A$ : girls are good at studies



Q India is going to win the world cup:

$H_0$ : Any team can WC.

$H_A$ : India is winning WC

Q I claim that avg salary of AE changed from \$ 150,000?

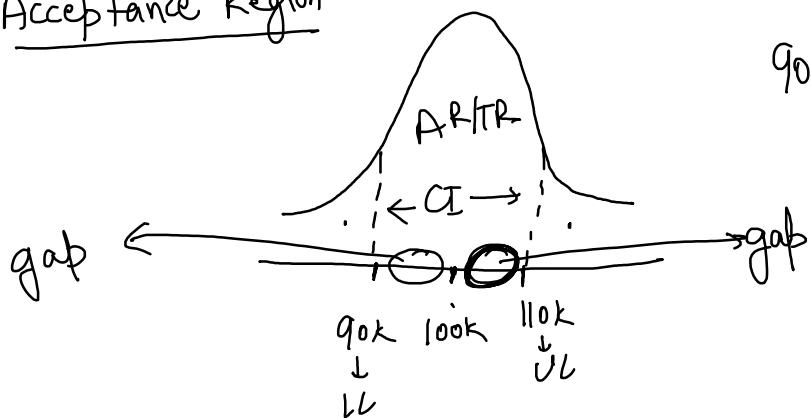
$H_0$ :  $\mu = \$150,000$

$$H_A: \mu \neq \$150,000$$

Built the criteria to test hypothesis:

Q Data scientists earn \$100,000 on avg.

I + Acceptance Region



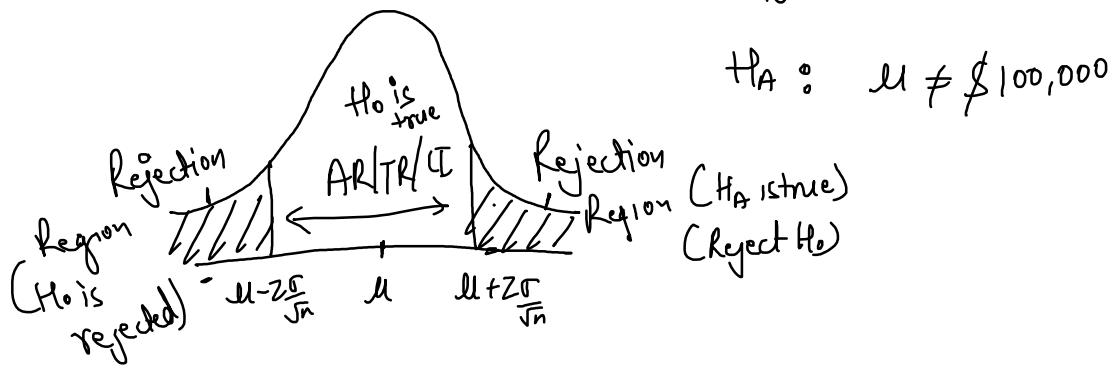
$$LL = \mu - gap$$

$$90k \text{ } \overset{100}{\textcircled{100}} \text{ } 110k$$

$$\begin{aligned} 110k &= \underline{100k} + \overline{gap} && \text{Margin} \\ UL &= \mu + \overline{gap} \Rightarrow \mu + \frac{Z \times \sigma}{\sqrt{n}} && \begin{array}{l} \text{Prob} \\ \text{Std error} \end{array} \end{aligned}$$

$$LL = \mu - Z \times \frac{\sigma}{\sqrt{n}}$$

$$H_0: \mu = \$100,000$$

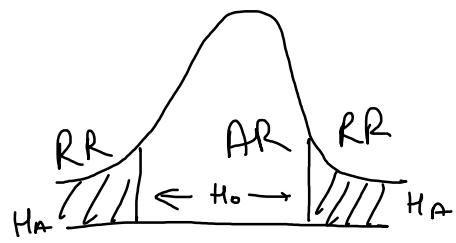


Tails

One tailed → Two tailed

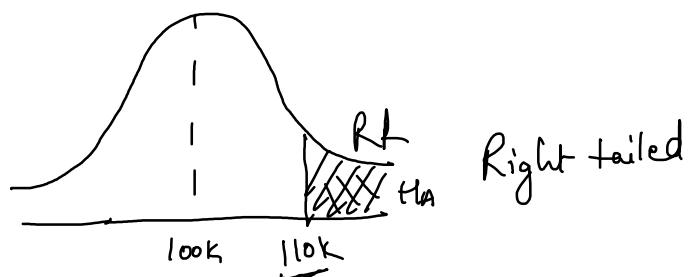
$$Q \quad H_0: \mu = \$100,000$$

$$H_A: \mu \neq \$100,000$$



$$Q \quad H_0: \mu \leq \$100,000$$

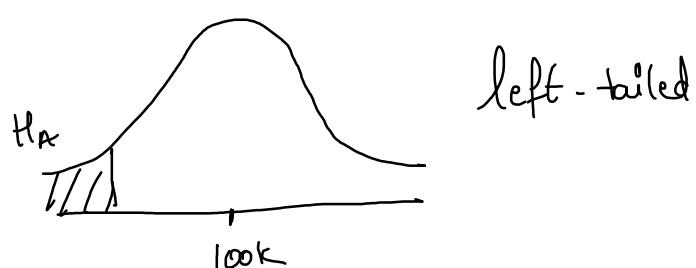
$$H_A: \mu > \$100,000$$



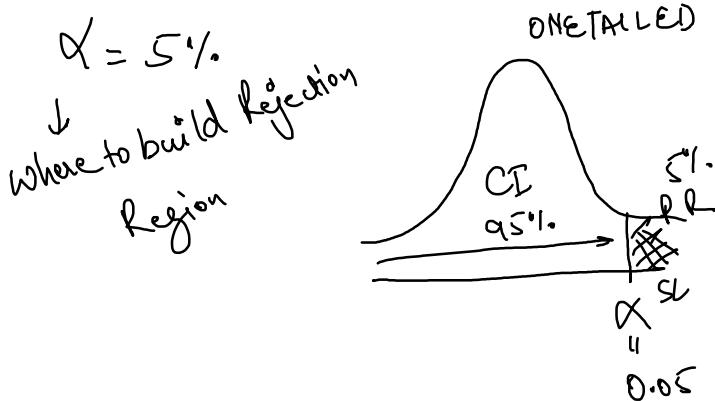
$$Q^3 \quad H_0: \mu \geq \$100,000$$

$$H_A: \mu < \$100,000$$

*Alternate hypothesis*

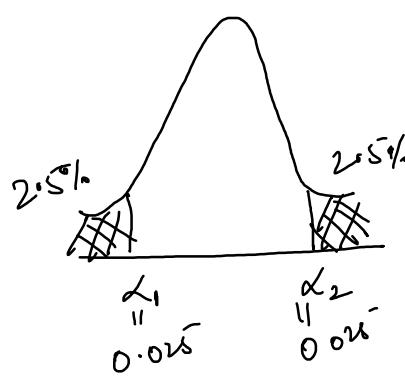


## ② Critical Value Method



$$CI + SL = 1$$

$$\dots \dots \quad CI = 1 - 0.95 = 0.05 = 5\%$$



$\alpha_1 = \text{Area to the left} \Rightarrow \text{prob.}$   
 $\alpha_2 = \text{Area to the right} \Rightarrow \text{prob.}$   
 $\dots \dots \quad 1 - \alpha_1 = 1 - \alpha_2$

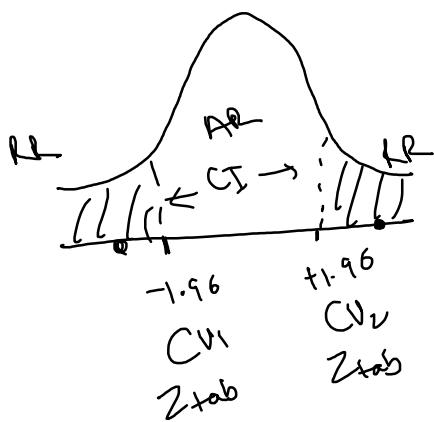
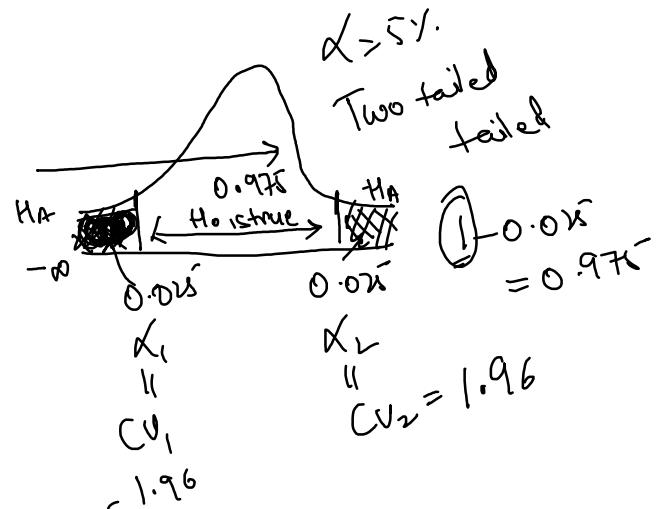
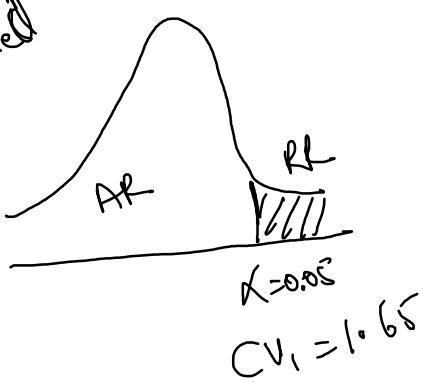
CI = 95%

$$SL = 1 - 0.95 = 0.05 = 5\%$$

$\alpha_2$  = Area to the right

Area to the left =  $1 - \alpha_2$

Ex:  $\alpha = 0.05$   
Test is one tailed  
(Right)



$$Z_{cal} \Rightarrow \frac{x - \mu}{\sigma / \sqrt{n}}$$

Compare  $Z_{cal}$  with  $Z_{tab}$

(if) ①  $Z_{cal} > Z_{tab}$  (for Right)

Reject  $H_0$

②  $Z_{cal} < Z_{tab}$  (for Left)

Reject  $H_0$

III

p-value method : p-value: prob. of null hypothesis to be true

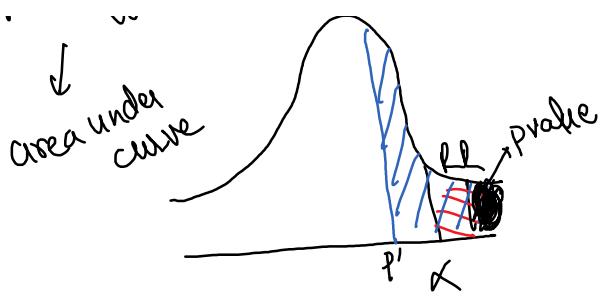
probability value  
 $\downarrow$   
...Area



pvalue  $\leftarrow \alpha$

Reject  $H_0$

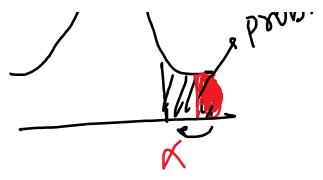




Reject  $H_0$

$p\text{value} > \alpha$

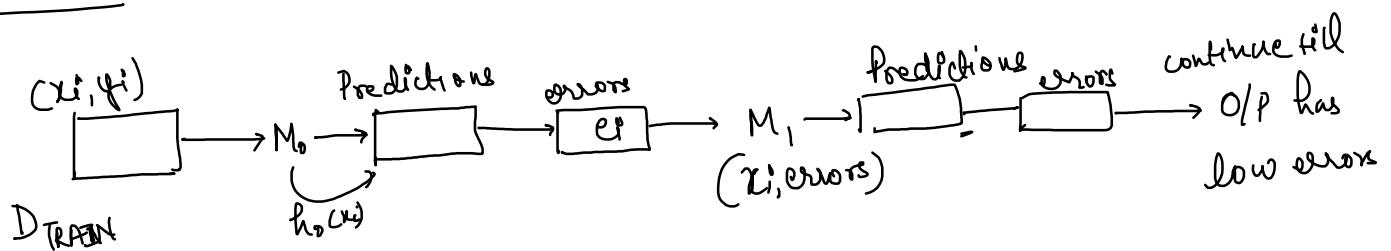
Failed to reject  $H_0$ .



Boosting  
↳ Sequentials

① → Boosting ⇒ high bias & low variance.

## Flowchart:



Steps

o)  $D_{Train} = \{x_i, y_i\}_{i=1}^n \rightarrow M_0 \rightarrow \text{predictions} \rightarrow \text{errors}$

$$e_i = y_i - \hat{y}_i$$

$f^n$   
Mathematical  
 $f^n$

$$e_i = y_i - f^n(x_i)$$

$$1) M_1 \rightarrow \{x_i^*, e_i^*\}_{i=1}^n \Rightarrow e_i^* = y_i^* - f_{\theta}(x)$$

$f_{\theta}(x)$

Model at the end of stage 1:

$$f_i(x) = \underbrace{\alpha_0 h_0(x)}_{\text{old predictions}} + \underbrace{\alpha_1 h_1(x)}_{\text{New predictions}}$$

error  $e_i = y_i - f_i(x)$

$$2) M_2 \rightarrow \{x_i, e_i\} \quad e_i = y_i - f_i(x)$$

$$2) M_2 \rightarrow \{x_i, e_i\} \quad e_i = y_i - f_1(x)$$

$\underbrace{h_2(x)}$

Model at end of stage 2,  $f_2(x)$

$$f_2(x) = \underbrace{\alpha_0 h_0(x) + \alpha_1 h_1(x)}_{f_1(x)} + \alpha_2 h_2(x)$$

New prediction  $\leftarrow$  
$$f_2(x) = f_1(x) + \alpha_2 h_2(x)$$
 Old prediction

for  $k^{\text{th}}$  stage;

$$f_k(x) = \sum_{i=0}^k \alpha_i h_i(x)$$

$K = \# \text{ models}$

~~#~~ Residuals & loss functions:

$$L(y_i, f_k(x)) = [y_i - f_k(x)]^2$$

$$\frac{\partial L}{\partial f_k(x)} = -2[y_i - f_k(x)]$$

pseudo-residual  $\downarrow$  Negative gradient  $\frac{\partial L}{\partial f_k(x)} = [y_i - f_k(x)]$   $\xrightarrow{\text{error}}$

Gradient Boosting

$\mathcal{G}|P \Rightarrow \{x_i, y_i\}_{i=1}^n + \text{differentiable loss function } L[y_i, f_k(x_i)]$

$$0) f_0 = \underset{\gamma}{\operatorname{argmin}} \sum_{i=0}^n L(y_i, \gamma) \quad \text{prediction} \Rightarrow \hat{y}_i$$

1) for  $m=1$  to  $M$

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

$m=1$

$$\alpha_1 = - \frac{\partial L(y_i, f_0(x))}{\partial f_{m-1}(x)}$$

2)  $h_m(x)$  that can fit pseudo-residuals, train  $h_m(x)$  with

$\{x_i, \alpha_m\}$

$$3) f_m = \underset{\gamma}{\operatorname{argmin}} \left[ L \left( y_i, f_{m-1}(x_i) + \gamma h_m(x_i) \right) \right] \quad \text{2nd} = m$$

$$f_{2-1}(x_i) + \gamma_2 h_2(x_i)$$

$$f_1(x_i) + \gamma_2 h_2(x_i)$$

$$\text{4) } f_m = \underbrace{f_{m-1}(x)} + \gamma_m h_m(x)$$

New prediction = old prediction + models (additive combined)

Hyperparameter:  $M \Rightarrow \# \text{models} \uparrow \rightarrow \text{bias} \downarrow \rightarrow \text{variance} \uparrow$

Shrinkage:  $f_m = f_{m-1}(x) + \gamma_m h_m(x)$

$\gamma_m$  learning rate  
 $0 < \gamma < 1$

$\gamma$  reduces  $f_m$  which in turn reduces overfitting.

If we keep all models as DJ (MSE)

GBDT  $\rightarrow$  very slow  $\rightarrow$  optimized (Taylor's Series)

XgBoost  $\rightarrow$  pip install xgboost

Example

|                   | $y_i^o$ |
|-------------------|---------|
| $\frac{58000}{4}$ | 12000   |
| $y$               | 16500   |
| $\downarrow$      | 15500   |
| $\frac{14500}{4}$ | 14000   |

$$\frac{\partial L}{\partial r} = - \sum_{i=0}^n (y_i - \hat{r}_i) \rightarrow \bar{y}_i$$

$$L = \frac{1}{2} (12000 - \hat{r}_i)^2 + \frac{1}{2} (16500 - \hat{r}_i)^2 + \frac{1}{2} (15500 - \hat{r}_i)^2 + \frac{1}{2} (14000 - \hat{r}_i)^2$$

$$0 = \frac{\partial L}{\partial r} = -\frac{1}{2} (12000 - \hat{r}_i) - \frac{1}{2} (16500 - \hat{r}_i) - \frac{1}{2} (15500 - \hat{r}_i) - \frac{1}{2} (14000 - \hat{r}_i)$$

$$-(12000 - \hat{r}_i) - (16500 - \hat{r}_i) - (15500 - \hat{r}_i) - (14000 - \hat{r}_i) = 0$$

$$\hat{r}_i - 12000 + \hat{r}_i - 16500 + \hat{r}_i - 15500 + \hat{r}_i - 14000 = 0$$

$$4\hat{r}_i - 58000 = 0$$

$$4\hat{r}_i = 58000$$

$$\hat{r}_i = \frac{58000}{4} = \frac{14500}{4} = \hat{y}_i$$

# HT - Numericals!

Tuesday, December 26, 2023 7:42 PM

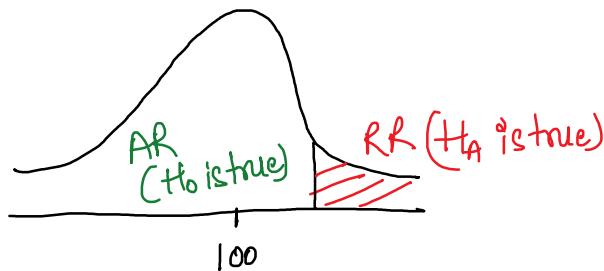
Q A principal of school claims that students have above average IQ. A random sample <sup>(30 students)</sup> is taken with a mean of 112.5. The mean & std dev of population is 100 & 15. Test your hypothesis!

Sol. ①  $H_0: \mu \leq 100$

$$H_A: \mu > 100$$

$$\frac{\mu <}{\text{Left}} \quad \frac{\mu >}{\text{Right}} \quad \frac{\mu \neq}{\text{two tailed}}$$

② Need to check whether test is one tailed & two tailed



③  $\mu = 100, \sigma = 15, \bar{x} = 112.5, \alpha = 0.05, CI = 95\%$

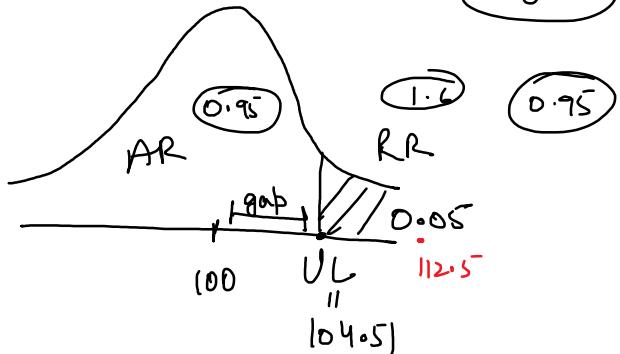
$$Z(\alpha = 0.05) = 1.65$$

$$\alpha = 0.05 \quad | \quad CI = 95\%$$

0.05

AR/TR

$$\begin{aligned} UL &= \mu + Z \times \frac{\sigma}{\sqrt{n}} \\ &= 100 + 1.65 \times \frac{15}{\sqrt{30}} \\ &= 104.5 \end{aligned}$$



$$112.5 > 104.5$$

Reject  $H_0$ ,

2) CRITICAL VALUE

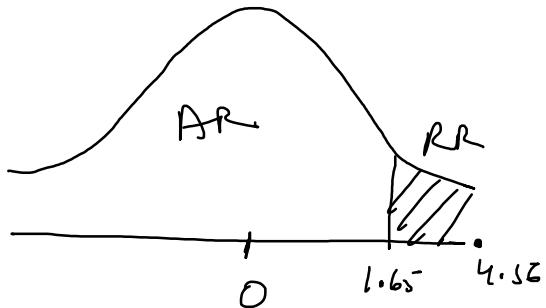
$$N - n - 1 \rightarrow \dots$$

$$\alpha = 0.05, Z_{\text{tab}} = 1.65$$

$$Z_{\text{cal}} = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}} = \frac{112.5 - 100}{\frac{15}{\sqrt{30}}} = 4.56$$

$$Z_{\text{cal}} > Z_{\text{tab}}$$

*Reject H<sub>0</sub>*



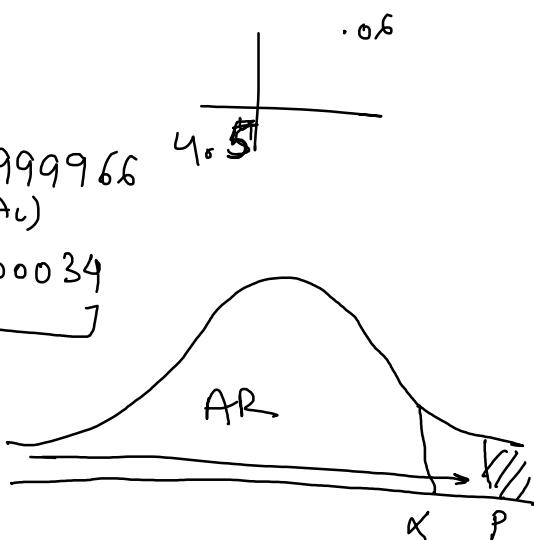
3) P-value  $\alpha = 0.05$

$$Z_{\text{cal}} = \frac{112.5 - 100}{\frac{15}{\sqrt{30}}} = 4.56$$

$$P(Z_{\text{cal}} = 4.56) = 1 - 0.9999966 \quad (A_U)$$

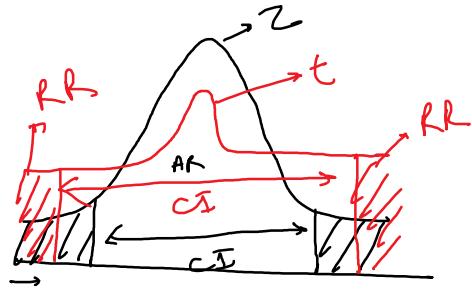
$$= 0.0000034$$

$$\alpha > P$$



*Reject H<sub>0</sub>.*

- Sample size < 30  $\Rightarrow$  Z-distribution is not applicable
  - $\downarrow$
  - T-distribution



$$t = \frac{x - \mu}{\sigma / \sqrt{n}} \quad t \approx Z$$

Degrees of freedom: logically independent values

$$\text{df} = n - 1$$

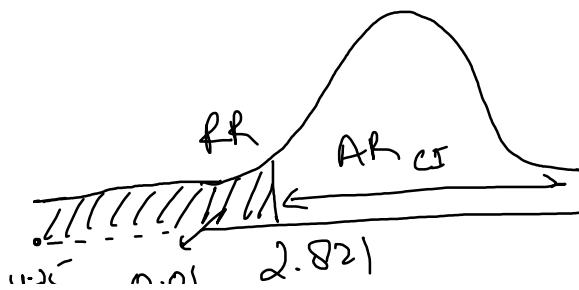
$$\Rightarrow \begin{cases} 2 \\ 3 \\ 5 \\ 8 \\ x \end{cases} \quad \text{avg} = 5 \quad \frac{2+3+5+8+x}{5} = 5 \\ x = 7 \quad 18 + x = 25 \\ x = 7$$

$$\boxed{\text{df} = n - 1}$$

- Q A company manufactures car batteries with avg life span of 2 years or more. An engineer believes this value to be less. Using 10 samples, he measured the life span & found it to be 1.8 years with a std dev. of 0.15. At 99% CI, is there is enough evidence to reject  $H_0$ ?

Sol:  $H_0: \mu \geq 2$

$H_A: \mu < 2$



$$\begin{aligned} CI &= 0.99 \\ x &= 1 - 0.99 \\ &= 0.01 \end{aligned}$$

$$n = 10$$

$$-4 \rightarrow 0.01$$

$$df = n - 1 = 10 - 1 = 9$$

$$t_{\text{cal}} = \frac{1.8 - 2}{\frac{0.15}{\sqrt{10}}} = -4.25$$

$$t_{\text{tab}} \Rightarrow 2.821 \\ (0.01)$$

Reject  $H_0$ .

Errors

Type I      Type II

|                         |  | <u>Actual</u>                      |                                    |
|-------------------------|--|------------------------------------|------------------------------------|
|                         |  | ( $H_A$ is false)<br>$H_0$ is true | ( $H_A$ is true)<br>$H_0$ is false |
| <u>Prediction value</u> | ( $H_A$ is true)<br>$H_0$ Rejected             | Type I error<br>TP                 | ✓                                  |
|                         | Failed to<br>Reject $H_0$<br>( $H_A$ is false) | ✓                                  | FN<br>Type II<br>error             |
|                         |  |                                    |                                    |

Type I  $\Rightarrow$  significance level  $\Rightarrow \alpha$

Type II  $\Rightarrow$   
( $\beta$ )

Power of test  $\uparrow$

$\hookrightarrow$  it's ability of test to make right decision

~ ~

large no. of  
samples

decision

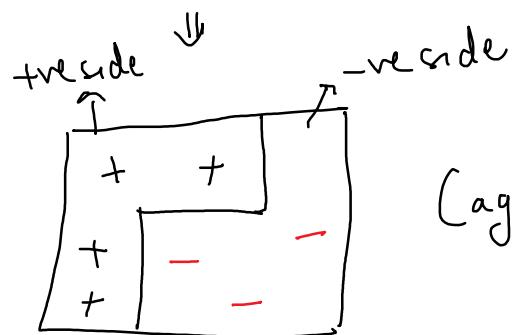
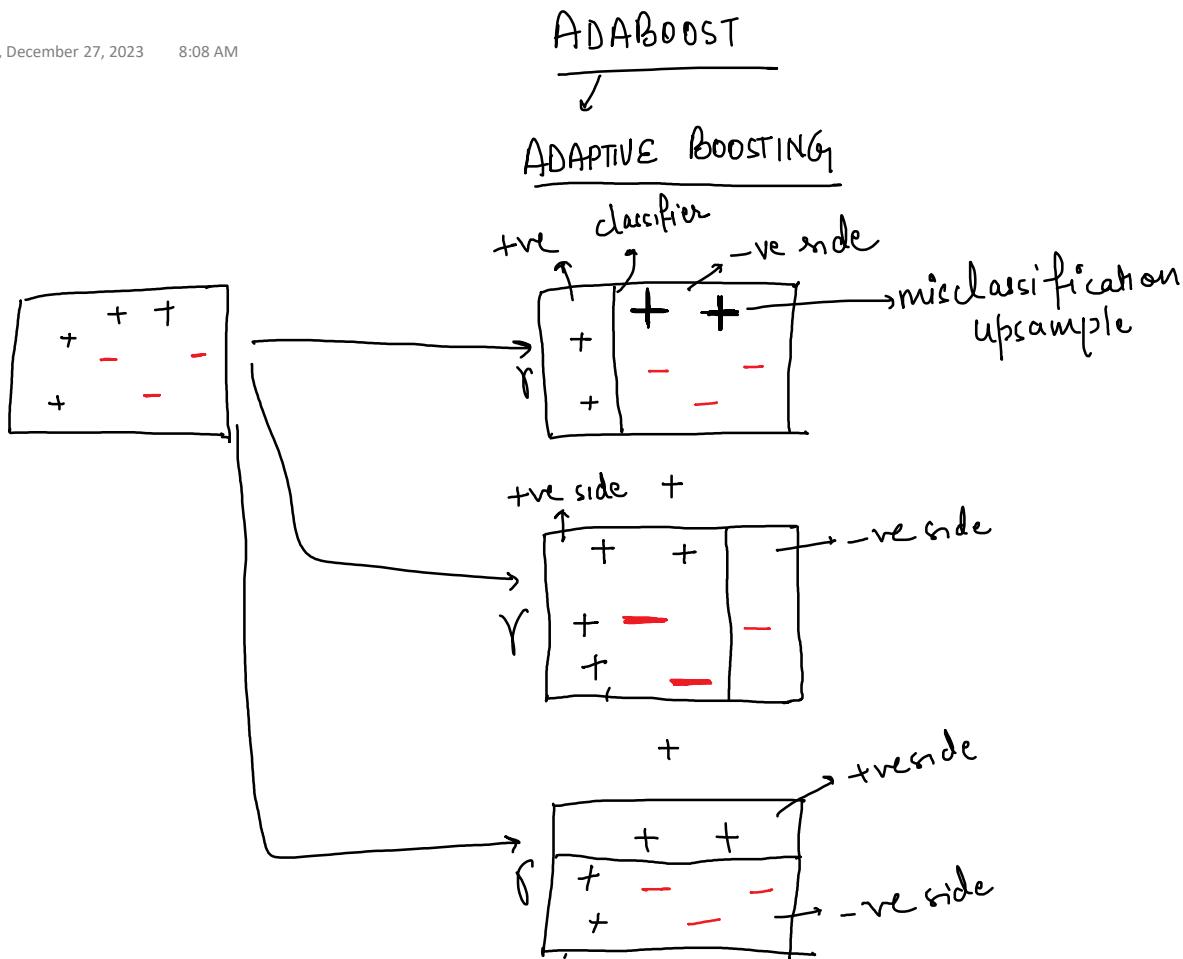
$$\text{Power} = 1 - \beta$$

$$\beta = 1 - \text{Power}$$

Relationship b/w Type I & Type II:

explored.

$$\text{Type I} \propto \frac{1}{\text{Type II}}$$



(aggregated classifier/model)

$$C = \gamma_1 C_1 + \gamma_2 C_2 + \gamma_3 C_3 + \dots$$

$n = \# \text{ rows}$

| $X_1$ | $X_2$ | $Y$ | $\hat{Y}$ | weight = $\gamma_n$ |
|-------|-------|-----|-----------|---------------------|
| 1     | 3     | 1   | 1         | $\gamma_5 = 0.2$    |
| 2     | 4     | 0   | 1         | $\gamma_5 = 0.2$    |

$\times$  = error rate

Error = algebraic sum of weights  
at misclassified points

|   |   |   |   |   |   |             |
|---|---|---|---|---|---|-------------|
| 1 | 3 | 1 | 4 | 0 | 1 | $y_5 = 0.2$ |
| 2 | 2 | 4 | 0 | 1 | * | $y_5 = 0.2$ |
| 3 | 1 | 5 | 1 | 0 | * | $y_5 = 0.2$ |
| 4 | 9 | 6 | 0 | 0 | 0 | $y_5 = 0.2$ |
| 5 | 5 | 7 | 0 | 0 | 0 | $y_5 = 0.2$ |

error = algebraic sum of weights  
of misclassified points

$$\text{error} = 0.2 + 0.2 = 0.4$$

$$\alpha = \frac{1}{2} \ln \left( \frac{1 - \text{error}}{\text{error}} \right) \quad \text{error} = 0.4$$

$$\alpha = \frac{1}{2} \ln \left( \frac{1 - 0.4}{0.4} \right) = \frac{1}{2} \ln \left( \frac{0.6}{0.4} \right) = 0.2$$

$$\begin{aligned} \text{new weights for correctly classified points} &= e^{-\alpha} \times \text{old weight} = e^{-0.2} \times 0.2 \\ &= 0.16 \end{aligned}$$

$$\begin{aligned} \text{new weights for misclassified points} &= e^{\alpha} \times \text{old weight} = e^{0.2} \times 0.2 \\ &= 0.24 \end{aligned}$$

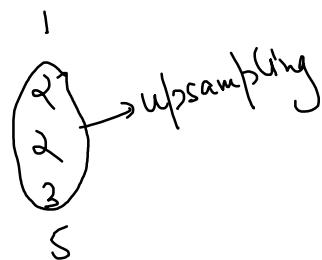
| $x_1$ | $x_2$ | $y$ | $\hat{y}$ | weights | new weights | Normalized weights  |
|-------|-------|-----|-----------|---------|-------------|---------------------|
| 3     | 9     | 1   | 1         | 0.2     | 0.16        | $0.16/0.96 = 0.167$ |
| 2     | 4     | 0   | *         | 0.2     | 0.24        | $0.24/0.96 = 0.25$  |
| 1     | 5     | 1   | *         | 0.2     | 0.24        | $0.24/0.96 = 0.25$  |
| 9     | 6     | 0   | 0         | 0.2     | 0.16        | $0.16/0.96 = 0.167$ |
| 5     | 7     | 0   | 0         | 0.2     | 0.16        | $0.16/0.96 = 0.167$ |
|       |       |     |           |         | <u>0.96</u> | <u>1</u>            |

$$5 + 0 - \frac{1}{\underline{\underline{0.96}}} = \underline{\underline{1}}$$

| $\hat{Y}$ | Norm Weight | Range         | Row No |
|-----------|-------------|---------------|--------|
| 0.16      | 0           | 0 - 0.167     | 1      |
| 0.25      | 0.167       | 0.167 - 0.417 | 2      |
| 0.25      | 0.417       | 0.417 - 0.667 | 3      |
| 0.16      | 0.667       | 0.667 - 0.834 | 4      |
| 0.16      | 0.834       | 0.834 - 1     | 5      |

Randomly pick 5 no b/w 0 & 1

$$0.1, 0.2, 0.9, 0.5, 0.9 \\ \downarrow \quad \downarrow \quad \downarrow \quad \downarrow \quad \downarrow \\ 1 \quad 2 \quad 2 \quad 3 \quad 5$$



CHI-SQUARE TEST

- Non-parametric → No distribution
- char-char situations  
(categorical)

$$\text{Degrees of freedom} = (\underbrace{r-1}_{\text{rows}}) (\underbrace{c-1}_{\text{columns}})$$

Q What is the relationship b/w gender and result?

|  |  | Result | Pass | Fail |
|--|--|--------|------|------|
|  |  | Gender |      |      |
|  |  | M      | 60   | 40   |
|  |  | F      | 24   | 32   |

Sol.  $H_0$ : There is no relationship b/w gender and result

$H_A$ : There is relationship b/w gender & result

|  |  | Result | Pass       | Fail      |            |
|--|--|--------|------------|-----------|------------|
|  |  | Gender |            |           |            |
|  |  | M      | 60         | 40        | = 100      |
|  |  | F      | 24         | 32        | = 56       |
|  |  |        | <u>n..</u> | <u>72</u> | <u>156</u> |

$$\begin{array}{r}
 F \\
 \hline
 84 \\
 \hline
 72 \\
 \hline
 156
 \end{array}$$

Total Males = 100 Total Females = 56 Total Pass = 84 Total Fail = 72

Expected Values:

$$\text{expected value} = \frac{\text{Total males} \times \text{total pass}}{\text{total no. of people}} = \frac{100 \times 84}{156} = 53.84$$

(total males who passed)

$$\text{expected value} = \frac{\text{Total females} \times \text{Total passed}}{\text{total no. of people}} = \frac{56 \times 84}{156} = 30.15$$

(total females who passed)

$$\text{expected value} = \frac{\text{Total males} \times \text{Total fail}}{\text{total no. of people}} = \frac{100 \times 72}{156} = 46.15$$

(total males who failed)

$$\text{expected value} = \frac{\text{Total females} \times \text{Total fail}}{\text{Total No. of people}} = \frac{56 \times 72}{156} = 25.84$$

(total females who failed)

$$EV_1 = 53.8 \quad EV_2 = 30.1 \quad EV_3 = 46.1 \quad EV_4 = 25.84$$

| Result | Pass                   | Fail                         | $df = (r-1)(c-1)$ |
|--------|------------------------|------------------------------|-------------------|
| Gender |                        |                              |                   |
| → M    | <u>53.8</u> ( $EV_1$ ) | <u>46.1</u> ( $EV_3$ ) = 100 |                   |
|        |                        |                              | = 56              |

$$\rightarrow F \quad \begin{array}{r} \underline{30.1} \\ \underline{87} \end{array} \quad \begin{array}{r} \underline{25.8} \\ \underline{72} \end{array} \stackrel{(EV_u)}{=} \begin{array}{r} 56 \\ \sqrt{156} \end{array}$$

Calculation of  $\chi^2$ :  $\chi^2 = \frac{(Actual - Expected)^2}{Expected}$

$$\textcircled{I} \quad \frac{(60 - 53.8)^2}{53.8} = 0.71$$

$$\textcircled{II} \quad \frac{(40 - 46.1)^2}{46.1} = 0.81$$

$$\textcircled{III} \quad \frac{(24 - 30.1)^2}{30.1} = 1.23$$

$$\textcircled{IV} \quad \frac{(32 - 25.8)^2}{25.8} = 1.48$$

$$\chi_{\text{cal}}^2 = 0.71 + 0.81 + 1.23 + 1.48 = 4.23$$

$$\text{Find } \chi_{\text{tab}}^2 : \chi = 0.05, df = (r-1)(c-1) \\ = (2-1)(2-1) = 1 \times 1 = 1$$

$$\chi_{\text{tab}}^2 = 3.841$$

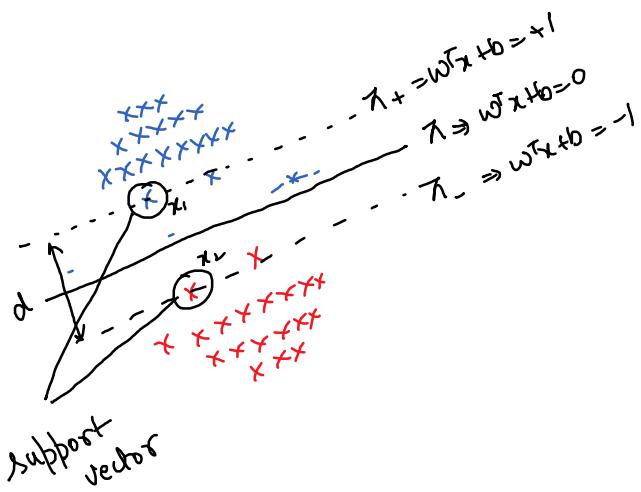
Compare:  $\chi_{\text{cal}}^2$  with  $\chi_{\text{tab}}^2$

Reject  $H_0$

$$\chi^2_{\text{cal}} > \chi^2_{\text{tab}}$$

$$4.23 > 3.84$$

## SVM (Support Vector Machine)



$$\begin{aligned} x_+ &= w^T x_1 + b = +1 \\ x_- &= w^T x_2 + b = -1 \\ \hline w^T (x_1 - x_2) &= 2 \end{aligned}$$

Let's Normalize,

$$\frac{w^T}{\|w\|} (x_1 - x_2) = \frac{2}{\|w\|}$$

$x_1 - x_2 = \frac{2}{\|w\|} = d$

$$MOF \Rightarrow f(x) = \underset{w}{\operatorname{argmax}} \frac{2}{\|w\|}$$

for each datapoint in negative zone,  
 $w^T x + b < 0$

for each datapoint in positive zone,  
 $w^T x + b > 0$

Simplify:  $y_i(w^T x_i + b)$

Case 1:  $y_i = +ve$   $w^T x_i + b > 0$

$$y_i(w^T x_i + b) > 0$$

Case 2:  $y_i = -ve$ ,  $w^T x_i + b < 0$

$$y_i(w^T x_i + b) > 0$$

*Correct  
Direction*

$$y_i(w^T x_i + b) > 0$$

Correct classification

Case 3:  $y_i = +ve, w^T x_i + b < 0$

$$y_i(w^T x_i + b) < 0$$

Incorrect classification

Case 4:  $y_i = -ve, w^T x_i + b > 0$

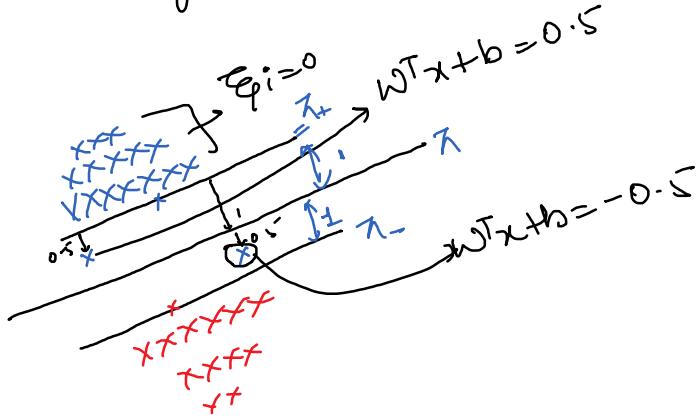
$$y_i(w^T x_i + b) < 0$$

For SVM  $\Rightarrow CC \Rightarrow y_i(w^T x_i + b) \geq 1$

MOF  $\Rightarrow \arg \max_w \frac{2}{\|w\|}$  such that  $y_i(w^T x_i + b) \geq 1$

↳ Hard margin

## Soft Margin



$$y_i(w^T x_i + b) = 0.5$$

$$\begin{aligned} y_i(w^T x_i + b) &= -0.5 \\ &= 1 - (1.5) \\ &\downarrow \\ \epsilon_i & \end{aligned}$$

$\epsilon_i$ : measure of how far a datapoint is in opp. direction from its correct plane.

$$\text{MoF} \Rightarrow f(x) = \underset{w, b}{\operatorname{argmax}} \frac{2}{\|w\|} + C \sum_{i=1}^n \epsilon_i =$$

↓  
Regularizer  
↓  
loss

$$\begin{aligned} \text{Loss} &= f(x) = \underset{w}{\operatorname{argmin}} \frac{\|w\|}{2} + C \sum_{i=1}^n \epsilon_i \\ f &\quad \text{regularizer} \quad \text{loss} \\ &\quad \downarrow \end{aligned}$$

$$f(x) = \sum_{i=1}^n \epsilon_i + \lambda \frac{\|w\|}{2}$$

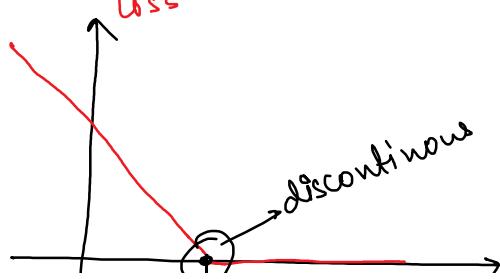
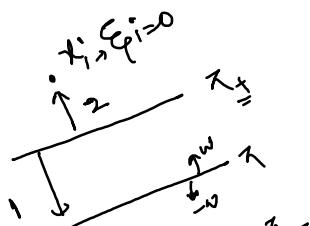
$C \uparrow \rightarrow$  More focus on errors  $\rightarrow$  less errors  $\rightarrow$  overfitting

$C \downarrow \rightarrow$  less focus on errors  $\rightarrow$  more errors  $\rightarrow$  underfitting

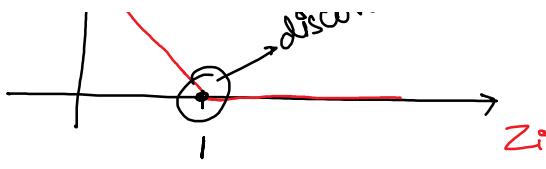
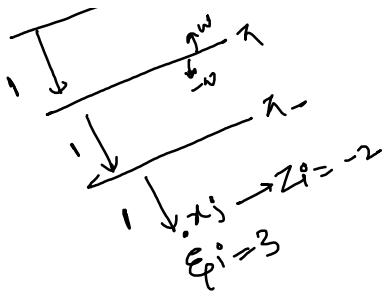
$$C \propto \frac{1}{\lambda}$$

### Loss Minimization (Hinge Loss)

$$z_i \rightarrow y_i (w^T x_i + b) \geq 1$$



$$1 - (-2) = \epsilon_i = 3$$



$$1 - (-2) = \xi_i = 3$$

$$1 - z_i = \xi_i$$

$$\text{Hinge loss} = \max(0, 1 - z_i)$$

① for correct classification,  $z_i \geq 1$ ,  $z_i = 2$  (assume)

$$\text{Loss} = \max(0, 1 - z) \Rightarrow \max(0, -1) = 0$$

② for incorrect classification,  $\xi_i = 1 - z_i = 1 - (-2) = 3$

$$\text{Loss} = \max(0, \frac{1 - z_i}{\xi_i}) = \max(0, 3) = 3$$

### Dual Form of SVM:

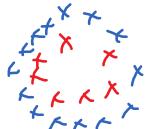
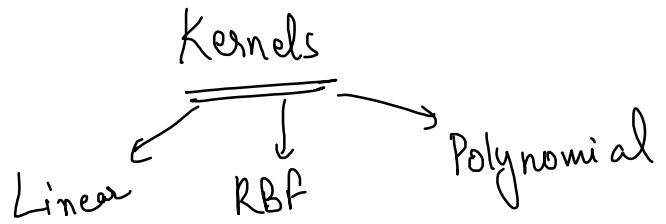
$$\text{Primal form: } \underset{w}{\operatorname{argmin}} \frac{\|w\|^2}{2} + C \sum_{i=1}^n \xi_i =$$

$$\text{Dual form: } \max_{\alpha_i} \sum_{i=1}^n \alpha_i + \frac{1}{2} \sum_{i=1}^n \sum_{j=1}^n \alpha_i \alpha_j y_i y_j (x_i^\top x_j)$$

similarity  
 kernel trick

$\alpha_i \Rightarrow$  support vectors       $\alpha_i > 0$  (for support vectors only)

$\alpha_i = 0$  (for other vectors)



"Kernel trick": apply kernels to transform data points to make linearly separable

Mercer's Theorem: Kernel converts the  $d$ -dimension dataset into  $d'$  dimension dataset such that  $d' > d$ .

Polynomial Kernel:  $k(x_1, x_2) = [x_1^T x_2 + c]^d$        $x_1 = \begin{bmatrix} x_{11} \\ x_{12} \end{bmatrix}$

Quadratic fn:  $d=2 \Rightarrow [1 + x_1^T x_2]^2$        $x_2 = \begin{bmatrix} x_{21} \\ x_{22} \end{bmatrix}$

$d \Rightarrow$  degrees

$$\Rightarrow \left[ 1 + [x_{11} \ x_{12}] \begin{bmatrix} x_{21} \\ x_{22} \end{bmatrix} \right]^2$$

$$\Rightarrow \left[ 1 + x_{11}x_{21} + x_{12}x_{22} \right]^2 = (a+b+c)^2$$

$$= \left[ 1 + x_{11}^2 x_{21}^2 + x_{12}^2 x_{22}^2 + 2x_{11}x_{21} + 2x_{11}x_{21}x_{12}x_{22} + 2x_{12}x_{22} \right]$$

$$x_1 \rightarrow [1, x_{11}^2, x_{12}^2, \sqrt[3]{2}x_{11}, \sqrt[5]{2}x_{12}, \sqrt[6]{2}x_{11}x_{12}] \rightarrow 6d$$

$$x_2' \rightarrow [1, x_{21}^2, x_{22}^2, \sqrt{2}x_{21}, \sqrt{2}x_{22}, \sqrt{2}x_{21}x_{22}] \rightarrow 6d$$

RBF (Radial Basis Function)

$$\text{Radial Basis Function) } RBF \quad K(x_1, x_2) = C e^{\frac{-||x_1 - x_2||^2}{2\sigma^2}} = e^{-\frac{d^2}{2\sigma^2}}$$

Calc $\|x_1 - x_2\|$  = d  $\rightarrow$  distance

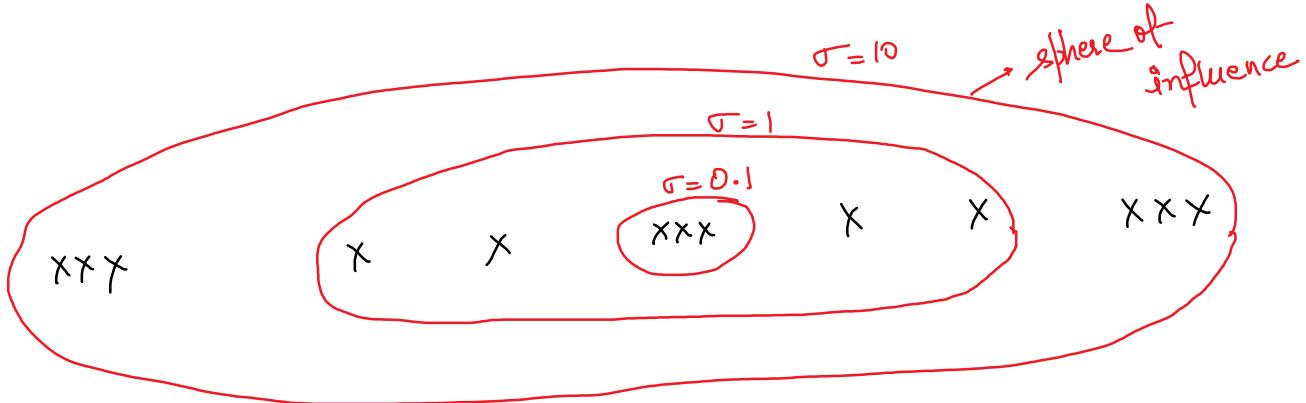
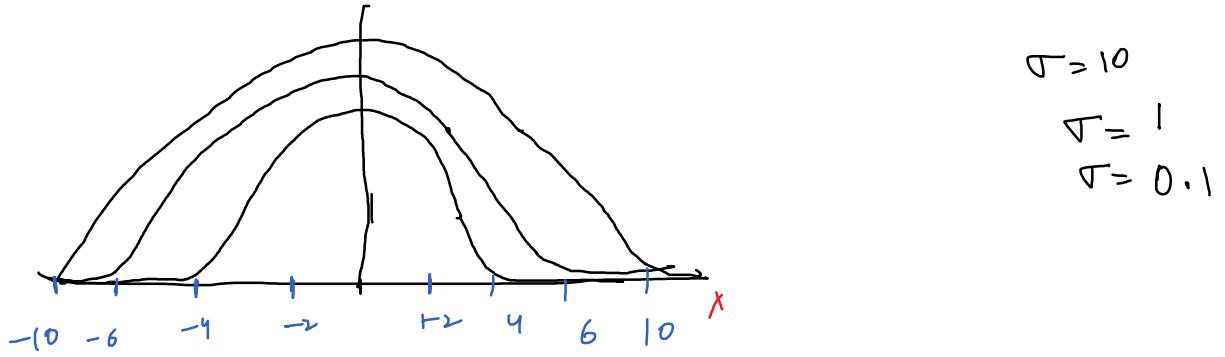
$$\pi = e^{-d^2/2r^2} = \frac{1}{e^{d^2/2r^2}}$$

$$d^{\uparrow} \rightarrow d^2 \uparrow \rightarrow \frac{d^2}{2\sigma^2} \uparrow \rightarrow e^{d^2/2\sigma^2} \uparrow \rightarrow \frac{1}{e^{d^2/2\sigma^2}} \downarrow \rightarrow k \downarrow$$

$$\underline{\underline{Calc}} \stackrel{2^{\circ}}{=} K = \frac{J}{d^2/2\sigma^2}$$

$$\tau \uparrow \rightarrow \tau^2 \uparrow \rightarrow \frac{d^2}{2\sigma^2} \downarrow \rightarrow e^{d^2/2\sigma^2} \downarrow \rightarrow \frac{1}{e^{d^2/2\sigma^2}} \uparrow \rightarrow kT$$

k



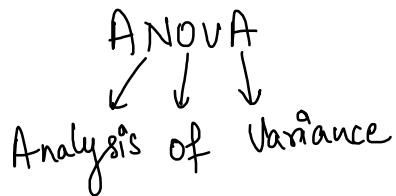
RBF  $\rightarrow$  hyperparameters

|                      |                       |
|----------------------|-----------------------|
| $C =$                | $\frac{1}{2\sigma^2}$ |
| $\gamma \Rightarrow$ | $\frac{1}{2\sigma^2}$ |

$$\sigma \uparrow \rightarrow \sigma^2 \uparrow \rightarrow 2\sigma^2 \uparrow \rightarrow \frac{1}{2\sigma^2} \downarrow \rightarrow \gamma \downarrow \rightarrow \frac{1}{e^{d^2/2\sigma^2}} \rightarrow \frac{1}{e^{d^2\gamma \downarrow}} \uparrow$$

$\downarrow$

$K \uparrow$



→ Extension of Z-test / t-test

Fischer  
F-statistic

→ It is used to compare variances among groups.

$$\rightarrow F = \frac{\text{variance of 1st group}}{\text{variance of 2nd group}} = \frac{SD_1^2}{SD_2^2} \Rightarrow SD_1 > SD_2$$

ANOVA:

One Way ANOVA

Q1: To assess the significance of possible variation in performance in a certain test between the convent schools of a city, a common test was given to a number of students taken at random from the 5<sup>th</sup> class of the 3 schools concerned. The result is given as follows:

| A  | B  | C  |
|----|----|----|
| 9  | 13 | 14 |
| 11 | 12 | 13 |
| 13 | 10 | 17 |
| 9  | 15 | 7  |
| 8  | 5  | 9  |

$n=15$        $\sum = 12$

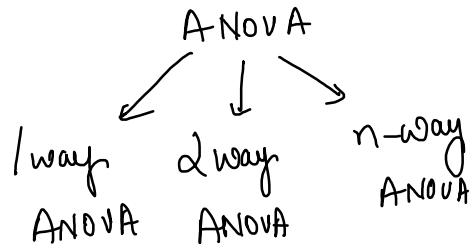
Make the Analysis of Variance of the given data. (Null Hypo: No Significance Variation in the schools).

Solution:

Null Hypothesis = No variation between schools

Alt. Hypothesis = There is variation between schools

| Source of Variation | Sum of Square       | Degrees of freedom  | Mean Square     | F              |
|---------------------|---------------------|---------------------|-----------------|----------------|
| Between the Sample  | $SSC = 10$          | $(c-1) = 3-1 = 2$   | $MSC = SSC/df1$ | $F = MSC/MSE$  |
| Within the sample   | $SSE = 15 - 10 = 5$ | $(n-c) = 15-3 = 12$ | $MSE = SSE/df2$ | $5/12 = 0.417$ |



$H_0$ : No variation in schools

$H_A$ : There is variation

$$n = 15 \quad c = 3$$

$$\bar{x}_A = \frac{50}{5} = 10 \quad \bar{\bar{x}} = \frac{10+11+12}{3}$$

$$\bar{x}_B = \frac{55}{5} = 11$$

$$\bar{x}_C = \frac{60}{5} = 12$$

SSC

$$\bar{x}_A - \bar{\bar{x}} \quad (\bar{x}_A - \bar{\bar{x}})^2 \quad (\bar{x}_B - \bar{\bar{x}}) \quad (\bar{x}_B - \bar{\bar{x}})^2$$

$$|10 - 11| = 1 \quad (10 - 11)^2 = 1 \quad |11 - 11| = 0 \quad 0$$

$$|10 - 11| = 1 \quad (10 - 11)^2 = 1 \quad |11 - 11| = 0 \quad 0$$

$$|10 - 11| = 1 \quad (10 - 11)^2 = 1 \quad |11 - 11| = 0 \quad 0$$

$$|10 - 11| = 1 \quad (10 - 11)^2 = 1 \quad |11 - 11| = 0 \quad 0$$

$$(\bar{x}_C - \bar{\bar{x}}) \quad (\bar{x}_C - \bar{\bar{x}})^2$$

$$|12 - 11| = 1$$

$$|12 - 11| = 1$$

$$|12 - 11| = 1$$

$$|12 - 11| = 1$$

$$|12 - 11| = 1$$

$$\begin{array}{rccccc}
 10 - 11 = -1 & (10 - 11)^2 = 1 & 11 - 11 = 0 & 0 & 12 - 11 = 1 & 1 \\
 10 - 11 = -1 & (10 - 11)^2 = 1 & 11 - 11 = 0 & 0 & 12 - 11 = 1 & 1 \\
 10 - 11 = -1 & \underline{(10 - 11)^2 = 1} & 11 - 11 = 0 & 0 & & \\
 & \underline{\underline{S}} & + & \underline{\underline{0}} & + & \underline{\underline{S}}
 \end{array}$$

$$SSC = S + 0 + S = 10$$

$$Df_1 = c - 1 = 3 - 1 = 2$$

$$MSC = \frac{SSC}{Df_1} = \frac{10}{2} = 5$$

SSE

$$\begin{array}{ccccccc}
 A - \bar{x}_A & (A - \bar{x}_A)^2 & (B - \bar{x}_B) & (B - \bar{x}_B)^2 & (C - \bar{x}_C) & (C - \bar{x}_C)^2 \\
 9 - 10 = -1 & 1 & 13 - 11 = 2 & 4 & 14 - 12 = 2 & 4 \\
 11 - 10 = 1 & 1 & 12 - 11 = 1 & 1 & 13 - 12 = 1 & 1 \\
 13 - 10 = 3 & 9 & 10 - 11 = -1 & 1 & 17 - 12 = 5 & 25 \\
 9 - 10 = -1 & 1 & 15 - 11 = 4 & 16 & 7 - 12 = -5 & 25 \\
 8 - 10 = -2 & \underline{9} & 5 - 11 = -6 & \underline{36} & 9 - 12 = 3 & \underline{\underline{9}} \\
 & \underline{\underline{16}} & + & \underline{\underline{58}} & + & \underline{\underline{64}} & = 138
 \end{array}$$

$$Df_2 = n - c = 15 - 3 = 12$$

$$MSE = \frac{138}{12} = 11.5$$

$$F_{cal} \Rightarrow 0.435$$

|                |                  |
|----------------|------------------|
| $Df_1 = 2 = 2$ | $f_{tab} = 2.81$ |
| $Df_2 = 12$    |                  |

$\therefore f_{tab} > f_{cal} \therefore$  Failed to Reject  $H_0$

## 2-way

The following data represents the number of Units of Tablet production (in thousands) per day by five different technicians by using 4 different machines.

- Tell whether the mean productivity of the different machines are same?
- Test whether the 5 technicians differ w.r.t. the mean productivity?

| Machines<br>.....<br>Technicians | A  | B  | C  | D  |
|----------------------------------|----|----|----|----|
| P                                | 54 | 48 | 57 | 46 |
| Q                                | 56 | 50 | 62 | 53 |
| R                                | 44 | 46 | 54 | 42 |
| S                                | 53 | 48 | 56 | 44 |
| T                                | 48 | 52 | 59 | 48 |

$$\text{Mid value} = \frac{62+42}{2} = 52$$

$$MSC = \frac{SSC}{df} = \frac{338}{3} = 112.93$$

$$MSR = \frac{SSR}{df} = \frac{158}{4} = 39.5$$

$$MSE = \frac{67.2}{12} = 5.6$$

Sol. ① → Calculate Grand total

$$\text{Mid value} = 50$$

|           | A              | B               | C               | D                | Total     |
|-----------|----------------|-----------------|-----------------|------------------|-----------|
| P         | $54 - 50 = 4$  | $48 - 50 = -2$  | $57 - 50 = 7$   | $46 - 50 = -4$   | 5         |
| Q         | $56 - 50 = 6$  | $50 - 50 = 0$   | $62 - 50 = 12$  | $53 - 50 = 3$    | 21        |
| R         | $44 - 50 = -6$ | $46 - 50 = -4$  | $54 - 50 = 4$   | $42 - 50 = -8$   | -14       |
| S         | $53 - 50 = 3$  | $48 - 50 = -2$  | $56 - 50 = 6$   | $44 - 50 = -6$   | 1         |
| T         | $48 - 50 = -2$ | $52 - 50 = 2$   | $59 - 50 = 9$   | $40 - 50 = -10$  | 7         |
| $T_{1-0}$ |                | $\overline{-6}$ | $\overline{58}$ | $\overline{-17}$ | <u>20</u> |

|       |        |                |                |        |      |        |      |        |       |        |                    |
|-------|--------|----------------|----------------|--------|------|--------|------|--------|-------|--------|--------------------|
| Total | $\sum$ | $48 - 50 = -2$ | $52 - 50 = -2$ | $\sum$ | $-6$ | $\sum$ | $88$ | $\sum$ | $-17$ | $\sum$ | $20$               |
|       |        |                |                |        |      |        |      |        |       |        | Grand total<br>(T) |

~~Explained~~ Correction factor =  $\frac{T^2}{N} = \frac{20^2}{20} = 20$

$$\underline{\underline{SSC}} \Rightarrow \left( \frac{\sum A}{n_A} \right)^2 + \left( \frac{\sum B}{n_B} \right)^2 + \left( \frac{\sum C}{n_C} \right)^2 + \left( \frac{\sum D}{n_D} \right)^2 - \text{correction factor}$$

$$\Rightarrow \frac{5^2}{5} + \frac{(-6)^2}{5} + \frac{38^2}{5} + \frac{(-17)^2}{5} - \frac{20^2}{20}$$

$$\Rightarrow 338.8$$

$$\underline{\underline{SSR}} \Rightarrow \left( \frac{\sum P}{n_P} \right)^2 + \left( \frac{\sum Q}{n_Q} \right)^2 + \left( \frac{\sum R}{n_R} \right)^2 + \left( \frac{\sum S}{n_S} \right)^2 + \left( \frac{\sum T}{n_T} \right)^2 - \frac{T^2}{N}$$

$$\Rightarrow \frac{5^2}{4} + \frac{21^2}{4} + \frac{(-14)^2}{4} + \frac{1}{4} + \frac{7^2}{4} - \frac{20^2}{20}$$

$$\Rightarrow 158$$

$SST \Rightarrow$  Sum of square of all observed residuals -  $\frac{T^2}{N}$

$$\Rightarrow 4^2 + 6^2 + (-6)^2 + 3^2 + (-2)^2 + \dots + (-6)^2 + (-2)^2 - \frac{20^2}{20}$$

$$\Rightarrow 564$$

$$\underline{\underline{SSE}} \Rightarrow SST - (SSC + SSR)$$

$$\Rightarrow 564 - (338.8 + 150) = 67.2$$

(a)  $f_{cal} = \frac{112.93 - 20.16}{5.6} \left\{ \begin{array}{l} J_1 = 3 \Rightarrow df_1 \\ J_2 = 12 \Rightarrow df_2 \end{array} \right.$

$$f_{tab} = 2.61$$

$$f_{tab} < f_{cal}$$

reject  $H_0$

b)  $f_{cal} = \frac{39.5}{5.6} = 7.05$

$$\left. \begin{array}{l} df_1 = 4 \\ df_2 = 12 \end{array} \right\}$$

$$f_{tab} = 2.41$$

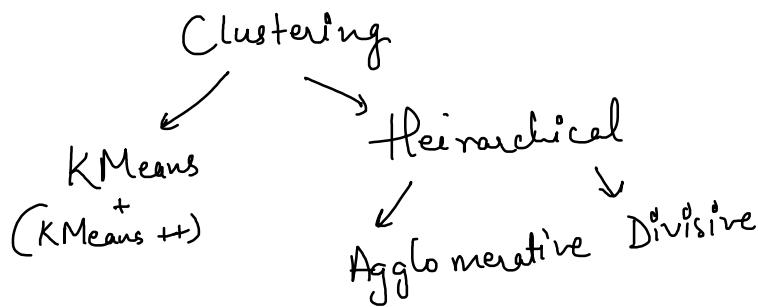
$$f_{tab} < f_{cal}$$

Reject  $H_0$

# Clustering

$D = \{x_i, y_i\}$        $y_i = f(x)$   
 ↓  
 supervised learning

$D = \{x_i\} \rightarrow$  No class label  $\Rightarrow$  Unsupervised learning



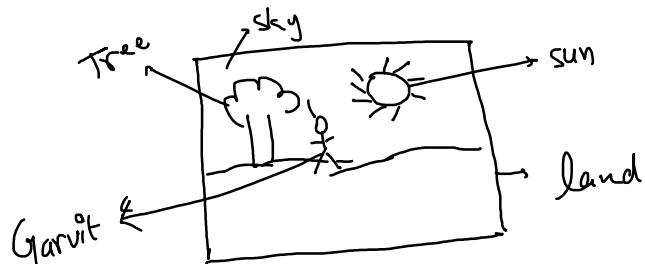
Applications: 1 → ecommerce : group customers on basis of location, gender, income level etc

2 → Review Analysis :

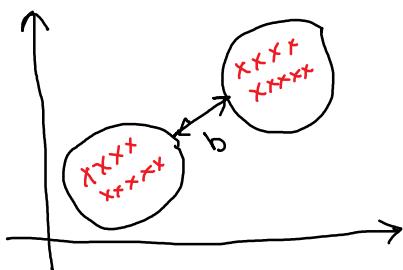
```

graph TD
    ReviewAnalysis --> AmazonReviews["Amazon Reviews"]
    AmazonReviews --> Positive["+ve"]
    AmazonReviews --> Negative["-ve"]
    
```

3 → Image Segmentation :  
 (object detect)



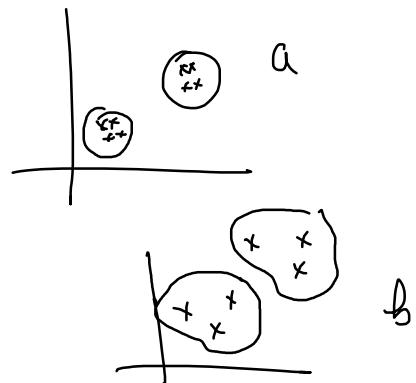
## Metric's



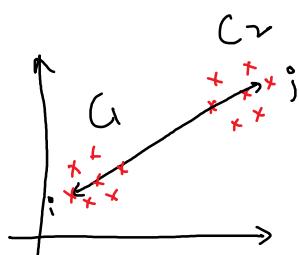
- \* Intracluster distance (a)
- \* Intercluster distance (b)

Characteristics of good cluster:

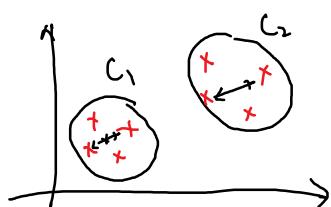
- Small intracluster distance
- large intercluster distance



$$\downarrow \text{Dunn's Index} \uparrow \Rightarrow \frac{\max d(i, j)}{\max d(k)} \rightarrow \begin{array}{l} \text{max intercluster distance} \\ (\text{0, } \infty) \end{array}$$



$\Rightarrow \max d(i, j) \Rightarrow$  distance b/w the farthest points in diff clusters



$\Rightarrow \max d(k) \Rightarrow$  distance b/w the farthest points within the cluster

Silhouette's Score  $\Rightarrow \frac{b - a}{b} \Rightarrow b \Rightarrow \text{avg intercluster distance}$

Silhouette's Score  $\Rightarrow \frac{b-a}{\max(b,a)}$   $b \Rightarrow$  avg intercluster distance  
 $a \Rightarrow$  avg intradcluster distance  
 $(-1, +1]$

Case 1:  $a \Rightarrow \min \Rightarrow 0$ ,  $b = \max \Rightarrow b$

$$SS = \frac{b-0}{\max(b,0)} = \frac{b}{b} = 1$$

Case 2:  $b < a$ ,  $b=0$ ,  $a=a$

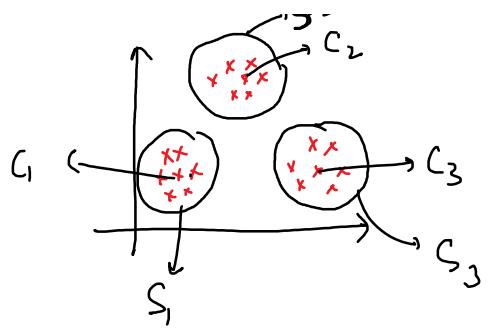
$$SS = \frac{b-a}{\max(b,a)} = \frac{0-a}{a} = -\frac{a}{a} = -1$$

Case 3:  $a=b$

$$SS = \frac{b-a}{\max(b,a)} = \frac{a-a}{\max(a,a)} = 0$$

$\Leftrightarrow$  K Means  $\rightarrow$  average  
# clusters





$$n = 5$$

$C_1, C_2, C_3 \Rightarrow$  Centroids

$S_1, S_2, S_3 \Rightarrow$  Sets

$$S_1 \cap S_2 = \emptyset$$

$$S_2 \cap S_3 = \emptyset$$

$$S_3 \cap S_1 = \emptyset$$

- 1) Randomly choose centroids
- 2) distance of pts from centroids & create sets
- 3) after creating group, recalculate & update centroids

$$\text{MOF} \Rightarrow C^* = \arg \min_{C_1, C_2, \dots, C_k} \sum_{i=1}^k \sum_{x \in S_i} \|x - C_i\|^2$$

$\hookdownarrow$  Intracluster distance  
 $\downarrow$   
 $S_i \cap S_j = \emptyset$  np hard problems

### Lloyd's Algorithm

- ① → Randomly choose  $k$  datapoints from dataset & call them centroids
- ② → Assignment: for each point, select the nearest centroid with the help of distance & add point to the corresponding cluster.

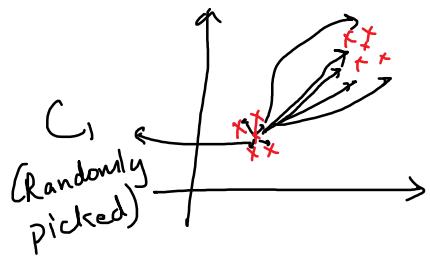
- ③ → Update: Recalculate centroids

$$C_j^* = \frac{1}{|S_j^*|} \sum_{i=1}^n x_i \quad x_i \in S_j^*$$

$$C_j = \frac{1}{S_j} \sum_{i=1}^n x_i \quad x_i \in S_j$$

(iv)  $\rightarrow$  Repeat (ii) & (iii) till convergence.

KMeans++  $\Rightarrow$  KMeans ( $\text{init} = \text{'Kmeans++'}$ )



| data points | distance | probability of distance |
|-------------|----------|-------------------------|
| $x_1$       | $d_1$    |                         |
| $x_2$       | $d_2$    |                         |
| $\vdots$    | $\vdots$ |                         |
| $x_n$       | $d_n$    |                         |

"The larger the distance b/w pt & Centroid , the greater the chance of it being picked as Centroid."

Thursday, January 25, 2024 8:38 PM



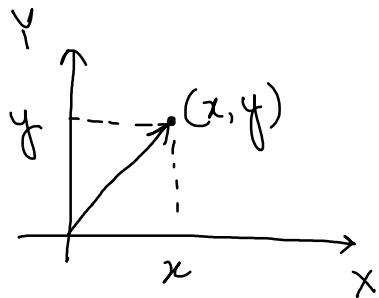






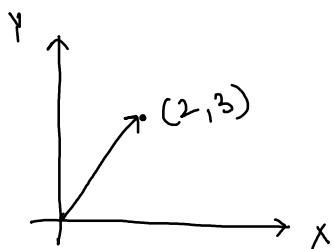
Linear Algebra

Quantities



Scalar  
↓  
Magnitude

Vector  
↓  
magnitude  
direction

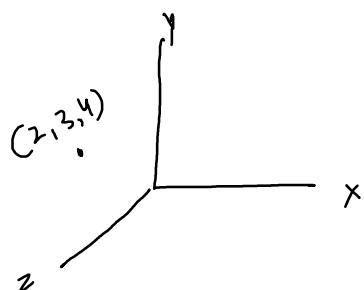


$$\begin{matrix} G \\ A \end{matrix} \xleftarrow[-1 \text{ km}]{\quad 1 \text{ km} \quad} \begin{matrix} B \\ G \end{matrix}$$

distance =  $\sqrt{2}$  km

$$\text{displacement} = +1 - 1 = 0$$

$$\text{vector} = [2 \ 3] \Rightarrow 2d$$



$$\text{vector} = [2 \ 3 \ 4] \Rightarrow 3d$$

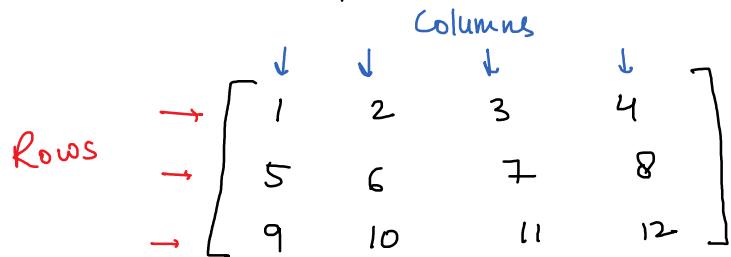
$$4d \Rightarrow \text{vector} = [2 \ 3 \ 4 \ 1]$$

$$6d \Rightarrow \text{vector} = [2 \ 3 \ 4 \ 1 \ 5 \ 7]$$

/  
/  
/

$$nd \Rightarrow \text{vector} = [2 \ 3 \ 4 \ 1 \ 5 \ 7 \ \dots \ n]$$

MATRIX  $\Rightarrow$  it is a table of numbers



Addition :

$$\begin{bmatrix} a & b \\ c & d \end{bmatrix} + \begin{bmatrix} e & f \\ g & h \end{bmatrix} = \begin{bmatrix} a+e & b+f \\ c+g & d+h \end{bmatrix}$$

$$\Rightarrow \begin{bmatrix} 1 & 3 & 5 \\ 2 & 4 & 6 \end{bmatrix} + \begin{bmatrix} 7 & 8 & 9 \\ 10 & 11 & 12 \end{bmatrix} = \begin{bmatrix} 7+1=8 & 3+8=11 & 5+9=14 \\ 10+2=12 & 4+11=15 & 6+12=18 \end{bmatrix}$$

Multiplication :

A diagram illustrating the multiplication of a 2x2 matrix by a 2x1 matrix. The first matrix has its second column circled and labeled "1x2". The second matrix has its second row circled and labeled "2x1". The result is a 1x1 matrix labeled "11".

$$\begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix} \begin{bmatrix} 3 \\ 4 \end{bmatrix} = \begin{bmatrix} 1 \times 3 + 2 \times 4 \\ 3 \times 1 \end{bmatrix} = \begin{bmatrix} 11 \end{bmatrix}_{1 \times 1}$$

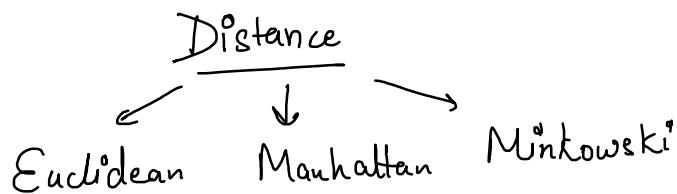
$$\begin{bmatrix} \overrightarrow{a} & b \\ c & d \end{bmatrix}_{2 \times 2} \times \begin{bmatrix} e & f \\ g & h \end{bmatrix}_{2 \times 2} = \begin{bmatrix} ae+bg & af+bh \\ ce+dg & cf+dh \end{bmatrix}_{2 \times 2}$$

\* In order to perform Matrix Multiplication,

no. of columns in first matrix = No of rows in second matrix

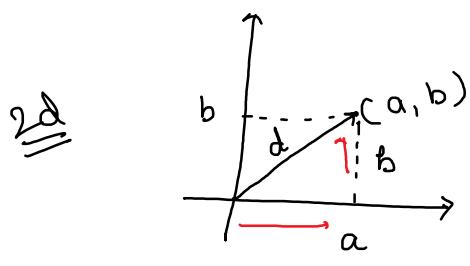
$$\Rightarrow a_{m \times n} \times b_{n \times q} \Rightarrow C_{m \times q}$$

$$\Rightarrow a_{m \times n} \times b_{q \times n} \Rightarrow \text{multiplication Not possible}$$



### Euclidean distance:

i) distance of a point from origin:



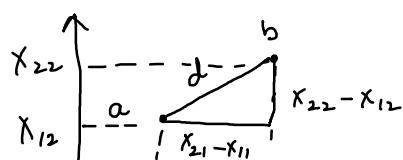
By Pythagoras Theorem,

$$d^2 = a^2 + b^2$$

$$d = \sqrt{a^2 + b^2}$$

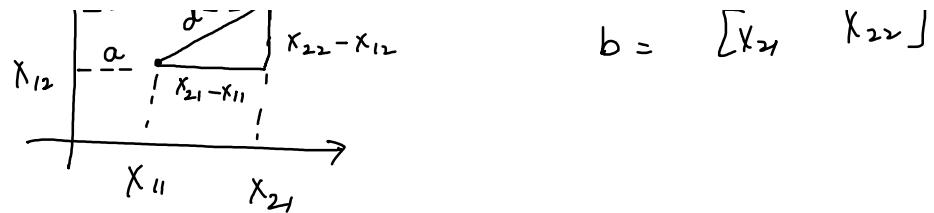
nd  $d = \sqrt{a^2 + b^2 + \dots + n^2}$

ii) Distance b/w two points.



$$a = [x_{11} \quad x_{12}]$$

$$b = [x_{21} \quad x_{22}]$$

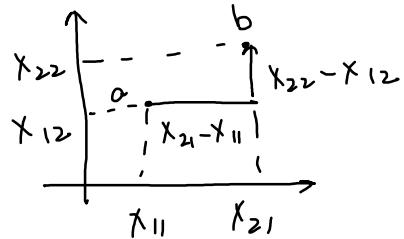
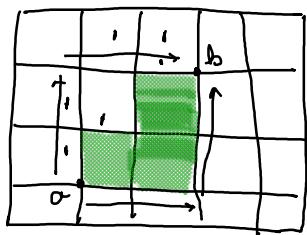


$$d = \sqrt{(x_{21} - x_{11})^2 + (x_{22} - x_{12})^2} = \sqrt{(x_{11} - x_{21})^2 + (x_{12} - x_{22})^2}$$

$$d = \left[ \sum_{i=1}^n |x_{1i} - x_{2i}|^2 \right]^{\frac{1}{2}} \rightarrow L_2 \text{ Norm}$$

2) Manhattan Distance:

$$a = [x_{11} \ x_{12}] , b = [x_{21} \ x_{22}]$$



$$d = |x_{21} - x_{11}| + |x_{22} - x_{12}|$$

$$d = |x_{11} - x_{21}| + |x_{12} - x_{22}|$$

$$d = \sum_{i=1}^n |x_{1i} - x_{2i}| \rightarrow L_1 \text{ Norm}$$

- when you have high dimensional dataset, use Manhattan distance.

## Minkowski Distance :

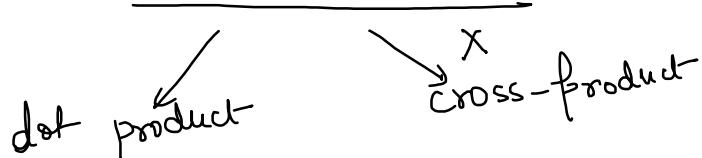
$$d = \left[ \sum_{i=1}^n |x_{1i} - x_{2i}|^p \right]^{\frac{1}{p}} \rightarrow L^p \text{ Norm}$$

$p = 1, 2, \dots, \infty$

Let  $p=1$ ,  $d = \left[ \sum_{i=1}^n |x_{1i} - x_{2i}| \right]^{\frac{1}{1}}$  → Manhattan distance

Let  $p=2$ ,  $d = \left[ \sum_{i=1}^n |x_{1i} - x_{2i}|^2 \right]^{\frac{1}{2}}$  → Euclidean distance

## Vectors Multiplication



### Vector Representation

default

← Column  
vector

$$= \begin{bmatrix} a_1 \\ a_2 \\ a_3 \end{bmatrix}$$

Row  
vector  
 $[a_1, a_2, a_3]$

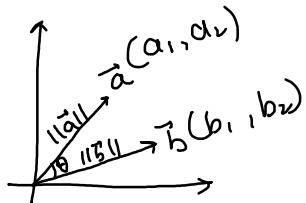
$$a = \begin{bmatrix} a_1 \\ a_2 \\ a_3 \\ \vdots \\ a_n \end{bmatrix}$$

$$b = \begin{bmatrix} b_1 \\ b_2 \\ b_3 \\ \vdots \\ b_n \end{bmatrix}$$

dot product  $\Rightarrow a \cdot b = a^T \cdot b = [a_1 \ a_2 \ a_3 \ \dots \ a_n]_{1 \times n} \begin{bmatrix} b_1 \\ b_2 \\ b_3 \\ \vdots \\ b_n \end{bmatrix}_{n \times 1}$

$$= [a_1 b_1 + a_2 b_2 + a_3 b_3 + \dots + a_n b_n]_{1 \times 1}$$

Angle b/w vectors:



$$a \cdot b = ||\vec{a}|| \cdot ||\vec{b}|| \cos \theta \rightarrow \text{Geometric way}$$

$$a \cdot b = [a_1 b_1 + a_2 b_2] = a^T \cdot b \rightarrow \text{LA way}$$

$$||a|| \cdot ||b|| \cos \theta = [a_1 b_1 + a_2 b_2]$$

$$\cos \theta = \frac{[a_1 b_1 + a_2 b_2]}{||a|| \ ||b||}$$

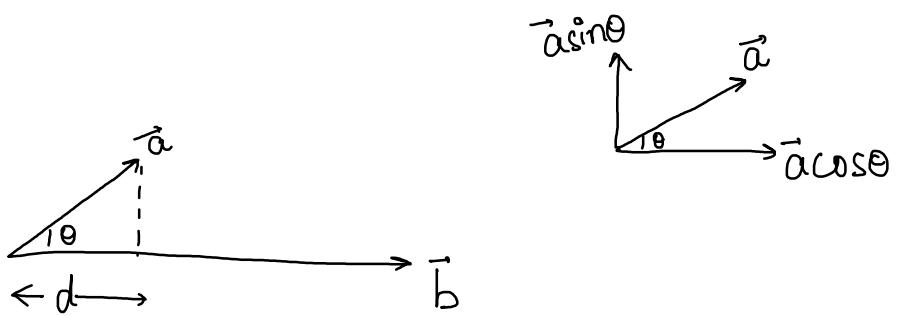
$$||a|| = \sqrt{a_1^2 + a_2^2}$$

Angle b/w vectors,  $\theta = \cos^{-1} \left[ \frac{(a_1 b_1 + a_2 b_2)}{||a|| \ ||b||} \right]$   $||b|| = \sqrt{b_1^2 + b_2^2}$

P

$\sin \theta$

## Projection



$$d \Rightarrow \|\vec{a}\| \cos \theta$$

$$\vec{a} \cdot \vec{b} = \|\vec{a}\| \|\vec{b}\| \cos \theta$$

$$\vec{a} \cdot \vec{b} = d \|\vec{b}\|$$

d =  $\frac{\vec{a} \cdot \vec{b}}{\|\vec{b}\|}$  Projection of  $\vec{a}$  on  $\vec{b}$

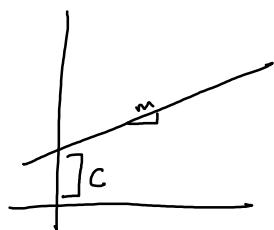
Unit Vector : vector with magnitude = 1

(^)  $\hookrightarrow$  direction

$$\hat{a} = \frac{\vec{a}}{\|\vec{a}\|}$$

## Lines & Planes

### Line



$$m = \frac{y_2 - y_1}{x_2 - x_1} = \tan \theta = \frac{dx}{dy}$$

slope  
 $y = mx + c \rightarrow y\text{-intercept}$

General Equation of line:  $Ax + By + C = 0$

$$By = -Ax - C$$

$$y = -\frac{A}{B}x - \frac{C}{B}$$

Planes

General Equation:  $Ax + By + Cz + D = 0$

$\downarrow$  change coeff

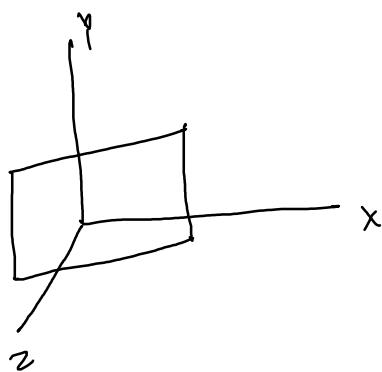
$$\omega_1 x + \omega_2 y + \omega_3 z + \omega_0 = 0$$

$\downarrow$  change axes names

$$\omega_1 x_1 + \omega_2 x_2 + \omega_3 x_3 + \omega_0 = 0$$

$\downarrow$  eqn of plane

$$\omega_0 + \omega_1 x_1 + \omega_2 x_2 + \omega_3 x_3 = 0$$



Hyperplane (plane for 4d & above)

$$\omega_0 + (\omega_1 x_1 + \omega_2 x_2 + \omega_3 x_3 + \omega_4 x_4 + \dots + \omega_n x_n) = 0$$

$$\omega_0 + [\omega_1 \ \omega_2 \ \omega_3 \ \omega_4 \ \dots \ \omega_n] \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \\ \vdots \\ x_n \end{bmatrix} = 0$$

$\omega_0 + \omega^T x = 0$

if hyperplane is passing through origin,  $\omega_0 = 0$

$$\boxed{\omega^T x = 0}$$

## Eigen Vector & Eigen Values

$$A\vec{x} = \lambda\vec{x}$$

Matrix      vectors      eigen value

## Curse of Dimensionality

binary classification (0, 1) =  $f_1, f_2, f_3$

$$\Rightarrow 2^3 = 8$$

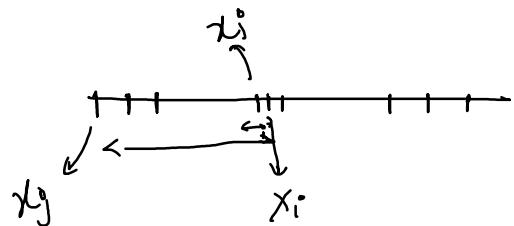
$$10 \text{ features} \Rightarrow 2^{10} = 1024$$

$$100 \text{ features} \Rightarrow 2^{100}$$

$$\text{for } n \text{ features} \Rightarrow 2^n$$

Hughes phenomenon: Whenever the dimension  $\uparrow$ , the performance of model decreases

Distance function : as the dimension  $\uparrow$ , distance function loses meaning



$$\text{dist}_{\min} = \min d[x_i, x_j], \quad \text{dist}_{\max} = \max d[x_i, x_j]$$

$$\frac{\text{dist\_max} - \text{dist\_min}}{\text{dist\_min}} > 0$$

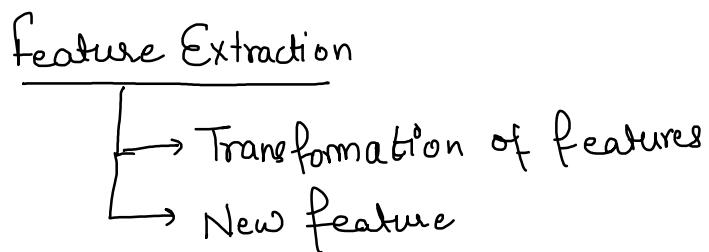
all points become equidistant, as you have higher dimension

$$\text{dist\_max} = \text{dist\_min}$$

$$\text{dist\_max} - \text{dist\_min} = 0$$

In NLP, in order to avoid this issue Hamming Distance is used!

3) As  $d \uparrow$ , chances of overfitting  $\uparrow$

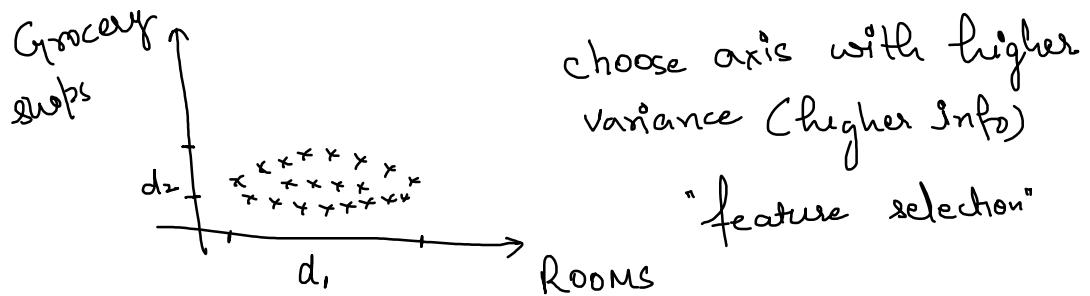


### Benefits

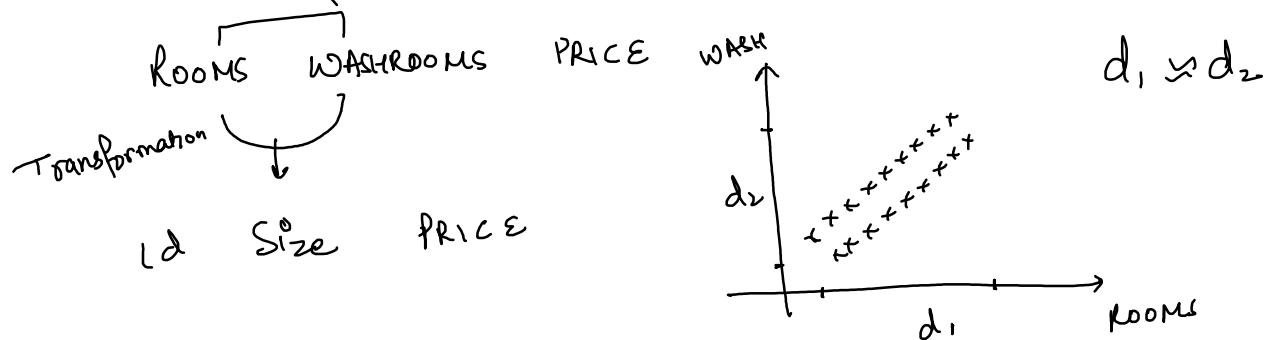
→ Faster Execution

### Basic Intuition:

| Rooms | Grocery-shops | Price of Flat |
|-------|---------------|---------------|
| 3     | 2             | 60            |
| 4     | 0             | 130           |
| 2     | 6             | 170           |
| 5     | 7             | 90            |

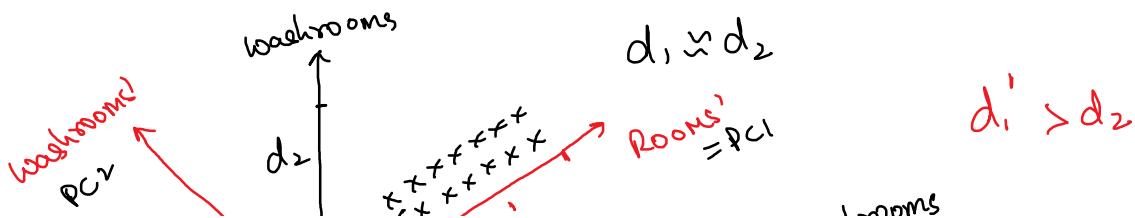


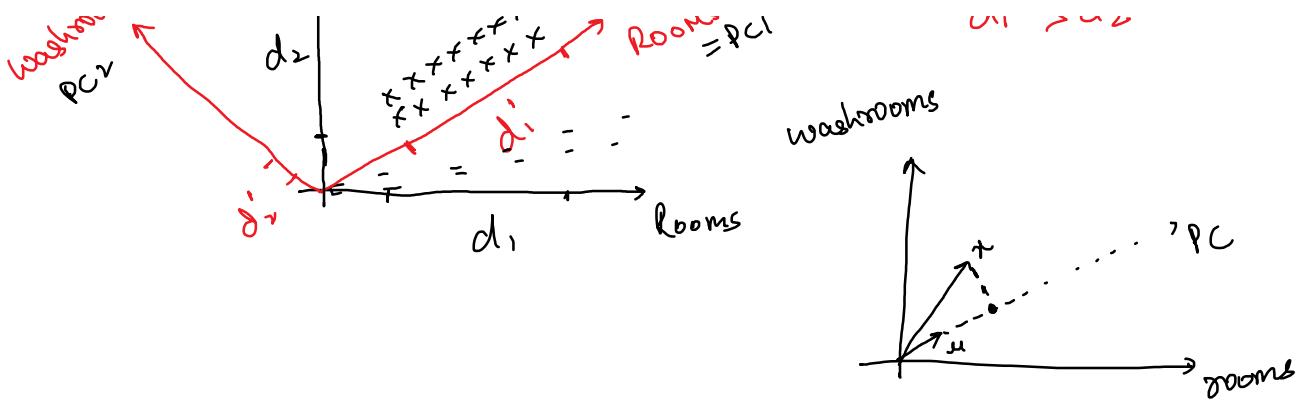
### Feature Extraction:



feature Extraction: Creates a set of new features from old features and choose a subset of features with higher variances

### Geometric Intuition of PCA:





Projection of  $\vec{x}$  on  $u \Rightarrow \frac{\vec{u} \cdot \vec{x}}{\|\vec{u}\|} \Rightarrow \vec{u} \cdot \vec{x} \xrightarrow{LP} u^T x$

The unit vector with higher variance will be chosen as the right axis!

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

↓  
Rayleigh Quotient  
(1950)

Covariances → build covariance Matrix

$$\begin{matrix} & X_1 & X_2 \\ X_1 & \left[ \begin{matrix} \text{Var}(X_1) & \text{Cov}(X_1, X_2) \\ \text{Cov}(X_1, X_2) & \text{Var}(X_2) \end{matrix} \right] \\ X_2 & \end{matrix}$$

↓ eigen decomposition  
eigen value, eigen vector

" largest eigen vector of covariance matrix always points in the direction of largest variance".

### Steps

1) Mean Centering  $\Rightarrow$  Standardization  $\Rightarrow$  Not a mandatory step but improves performance

2) Build Covariance Matrix

|       | $f_1$                  | $f_2$                  | $f_3$                  |
|-------|------------------------|------------------------|------------------------|
| $f_1$ | $\text{var}(f_1)$      | $\text{cov}(f_1, f_2)$ | $\text{cov}(f_1, f_3)$ |
| $f_2$ | $\text{cov}(f_1, f_2)$ | $\text{var}(f_2)$      | $\text{cov}(f_2, f_3)$ |
| $f_3$ | $\text{cov}(f_1, f_3)$ | $\text{cov}(f_2, f_3)$ | $\text{var}(f_3)$      |

3) eigen decomposition of covariance matrix

$PC_1 > PC_2 > PC_3$

$f_1 \downarrow \quad f_2 \downarrow \quad f_3 \downarrow$   
 $\lambda_1 \quad \lambda_2 \quad \lambda_3 \Rightarrow$  eigen values  
 $\downarrow \quad \downarrow \quad \downarrow$   
 $PC_1 \quad PC_2 \quad PC_3 \Rightarrow$  info components

if you choose  $\lambda_1$ ; it will be 1d  
" " " $\lambda_1, \lambda_2$ ; it will be 2d

" " " $\lambda_1, \lambda_2, \lambda_3$ "; " " " 3d

How to transform into 1d from 3d?

lets say dataset  $\rightarrow 1000 \text{ row} \times \underline{3 \text{ columns}}$

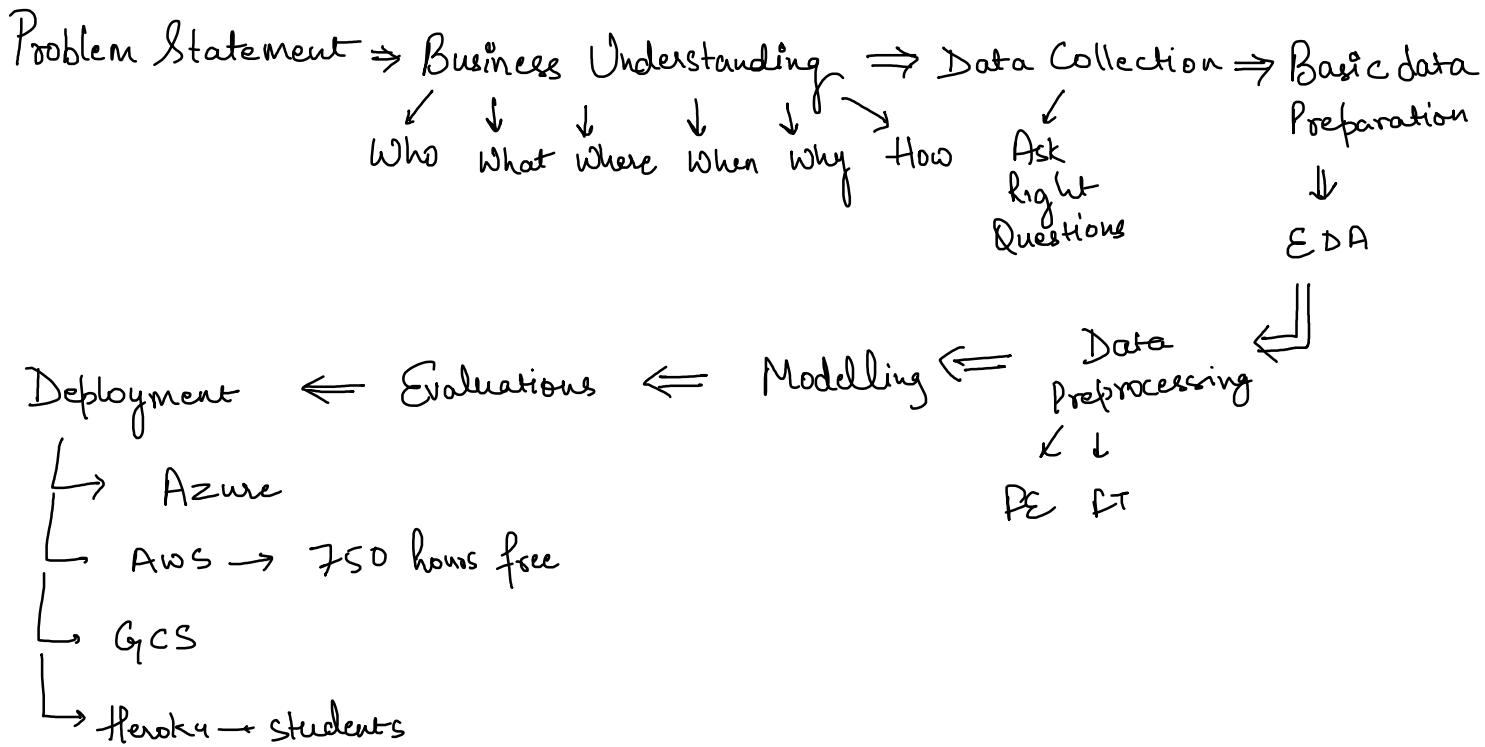
Shape of unit vector  $\Rightarrow [ ]_{1 \times 3}$

$$X \cdot u^T \Rightarrow [ ]_{1000 \times 3} [ ]_{3 \times 1}$$
$$\Rightarrow [ ]_{(1000 \times 1)}$$

Thursday, January 11, 2024 8:04 AM

# Lifecycle of Data Science

Project (3 months - 1 year)  
vague

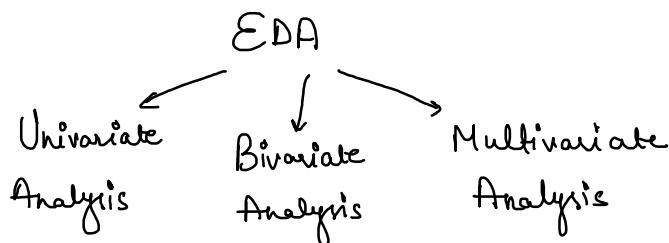


## Starting Project

- $\rightarrow$  Import libraries & load the data
- $\rightarrow$  Basic Info about data  $\Rightarrow$  `df.info()`
  - $\rightarrow$  metadata
  - $\rightarrow$  Rows
  - $\rightarrow$  columns
  - $\rightarrow$  nulls & non-nulls
  - $\rightarrow$  datatypes
- $\rightarrow$  Basic description of the data  $\Rightarrow$  `df.describe()`
- $\rightarrow$  Basic data study  $\Rightarrow$  unique, nunique & value-counts()

→ Basic Data Preparation → data quality check / data assessment

⇒ EDA

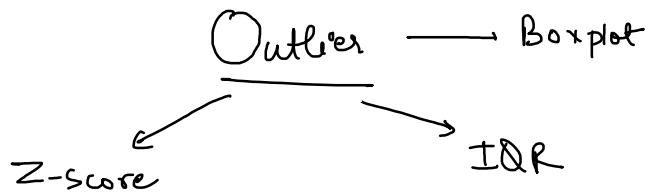


Univariate ⇒ histogram, countplot, boxplot, kdeplot

Bivariate ⇒ line, lmplot, scatter, bar, pie

Multivariate ⇒ "hue", heatmap, pairplot

UNIVARIATE → BIVARIATE → MULTIVARIATE (Project EDA flow)



$$Z = \frac{x - \mu}{\sigma}$$

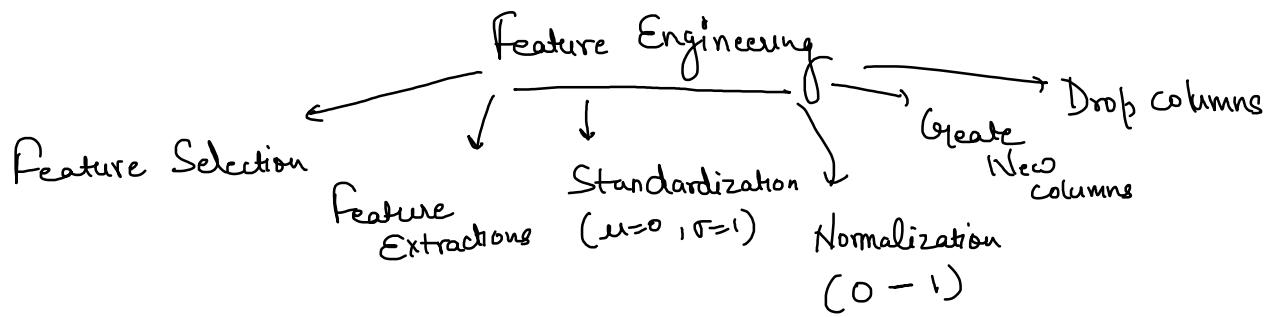
$$UL = Q3 + 1.5 \text{ IQR}$$

$$-3 > Z \quad \& \quad Z > 3$$

$$LL = Q1 - 1.5 \text{ IQR}$$

→ Missing Values → categorical → mode

numerical → Skewness → median  
no skewness ⇒ mean



Encoding  $\Rightarrow$  OHE (Nominal)  
 (One hot encoding)  
`pd.get_dummies`

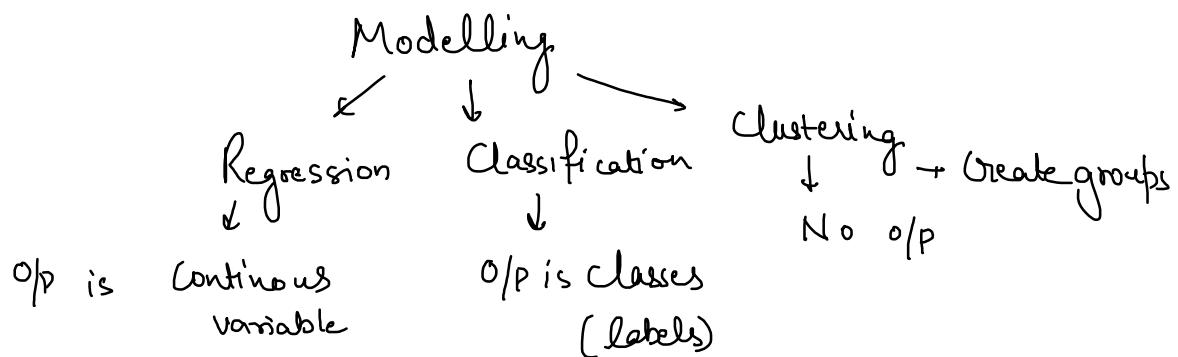
Ordinal Encoding, Label Encoding  
 (o/p variables)  $\downarrow$  (o/p variables)  
 ordinal category

| Grade | Encoding |
|-------|----------|
| A     | 1        |
| B     | 2        |
| C     | 3        |

### One hot Encoding

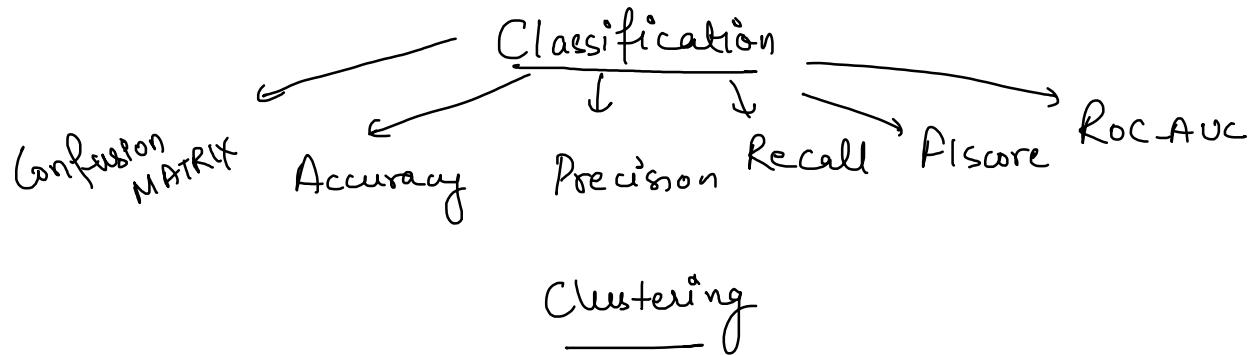
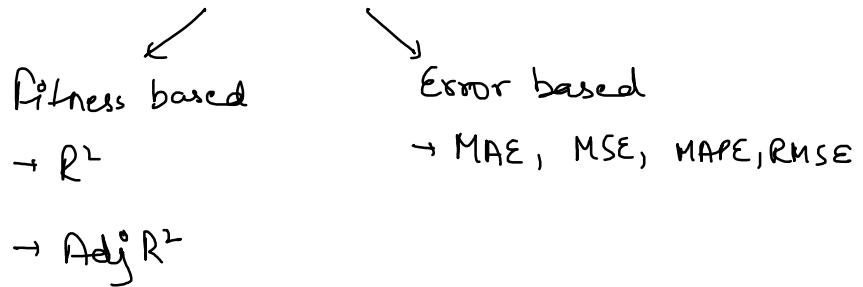
|   | A | B | C | D | E |  |
|---|---|---|---|---|---|--|
| A | 0 | 0 | 0 | 0 | 0 |  |
| B | 0 | 1 | 0 | 0 | 0 |  |
| C | 0 | 0 | 1 | 0 | 0 |  |
| D | 0 | 0 | 0 | 1 | 0 |  |
| E | 0 | 0 | 0 | 0 | 1 |  |

drop first = True



$\Rightarrow$  Evaluations

Regression



$$\text{Silhouette's score} = \frac{b - a}{\max(b, a)}$$

[-1, 1]

### INDO - PAK Relations

- Tanya
- $\rightarrow$  India  $\rightarrow$  Hindu Majority
  - Pak  $\rightarrow$  Muslim Majority
  - $\rightarrow$  Bigger Land area of India.

Bharath

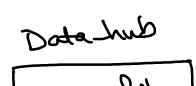
- $\rightarrow$  Health care system is better in India
- $\rightarrow$  Better Education in India

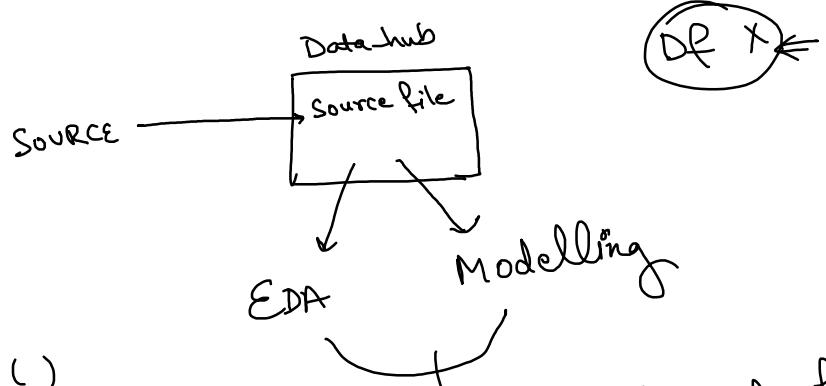
Monisha

- $\rightarrow$  Better foreign relations of India
- $\rightarrow$  GDP is better (India)

### PROJECT

#### AIR TICKET PRICE PREDICTION



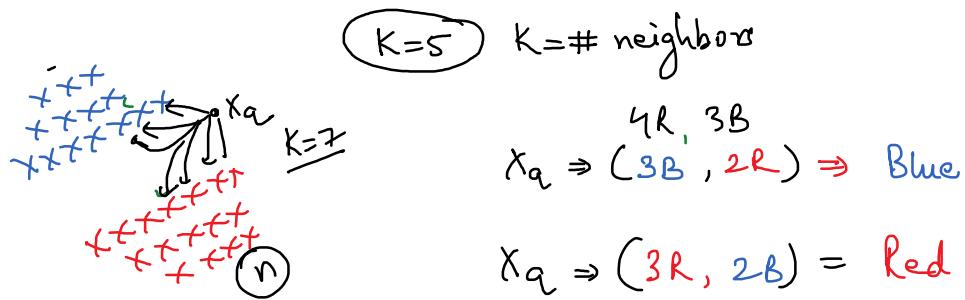


```

def preprocess():
    { to be filled by you }
    return data_EDA, data_modelling
  
```

## K-Nearest Neighbors

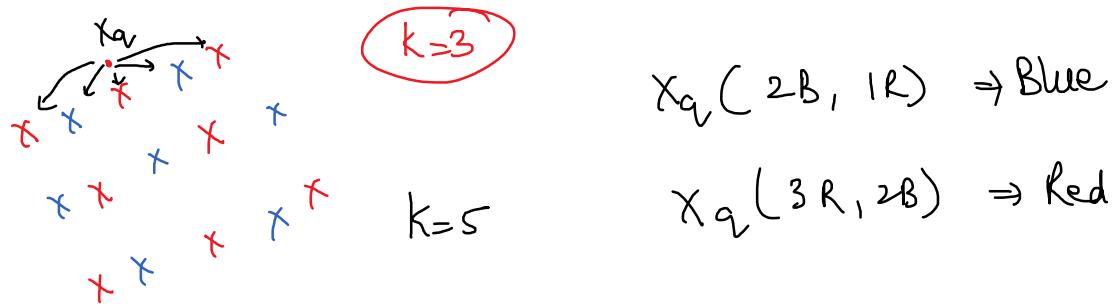
↳ you are like your Neighbors



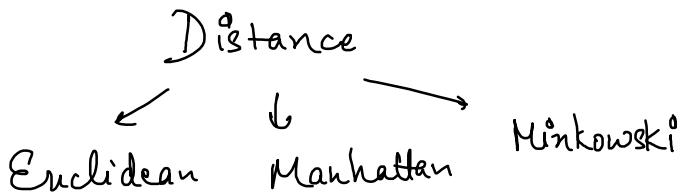
$K \Rightarrow \# \text{ neighbours} \Rightarrow \text{Hyperparameter}$

Can I choose even numbers? always keep value of  $K$  as odd!

NO!

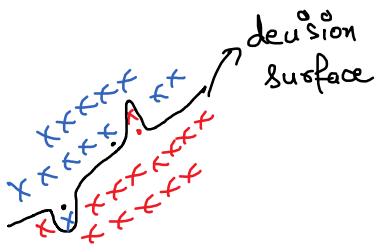


$\cancel{x} \uparrow / K \downarrow \rightarrow \text{to improve the model}$



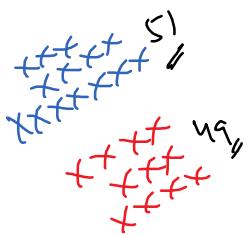
Effects of  $K$  ( $K = 3, 5, \dots$ )

$K=1$ :



- ⇒ decision surface is smooth
- ⇒ No mistakes
- ⇒ working too perfectly → overfitting
- ⇒ Training accuracy ↑
- ⇒ Test accuracy ↓

$\underline{K=n}$



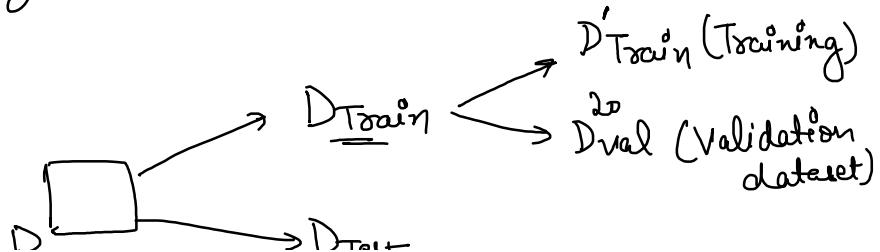
$$x_q = \text{Blue} \ (\text{blue} > \text{red})$$

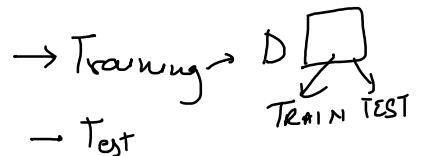
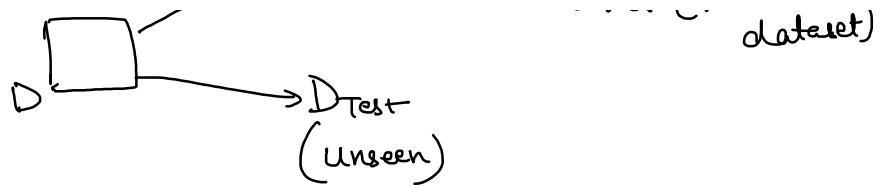
⇒ working too poorly

Curve of  $K$  with accuracy



Choose the right value of  $K$ :





CROSS-VALIDATION ⇒ "K'-Fold"

$$k' = \underline{10}$$

Train  
Test

$$k' = 4$$

|  | D <sub>TRAIN</sub> |                |                |                |
|--|--------------------|----------------|----------------|----------------|
|  | D <sub>1</sub>     | D <sub>2</sub> | D <sub>3</sub> | D <sub>4</sub> |

| (Neighbors) K | Training   | Validation     | Accuracy                    |
|---------------|--|----------------|-----------------------------|
| 1             | D <sub>1</sub> , D <sub>2</sub> , D <sub>3</sub> | D <sub>4</sub> | a' <sub>1</sub>             |
| 1             | D <sub>2</sub> , D <sub>3</sub> , D <sub>4</sub> | D <sub>1</sub> | a' <sub>2</sub>             |
| 1             | D <sub>1</sub> , D <sub>3</sub> , D <sub>4</sub> | D <sub>2</sub> | a' <sub>3</sub>             |
| 1             | D <sub>1</sub> , D <sub>2</sub> , D <sub>4</sub> | D <sub>3</sub> | a' <sub>4</sub>             |
| <hr/>         |  |                |                             |
| 2             | D <sub>1</sub> , D <sub>2</sub> , D <sub>3</sub> | D <sub>4</sub> | a <sup>2</sup> <sub>1</sub> |
| 2             | D <sub>2</sub> , D <sub>3</sub> , D <sub>4</sub> | D <sub>1</sub> | a <sup>2</sup> <sub>2</sub> |
| 2             | D <sub>1</sub> , D <sub>3</sub> , D <sub>4</sub> | D <sub>2</sub> | a <sup>2</sup> <sub>3</sub> |
| 2             | D <sub>1</sub> , D <sub>2</sub> , D <sub>4</sub> | D <sub>3</sub> | a <sup>2</sup> <sub>4</sub> |

$$\text{avg} \Rightarrow [a'^{\text{avg}}, a^2_{\text{avg}}, a^3_{\text{avg}}, a^4_{\text{avg}}] \rightarrow \text{Val-acc.}$$

$\downarrow$        $\downarrow$        $\downarrow$        $\downarrow$   
 K=1      K=2      K=3      K=4

Choose the value of  $K$  with highest accuracy!

### Advantages:

- Very easy to understand
- No assumptions (lazy learning)

### Disadvantages

- Space issues → high computation cost
- slow algorithm
- Imbalanced data

Application  $\Rightarrow$  Healthcare

### Evaluation Metrics

#### CONFUSION MATRIX

|           |         | Actual  |                     |
|-----------|---------|---------|---------------------|
|           |         | 0 (neg) | 1 (pos)             |
| Predicted | 0 (neg) | TN      | FN                  |
|           | 1 (pos) | FP      | TP                  |
| Total -ve |         |         | ↓ Total actual +ves |

⇒ Total predicted -ves  
⇒ Total predicted +ves

Accuracy  $\Rightarrow$

$$\frac{TP + TN}{TP + FP + TN + FN}$$

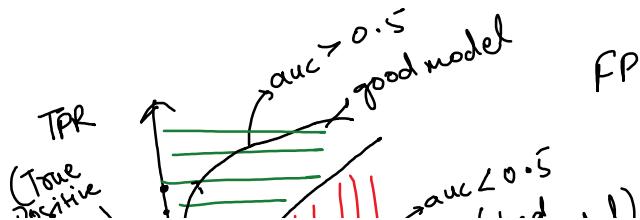
$$\uparrow \downarrow \text{Precision} \Rightarrow \frac{TP}{TP + FP} \uparrow \downarrow \rightarrow \text{Prediction.}$$

$\uparrow$  Recall  $\Rightarrow$   
(True Positive Rate) & Sensitivity

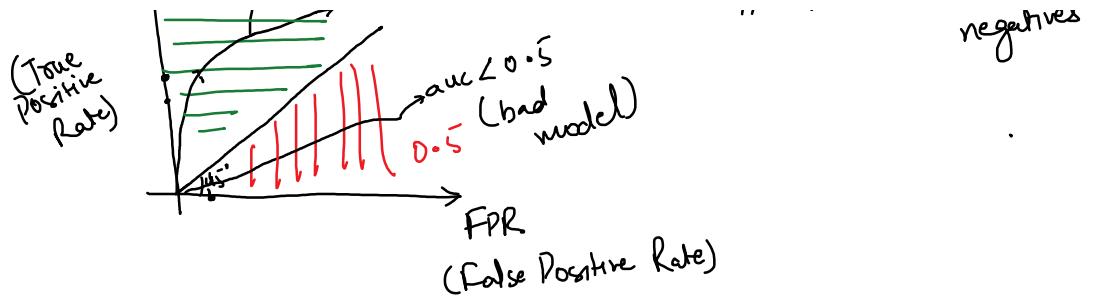
$$\frac{TP}{TP + FN} \rightarrow \text{Actual}$$

$$F1\text{-Score} \Rightarrow \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

#### ROC AUC Curve



$$FPR = \frac{FP}{TN + FP} \Rightarrow \text{Actual negatives}$$



$$\underline{\text{SPECIFICITY}} \Rightarrow 1 - \text{FPR} \Rightarrow \frac{TN}{TN + FP}$$

## Bayes Theorem

$$P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{P(B \cap A)}{P(B)}$$

$$P(A|B) = \frac{\underset{\substack{\text{evidence} \\ \uparrow}}{P(B|A) \times P(A)}}{\underset{\substack{\text{Posterior} \\ \uparrow}}{P(B)}} \rightarrow \text{Prior}$$

←  $P(B)$  → Law of Total Probability

\*<sup>RR</sup> Probability & likelihood → Difference?

↙ Naive Bayes → Bayes Theorem

"ignorant"  
+  
"innocent"

→ NB assumes that all features are independent of each other.

$$P(C_x/x_i) = \frac{P(x_i/C_x) \times P(C_x)}{P(x_i)}$$

$C_x \Rightarrow$  class labels

$X_i \Rightarrow$  g/p variables.

$$P(\text{Yes}/\text{outlook}) = P(\text{outlook}/\text{Yes}) \times P(\text{Yes})$$

| Outlook  | Temperature | Humidity | Windy | PlayTennis |
|----------|-------------|----------|-------|------------|
| Sunny ☀  | Hot         | High     | False | No         |
| Sunny ☁  | Hot         | High     | True  | No         |
| Overcast | Hot         | High     | False | Yes        |
| Rainy    | Mild        | High     | False | Yes        |
| Rainy    | Cool        | Normal   | False | Yes        |
| Rainy    | Cool        | Normal   | True  | No         |
| Overcast | Cool        | Normal   | True  | Yes        |
| Sunny ☀  | Mild        | High     | False | No         |
| Sunny ☁  | Cool        | Normal   | False | Yes        |
| Rainy    | Mild        | Normal   | False | Yes        |
| Sunny ☁  | Mild        | Normal   | True  | Yes        |
| Overcast | Mild        | High     | True  | Yes        |

$$P(\text{Yes}/\text{outlook}) = \frac{P(\text{outlook}/\text{Yes}) \times P(\text{Yes})}{P(\text{outlook})}$$

|          |      |        |       |     |
|----------|------|--------|-------|-----|
| Sunny    | Mild | Normal | True  | Yes |
| Overcast | Mild | High   | True  | Yes |
| Overcast | Hot  | Normal | False | Yes |
| Rainy    | Mild | High   | True  | No  |

$$P(\text{No}/\text{outlook}) = \frac{P(\text{outlook}/\text{No}) \times P(\text{No})}{P(\text{outlook})}$$

$$P(\text{Yes}/O, T, H, W) = ?$$

$$P(\text{No}/O, T, H, W) = ?$$

if  $P(\text{Yes}/O, T, H, W) > P(\text{No}/O, T, H, W)$

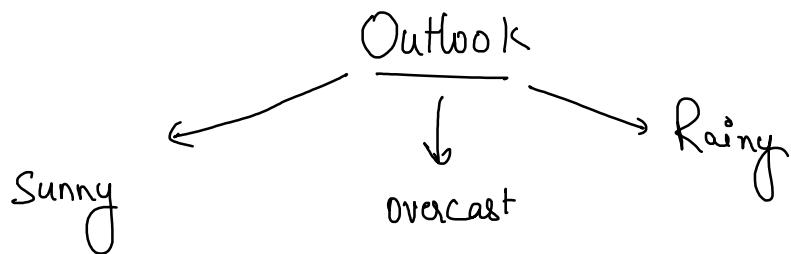
then will play

$$P(\text{Yes}/\text{outlook}) = P(\text{outlook}/\text{Yes}) P(\text{Yes})$$

$$P(\text{No}/\text{outlook}) = P(\text{outlook}/\text{No}) P(\text{No})$$

$$P(\text{Yes}) = ? \quad P(\text{No}) = ? \quad P(\text{outlook}/\text{Yes}) = ? \quad P(\text{outlook}/\text{No}) = ?$$

Working:  $P(\text{Yes}) = \frac{9}{14}$   $P(\text{No}) = \frac{5}{14}$



$$P(\text{Sunny}/\text{Yes}) = \frac{2}{9}$$

$$P(\text{Sunny}/\text{No}) = \frac{3}{5}$$

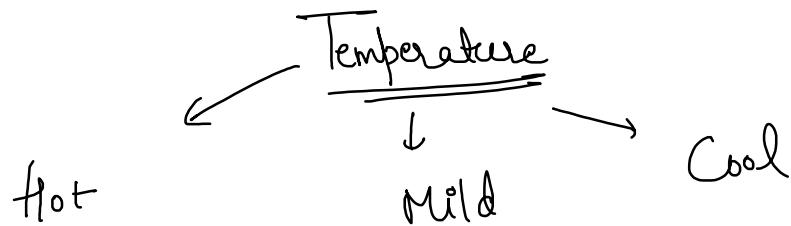
$$P(\text{Overcast}/\text{Yes}) = \frac{4}{9}$$

$$P(\text{Overcast}/\text{No}) = \frac{0}{5} = 0$$

- 10 • ... -

$$P(\text{Rainy} | \text{Yes}) = \frac{3}{9}$$

$$P(\text{Rainy} | \text{No}) = \frac{2}{5}$$



$$P(\text{Hot} | \text{Yes}) = \frac{2}{9}$$

$$P(\text{Hot} | \text{No}) = \frac{2}{5}$$

$$P(\text{Mild} | \text{Yes}) = \frac{4}{9}$$

$$P(\text{Mild} | \text{No}) = \frac{2}{5}$$

$$P(\text{Cool} | \text{Yes}) = \frac{3}{9}$$

$$P(\text{Cool} | \text{No}) = \frac{1}{5}$$



$$P(\text{High} | \text{Yes}) = \frac{3}{9}$$

$$P(\text{High} | \text{No}) = \frac{4}{5}$$

$$P(\text{Normal} | \text{Yes}) = \frac{6}{9}$$

$$P(\text{Normal} | \text{No}) = \frac{1}{5}$$

Windy



$$P(\text{True} | \text{Yes}) = \frac{3}{9}$$

$$P(\text{True} | \text{No}) = \frac{3}{5}$$

$$P(\text{False} | \text{Yes}) = \frac{6}{9}$$

$$P(\text{False} | \text{No}) = \frac{2}{5}$$

Q Outlook is overcast, temp is cool, humidity is high and wind is true(strong). Will I play?

Sol.

$$\begin{aligned}
 P(\text{Yes} | \text{overcast, cool, high, true}) &= P(\text{Yes}) \times P(\text{overcast} | \text{Yes}) \times P(\text{cool} | \text{Yes}) \\
 &\quad \times P(\text{high} | \text{Yes}) \times P(\text{true} | \text{Yes}) \\
 &= \frac{9}{14} \times \frac{4^2}{9} \times \frac{3}{9} \times \frac{3}{7} \times \frac{5}{9} = \frac{2}{189} \\
 &= 0.0105
 \end{aligned}$$

$$\begin{aligned}
 P(\text{No} | \text{overcast, cool, high, true}) &= P(\text{No}) \times P(\text{overcast} | \text{No}) \times P(\text{cool} | \text{No}) \times P(\text{high} | \text{No}) \times \\
 &\quad P(\text{true} | \text{No})
 \end{aligned}$$

$$\begin{aligned}
 &= \frac{5}{14} \times 0 \times \frac{4}{9} \times \frac{2}{7} \times \frac{4}{5} = 0
 \end{aligned}$$

$$P(\text{Yes} | \text{overcast, cool, high, true}) > P(\text{No} | \text{overcast, cool, high, true})$$

$$0.0105 > 0$$

$\therefore$  g will play!

Q Outlook is sunny, temp is cool, humidity is high & wind is strong?  
Will g play?

$$P(\text{Yes} | \begin{matrix} \text{outlook} \\ \text{sunny} \end{matrix}, \begin{matrix} \text{temp} \\ \text{cool} \end{matrix}, \begin{matrix} \text{humidity} \\ \text{high} \end{matrix}, \begin{matrix} \text{wind} \\ \text{strong} \end{matrix}) = 0.0052$$

$$P(\text{No} | \text{sunny, cool, high, strong}) = 0.02 = \frac{5}{14} \times \frac{3}{5} \times \frac{1}{5} \times \frac{4}{5} \times \frac{3}{5}$$

$$P(\text{No} | S, C, H, S) \quad P(\text{Yes} | S, C, H, S)$$

$$0.02 > 0.0052$$

No, g will not play

\* if you have a lot of dimensions, prob value will be very low.

$\hookrightarrow$  Solution Take log

$$P(\text{Yes} / \underset{\text{outlook}}{\text{cloudy, cool, high, weak}}) = P(\text{Yes}) P(\overset{O}{\text{cloudy}} | \text{Yes}) P(\overset{O}{\text{cool}} | \text{Yes}) \\ P(\overset{O}{\text{high}} | \text{Yes}) P(\overset{O}{\text{weak}} | \text{Yes})$$

$$P(\text{No} / \text{cloudy, cool, high, weak}) = P(\text{No}) P(\overset{O}{\text{cloudy}} | \text{No}) P(\overset{O}{\text{cool}} | \text{No}) \\ P(\overset{O}{\text{high}} | \text{No}) P(\overset{O}{\text{weak}} | \text{No})$$

$\Rightarrow$

The problem of zero probability:

$$P(\text{Yes} / \text{cloudy}) = P(\text{Yes}) P(\overset{O}{\text{cloud}} | \text{Yes}) = 0$$

$$P(\text{No} / \text{cloudy}) = P(\text{No}) P(\overset{O}{\text{cloudy}} | \text{No}) = 0$$

But cloudy isn't present, hence all cloudy probabilities are 0.

① Ignore cloudy

$$P(\text{Yes} / \text{cloudy, cool, high, weak}) = P(\text{Yes}) P(\overset{O}{\text{cool}} | \text{Yes}) P(\overset{O}{\text{high}} | \text{Yes}) \\ P(\overset{O}{\text{weak}} | \text{Yes})$$

$= 0.66$

$$P(\text{cloudy} | \text{Yes}) =$$



Absurd logic

$$P(\text{No}/\text{cloudy, cool, high, weak}) = P(\text{No}) P(\text{cool}/\text{No}) P(\text{high}/\text{No}) P(\text{weak}/\text{No}) \times 1$$

$$P(\text{Cloudy}/\text{No}) = 1$$

Laplace Smoothing

(var\_smoothing)

$$\therefore \alpha = 1$$

$$P(\text{Yes}/\text{cloudy}) = \frac{0 + \alpha}{n + \alpha k}$$

↑  
smoothing parameter  
# data points ( $y = \text{Yes}$ )

# distinct values your feature can take

Effect of  $\alpha$ :

$$10;00 \xrightarrow{Y=1 \rightarrow \text{class label}} 2 \text{ rare words}$$

①  $\alpha = 0$

$$P(R^W/Y=1) = \frac{P(R^W \wedge Y=1)}{P(Y=1)} = \frac{2}{1000} \rightarrow \text{overfitting}$$

②  $\alpha = 10,000$

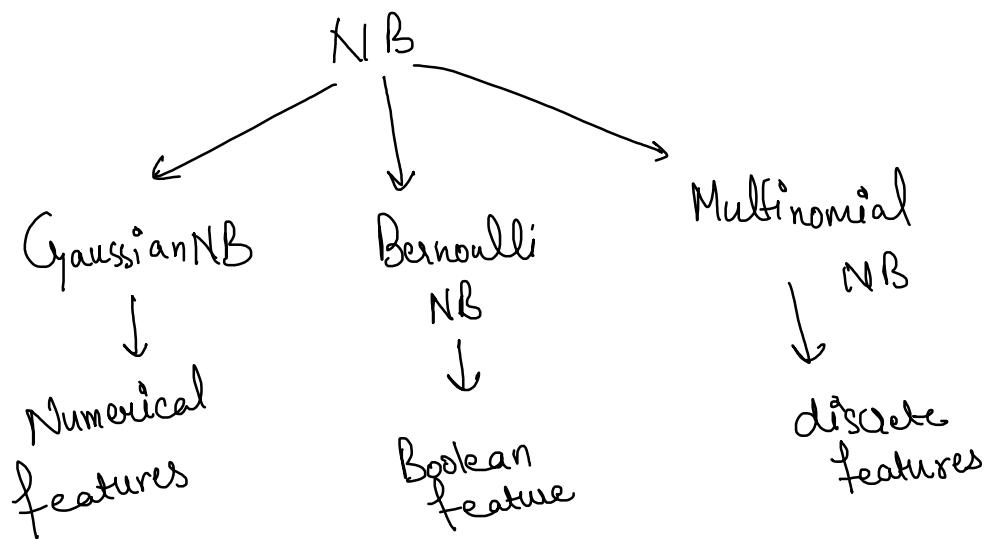
$$P(W/Y=1) = \frac{2 + 10000}{1000 + 2 \times 1000} = \frac{10002}{21000} \approx \frac{1}{2}$$

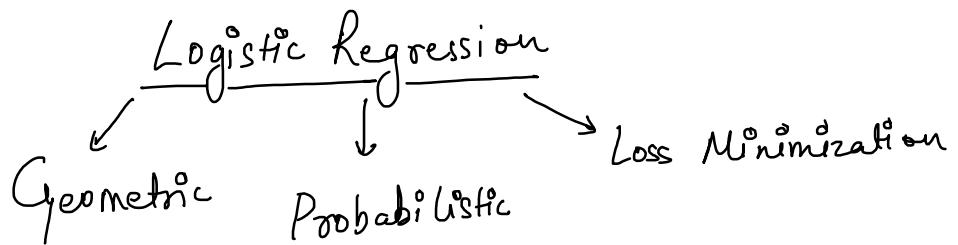
↳ underfitting

$\therefore \alpha=1$ , Hyperparameter  $\Rightarrow$  Cross-validation



GSCV      RSCV



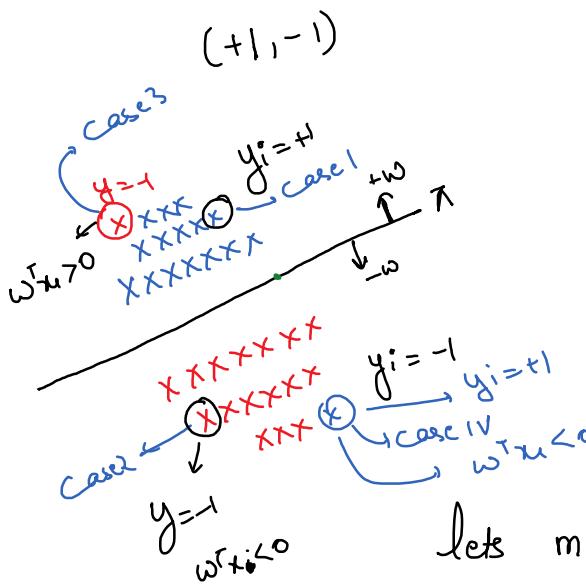


→ It is used for binary classification

→ Data is linearly separable.

Equation of plane :  $w_0 + \sum_{i=1}^n w_i x_i = 0$

if it is passing through origin



$$\sum_{i=1}^n w_i x_i = 0$$

①  $w^T x_i > 0$  for +ve class

②  $w^T x_i < 0$  for -ve class

lets multiply  $y_i$  with  $w^T x_i$

$$y_i w^T x_i$$

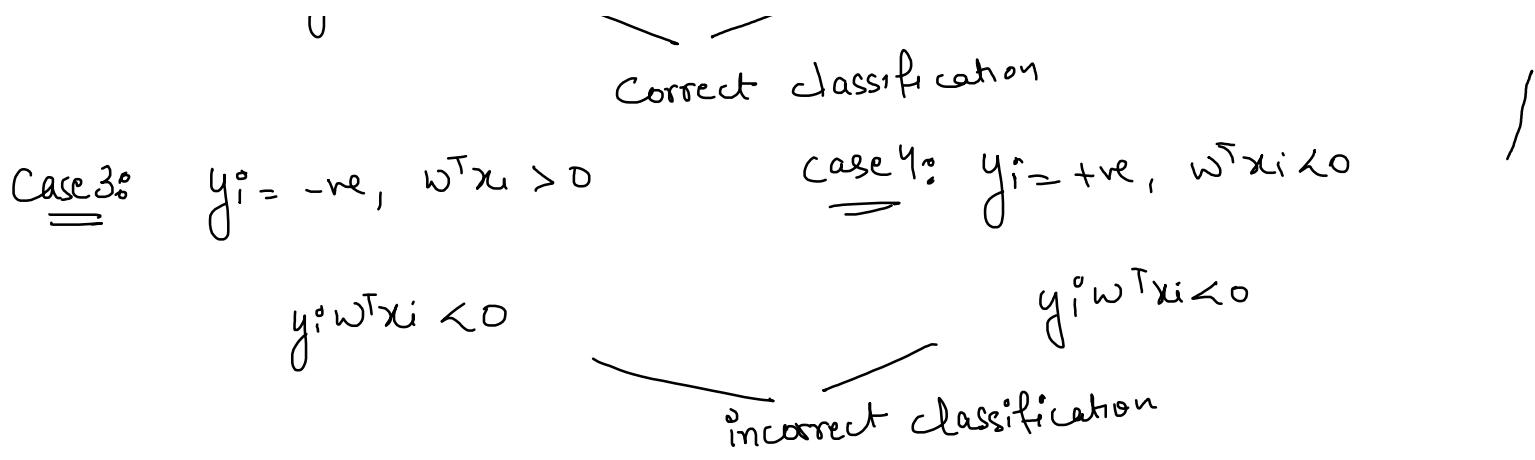
Case 1:  $y_i = +1$ ,  $w^T x_i > 0$

Case 2:  $y_i = -1$ ,  $w^T x_i < 0$

$$y_i w^T x_i > 0$$

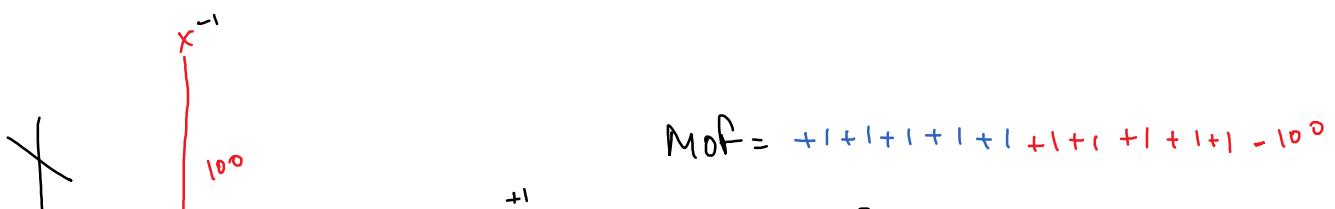
$$y_i w^T x_i > 0$$

Correct classification

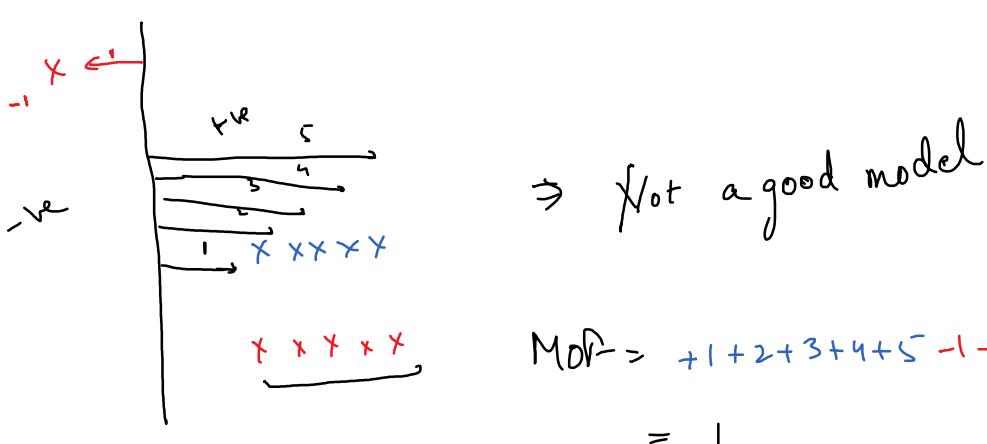


if  $\begin{cases} z_i \\ y_i w^T x_i > 0 ; \text{ correct classification} \\ z_i \\ y_i w^T x_i \leq 0 ; \text{ incorrect classification} \end{cases}$

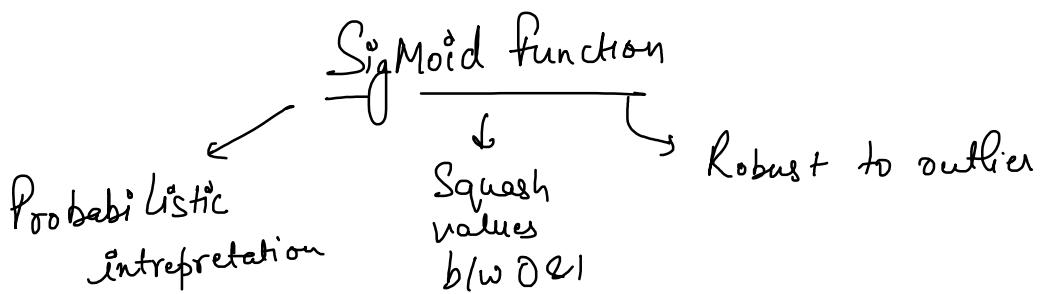
Mathematical Objective Function =  $\underset{w}{\operatorname{argmax}} \left( \sum_{i=1}^n y_i w^T x_i \right) = w^*$



$$= -90$$



$$= 1$$



$$\pi(x) = \frac{1}{1+e^{-x}} \Rightarrow \text{expression}$$

$$\pi(y_i^* \omega^T x_i) = \frac{1}{1+e^{-y_i^* \omega^T x_i}}$$

$$M o F \Rightarrow \arg \max [\pi(y_i^* \omega^T x_i)]$$

$$\Rightarrow \arg \max \left[ \frac{1}{1+e^{-y_i^* \omega^T x_i}} \right]$$

$$\Rightarrow \arg \max \left[ \log \left( \frac{1}{1+e^{-y_i^* \omega^T x_i}} \right) \right]$$

$$\Rightarrow \arg \max_x \left[ -\log (1+e^{-y_i^* \omega^T x_i}) \right]$$

$$\Rightarrow \arg \min \left[ \log (1+e^{-y_i^* \omega^T x_i}) \right] \Rightarrow \text{Logistic Loss}$$

Squashes b/w 0 & 1

$$\log(a^{-1}) = -\log a$$

$$\log \frac{1}{a} = -\log a$$

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

↓  
 $-\infty = x$        $+ \infty = x$   
 $\frac{1}{1+e^{\infty}}$        $\frac{1}{1+e^{-\infty}}$   
 ↓                          ↓  
 $\frac{1}{1+\infty}$        $\frac{1}{1+0}$   
 ↓                          ↓  
 $\frac{1}{\infty} = 0$        $\frac{1}{1} = 1$

Probabilistic Interpretation (1, 0)

$$P = \frac{1}{1+e^{-y}}$$

$$P(1+e^{-y}) = 1$$

$$P + P e^{-y} = 1$$

$$P e^{-y} = 1 - P$$

$$e^{-y} = \frac{1-P}{P}$$

$$\bar{e}^y = \frac{P}{1-P}$$

$$e^y = \left(\frac{P}{1-P}\right) \rightarrow \text{odd's ratio}$$

Take  $\ln$  on both sides

$$\ln(e^t) = \ln\left(\frac{P}{1-P}\right)$$

$y = \ln\left(\frac{P}{1-P}\right)$

→ Logit function

1, 0

$$\text{log loss} = - \left[ y_i \log p(y_i) + (1-y_i) \log p(1-y_i) \right]$$

$\downarrow$  Class 1       $\downarrow$  Class 0

Case 1:  $y_i = \begin{cases} \text{prob. } = 1 \\ \text{geo. } = 1 \end{cases}$

$$\text{geo. } \log(1+e^{-y_i w^T x_i}) \Rightarrow \log(1+e^{-w^T x_i})$$

$$\text{prob. } - [y_i \log p(y_i) + (1-y_i) \log p(1-y_i)]$$

$\hookrightarrow -\log p(y_i) \Rightarrow -\log \left( \frac{1}{1+e^{-w^T x_i}} \right)$

$$\Rightarrow \log(1+e^{-w^T x_i})$$

Derive Sigmoid from MoF

$$\ln\left(\frac{p}{1-p}\right) = y_i^* \omega^\top x_i^*$$

Take exp on both sides,

$$\exp\left[\ln\left(\frac{p}{1-p}\right)\right] = e^{y_i^* \omega^\top x_i^*}$$

$$\frac{p}{1-p} = e^{y_i^* \omega^\top x_i^*}$$

$$p = (1-p) e^{y_i^* \omega^\top x_i^*}$$

$$p = e^{y_i^* \omega^\top x_i^*} - p e^{y_i^* \omega^\top x_i^*}$$

$$p + p e^{y_i^* \omega^\top x_i^*} = e^{y_i^* \omega^\top x_i^*}$$

$$p(1 + e^{y_i^* \omega^\top x_i^*}) = e^{y_i^* \omega^\top x_i^*}$$

$$p = \frac{e^{y_i^* \omega^\top x_i^*}}{1 + e^{y_i^* \omega^\top x_i^*}}$$

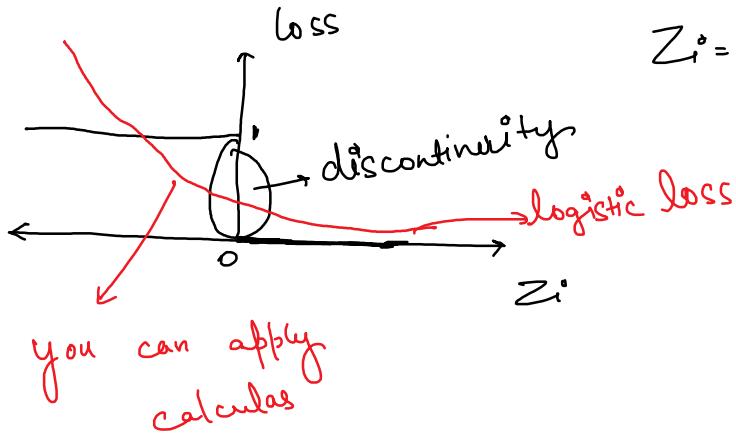
divide above eqn by  $e^{y_i^* \omega^\top x_i^*}$

$$p = \frac{e^{y_i^* \omega^\top x_i^*} / e^{y_i^* \omega^\top x_i^*}}{\frac{1 + e^{y_i^* \omega^\top x_i^*}}{e^{y_i^* \omega^\top x_i^*}}} = \frac{1}{\frac{1}{e^{y_i^* \omega^\top x_i^*}} + 1}$$

$$P = \frac{1}{1 + e^{-y_i^* \omega^\top x_i^*}} \Rightarrow \text{Sigmoid function}$$

## Loss Minimization

$\hookrightarrow \text{0-1 loss}$



## Overfitting & Underfitting

$$w^* = \underset{w}{\operatorname{argmin}} \sum_{i=1}^n \log(1 + e^{-y_i w^T x_i})$$

## Regularization: ① RIDGE REGULARIZATION:

$$\text{Loss}_{f^n} = \underset{w}{\operatorname{argmin}} \sum_{i=1}^n \underbrace{\log(1 + e^{-y_i w^T x_i})}_{\text{O}} + \underbrace{\lambda [w^T w]}_{\substack{\text{l}^2 \text{ Norm} \\ \text{hyperparameter}}} \rightarrow w^*$$

## ② LASSO REGULARIZATION:

$$\text{Loss}_{f^n} = \underset{w}{\operatorname{argmin}} \sum_{i=1}^n \log(1 + e^{-y_i w^T x_i}) + \lambda |w|$$

LASSO  $\rightarrow$  creates sparsity  $\rightarrow$  sparse vectors  $\Rightarrow [1, 0, 0, 0, 0, 1, 0, 0, 1]$

↳ extract out important feature

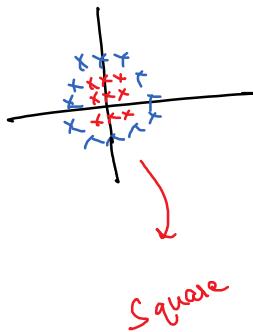
$\lambda \Rightarrow$  Hyperparameter  $\Rightarrow \lambda=0 \Rightarrow$  overfitting

$\lambda = \text{high} \Rightarrow$  underfitting

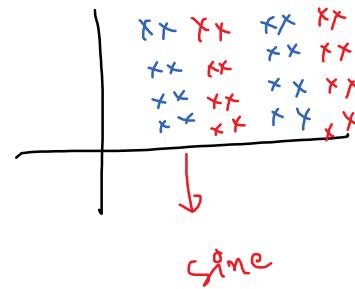
Feature  $\rightarrow$  standardize

↳ linearly separable

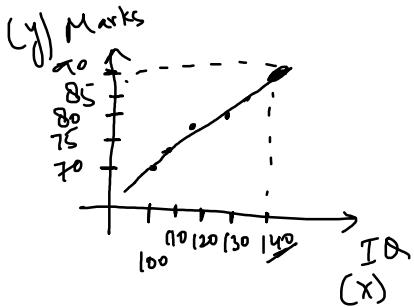
Exception  
Ex - 1



⑪



# Linear Regression

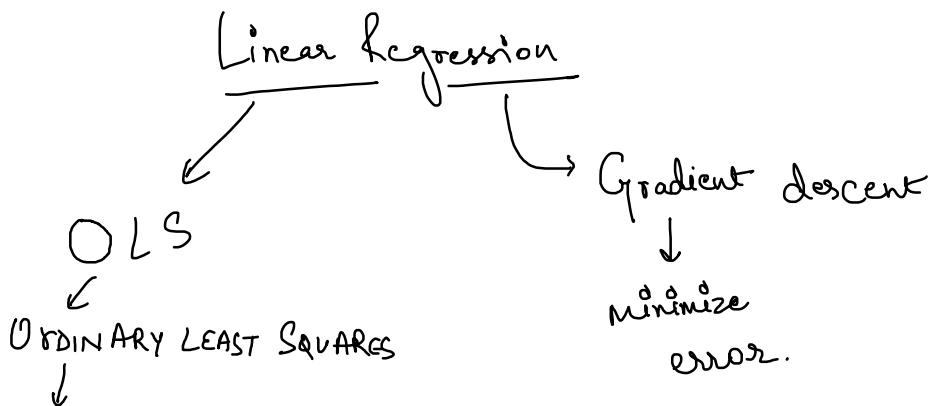
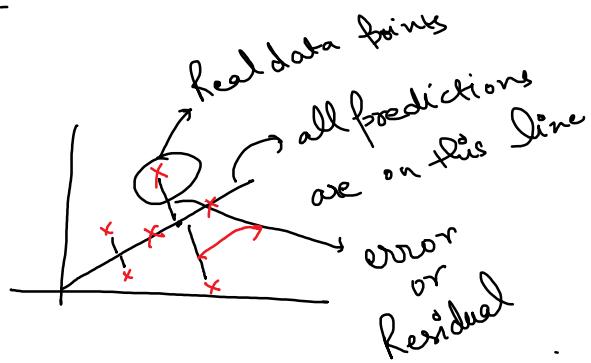
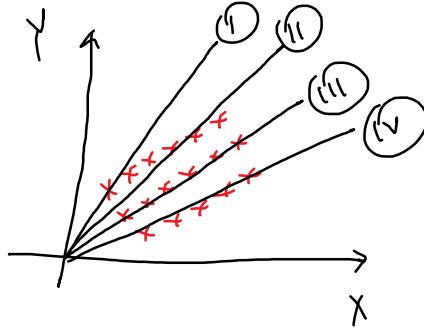


$$y = \underline{m} x$$

140

$$m = \text{slope} = \frac{y_2 - y_1}{x_2 - x_1} = \tan\theta = \frac{d}{dx}$$

Reality:

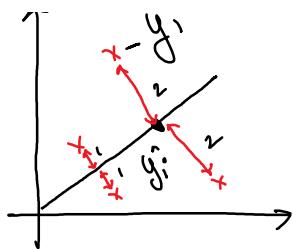


OLS

$$y = w_0 + w_1 x_1 + w_2 x_2 + \dots + w_n x_n \Rightarrow \text{eqn of hyperplane}$$

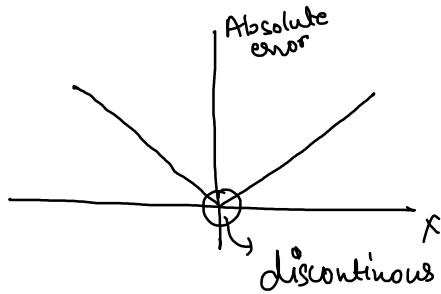


Absolute Error =  $|1| + |2| + |-1| + |-2| = 1 + 2 + 1 + 2 = 6$



$$\text{Error} = |1| + |2| + |-1| + |-4| = 1+2+1+4 = 8$$

$$\text{Absolute error} = \sum_{i=1}^n |y_i - \hat{y}_i|$$



$$\text{Squared error} = \sum (y_i - \hat{y}_i)^2$$

$$E(m, b) = (y_i - \hat{y}_i)^2 = 0$$

$$\hat{y}_i = mx_i + b = f(x)$$

↳ prediction

$$E = [y_i - (mx_i + b)]^2 = 0$$

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

$$\Rightarrow 2 \sum (y_i - (mx_i + b)) (-1) = 0$$

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

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

Divide by n

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

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

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

intercept of your best fit line

m

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

Assignment

value of m

$$m = \frac{\sum (y_i - \bar{y}_i)(\bar{x}_i - x_i)}{\sum (\bar{x}_i - \bar{\bar{x}}_i)^2} \rightarrow \text{slope of best fit line}$$

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

& m

MODEL

$$x_i \rightarrow \boxed{mx_i + b} \rightarrow y_i$$

$$\boxed{y_i = mx_i + b}$$

With the help of m & b, you can directly create best fit line.

Evaluation

METRICS

Goodness of fit

$$R^2 = 1 - \frac{\sum (y_i - \hat{y}_i)^2}{\sum (y_i - \bar{y}_i)^2}$$

Error based

Metric

↓

MAE

$$\begin{array}{c}
 \text{Coeff. of determination} \leftarrow R^2 \\
 \downarrow \\
 \text{Adj } R^2
 \end{array}
 \quad
 \begin{array}{c}
 \checkmark \\
 \text{MAE} \\
 \text{MSE} \\
 \text{RMSE} \\
 \text{MAPE}
 \end{array}$$

MAE: Mean Absolute Errors =  $\frac{1}{n} \sum_{i=1}^n |y_i - \hat{y}_i|$

Advantages:

Disadvantage:

- 1 → same scale as that of data → can't be differentiated.
- 2 → less sensitive to outliers

MSE: Mean Squared Error:  $\frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2$

Advantages

Disadvantage:

- optimizable
- sensitive to outlier
- not interpretable

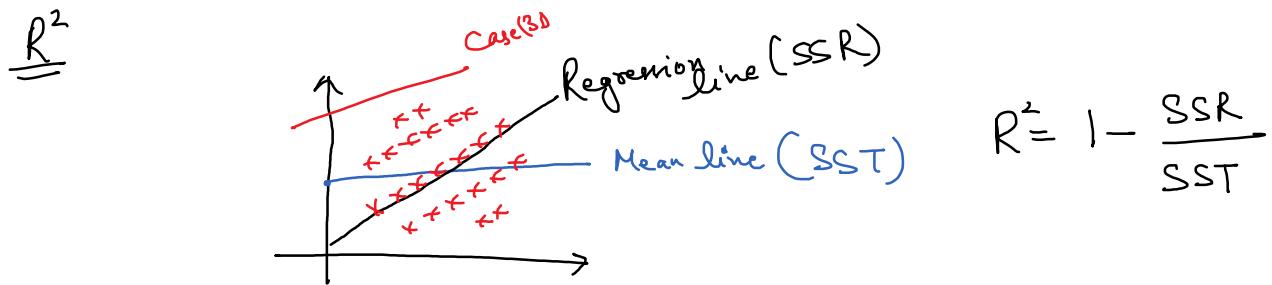
RMS E: Root Mean Squared Errors:  $\sqrt{\text{MSE}}$

$$\text{RMS E} = \sqrt{\frac{1}{n} \sum (y_i - \hat{y}_i)^2}$$

MAPE: Homework!

Goodness of fit

$$R^2 \quad \downarrow \quad \overbrace{\quad}^{\rightarrow} \quad \text{Adj } R^2$$



Case 1:  $SSR=0$ ,  $SST=SST$

$$R^2 = 1 - \frac{0}{SST} = 1 - 0 = 1 \quad (\text{Overfitting})$$

Case 2:  $SSR = SST$

$$R^2 = 1 - \frac{SSR}{SST} = 1 - \frac{SST}{SST} = 1 - 1 = 0$$

Case 3:  $SSR > SST$   $\frac{SSR}{SST} > 1$

$$R^2 = 1 - \frac{SSR}{SST} = -\text{ve}$$

Problem with  $R^2$ ?  $\rightarrow$  As your dimensions  $\uparrow$  (columns),  $R^2 \uparrow$

## Adjusted R<sup>2</sup>

$$\text{Adj } R^2 = 1 - \left[ \frac{(1-R^2)(n-1)}{(n-p-1)} \right]$$

↗ # datapoints  
↙ # features/dimensions/columns

## Multicollinearity:

$y = \underbrace{1f_1 + 2f_2 + 3f_3}_{\text{slopes}}$

Column names

|       |       |       |
|-------|-------|-------|
| $f_1$ | $f_2$ | $f_3$ |
| 1     | 2     | 3     |
| 0     | 3.5   | 3     |

$f_1 = 1.5 f_2$

$$y = 0f_1 + 3.5f_2 + 3f_3$$

## Detect of Multicollinearity

Correlation  
MATRIX

One of highly  
correlated  
columns  
(corr < -0.7 & corr > 0.7)

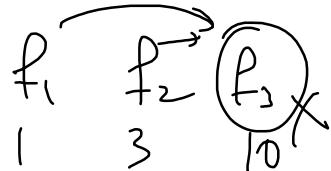
VIF  $\Rightarrow$  variance inflation factor

$$\frac{1}{1-R^2}$$

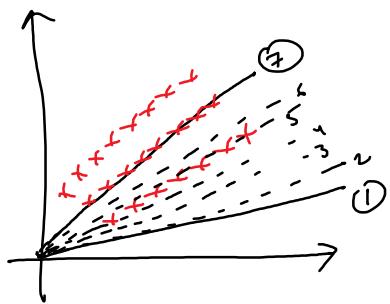
$[1, \infty]$

$\Rightarrow VIF > 5$

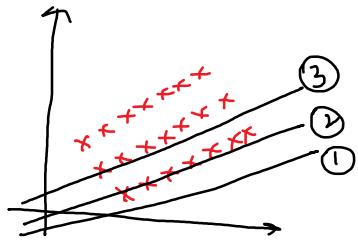
No multicollinearity



## Gradient Descent



$y = mx + b$  (line rotation is actually done by slope)



$y = mx + b$  (by changing  $b$ , I can move the line up & down)

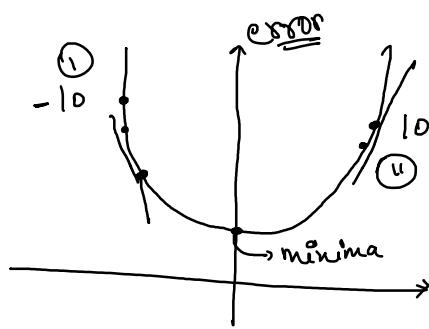
Steps :  $m = \text{constant}$ ,  $b = \text{variable}$

1) choose any random value of  $b$

$$2) \frac{dL}{db} = \frac{d(y_i - mx_i - b)^2}{db} = -2(y_i - mx_i - b)$$

slope

$$3) b_{\text{next}} = b_{\text{old}} - \frac{\text{slope}}{\text{slope}}$$



slope = -1

$$\textcircled{1} b_{\text{next}} = -10 - (-1) = -10 + 1 = -9$$

$$\textcircled{11} b_{\text{old}} = 10, \text{slope} = 1$$

$$b_{\text{next}} = 10 - 1 = 9$$

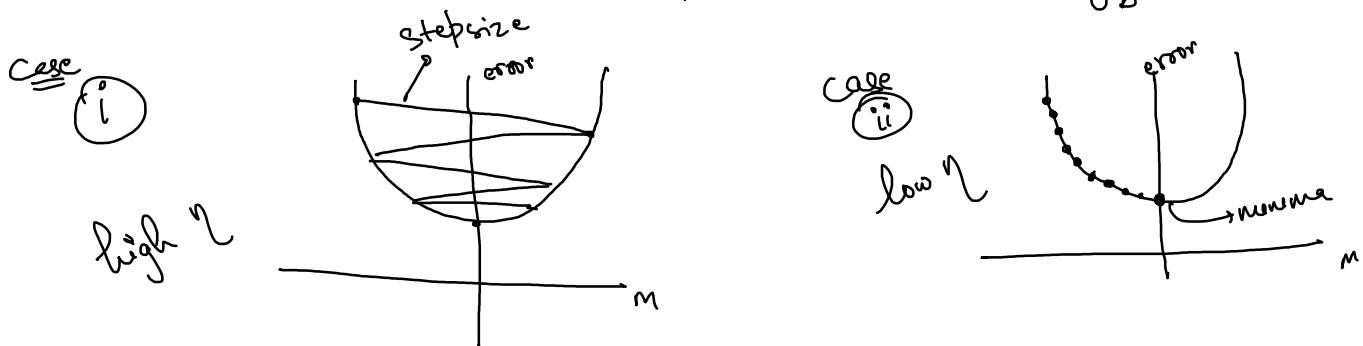
Actual steps:

## Actual steps:

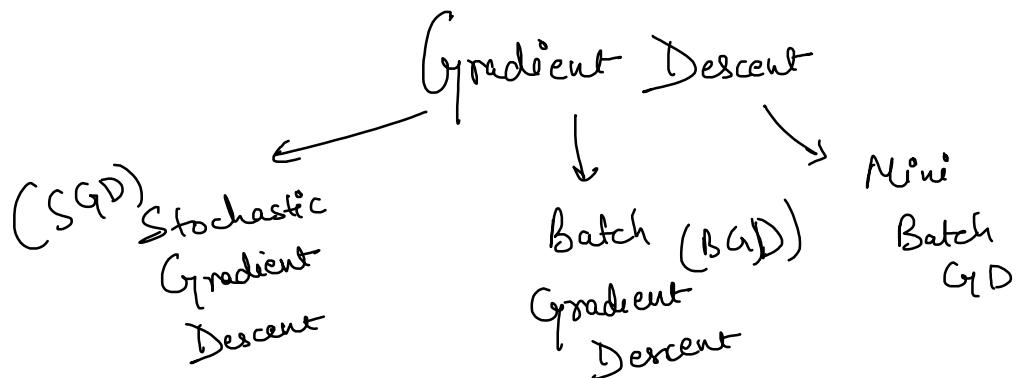
1) Take random values of  $m$  and  $b$

2) Find  $\frac{\partial L}{\partial m}$  &  $\frac{\partial L}{\partial b}$  gradient

3)  $m_{\text{next}} = m_{\text{old}} - \eta \frac{\partial L}{\partial m}$       |       $b_{\text{next}} = b_{\text{old}} - \eta \frac{\partial L}{\partial b}$



$\eta$ : step-size / learning rate

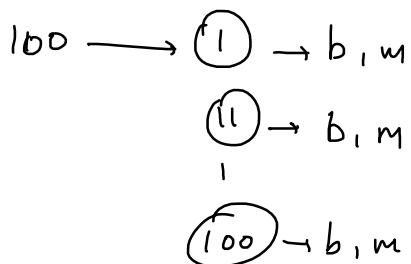
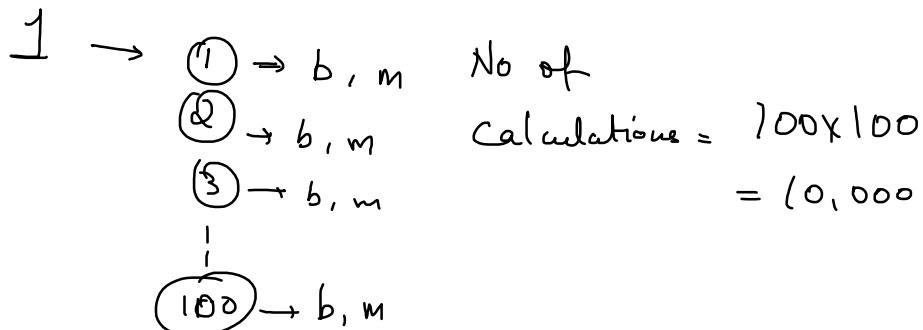


## Stochastic Gradient Descent:

$\rightarrow D_1, D_2, \dots, D_n$

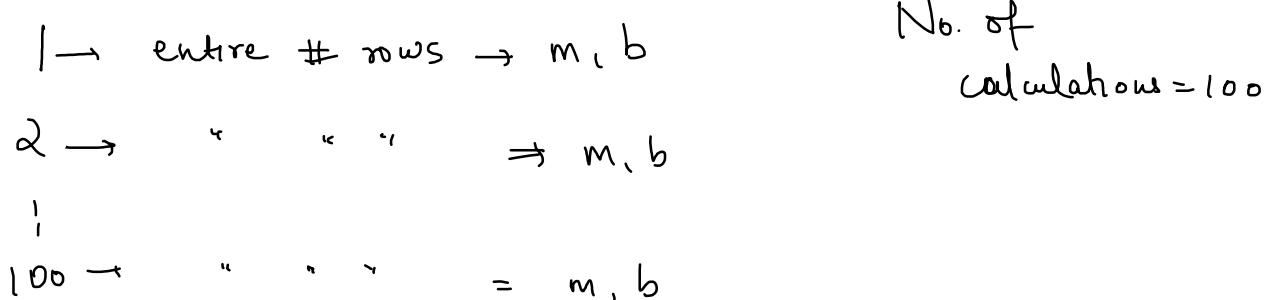
faster  $\rightarrow$  row wise operation

100 rows, 100 iterations

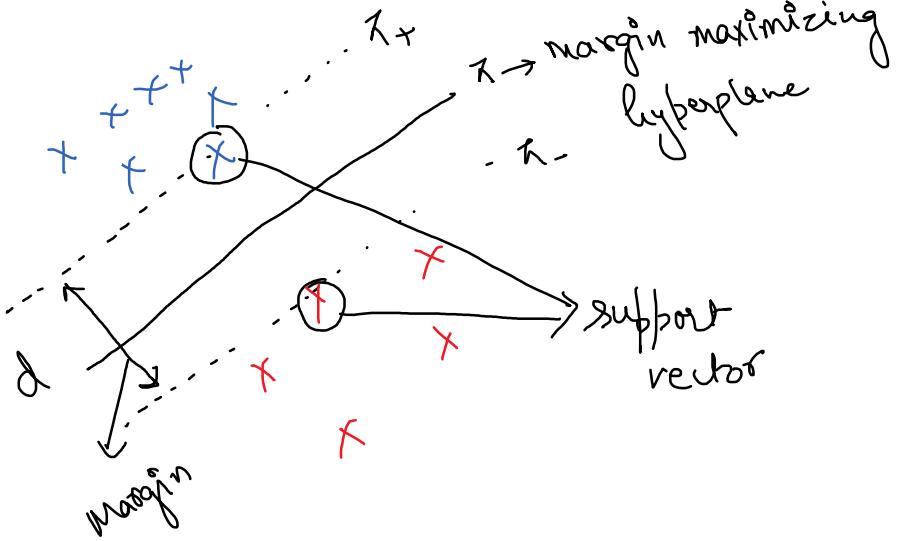
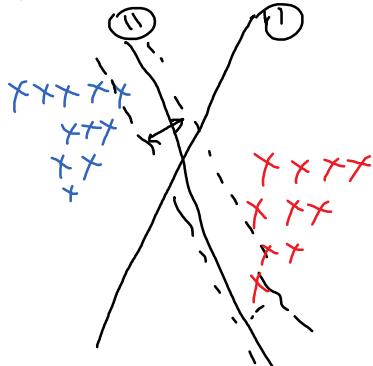


Batch Gradient Descent  $\Rightarrow$  for small dataset

100 rows, 100 iteration



Thursday, March 14, 2024 9:14 PM

SVM

## Mathematical Formulation:

Given data points  $(+1) = \mathbf{x}^i$ ,  $(-1) = \mathbf{y}^i$

Margin boundary conditions:

$$\mathbf{w}^\top \mathbf{x}_1 + b = +1$$

$$\mathbf{w}^\top \mathbf{x}_2 + b = -1$$

$$\mathbf{w}^\top \mathbf{x}_1 + b = 0$$

$$\mathbf{w}^\top \mathbf{x}_2 + b = 0$$

$$d = \frac{\|\mathbf{x}_1 - \mathbf{x}_2\|}{\|\mathbf{w}\|}$$

Normalizing,

$$\frac{\mathbf{w}^\top (\mathbf{x}_1 - \mathbf{x}_2)}{\|\mathbf{w}\|} = \frac{2}{\|\mathbf{w}\|}$$

$$d = \frac{(\mathbf{x}_1 - \mathbf{x}_2)}{\|\mathbf{w}\|} = \frac{2}{\|\mathbf{w}\|}$$

Find  $\mathbf{w}^*, b^* = \underset{(\mathbf{w}, b)}{\operatorname{argmax}} \left( \frac{2}{\|\mathbf{w}\|} \right) \rightarrow \text{Hard Margin SVM}$

→ Building constraints,  $y_i(w^T x_i + b)$

① For +ve support vector,

$$y_i(w^T x_i + b) = 1$$

② for -ve support vector

$$y_i(w^T x_i + b) = -1$$

③ for any blue point,

$$y_i(w^T x_i + b) > 1$$

④ for any red point

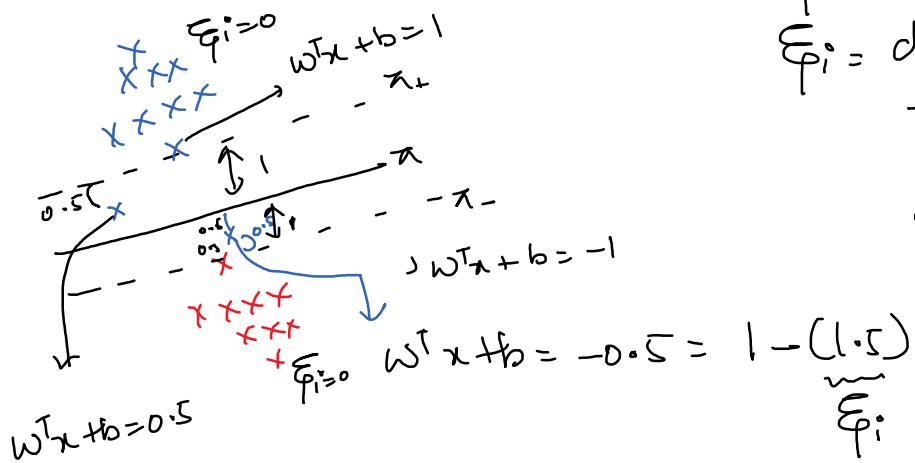
$$y_i(w^T x_i + b) < -1$$

MoF

$$w^*, b^* = \underset{(w,b)}{\operatorname{argmax}} \frac{2}{\|w\|} \text{ such that}$$

$$y_i(w^T x_i + b) \geq 1$$

Reality



represent misclassifications  
 $\epsilon_i$  = distance of point from your hyperplane in incorrect direction

$$\sum \epsilon_i$$

"Hinge loss"

$\omega^*, b^* = \underset{(\omega, b)}{\operatorname{arg\min}} \frac{\|\omega\|}{2} + C \sum_{i=1}^n \xi_i$  such that  
 $y_i (\omega^T x_i + b) \geq 1 - \xi_i$   
 Soft margin SVM  
 Regularizer  
 hyperparameter

- 1)  $C \uparrow \rightarrow$  overfitting
- 2)  $C \downarrow \rightarrow$  underfitting
- $\lambda \uparrow \rightarrow$  underfitting
- $\lambda \downarrow \rightarrow$  overfitting

$$C \propto \frac{1}{\lambda}$$

Loss minimization  $\rightarrow$  hinge loss  $\Rightarrow \max(0, 1 - z_i)$



$\xi_i = 0$   
 $x_i \text{ on } w^T x_i + b > 1$   
 $\xi_i = 1$   
 $x_i \text{ on } w^T x_i + b = 1$   
 $\xi_i = 2$   
 $x_i \text{ on } w^T x_i + b = -1$   
 $\xi_i = 3$

for  $x_i \Rightarrow y_i(w^T x_i + b) > 1$

$z_i > 1$ , assume  $z_i = 2$

$\Rightarrow$  hinge loss  $\Rightarrow \max(0, 1 - z_i)$

$\Rightarrow \max(0, 1 - 2) = 0$

$\Rightarrow \max(0, -1) = 0$

for  $\underline{x}_i$ : Hinge loss  $\Rightarrow \max(0, 1 - z_i)$ ,  $\xi_i = 1 - z_i = 3$   
 $\Rightarrow \max(0, 1 - (-2)) = \max(0, 3) = 3$

## Dual form of SVM:

Primal:  $\underset{w, b}{\arg \min} \frac{\|w\|}{2} + C \sum_{i=1}^n \xi_i$

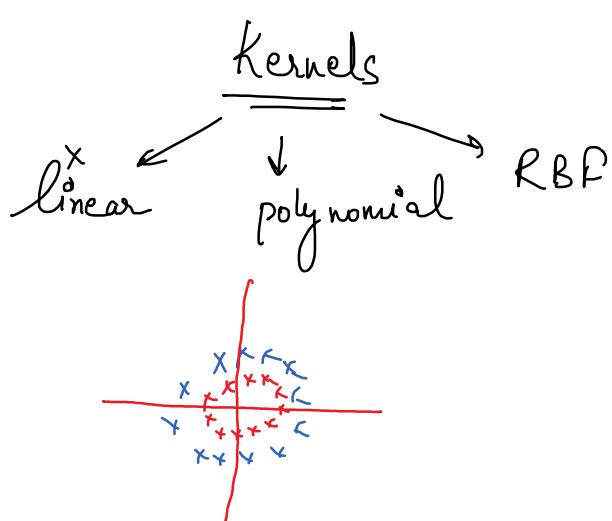
equivalent

Dual:  $\underset{\alpha_i}{\max} \sum_{i=1}^n \alpha_i - \frac{1}{2} \sum_{i=1}^n \sum_{j=1}^n \alpha_i \alpha_j y_i y_j (x_i^T x_j)$

$x_i^T x_j \downarrow$  similarity  
 $x_i \cdot x_j \downarrow$   
 $k(x_i, x_j)$

For  $x_i \rightarrow \alpha_i$

$\alpha_i \geq 0$  (for support vector)



"Kernel trick": transform your datapoints by applying kernels to make them linearly separable.

$$\text{Polynomial Kernel: } K(x_1, x_2) = (x_1^T x_2 + c)^d$$

## Quadratic Equations, $d=2$

$$X_1 = \begin{bmatrix} X_{11} & X_{12} \end{bmatrix}$$

$$X_2 = \begin{bmatrix} X_{21} & X_{22} \end{bmatrix}$$

$$\Rightarrow \left( 1 + \chi_1^T \chi_2 \right)^2 \Rightarrow \left( 1 + \begin{bmatrix} \chi_{11} & \chi_{12} \end{bmatrix} \begin{bmatrix} \chi_{21} \\ \chi_{22} \end{bmatrix} \right)^2$$

$$\Rightarrow \left( 1 + x_{11}x_{21} + x_{12}x_{22} \right)^2$$

$$\Rightarrow \left[ r^2 + x_{11}^2 x_{21}^2 + x_{12}^2 x_{22}^2 + 2x_{11}x_{21}x_{12}x_{22} + 2x_{11}x_{21} + 2x_{12}x_{22} \right]$$

$$X_1 = \begin{bmatrix} 1, X_{11}^2, X_{12}^2, \sqrt{2}X_{11}X_{12}, \sqrt{2}X_{11}, \sqrt{2}X_{12} \end{bmatrix} \quad 6d$$

\* Mercer's Theorem: Kernel converts the d-dim dataset into  $d'$  dim dataset such that  $d' \geq d$ .

## RBF (Radial Basis Function)

$$\text{RBF (Radial Basis Function)} \\ \kappa(x_1, x_2) = e^{-\frac{\|x_1 - x_2\|^2}{2\sigma^2}}$$

distance  
variance

$$= e^{-\alpha^2/2r^2}$$

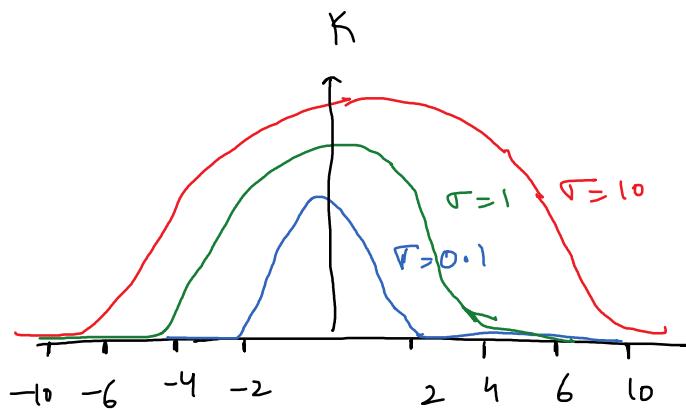
Case 1:  $\|x_1 - x_2\| = d$

$$K = \frac{1}{e^{d^2/2\sigma^2}}$$

$$d \uparrow \rightarrow d^2 \uparrow \rightarrow e^{d^2/2\sigma^2} \uparrow \rightarrow \frac{1}{e^{d^2/2\sigma^2}} \downarrow \rightarrow K \downarrow$$

Case 2:  $K = \frac{1}{e^{d^2/2\sigma^2}}$

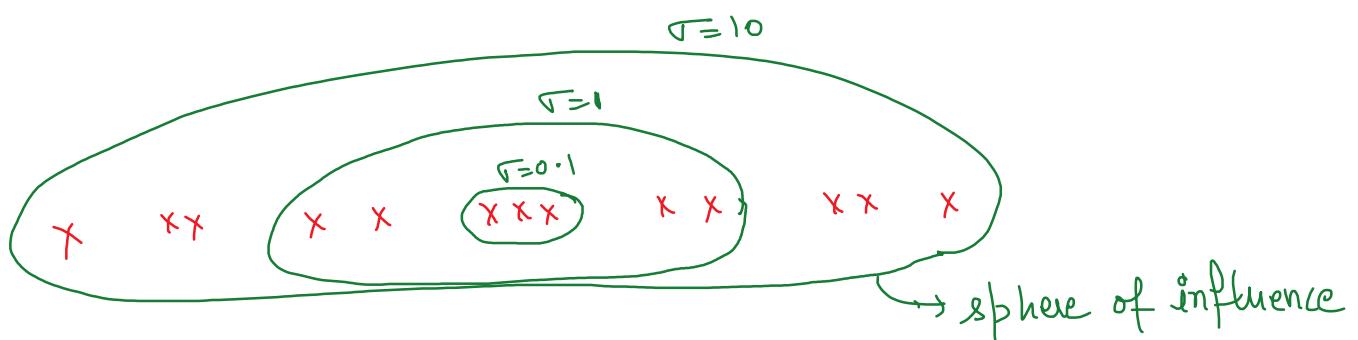
$$\sigma \uparrow \rightarrow 2\sigma^2 \uparrow \rightarrow e^{d^2/2\sigma^2} \downarrow \rightarrow \frac{1}{e^{d^2/2\sigma^2}} \uparrow \rightarrow K \uparrow$$



$\sigma = 0.1$

$\sigma = 1$

$\sigma = 10$



# hyperparameter is gamma in SVM (RBF)

$$V = \frac{1}{2\sigma^2} \Rightarrow \frac{1}{e^{d^2\gamma}} \quad K \propto \underline{\perp}$$

$$\tau \uparrow \rightarrow 2\sigma^2 \uparrow \rightarrow \gamma \downarrow \rightarrow \frac{1}{e^{d^2\gamma}} \downarrow \rightarrow K \uparrow$$

# Decision Trees

Dataset =  $[SL, SW, PL, PW]$

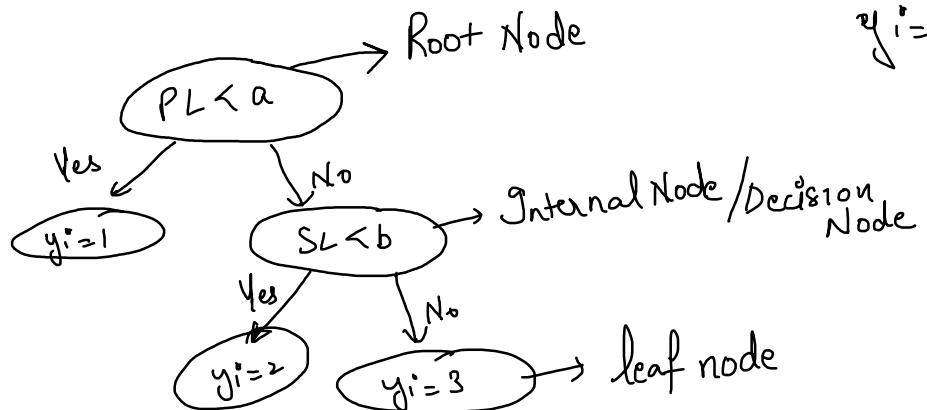


```

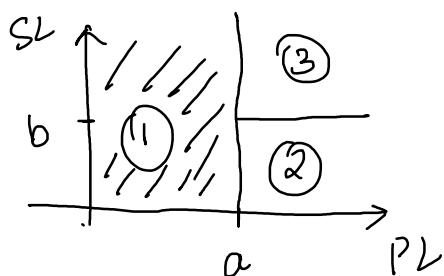
if  $PL < a$ :
     $y_i = 1$ 
else:
    if  $SL < b$ :
         $y_i = 2$ 
    else:
         $y_i = 3$ 

```

## Flowchart



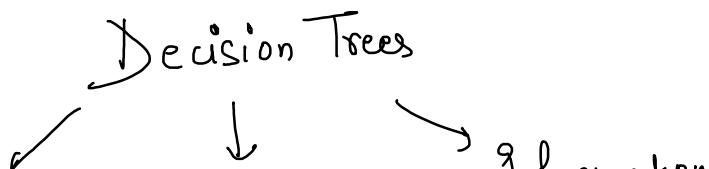
## Recursive Partitioning (Axis Parallel Hyperplanes)

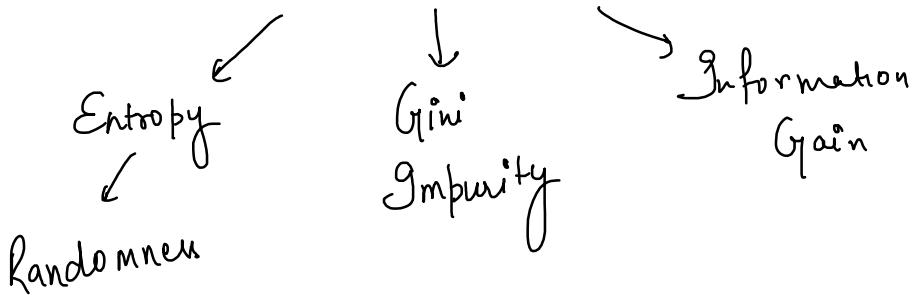


```

if  $PL < a$ :
     $y_i = 1$ 
else:
    if  $SL < b$ :
         $y_i = 2$ 
    else:
         $y_i = 3$ 

```





Entropy:  $H_D(Y) = - \sum_{i=1}^n p_i \lg(p_i) \Rightarrow \lg \Rightarrow \log_2$

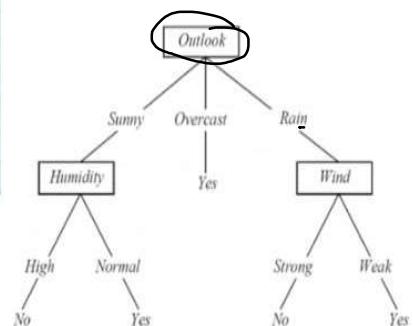
$$H_D(Y) = -P(Y_s) \lg(P(Y_s)) - P(Y_n) \lg(P(Y_n))$$

(Parent's Entropy)

$$= - \left[ \frac{9}{14} \lg \frac{9}{14} + \frac{5}{14} \lg \frac{5}{14} \right]$$

$$= 0.94$$

| Outlook  | Temperature | Humidity | Windy | PlayTennis |
|----------|-------------|----------|-------|------------|
| Sunny    | Hot         | High     | False | No         |
| Sunny    | Hot         | High     | True  | No         |
| Overcast | Hot         | High     | False | Yes        |
| Rainy    | Mild        | High     | False | Yes        |
| Rainy    | Cool        | Normal   | False | Yes        |
| Rainy    | Cool        | Normal   | True  | No         |
| Overcast | Cool        | Normal   | True  | Yes        |
| Sunny    | Mild        | High     | False | No         |
| Sunny    | Cool        | Normal   | False | Yes        |
| Rainy    | Mild        | Normal   | False | Yes        |
| Sunny    | Mild        | Normal   | True  | Yes        |
| Overcast | Mild        | High     | True  | Yes        |
| Overcast | Hot         | Normal   | False | Yes        |
| Rainy    | Mild        | High     | True  | No         |



entropy for each column:

$$\begin{aligned}
 & \text{Outlook} \rightarrow \\
 & \quad \text{Sunny } \xrightarrow{5} (2Y, 3N) \rightarrow -\frac{2}{5} \lg \frac{2}{5} - \frac{3}{5} \lg \frac{3}{5} = 0.97 \\
 & \quad \text{overcast } \xrightarrow{4} (4Y, 0N) \rightarrow -\frac{4}{4} \lg \frac{4}{4} - 0 \lg 0 = 0 \\
 & \quad \text{Rainy } \xrightarrow{5} (3Y, 2N) \rightarrow -\frac{3}{5} \lg \frac{3}{5} - \frac{2}{5} \lg \frac{2}{5} = 0.97
 \end{aligned}$$

weighted entropy  $\Rightarrow H_D(Y, \text{outlook}) = \frac{5}{14} \times 0.97 + \frac{4}{14} \times 0 + \frac{5}{14} \times 0.97$

$$= 0.97 \times \frac{5}{7} = 0.69$$

Temperature

|   |  |  |
|---|--|--|
| $\begin{cases} 4 \\ 6 \\ 4 \end{cases}$ | $\begin{cases} (2Y, 2N) \\ (4Y, 2N) \\ (3Y, 1N) \end{cases}$ | $\begin{aligned} \text{Hot} &\Rightarrow -\frac{2}{4} \lg \frac{2}{4} - \frac{2}{4} \lg \frac{2}{4} = 1 \\ \text{Mild} &\Rightarrow -\frac{4}{6} \lg \frac{4}{6} - \frac{2}{6} \lg \frac{2}{6} = 0.91 \\ \text{Cool} &\Rightarrow -\frac{3}{4} \lg \frac{3}{4} - \frac{1}{4} \lg \frac{1}{4} = 0.81 \end{aligned}$ |
|---|--|--|

weighted entropy  $= H_D(Y, \text{temp}) = \frac{4}{14} \times 1 + \frac{6}{14} \times 0.91 + \frac{4}{14} \times 0.81$

$$= 0.91$$

Humidity

|                                    |  |  |
|------------------------------------|--|--|
| $\begin{cases} 7 \\ 7 \end{cases}$ | $\begin{cases} (3Y, 4N) \\ (6Y, 1N) \end{cases}$ | $\begin{aligned} \text{High} &\Rightarrow -\frac{3}{7} \lg \frac{3}{7} - \frac{4}{7} \lg \frac{4}{7} = 0.98 \\ \text{Normal} &\Rightarrow -\frac{6}{7} \lg \frac{6}{7} - \frac{1}{7} \lg \frac{1}{7} = 0.59 \end{aligned}$ |
|------------------------------------|--|--|

weighted entropy  $\Rightarrow \frac{7}{14} \times 0.98 + \frac{7}{14} \times 0.59$

$$\Rightarrow \frac{7}{14} (0.98 + 0.59) = 0.78$$

$$\begin{array}{l}
 \text{windy} \rightarrow \\
 \left\{ \begin{array}{ll}
 \text{True} & \Rightarrow -\frac{3}{6} \lg \frac{3}{6} - \frac{3}{6} \lg \frac{3}{6} = 1 \\
 \text{False} & \Rightarrow -\frac{6}{8} \lg \frac{6}{8} - \frac{2}{8} \lg \frac{2}{8} = 0.81
 \end{array} \right.
 \end{array}$$

$$\begin{aligned}
 \text{weighted entropy} &\Rightarrow H_D(Y, \text{windy}) = \frac{6}{14} \times 1 + \frac{8}{14} \times 0.81 \\
 &= 0.89
 \end{aligned}$$

choose column for the split:

- a) compare weighted entropies & choose column with least entropy:

|                |             |          |       |
|----------------|-------------|----------|-------|
| <u>outlook</u> | temperature | humidity | windy |
| ↓              | ↓           | ↓        | ↓     |
| 0.69           | 0.91        | 0.78     | 0.89  |

- (b) Information Gain:

$$IG(Y) = \text{Parent entropy} - \text{column entropy} \text{ (weighted)}$$

$$= IG(Y, \text{outlook}) = 0.94 - 0.69 = 0.25$$

$$IG(Y, \text{temperature}) = 0.94 - 0.91 = 0.03$$

$$IG(Y, \text{humidity}) = 0.94 - 0.78 = 0.16$$

$$IG(Y, \text{windy}) = 0.94 - 0.81 = 0.13$$

Since, outlook has highest  $IG$ , choose this column for split!

Properties:  $H_D(Y) = -P(y_+) \lg(P(y_+)) - P(y_-) \lg(P(y_-))$

Case 1:  $P(y_+) = 0.99 \quad P(y_-) = 0.01$

$$H_D(Y) = -0.99 \lg 0.99 - 0.01 \lg 0.01 = 0.08$$

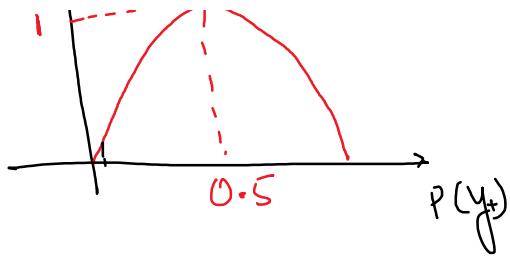
Case 2:  $P(y_+) = 0.5 \quad P(y_-) = 0.5$

$$H_D(Y) = -0.5 \lg 0.5 - 0.5 \lg 0.5 = 1$$

Case 3:  $P(y_+) = 1 \quad P(y_-) = 0$

$$H_D(Y) = -1 \lg 1 - 0 \lg 0 = 0$$





$$\frac{\text{Gini Impurity}}{(I_G)}$$

$$I_G(Y) = 1 - \sum_{i=1}^n (P_i^+)^2$$

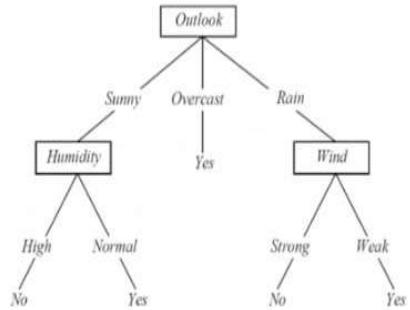
For binary classification,

$$I_G(Y) = 1 - [P(y_+)^2 + P(y_-)^2]$$

| Outlook  | Temperature | Humidity | Windy | PlayTennis |
|----------|-------------|----------|-------|------------|
| Sunny    | Hot         | High     | False | No         |
| Sunny    | Hot         | High     | True  | No         |
| Overcast | Hot         | High     | False | Yes        |
| Rainy    | Mild        | High     | False | Yes        |
| Rainy    | Cool        | Normal   | False | Yes        |
| Rainy    | Cool        | Normal   | True  | No         |
| Overcast | Cool        | Normal   | True  | Yes        |
| Sunny    | Mild        | High     | False | No         |
| Sunny    | Cool        | Normal   | False | Yes        |
| Rainy    | Mild        | Normal   | False | Yes        |
| Sunny    | Mild        | Normal   | True  | Yes        |
| Overcast | Mild        | High     | True  | Yes        |
| Overcast | Hot         | Normal   | False | Yes        |
| Rainy    | Mild        | High     | True  | No         |

for multiclass classification,

$$I_G(Y) = 1 - [P(y_1)^2 + P(y_2)^2 + \dots + P(y_n)^2]$$



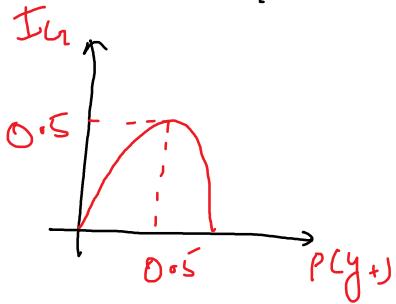
Properties of Gini Impurity:

$$\underline{\text{Case 1:}} \quad P(y_+) = 0.5 \quad , \quad P(y_-) = 0.5$$

$$\begin{aligned}
 I_G &= 1 - [P(y_+)^2 + P(y_-)^2] = 1 - [0.5^2 + 0.5^2] \\
 &= 1 - [0.25 + 0.25] \\
 &= 0.5
 \end{aligned}$$

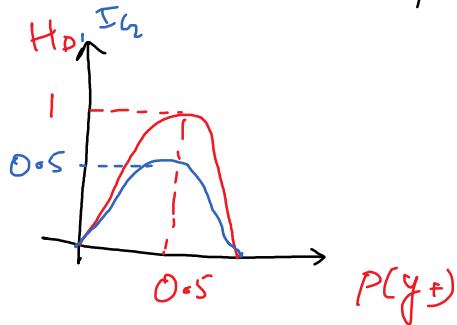
Case 2:  $P(y_+) = 1$ ,  $P(y_-) = 0$

$$I_G = 1 - [P(y_+)^2 + P(y_-)^2] = 1 - [1 + 0] = 0$$



Comparison b/w Gini Impurity and Entropy:

(I)



(II)  $I_G$  is easier to calculate, hence more computationally efficient than entropy.

Calculate  $I_G$  for our dataset:

$$I_G(Y) = 1 - \left[ P(\text{Yes})^2 + P(\text{No})^2 \right] = 1 - \left[ \left( \frac{9}{14} \right)^2 + \left( \frac{5}{14} \right)^2 \right]$$

$$= 0.459$$

$I_G$  for each column:

$$\underbrace{5}_{\text{columns}} \xrightarrow{(2Y, 3N)} 1 - \left[ \left( \frac{2}{5} \right)^2 + \left( \frac{3}{5} \right)^2 \right] = 0.48$$

$\overline{I_G(\text{Outlook})} = \frac{1}{3} \left[ 1 - \left[ \left( \frac{2}{5} \right)^2 + \left( \frac{3}{5} \right)^2 \right] \right] = 0.48$   
 Outlook →  $\begin{cases} 5 & (\text{Sunny}, 3N) \\ 4 & (\text{Overcast}, 0N) \\ 5 & (\text{Rainy}, 2N) \end{cases}$ 
 $\Rightarrow 1 - [1 - 0] = 0$   
 $\Rightarrow 1 - \left[ \left( \frac{2}{5} \right)^2 + \left( \frac{3}{5} \right)^2 \right] = 0.48$

$\text{Weighted Impurity} \Rightarrow \frac{5}{14} \times 0.48 + \frac{4}{14} \times 0 + \frac{5}{14} \times 0.48 = \frac{5}{7} \times 0.48 = 0.342$

$\overline{I_G(\text{Temp})} = \frac{4}{14} \left[ 1 - \left[ \left( \frac{2}{4} \right)^2 + \left( \frac{2}{4} \right)^2 \right] \right] = 0.5$   
 Temperature →  $\begin{cases} 4 & (\text{Hot}, 2N) \\ 6 & (\text{Mild}, 2N) \\ 4 & (\text{Cold}, 1N) \end{cases}$ 
 $\Rightarrow 1 - \left[ \left( \frac{4}{6} \right)^2 + \left( \frac{2}{6} \right)^2 \right] = 0.444$   
 $\Rightarrow 1 - \left[ \left( \frac{3}{4} \right)^2 + \left( \frac{1}{4} \right)^2 \right] = 0.375$

$\text{Weighted } I_G(\text{Temp}) = \frac{4}{14} \times 0.5 + \frac{6}{14} \times 0.444 + \frac{4}{14} \times 0.375 = 0.44$

$\overline{I_G(\text{Humidity})} = \frac{7}{14} \left[ 1 - \left[ \left( \frac{3}{7} \right)^2 + \left( \frac{4}{7} \right)^2 \right] \right] = 0.489$   
 Humidity →  $\begin{cases} 7 & (\text{High}, 4N) \\ 7 & (\text{Normal}, 1N) \end{cases}$ 
 $\Rightarrow 1 - \left[ \left( \frac{6}{7} \right)^2 + \left( \frac{1}{7} \right)^2 \right] = 0.244$

$\text{Weighted } I_G(\text{Humidity}) = \frac{7}{14} \times 0.489 + \frac{7}{14} \times 0.244$   
 $\Rightarrow 0.367$

$\overline{I_G(\text{Temp})} = \frac{6}{14} \left[ 1 - \left[ \left( \frac{3}{3} \right)^2 + \left( \frac{3}{3} \right)^2 \right] \right] = 0.5$   
 $\Rightarrow 1 - \left[ \left( \frac{3}{3} \right)^2 + \left( \frac{3}{3} \right)^2 \right] = 0.5$

$$\begin{array}{l}
 \text{Windy} \\
 \left| \begin{array}{l}
 \xrightarrow{6} \text{True} \quad \Rightarrow 1 - \left[ \left( \frac{3}{6} \right)^2 + \left( \frac{3}{6} \right)^2 \right] = 0.5 \\
 \xrightarrow{8} \text{False} \quad \Rightarrow 1 - \left[ \left( \frac{6}{8} \right)^2 + \left( \frac{2}{8} \right)^2 \right] = 0.375
 \end{array} \right.
 \end{array}$$

$$\begin{aligned}
 \text{weighted } I_b(\text{Windy}) &= \frac{6}{14} \times 0.5 + \frac{8}{14} \times 0.375 \\
 &= 0.429
 \end{aligned}$$

Two ways to choose:

- ① choose column with least Gini Impurity  
(weighted)

Outlook, temp, humidity, Windy  
 0.342, 0.44, 0.367, 0.429

- ② Information Gain  $\Rightarrow$  Parent's Gini Impurity - weighted Gini Impurity

$$I_{G_O}(Y) \Rightarrow 0.459 - 0.342 = 0.117$$

$$I_{G_T}(Y) \Rightarrow 0.459 - 0.440 = 0.019$$

$$I_{G_H}(\gamma) \Rightarrow 0.459 - 0.367 \Rightarrow 0.092$$

$$I_{G_W}(\gamma) \Rightarrow 0.459 - 0.429 = 0.030$$

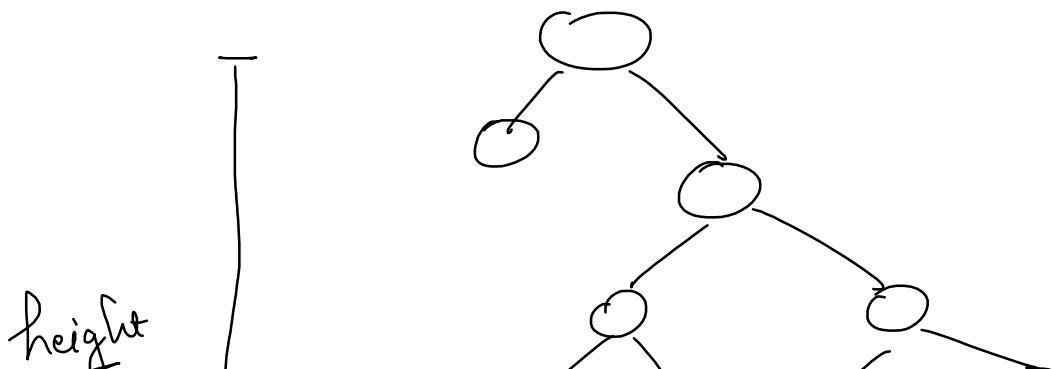
Since, information gained is largest for outlook, we will choose outlook for the split!

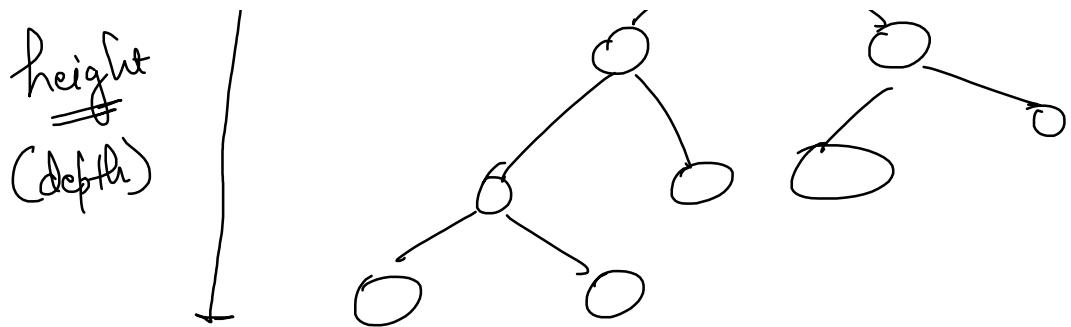
When to stop a tree?

- a) Pure Node
- b) If you have very few points left in your leaf node



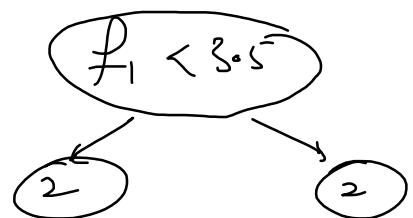
\* Hyperparameter  $\Rightarrow$  depth  $\Rightarrow$  max-depth  $\Rightarrow$  height of your tree



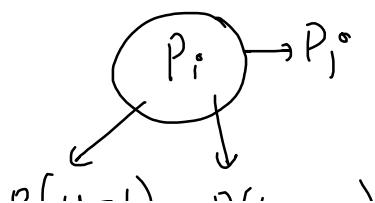


### Splitting of Numerical features

| $f_1$ | $y$ | ① Sort the variable           |
|-------|-----|-------------------------------|
| 2.2   | 1   | ② $f_1 < 2.2 \quad f_1 < 4.6$ |
| 2.6   | 1   | $f_1 < 2.6 \quad f_1 < 5.3$   |
| 3.5   | 0   | $f_1 < 3.5$                   |
| 4.6   | 1   |                               |
| 5.3   | 0   |                               |



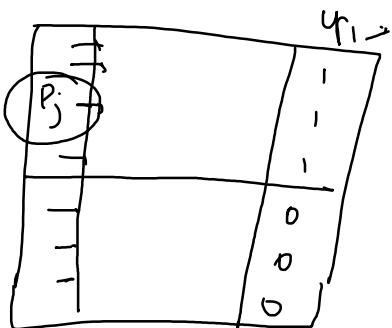
feature Engineering: Categorical column : PINCODE  
with a lot of categories



$$P(y_{i=1} / P_j) = \frac{P_j \cap y=1}{P_j}$$

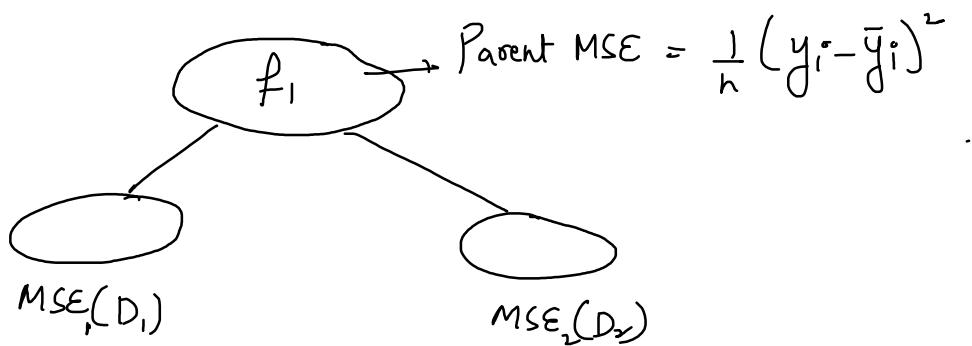
$$P(y=1) \quad P(y=0)$$

$$P(y_i=1/P_j^*) = \frac{r_j \cap y=1}{P_j}$$



$$P(y=1/P_j^*) = \frac{\# P_j \text{ when } y_i=1}{\# P_j} = \text{Numerical column}$$

Regression in DT:

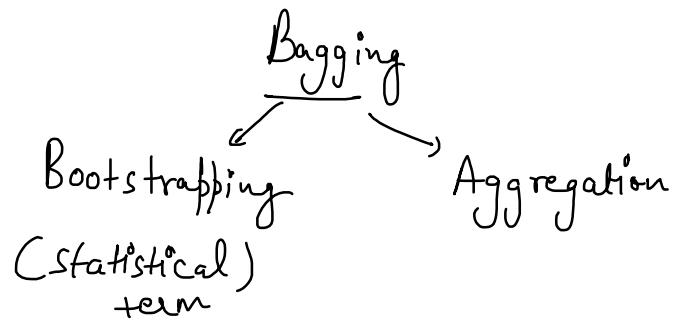
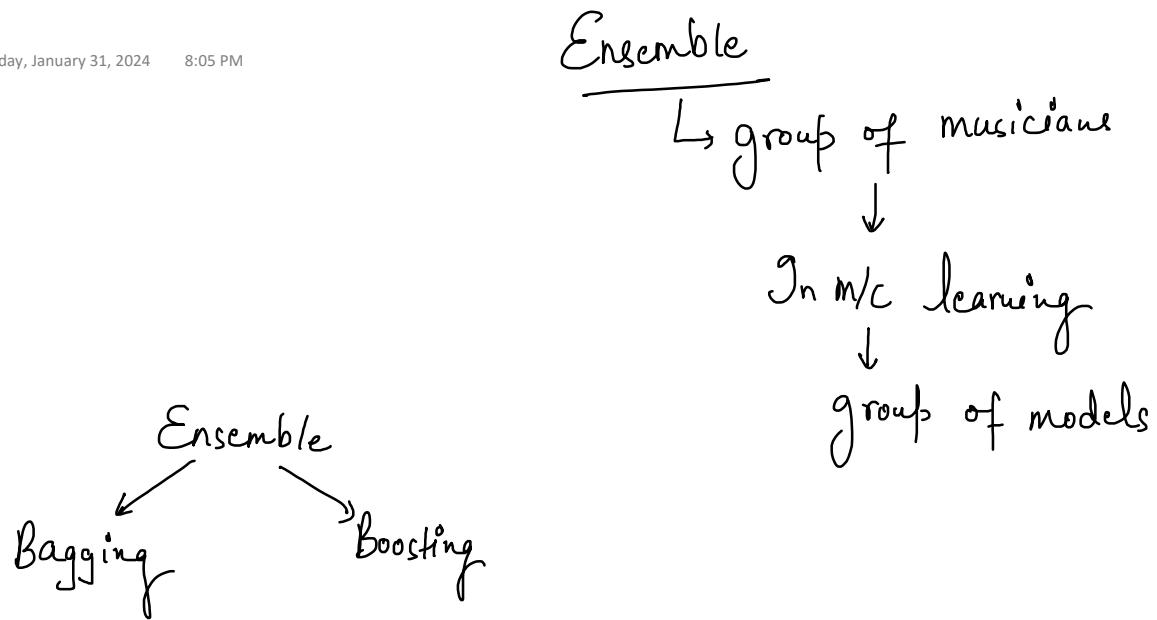


ADVANTAGES:

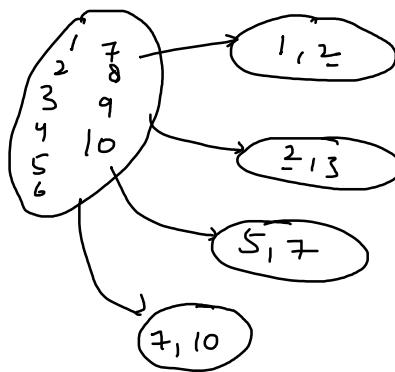
→ It is very easily interpretable

→ Important features

→ No need to standardize



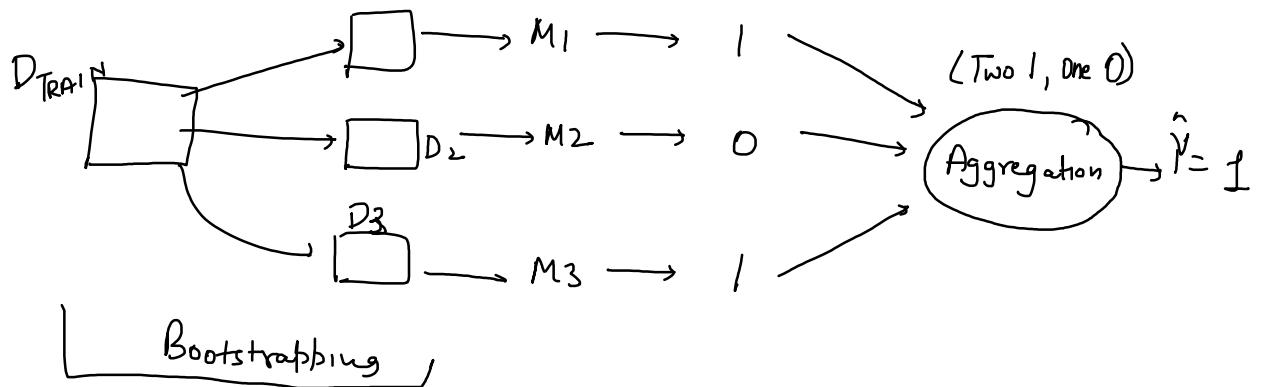
Bootstrapping  $\Rightarrow$  sampling with replacement



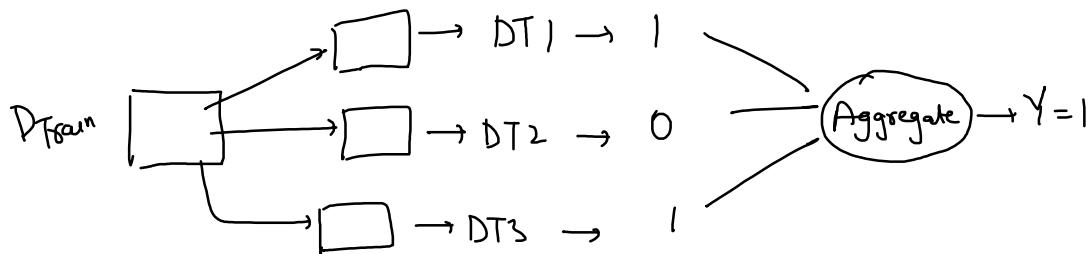
Bagging Algo.

Samples      Models      O/P

$D_1$



Random Forest: DT should be reasonable depth (low bias & high variance)



# Models should be different from each other

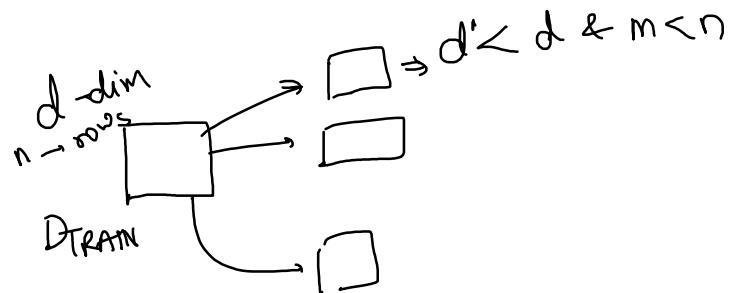
|   | CGPA | IQ  | EXTRA | SOCIAL | PLACED |
|---|------|-----|-------|--------|--------|
| ① | 7    | 110 | 10    | 9      | 1      |
| ② | 8    | 112 | 9     | 8      | 0      |
| ③ | 9    | 120 | 8     | 7      | 0      |
| ④ | 10   | 125 | 7     | 6      | 1      |

Take ① & ② rows, 50% of columns

| CGPA | $D_1$ |        | $D_2$ |        |
|------|-------|--------|-------|--------|
|      | Extra | Placed | IQ    | Social |
| 7    | 10    | 1      | 110   | 9      |
| 8    | 9     | 0      | 112   | 8      |

In order to create different samples, we are doing column sampling & row sampling.

$RF = \underbrace{\text{low bias \& high variance DTs} + \text{Row Sampling} + \text{Column Sampling}}_{\text{max\_depth}} + \text{Aggregation}$



- Hyperparameters:
  - $\Rightarrow n_{\text{estimators}} \Rightarrow [100 - 500]$
  - $\Rightarrow \text{Max\_depth} \uparrow \rightarrow \text{chances of overfitting} \uparrow$
  - $\Rightarrow \text{column sampling} \xrightarrow{\text{max\_features}} ["\text{auto}", "sqrt", "log", 0.7]$
  - $\Rightarrow \text{column sampling rate} \Rightarrow \frac{d'}{d} \quad \left. \right] \uparrow \alpha \text{ overfitting} \uparrow$
  - $\Rightarrow \text{row sampling rate} \Rightarrow \frac{m}{n}$
  - $\Rightarrow n_{\text{jobs}} = -1$

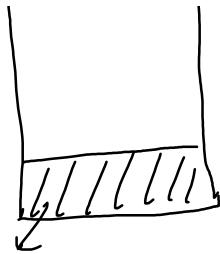
OOB score:

$\downarrow$   
out of  
0



$RF(\text{OOB\_score} = \text{True})$   
 $\hookrightarrow$  accuracy

out of bag



↳ accuracy

oob-score & accuracy

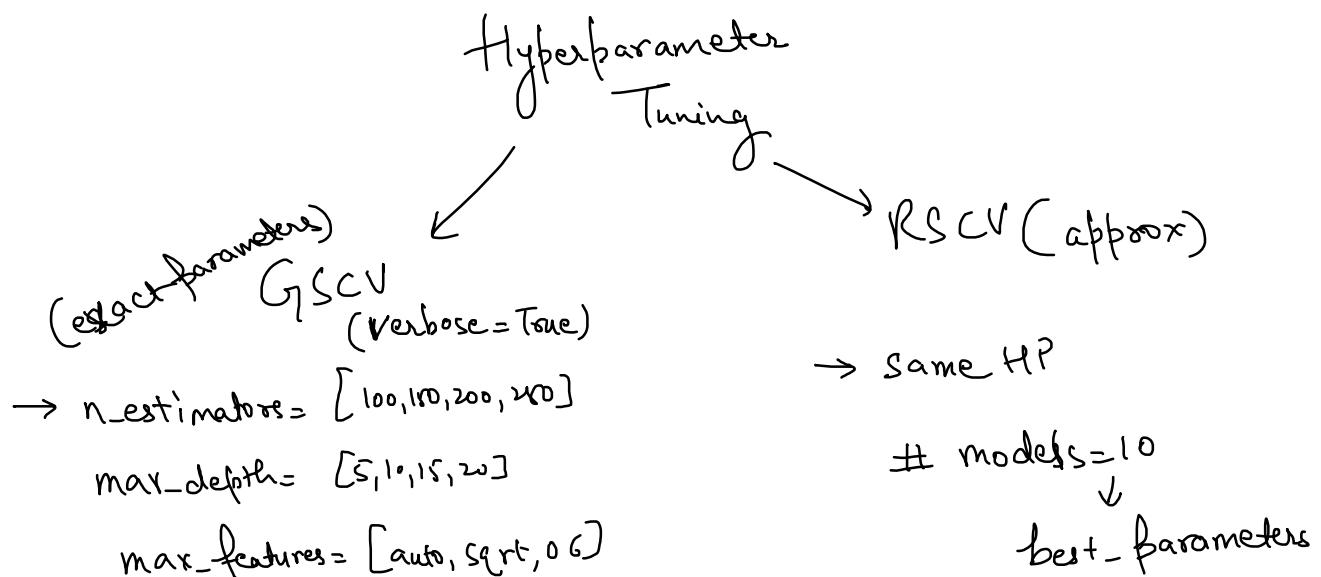
out of bag (37.1%)  
sample

Advantages:

→ feature importances

Disadvantages:

→ black box

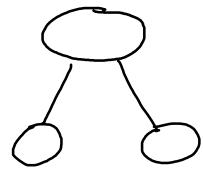


$$GSCV = 4 \times 4 \times 3 = 48 \text{ models}$$

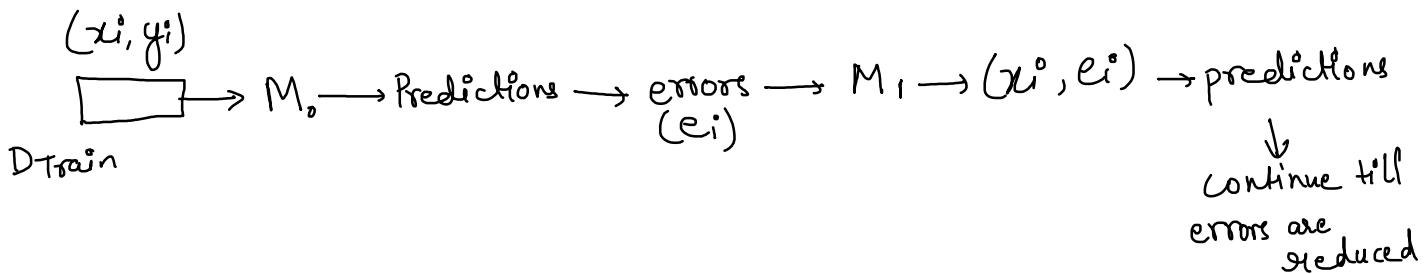
# Boosting

① Boosting Models  $\Rightarrow$  high bias & low variance

[Decision Tree of depth  $\rightarrow 1$ ]  
 ↘ Decision  
 Stump



## Boosting Flowchart:



o)  $D_{Train} = \{x_i^o, y_i^o\}_{i=1}^n \rightarrow M_0 \rightarrow$  predictions  $\rightarrow$  errors  
 $\downarrow$   
 $e_i = y_i - \hat{y}_i$   
 $= y_i - h_0(x)$

1)  $M_1 \rightarrow \{x_i^o, e_i\}_{i=1}^n \Rightarrow e_i = y_i - h_0(x)$   
 $\downarrow$   
 $h_1(x)$

Model at end of stage 1:

$f_1(x) = \underbrace{\lambda_0 h_0(x)}_{\text{New predictions}} + \underbrace{\lambda_1 h_1(x)}_{\text{predictions at previous stage}} \quad [\text{additive combine}]$

$$\text{errors} = y_i - f_1(x)$$

$$2) M_2 \rightarrow \{x_i^i, e_i\}_{i=1}^n, e_i = y_i - f_1(x)$$

$\underbrace{h_2(x)}$

Model at end of 2<sup>nd</sup> stage:

$$f_2(x) = \underbrace{\alpha_0 h_0(x) + \alpha_1 h_1(x)}_{f_1(x)} + \alpha_2 h_2(x)$$

$$f_2(x) = f_1(x) + \alpha_2 h_2(x)$$

New predictions       $\downarrow$  Previous predictions

for k<sup>th</sup> stage:

$$f_k(x) = \alpha_0 h_0(x) + \alpha_1 h_1(x) + \dots + \alpha_k h_k(x)$$

$$f_k(x) = \sum_{i=0}^k \alpha_i h_i(x)$$

$$f_k(x) = f_{k-1}(x) + \alpha_k h_k(x)$$

$k = \# \text{ Models} \rightarrow \text{hyperparameters}$

~~Residuals & loss functions:~~

$$L(y_i, f_k(x)) = [y_i - f_k(x)]^2$$

$$\frac{\partial L}{\partial f_k(x)} = \frac{\partial [y_i - f_k(x)]^2}{\partial f_k(x)} = -2[y_i - f_k(x)]$$

$$-\frac{\partial L}{\partial f_k(x)} = [y_i - f_k(x)] \quad \text{error}$$

negative gradient  
or  
pseudo-residual

### Gradient Boosting

$\mathcal{G}(P \Rightarrow \{x_i^*, y_i^*\}_{i=1}^n + \text{differentiable loss function})$

0)  $F_0 = \arg \min_r \sum_{i=0}^n L(y_i^*, r) \Rightarrow r = \bar{y}^*$

1) for  $m=1$  to  $M$

$$\eta_m = - \left[ \frac{\partial L(y_i^*, f_{m-1}(x))}{\partial f_{m-1}(x)} \right]$$

for  $m=1$ ,

$$\eta_1 = - \left[ \frac{\partial L(y_i^*, F_0(x))}{\partial F_0(x)} \right]$$

2)  $f_m(x)$  that can fit pseudo-residuals, train  $f_m(x)$  on  $\{x_i^*, \eta_i^*\}$

$$3) \quad r_m = \underset{r}{\operatorname{argmin}} \quad [L(y_i, \underbrace{f_{m-1}(x_i)}_{f_m} + \underbrace{r_m h_m(x)}_{h_m})]$$

$$4) \quad f_m(x) = f_{m-1}(x_i) + r_m h_m(x)$$

New predictions = old predictions + models (additive combine)

Hyperparameters:  $M = \# \text{models} \uparrow \rightarrow \text{bias} \downarrow \rightarrow \text{variance} \uparrow \rightarrow \text{overfitting} \uparrow$

Shrinkage:  $f_m = f_{m-1}(x) + \begin{cases} r_m h_m(x) \\ \text{learning rate} \rightarrow 0 < \gamma < 1 \end{cases}$

$\gamma$  reduces  $r_m h_m(x)$  which in turn reduces overfitting

GBDT  $\Rightarrow$  models are DTs. (MSE)

GBDT  $\rightarrow$  very slow  $\rightarrow$  optimized

- Taylor's Series
- Xgboost
- ↳ pip install Xgboost

Example:  $y_i$

|       |  |
|-------|--|
| 12000 | $\frac{\partial L}{\partial r} = - \sum (y_i - r)$ |
| 16500 | $r^2 - 16500r + 12000^2$                           |
| 14000 | $r^2 - 14000r + 12000^2$                           |
| 11700 | $r^2 - 11700r + 12000^2$                           |

14000

15500

$$L = -\frac{1}{2} (12000 - r)^2 - \frac{1}{2} (16500 - r)^2 - \frac{1}{2} (14000 - r)^2 - \frac{1}{2} (15500 - r)^2$$

$$\frac{\partial L}{\partial r} = -\frac{1}{2} (12000 - r) - \frac{1}{2} (16500 - r) - \frac{1}{2} (14000 - r) - \frac{1}{2} (15500 - r) = 0$$

$$\Rightarrow 12000 + r - 16500 + r - 14000 + r - 15500 + r = 0$$

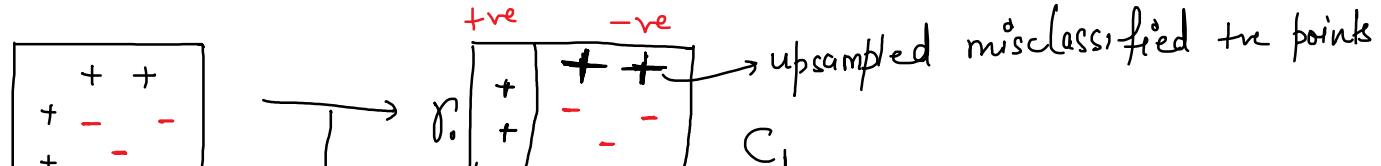
$$\Rightarrow 4r = 58000$$

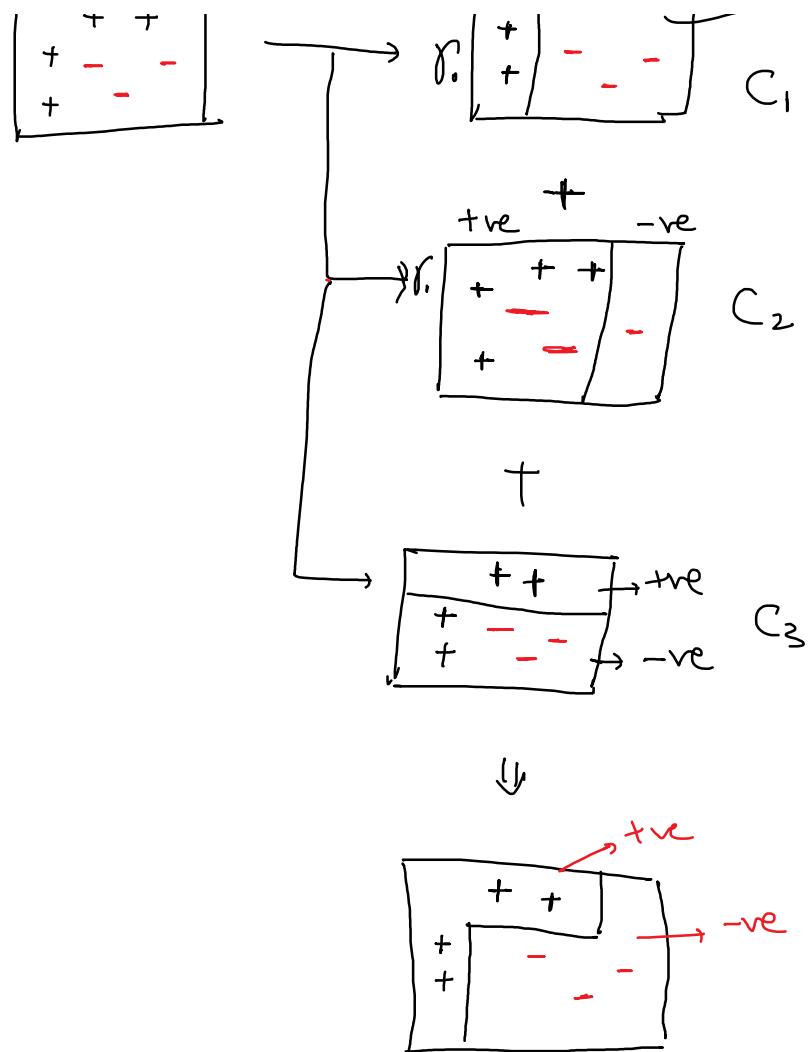
$$\Rightarrow r = \frac{58000}{4} = 14500$$

Simply calculating mean of  $y_i^*$   $\Rightarrow \bar{y}_i^* = \frac{12000 + 15500 + 16500 + 14000}{4}$

$$\boxed{\bar{y}_i^* = 14500}$$

ADA BOOST  
Adaptive Boosting





$$C = f_1 C_1 + f_2 C_2 + f_3 C_3 + \dots \quad n = \# \text{ rows}$$

| $X_1$ | $X_2$ | $y$ | $\hat{y}$  | weight( $f_n$ )  |
|-------|-------|-----|------------|------------------|
| 3     | 9     | 1   | 1          | $\gamma_5 = 0.2$ |
| 2     | 4     | 0   | 1 $\times$ | $\gamma_5 = 0.2$ |
| 1     | 5     | 1   | 0 $\times$ | $\gamma_5 = 0.2$ |
| 9     | 6     | 0   | 0          | $\gamma_5 = 0.2$ |
| 5     | 7     | 0   | 0          | $\gamma_5 = 0.2$ |

$\chi$  = error rate

error = algebraic sum of weights of misclassified points

$$\text{error} = 0.2 + 0.2 = 0.4$$

$$\chi = \frac{1}{2} \ln \left( \frac{1 - \text{error}}{\text{error}} \right) = \frac{1}{2} \ln \left( \frac{1 - 0.4}{0.4} \right)$$

~ ~ error

$$= \frac{1}{2} \ln \left( \frac{0.6}{0.4} \right)$$

$$\alpha = 0.2$$

$$e = 2.78$$

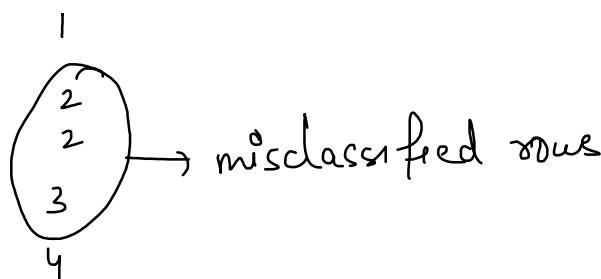
new weight for correctly classified points =  $e^{-\alpha} \times \text{old weight} = e^{0.2} \times 0.2$   
 $= 0.16$

new weight for incorrectly classified points =  $e^{\alpha} \times \text{old weight} = e^{0.2} \times 0.2$   
 $= 0.24$

|   | $x_1$ | $x_2$ | $y$ | $\hat{y}$ | Old weights | New weights              | Normalized weights                    | Range            |
|---|-------|-------|-----|-----------|-------------|--------------------------|---------------------------------------|------------------|
| ① | 3     | 9     | 1   | 1         | 0.2         | 0.16                     | $0.16/0.96 = 0.167$                   | 0 - 0.167        |
| ② | 2     | 4     | 0   | 1         | 0.2         | 0.24                     | $0.24/0.96 = 0.247$                   | 0.167 - 0.417    |
| ③ | 1     | 5     | 1   | 0         | 0.2         | 0.24                     | $0.24/0.96 = 0.25$                    | 0.417 - 0.667    |
| ④ | 9     | 6     | 0   | 0         | 0.2         | 0.16                     | $0.16/0.96 = 0.167$                   | 0.667 - 0.834    |
| ⑤ | 5     | 7     | 0   | 0         | <u>0.2</u>  | <u>0.16</u>              | <u><math>0.16/0.96 = 0.167</math></u> | <u>0.834 - 1</u> |
|   |       |       |     |           | <u>1</u>    | <u><math>0.96</math></u> | <u>1</u>                              |                  |

Randomly choose 5 no. b/w 0 & 1

① ② ③ ④  
 $0.1, 0.2, 0.4, 0.6, 0.8$





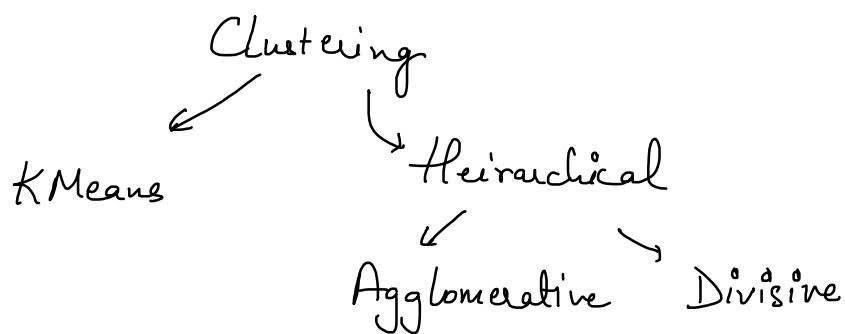
## Clustering

Friday, February 2, 2024 9:04 PM

Clustering → Unsupervised learning  
↓

$D = \{x_i, y_i\}$  → supervised learning      No O/P variable

$D = \{x_i\}$  → unsupervised learning



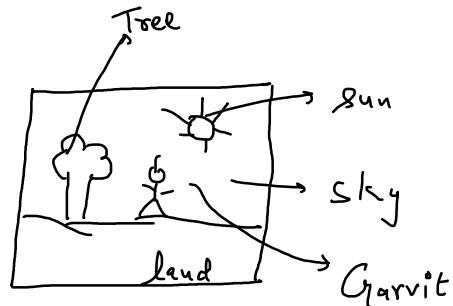
Applications → e-commerce: group customers on the basis of income, gender, locations etc.

2 → Review Analysis:

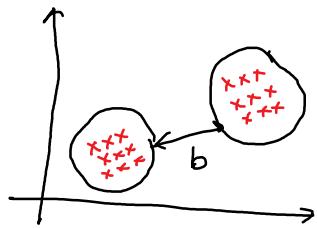
Amazon Reviews:



→ Image Segmentation:



## Metric's

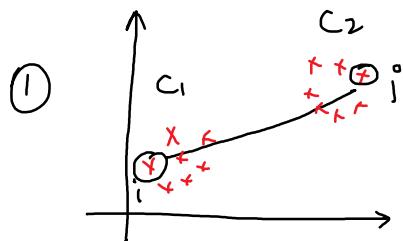


⇒ Intercluster distance (b)

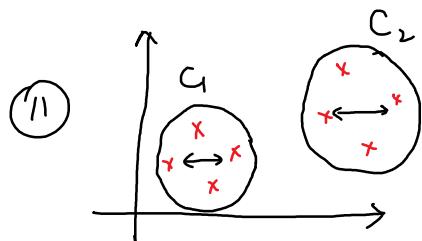
⇒ Intracluster distance (a)

Characteristics : 1 → intracluster distance should be small  
of good cluster      2 → intercluster distance should be large

$$\uparrow \text{Dunn's Index} \downarrow = \frac{\uparrow \max d(i, j)}{\uparrow \max d'(k)} \rightarrow \begin{array}{l} \text{intercluster distance} \\ \text{intracluster distance} \end{array}$$



⇒  $\max d(i, j) \Rightarrow$  distance b/w the farthest points in diff. clusters



⇒  $\max d'(k) \Rightarrow$  distance b/w the farthest points within the cluster

$$\Rightarrow \text{Silhouette's Score} := \frac{b - a}{\max(b, a)}$$

$b \Rightarrow$  avg intercluster distance  
 $a \Rightarrow$  avg intracluster distance

Case 1:       $a \Rightarrow \min \Rightarrow 0$       ,       $b = \max = b$

best result

$$S \cdot S = \frac{b - 0}{\max(b, 0)} = \frac{b}{b} = 1$$

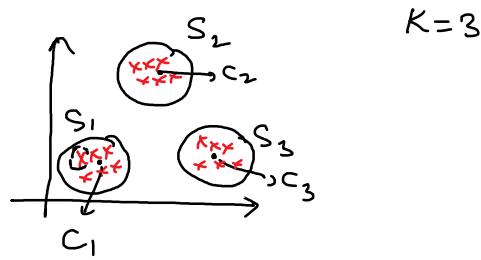
Case 2:  $b < a$ ,  $b=0$ ,  $a=a$

$$S \cdot S = \frac{0 - a}{\max(0, a)} = -\frac{a}{a} = -1$$

Case 3:  $a=b$

$$S \cdot S = \frac{a - a}{\max(a, a)} = 0$$

# clusters  $\leftarrow$  K-Means  $\rightarrow$  Mean (average)  
or  
Centroid



$c_1, c_2, c_3 \Rightarrow$  centroids

$S_1, S_2, S_3 \Rightarrow$  sets

$$S_1 \cap S_2 = \emptyset$$

$$S_2 \cap S_3 = \emptyset$$

$$S_3 \cap S_1 = \emptyset$$

$$C = \frac{1}{n} \sum_{i=1}^n x_i$$

$$x_i \in S_i$$

$$MOF \Rightarrow C^* = \underset{c_1, c_2, c_3, \dots, c_k}{\operatorname{argmin}} \sum_{i=1}^k \sum_{x \in S_i} \|x - c_i\|^2$$

st  $x \in S_i$

$$S_i \cap S_j = \emptyset$$

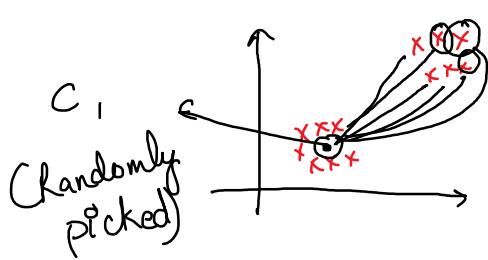
↓  
np hard problem

### Lloyd's Algorithm:

- 1) Randomly choose  $k$  datapoints from datasets & call them centroids.
  - 2) Assignment: For each point, select the nearest centroid with the help of distance & add that point to the corresponding cluster
  - 3) Update: Recalculate the centroids
- $$C_j^* = \frac{1}{|S_j|} \sum_{i=1}^n x_i \quad x_i \in S_j$$
- 4) Repeat Step ② & ③ till convergence.

KMeans++  $\Rightarrow$  KMeans (`init = 'Kmeans++'`)

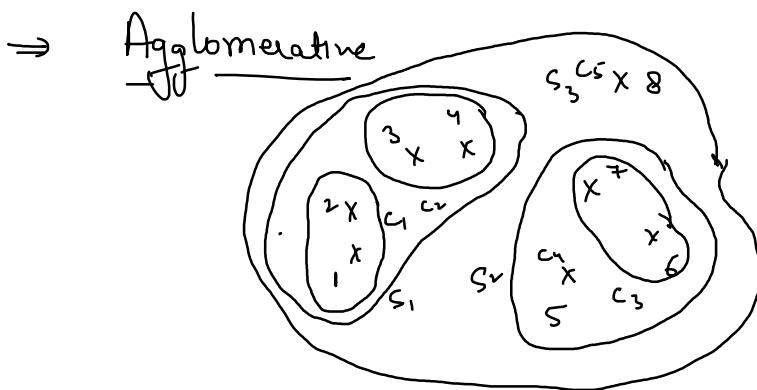
$\uparrow$   $\times$  ~~XX~~ datapoint distance (Not deterministic)



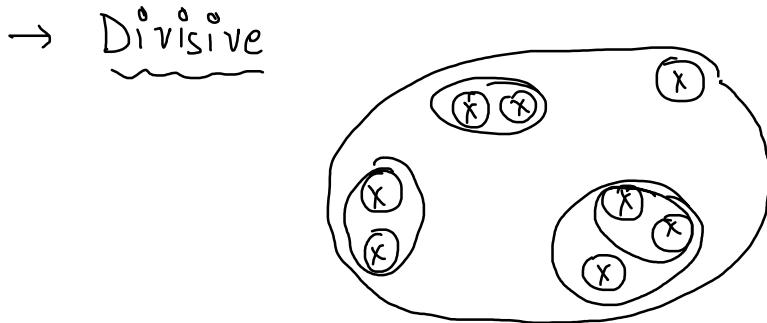
|            |          |          |  |
|------------|----------|----------|--|
| data point | $x_1$    | distance | $d_1$  |
|            | $x_2$    | $d_2$    | probability of being picked                          |
|            | $\vdots$ | $\vdots$ | out as centroid is directly proportional to distance |
|            | $x_n$    | $d_n$    |  |

(Not deterministic)

## Heirarchical Clustering



8 clusters  
↓  
5 clusters ( $c_1 c_2 c_3 c_4 c_5$ )  
↓  
3 clusters ( $s_1 s_2 s_3$ )  
↓  
1 cluster



1 cluster  
↓  
3 clusters  
↓  
5 clusters

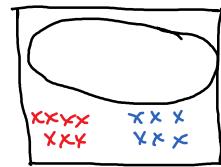
+  
8 clusters

Dendrogram: A tree like structure that merges & splits.

## Curse of Dimensionality

binary features  $\Rightarrow f_1 \ f_2 \ f_3 = \# \text{ datapoints} = 2^3$

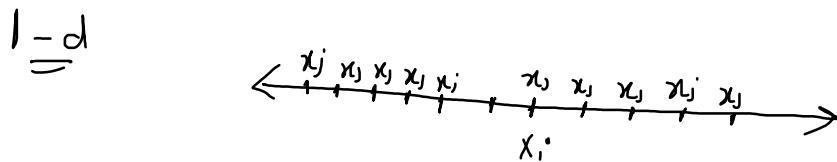
$(10 - d) \ f_1 \ f_2 \ f_3 \dots f_{10} = \# \text{ datapoints} = 2^{10}$



lets say, you have fixed 10k rows, keep on adding dimensions.

$\Rightarrow$  Hughes Phenomenon: as the no of dimensions  $\uparrow$ , the model performance.

### 2) Distance function (Euclidean Distance)



$$\text{dist\_min} = \min d(x_i, x_j)$$

$$\text{dist\_max} = \max d(x_i, x_j)$$

$$\frac{\text{dist\_max} - \text{dist\_min}}{\text{dist\_min}} > 0$$

if  $d \rightarrow \infty$   $\frac{\text{dist\_max} - \text{dist\_min}}{\text{dist\_min}} \rightarrow 0 \Rightarrow \underbrace{\text{dist\_max} \asymp \text{dist\_min}}$   
points are getting equidistant.

NLP  $\rightarrow$  BOW, Word2Vec  $\rightarrow$  a lot of dimensions

NLP  $\rightarrow$  BOW, Word2Vec  $\longrightarrow$  a lot of dimensions

↓  
how do you deal with this?

↓  
cosine similarity / hamming distance

high dim & sparse data

↳ similarity will work in better fashion

3)  $d \uparrow \propto$  overfitting  $\Rightarrow$  Forward Selection Method

feature Extraction

↳ Transform your features

↳ new features

PCA: Principal Component Analysis

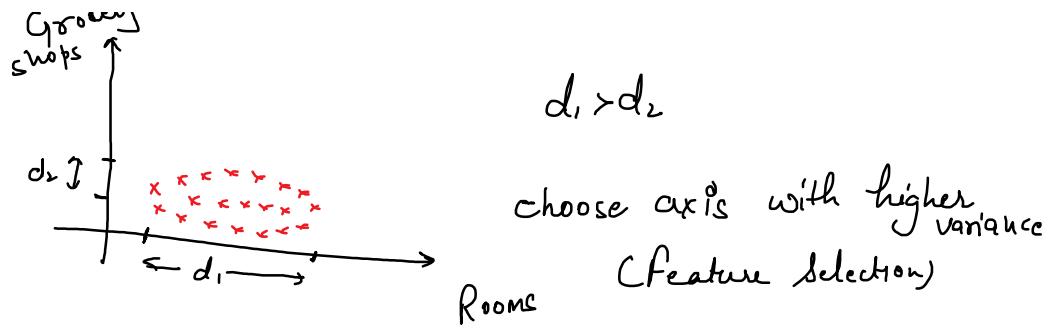
↳ reduces the dimensions to the best possible lowest dimension to capture the essence of data

Basic Intuition:

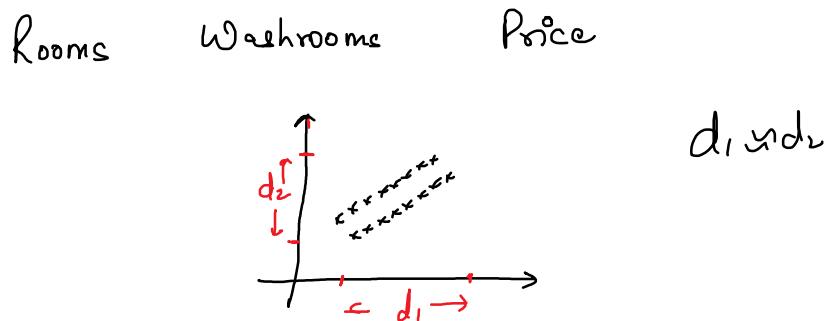
| Rooms | Grocery-shops | Price of flat |
|-------|---------------|---------------|
| 3     | 2             | 60            |
| 4     | 0             | 130           |
| 2     | 6             | 170           |
| 5     | 7             | 90            |

Grocery  
shops ↑

d. > d.



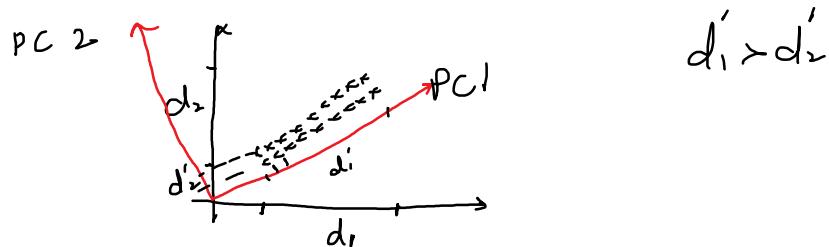
$\Rightarrow$  More variance means more information

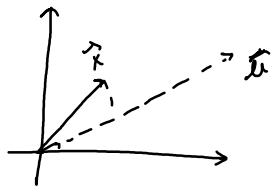


Rooms & washrooms  $\Rightarrow$  size of flat  $\Rightarrow$  Price  
transformation

Feature Extraction  $\Rightarrow$  Create new features from old features & choose a subset of features with higher importances.

### Geometric Intuition of PCA:





$$\text{Projection of } \vec{x} \text{ on } u \Rightarrow \frac{\vec{u} \cdot \vec{x}}{\|\vec{u}\|} = \underbrace{\vec{u} \cdot \vec{x}}_{\perp A}$$

$$u^T x$$

The unit vector with higher variance is chosen as the right axis

$$\text{MDF} = \frac{\sum_{i=1}^n (x_i - \mu)^2}{n} \Rightarrow \frac{\sum_{i=1}^n (u^T x_i - \mu^T u)^2}{n}, \text{ mean}$$

### PCA Steps

1 → Mean centering (Standardization) → not a mandatory step but improves performance

2 → Co-variance Matrix (square and symmetric)

$$\begin{matrix} & f_1 & f_2 & f_3 \\ f_1 & \left[ \begin{array}{ccc} \text{var}(f_1) & \text{cov}(f_2, f_1) & \text{cov}(f_3, f_1) \\ \text{cov}(f_1, f_2) & \text{var}(f_2) & \text{cov}(f_3, f_2) \\ \text{cov}(f_1, f_3) & \text{cov}(f_2, f_3) & \text{var}(f_3) \end{array} \right] & \text{cov}(f_a, f_b) \\ f_2 & & & \text{cov}(f_b, f_a) \end{matrix}$$

3) → Eigen decomposition is applied on above covariance Matrix

$$\begin{matrix} f_1 & f_2 & f_3 \\ \downarrow & \downarrow & \downarrow \end{matrix}$$

$$\begin{matrix} \lambda_1 & \lambda_2 & \lambda_3 \end{matrix} \Rightarrow \text{eigen values}$$

$$\downarrow \quad \downarrow \quad \downarrow$$

$$PC_1 \quad PC_2 \quad PC_3$$

Info Comparison:

$$PC_1 > PC_2 > PC_3$$

If you choose  $\lambda_1$ :  $PC_1$ : 1d dataset

" " "  $\lambda_1 \& \lambda_2$ :  $PC_1 \& PC_2$ : 2d dataset

How to transform from 3d to 1d?

lets say, dataset  $\neq 1000$  rows 3 columns  
shape of unit vector =  $[1, 3]$

$$X \cdot u^T = \left[ \begin{array}{c} \quad \\ \quad \\ \quad \end{array} \right]_{1000 \times 3} \left[ \begin{array}{c} \quad \\ \quad \\ \quad \end{array} \right]_{3 \times 1}$$

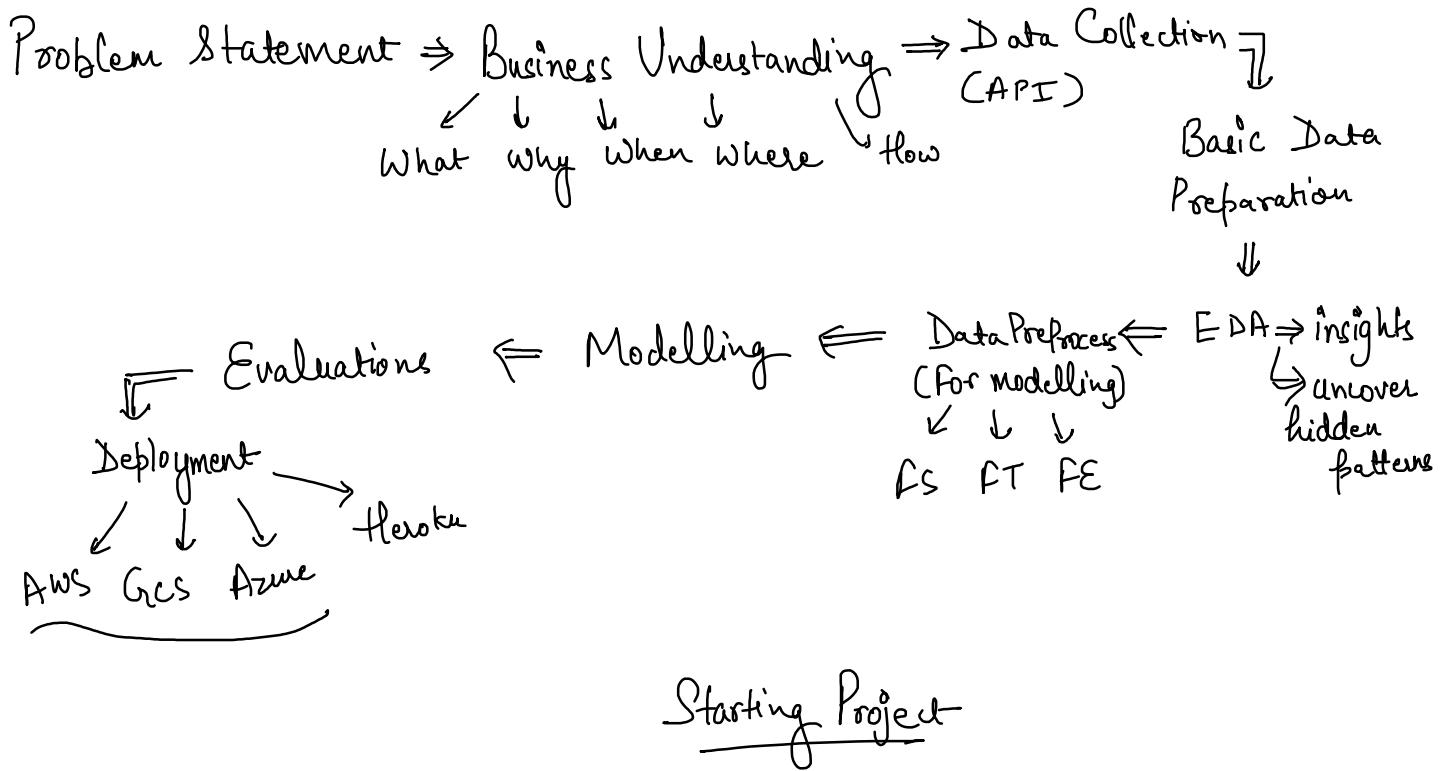
$$= \left[ \begin{array}{c} \quad \\ \quad \\ \quad \end{array} \right]_{1000 \times 1}$$

Assignments

Transform 3d to 2d?

Sunday, February 25, 2024 10:02 AM

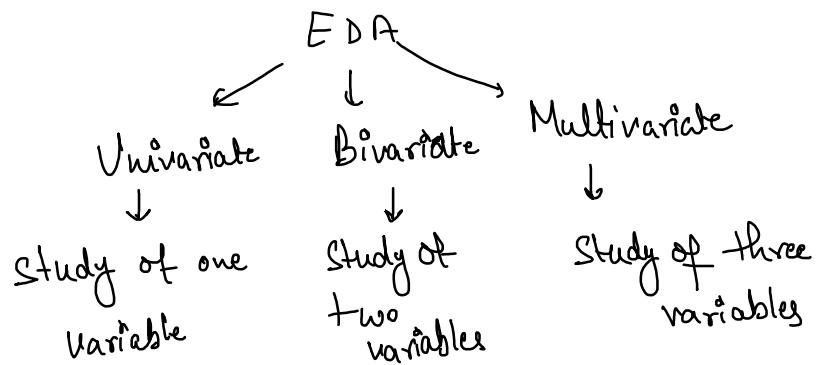
# LifeCycle of Data Science Project



- Import libraries & load the data
- Basic info regarding dataset  $\Rightarrow \text{df.info() \rightarrow datatypes}$   
non-nulls      metadata      rows & columns
- Basic Description of dataset  $\Rightarrow \text{df.describe() \rightarrow count}$   
Quartiles      mean      max      min  
                  |  
                  std dev
- \*  $\text{include} = \text{'all'}$  |  $\text{include} = \text{'object'}$

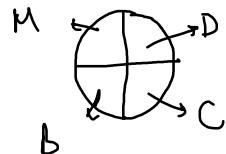
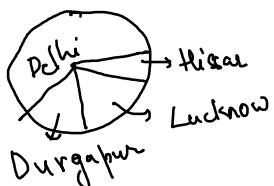
→ Basic Data preparation → data quality check / data assessment

⇒ EDA

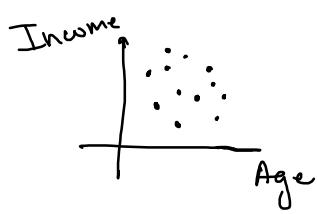


Univariate Analysis ⇒ histogram, kde plot, count plot, distplot, boxplot

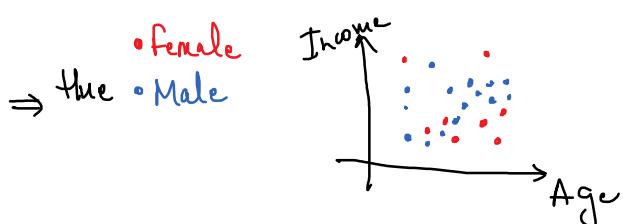
Bivariate Analysis ⇒ scatterplot, LMplot, bar graph, pie graph, line



Multivariate ⇒ heatmap, pairplot. (Bivariate plot ⇒ hue, marker, size), clustered bargraph

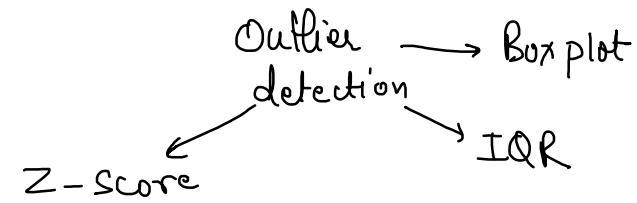


⇒ hue  
• Female  
• Male



Outliers

$$0-10K \quad | \quad 10K-1 lac \quad | \quad 1 lac - 10 lac \quad | \quad 10 lac - 50 lac \quad | \quad 50 lac +$$



$$Z = \frac{x - \mu}{\sigma} \quad | \quad \frac{x - \bar{x}}{\text{std error } (\sigma/\sqrt{n}) \text{ or } (\delta/\sqrt{n})}$$

$IQR = Q_3 - Q_1$

$LL = Q_1 - 1.5IQR$

$UL = Q_3 + 1.5IQR$

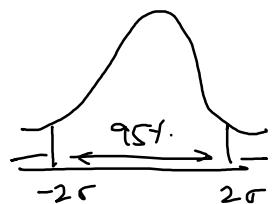
$Z = -3\sigma \quad (approx) \quad Z = +3\sigma \quad (approx)$

$Q_1 \Rightarrow -0.675\sigma$   
 $Q_3 \Rightarrow +0.675\sigma$

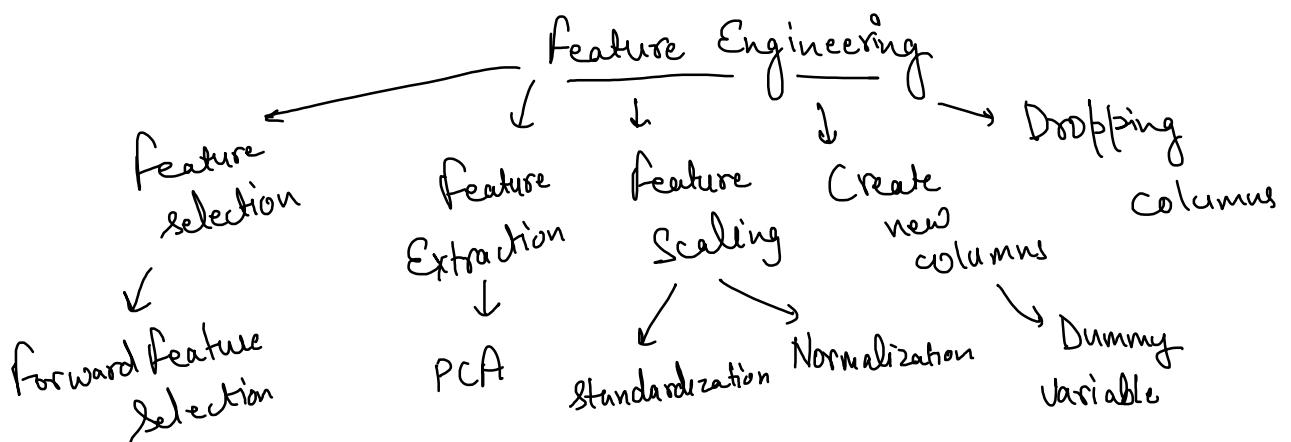
IQR       $LL = Q_1 - 1[Q_3 - Q_1]$

$$= -0.675\sigma - [0.675\sigma - (-0.675\sigma)]$$

$$= -2.025\sigma = -2\sigma$$



$$UL = 2\sigma$$



Encoding  $\Rightarrow$  OHE/dummy variable  $\Rightarrow$  drop\\_first = True

| Category | A | B | C | D | E |
|----------|---|---|---|---|---|
| A        | 1 | 0 | 0 | 0 | 0 |
| B        | 0 | 1 | 0 | 0 | 0 |
| C        | 0 | 0 | 1 | 0 | 0 |
| D        | 0 | 0 | 0 | 1 | 0 |
| E        | 0 | 0 | 0 | 0 | 1 |

Label / ordinal Encoding  $\Rightarrow$  ordered category

O/P  
variable

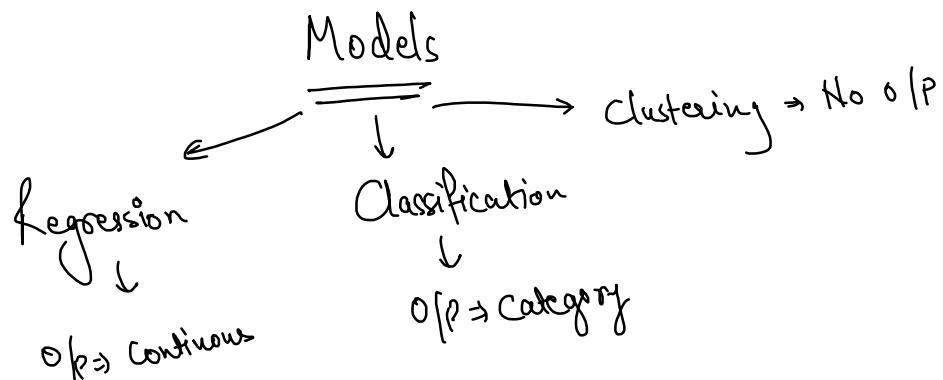
G/P  
variable

Age groups  
 $0 \rightarrow 0 \rightarrow 1$

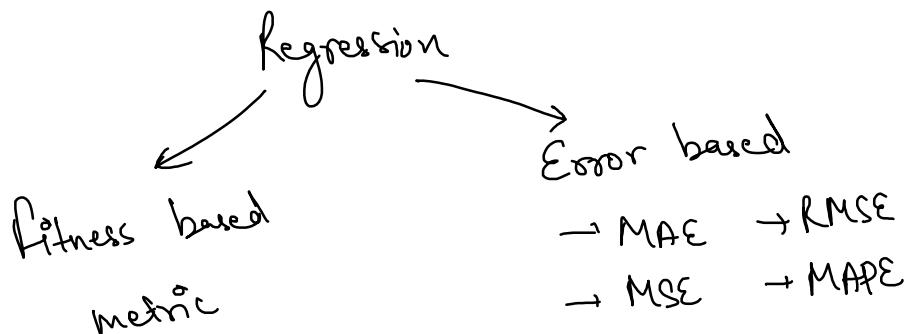
$10 - 20 \rightarrow 2$

$20 - 30 \rightarrow 3$

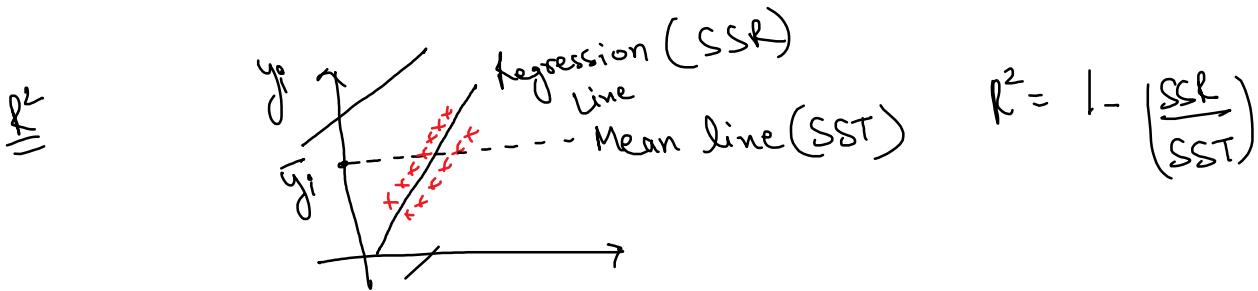
y  
Male  $\rightarrow 0$   
Female  $\rightarrow 1$



Evaluations



$$\rightarrow R^2 \rightarrow \text{Adj } R^2$$



Case 1:  $SSR=0$ ,  $SST=SST$

$$R^2 = 1 - \frac{0}{SST} = 1 - 0 = 1 \quad (\text{overfitting})$$

Case 2:  $SSR=SST$

$$R^2 = 1 - \frac{SST}{SST} = 1 - 1 = 0 \quad (\text{underfitting})$$

Case 3:  $SSR > SST$

$$R^2 = 1 - \frac{SSR}{SST} = 1 - (0(p>1)) = -ve$$

### Classifications

→ Accuracy  $\Rightarrow \frac{TP+TN}{TP+FP+TN+FN}$  → Not reliable: Class Imbalance  
 $P(y)=0.5$

$$\rightarrow \text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}$$

$$\rightarrow \text{Recall} \rightarrow \frac{\text{TP}}{\text{TP} + \text{FN}}$$

$$\rightarrow F1\text{-score} \rightarrow \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

### Indo-Pak

Madan

$\rightarrow$  India  $\rightarrow$  larger in area

Arun

$\rightarrow$  India  $\rightarrow$  better edu.

Neha

$\rightarrow$  India  $\rightarrow$  better economy

$\rightarrow$  India  $\rightarrow$  greater diversity

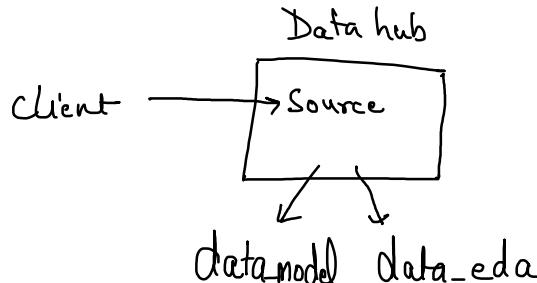
$\rightarrow$  India  $\rightarrow$  safer

$\rightarrow$  India  $\rightarrow$  more developed

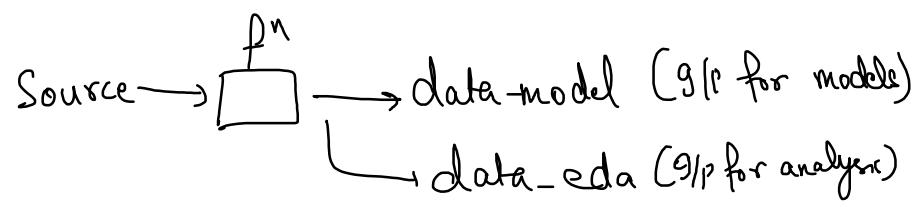
### Project

Air Ticket Price Prediction

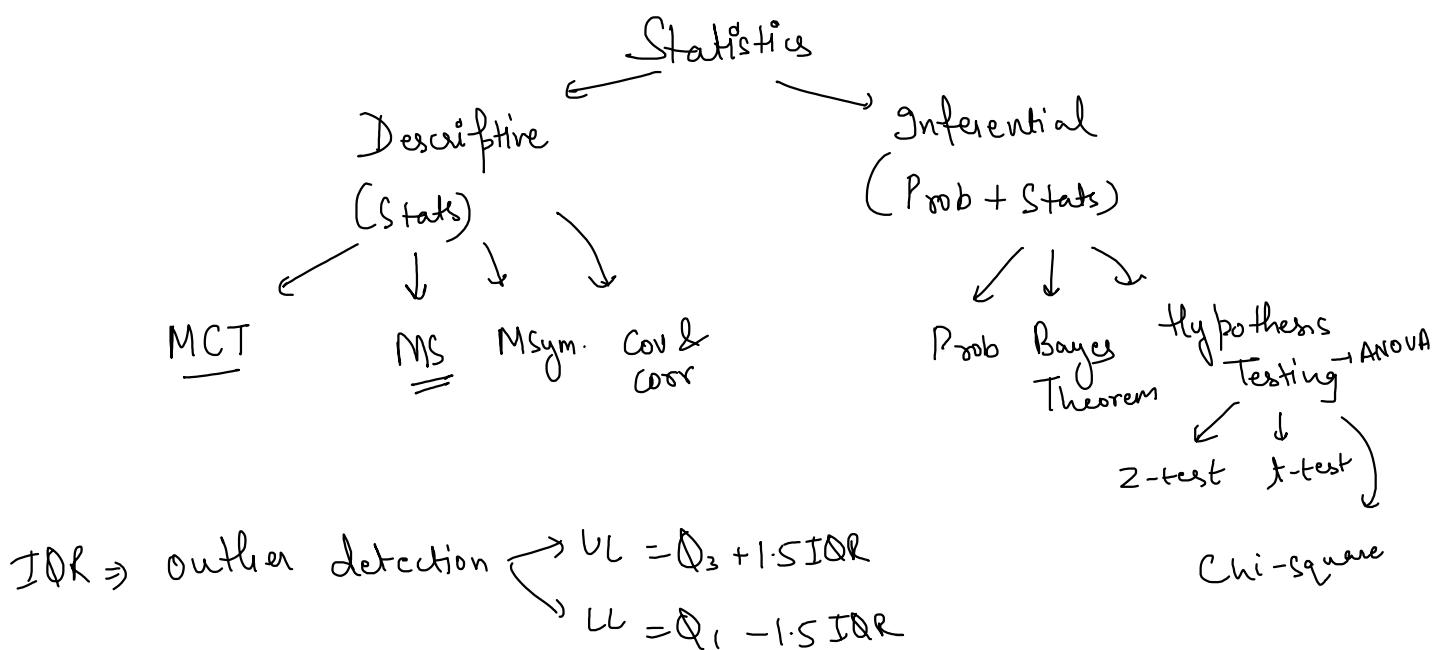
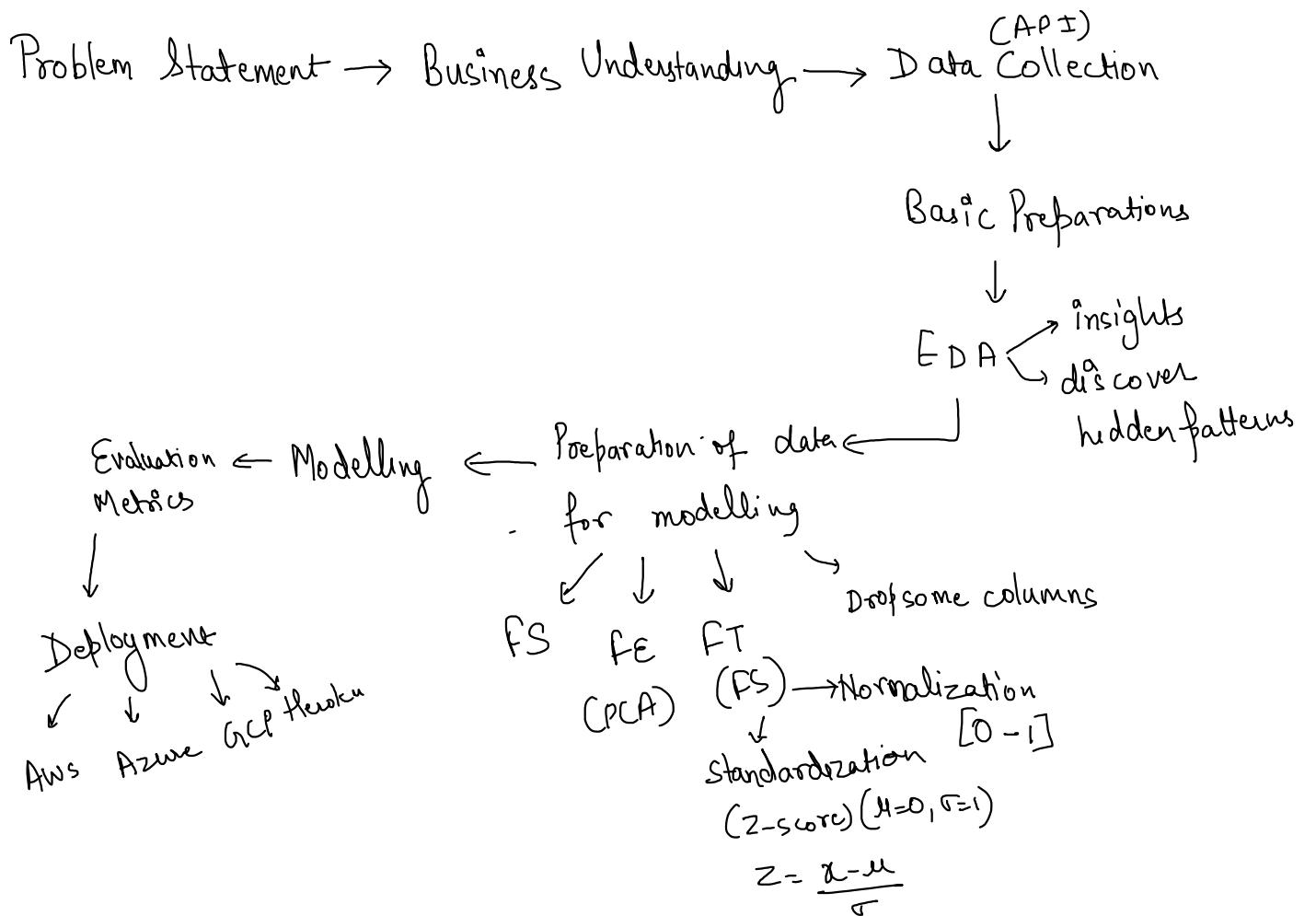
↳ Regression



$\Rightarrow$  a single function to create data-model & data-eda?



# Lifecycle of DataScience Project

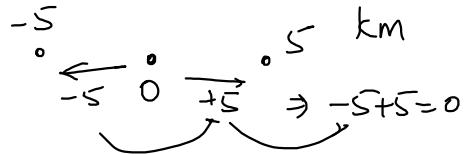




$$IQR = 6 - 2 = 4$$

Variance  $\Rightarrow ?$

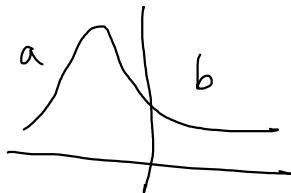
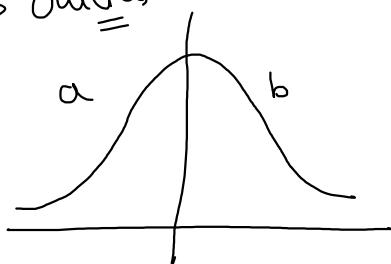
$$\frac{1}{n} \sum (x - \bar{x})^2 \Rightarrow \text{continuous}$$



$$\frac{(-5-0)^2 + (0-0)^2 + (5-0)^2}{3} = \frac{50}{3} \text{ km}^2$$

$$\sqrt{\frac{50}{3} \text{ km}^2} \Rightarrow \text{std dev.}$$

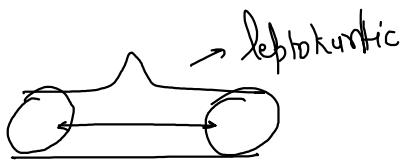
Skewness  $\Rightarrow$  Outliers



Kurtosis

Mesokurtic

Mesokurtic



platykurtic  
PPT, Big Firm Take

Z-score

$q_1 \rightarrow q_0$  in Maths  $\rightarrow 45$

$\rightarrow \sim 10$

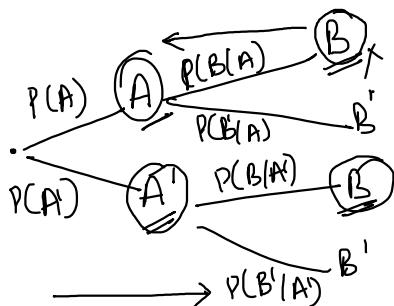
Z-score

$$G \rightarrow 90 \text{ in Maths} \rightarrow 45 \quad D \rightarrow 90 \text{ in English} \rightarrow 80 \quad \Rightarrow \sigma = 10$$

$$Z = \frac{90 - 45}{10} = 4.5$$

$$Z = \frac{90 - 80}{10} = 1$$

"Bayes Theorem"

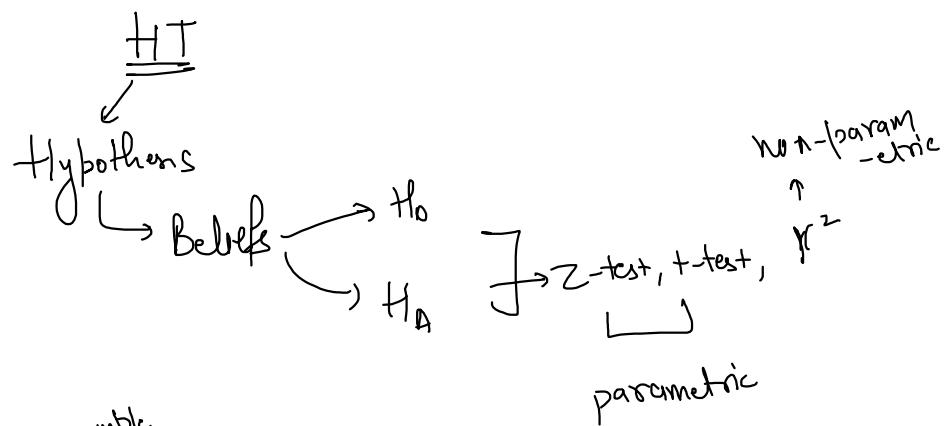


$$P(A|B) \Rightarrow \frac{P(A \cap B)}{P(B)}$$

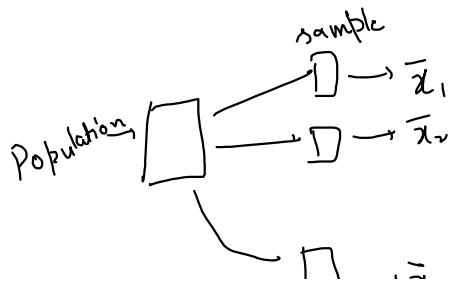
$$P(A|B) \Rightarrow \frac{P(A) \times P(B|A)}{P(B)} \Rightarrow \text{evidence}$$

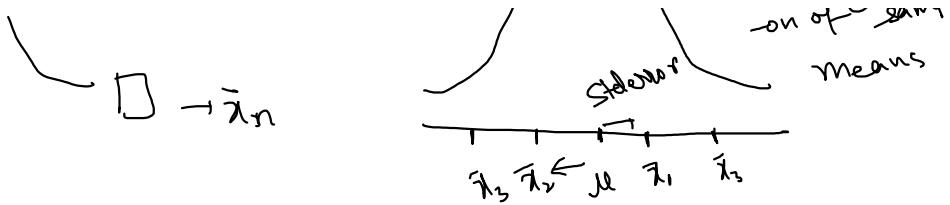
$\hookrightarrow L + P$

$$P(B) = P(A) \times P(B|A) + P(A') \times P(B|A')$$

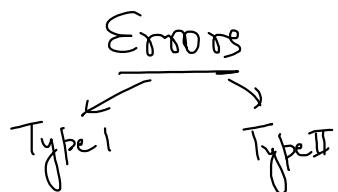
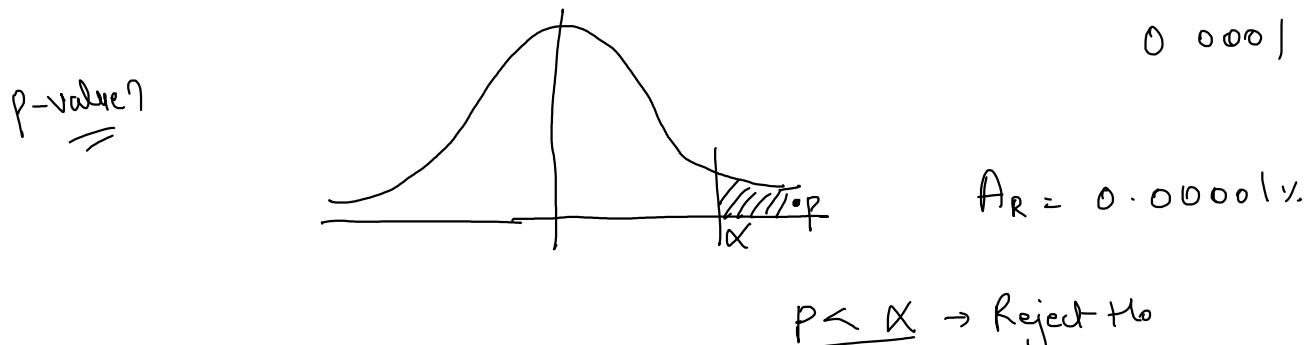


CLT





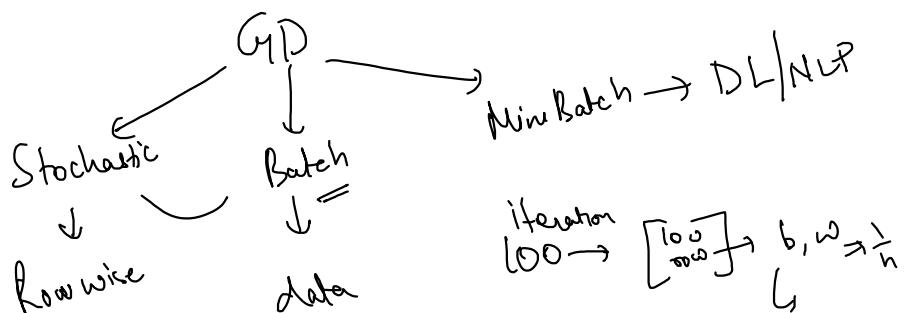
$$\text{Std Err} \Rightarrow \frac{\sigma}{\sqrt{n}}$$



$X \rightarrow \text{Rejection region}$

|            |           | Actual                               |   |
|------------|-----------|--------------------------------------|---|
|            |           | $H_0$ True                           | $H_0$ False   |
| Prediction | $H_0$ Rej | Type I error<br>$\hookrightarrow FP$ | ✓   |
|            | $H_0$ Acc | ✓                                    | Type II<br>$\hookrightarrow \beta = 1 - \text{Power}$ |

Type I  $\alpha$       Type II



100 → 100 rows

①      1 →  $b_1 w$   
           2 →  $b_1 w$   
           3 →  $b_1 w$

100 →  $b_1 w$

100 →  $b_{\text{next}} w_{\text{next}}$

100 (  $b_1 w$  )      100 (  $b_{\text{next}} w_{\text{next}}$  )

→  $100 \times 100 = 10,000$

SG, D Regressor(1).

( $\pm$ ) Covariance & correlation → Strength + directional relationship

↓

directional relationship

↑  $S = \frac{D}{T}$

Corr =  $\frac{\text{Cov}(x, y)}{S_x S_y} = \frac{\sum (x - \bar{x})(y - \bar{y})}{n S_x S_y}$

Pearson

non-linear

Spearmann coeff =  $1 - \frac{6 \sum d^2}{n(n^2 - 1)}$  → diff. in ranks

linear

$\underline{60\% R^2}$

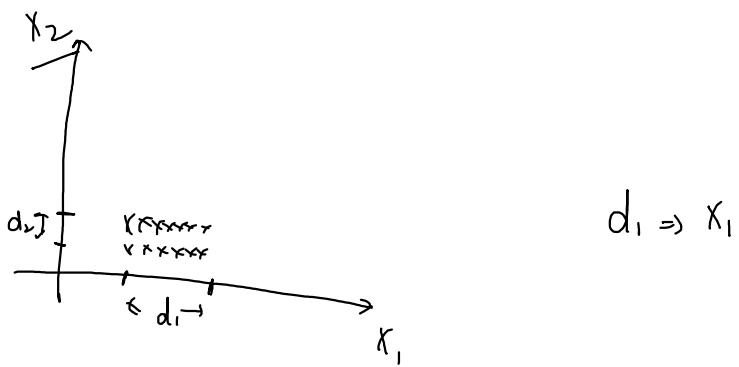
Free 50 + 40%

$X \& Y$  are not linearly related  
 $\downarrow$

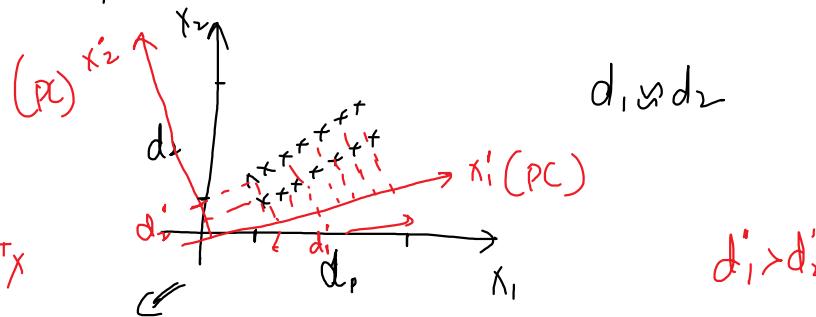
$X \rightarrow$  apply Maths  $\rightarrow X \& Y$  linearly related  
 (log, tri, )

PCA  $\rightarrow$  feature Extraction  $\rightarrow$  Dimensional Reduction

Old features  $\rightarrow$  Transform  $\rightarrow$  New features (PC)



$$d_1 \Rightarrow X_1$$



$$d_1' > d_2'$$

Projection of  $x$  on  $u$

$$= \frac{\bar{u} \cdot \bar{x}}{\|\bar{u}\|} = \bar{u}^T \bar{x}$$



$\bar{u}' (PC)$

$$d_1' > d_2'$$

$$MDF = \sum_{i=1}^n \left( \frac{(x_i - \bar{u})^2}{n} \right) \Rightarrow \sum \frac{(u^T x - u^T \bar{u})^2}{n}$$

Steps

1. Do standardization

2. Covariance Matrix

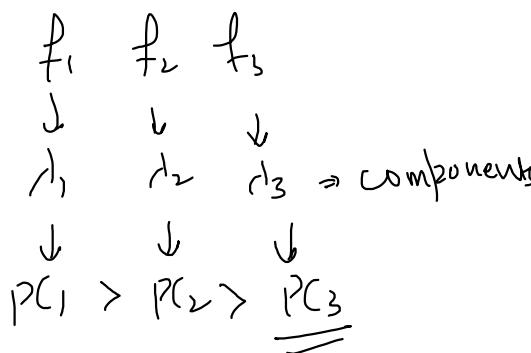
$\rightarrow$  Eigen decomp. C.M



eigen values & eigen vector

$$(\lambda_1, \lambda_2, \lambda_3)$$

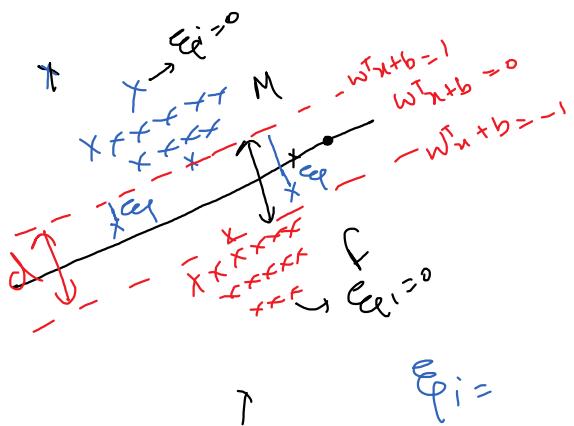
(70%)



PCA (n-components = 6)

## SVM



$$d = \frac{2}{\|w\|}$$

$$\text{MOF} = \underset{w}{\operatorname{argmax}} \left( \frac{2}{\|w\|} \right) \quad \text{margin}$$

$$\text{MOF} = \underset{w}{\operatorname{argmin}} \left( \frac{\|w\|}{2} \right) + C \sum \epsilon_i \rightarrow \text{loss}$$

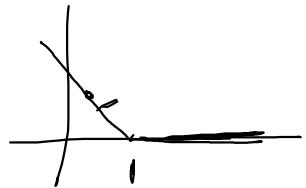
Reg:

$$\rightarrow u^T w$$

hyper

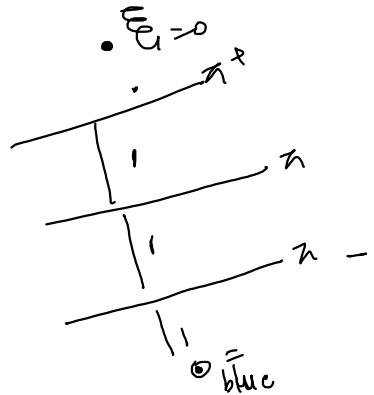
Kern:

$$\text{Hinge loss} = \max(0, \underline{\xi_i})$$



$$= \max(0, 3)$$

$$\underline{\xi_i} = 3$$



Dual form of SVM:

$$(I) \underset{w_0}{\operatorname{argmin}} \frac{\|w\|}{2} + C \sum_{i=1}^n \xi_i =$$

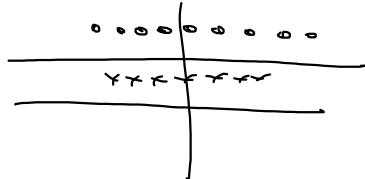
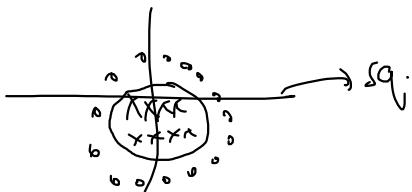
linear [ ] [ ]  
cosine similarity

$$(II) \max_{x_i} \sum_{i=1}^n x_i + \frac{1}{2} \sum_{i=1}^n \sum_{j=1}^n x_i x_j y_i y_j k(x_i^\top x_j) =$$

Kernel  
Trick  
similarity  
or  
Kernel

for  $x_i \rightarrow x_i$

$x_i > 0$  (for support vectors)



$$RBF \rightarrow f \Rightarrow \frac{1}{2\sigma^2} e^{-\left[\frac{\|x_1 - x_2\|^2}{2\sigma^2}\right]} = k(x_1, x_2)$$

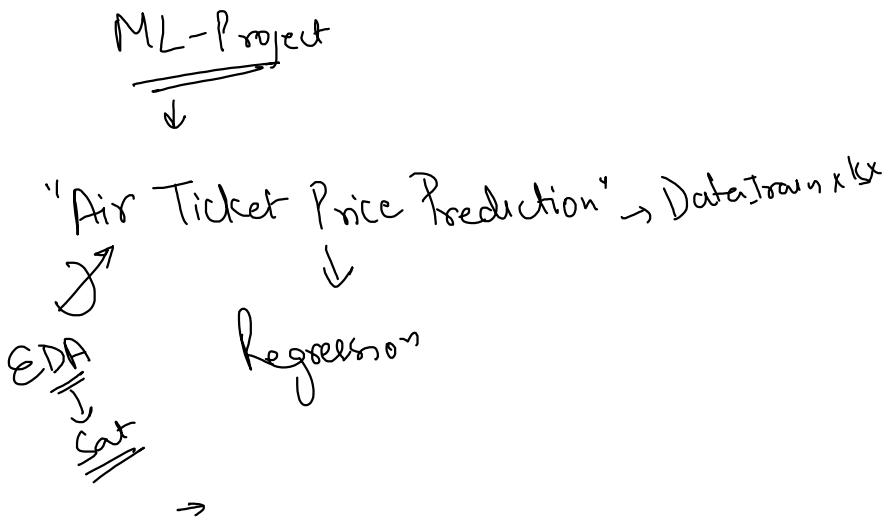
$\bullet \left[ \frac{1}{2\sigma^2} \right] \rightarrow \text{gamma}$

$$\downarrow K(x_1, x_2) \Rightarrow e^{-\frac{d^2}{2\sigma^2}}$$

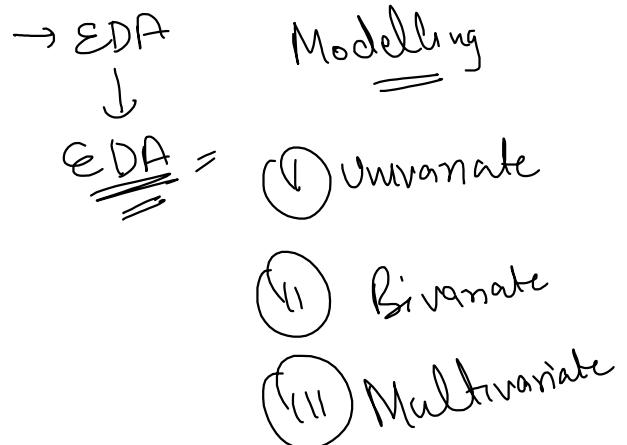
$$[x_1 - x_2] = d \Rightarrow e^{-\frac{d^2}{2\sigma^2}}$$

$$\downarrow K = \frac{1}{e^{d^2/2\sigma^2}} \quad \text{---} \quad \circled{d}$$

$$\sigma^2 \uparrow \rightarrow r^2 \uparrow \Rightarrow e^{d^2/2\sigma^2} \downarrow \Rightarrow \frac{1}{e^{d^2/2\sigma^2}} \uparrow \Rightarrow K \uparrow$$



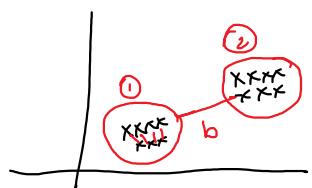
- Read
- Check null
- Unique → value-count



## India - Park

- | F                            | G                                 | S                      |
|------------------------------|-----------------------------------|------------------------|
| - Larger Navy of India       | - higher pop than Pak             | - Bigger area than Pak |
| - Economy of India is better | - Bigger maritime border than Pak | - More airports Park   |

## Clustering



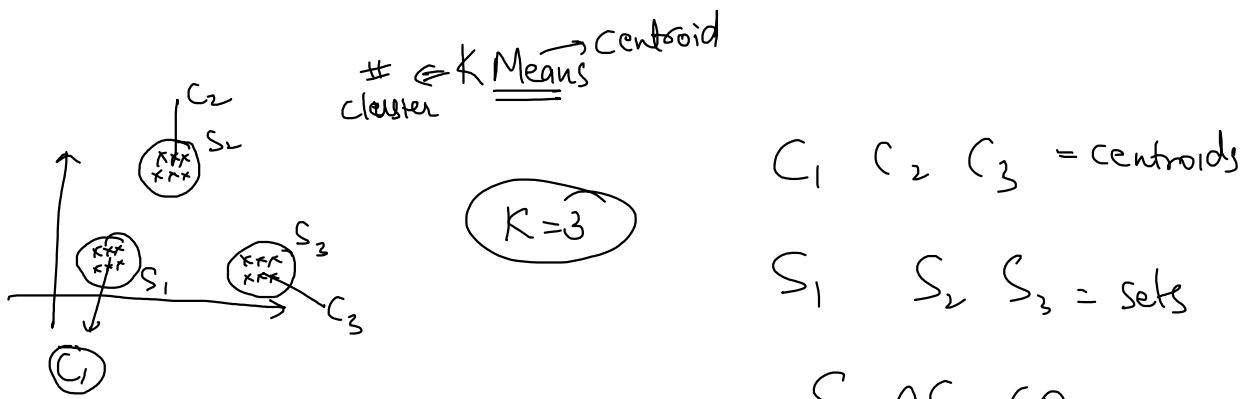
$$[-1, +1]$$

Silhouette Score

$$\Rightarrow \frac{b - a}{\max(b, a)} = \underline{\underline{-ve}}$$

$b \Rightarrow$  avg intercluster distance

$a \Rightarrow$  " intra "



$$C_i = \frac{1}{n} \sum_{i=1}^n x_i \quad x_i \in S_i$$

$C_1, C_2, C_3$  = centroids

$S_1, S_2, S_3$  = sets

$S_1 \cap S_2 \neq \emptyset$

$S_2 \cap S_3 \neq \emptyset$

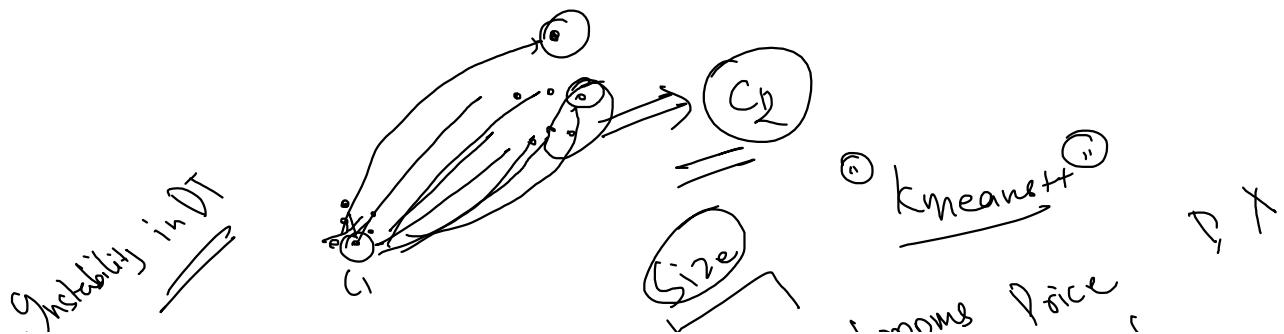
$S_1 \cap S_3 \neq \emptyset$

$$\text{MOF} = C^* = \underset{C_1, C_2, \dots, C_k}{\operatorname{argmin}} \sum_{i=1}^k \sum_{x \in S_i} \|x_i - C_i\|^2 \quad O(2^n)$$

intracluster  
NP hard

Lloyd's Alg. -

- Randomly choose centroids  $\Rightarrow$  Kmeans++
  - Assignments  $\Rightarrow S_1, S_2, S_3$
  - Recalculate Centroids.
- 



Instability

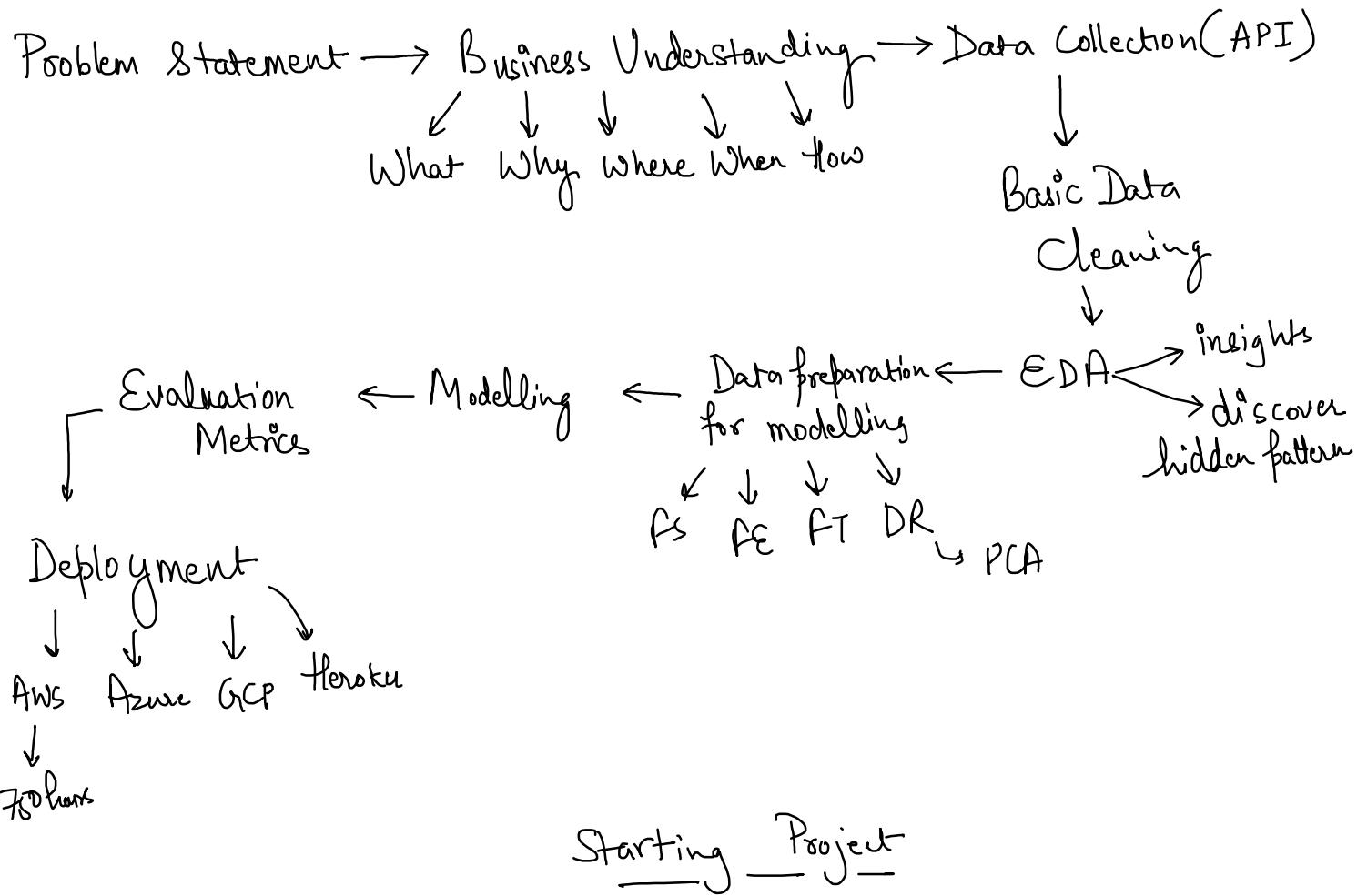
(1)

0

~~Short Party~~

A hand-drawn diagram illustrating factors influencing house price. Three arrows point from the words "Size", "Rooms", and "Washrooms" towards a single horizontal line labeled "Price".

# Lifecycle of Data Science Project



## Starting Project

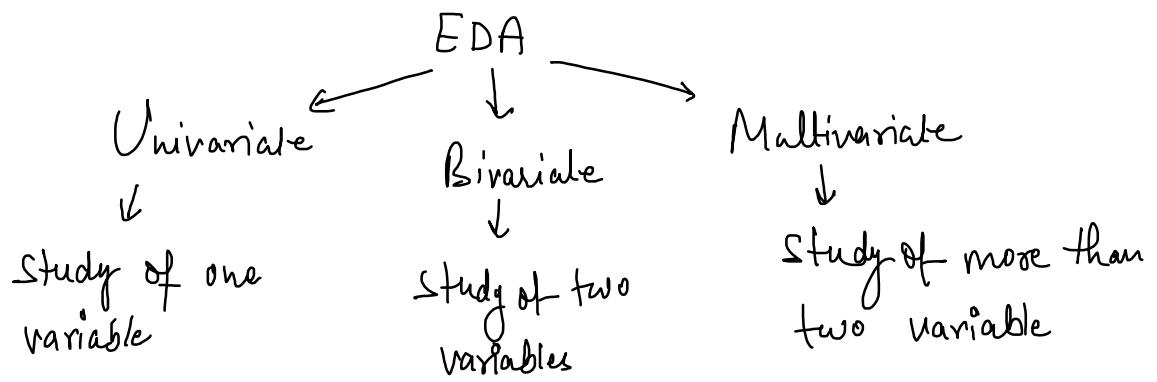
- 1) Import lib & load the data
- 2) Basic information  $\Rightarrow$  `df.info()`  $\rightarrow$  #columns  
 non-nulls    datatype of columns    meta data    #rows
- 3) Basic description  $\Rightarrow$  `df.describe()`  $\rightarrow$  count  
 ...    ...    ...    ...    Quartiles

| Min | Max | Mean | $\sigma$ |
|-----|-----|------|----------|
|-----|-----|------|----------|

# for categorical  $\Rightarrow$  include = 'all'

4) Basic Data Cleaning  $\rightarrow$  data quality check / data assessment

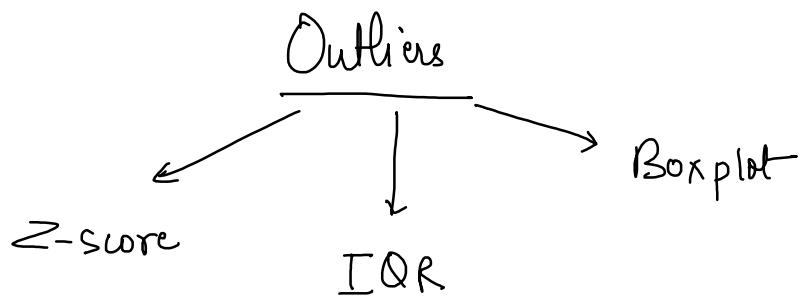
5) EDA



Univariate  $\Rightarrow$  histogram, kdeplot, countplot, box plot

Bivariate  $\Rightarrow$  scatter, bar, pie, line

Multivariate  $\Rightarrow$  heatmap, pairplot



$$z = \frac{x - \mu}{\sigma}$$

$$\frac{-3}{z} < z \\ z > 3$$

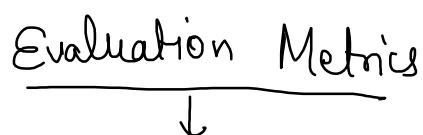
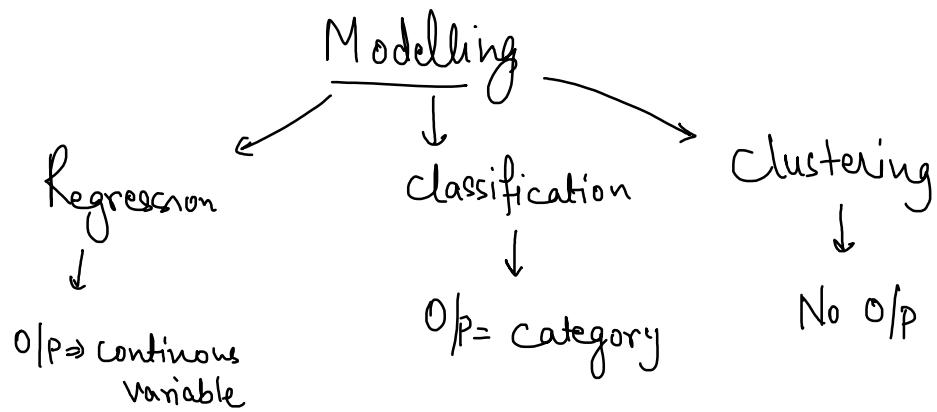
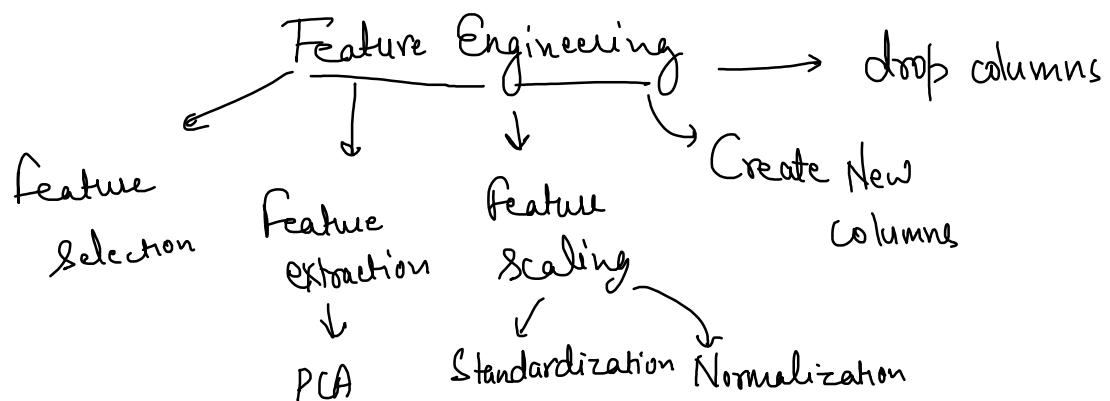
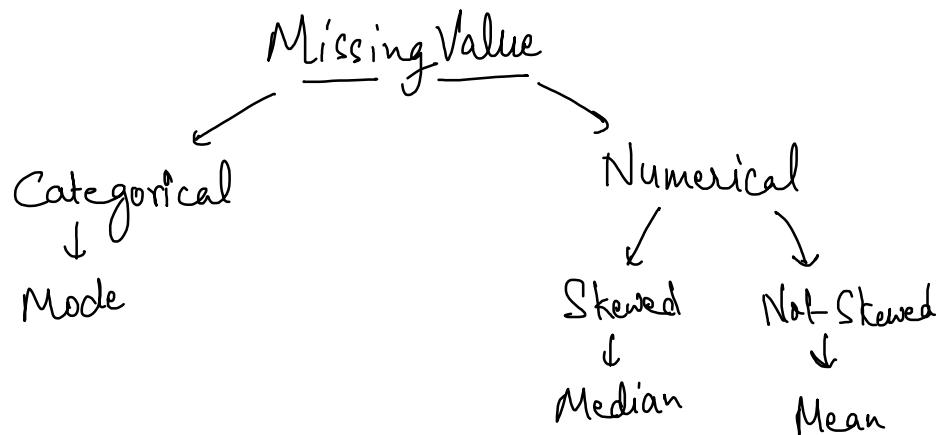
$$IQR = Q_3 - Q_1$$

$$UL = Q_3 + 1.5 \times IQR \approx \underline{3\sigma}$$

$$Q_3 = \frac{0.675\sigma}{\sigma} \\ Q_1 = \frac{-0.675\sigma}{\sigma}$$

$\approx > 3$

$$LL = Q_1 - 1.5 \cdot IQR - 3r$$



## Regression

$$\rightarrow R^2 \Rightarrow 1 - \frac{SSR}{SST}$$

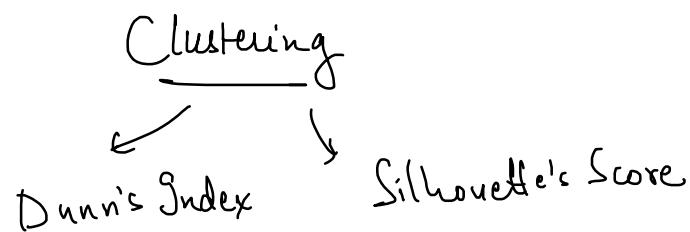
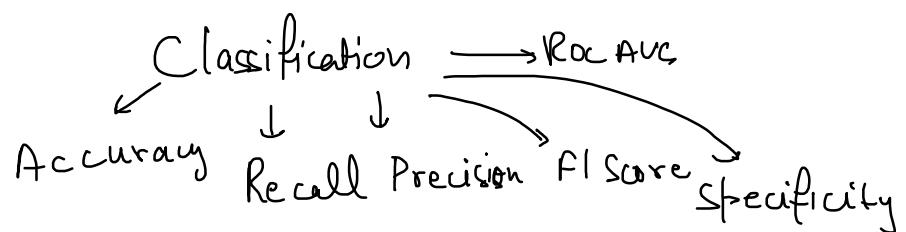
$\rightarrow MAE$

$\rightarrow MAPE$

$$\rightarrow \text{Adj} R^2$$

$\rightarrow MCE$

$\rightarrow RMSE$



## India Park

Jennifer

Sayan

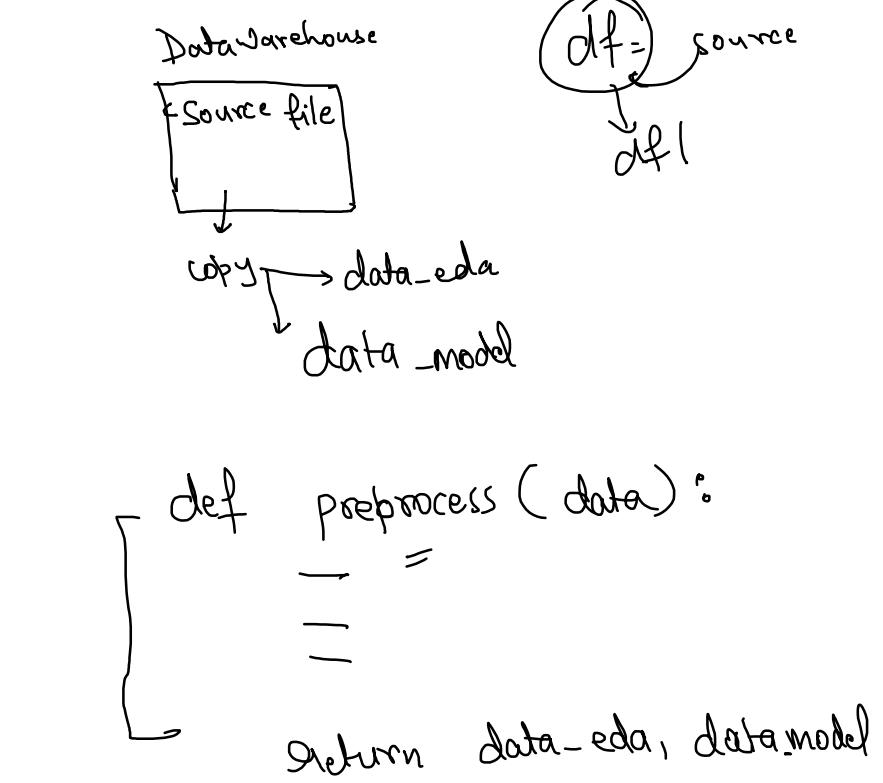
Routh

- Better Military of India
- India - larger population
- India - larger demographies of India
- India - stronger economy
- education  $\Rightarrow$  India
- employment  $\Rightarrow$  India

## Project

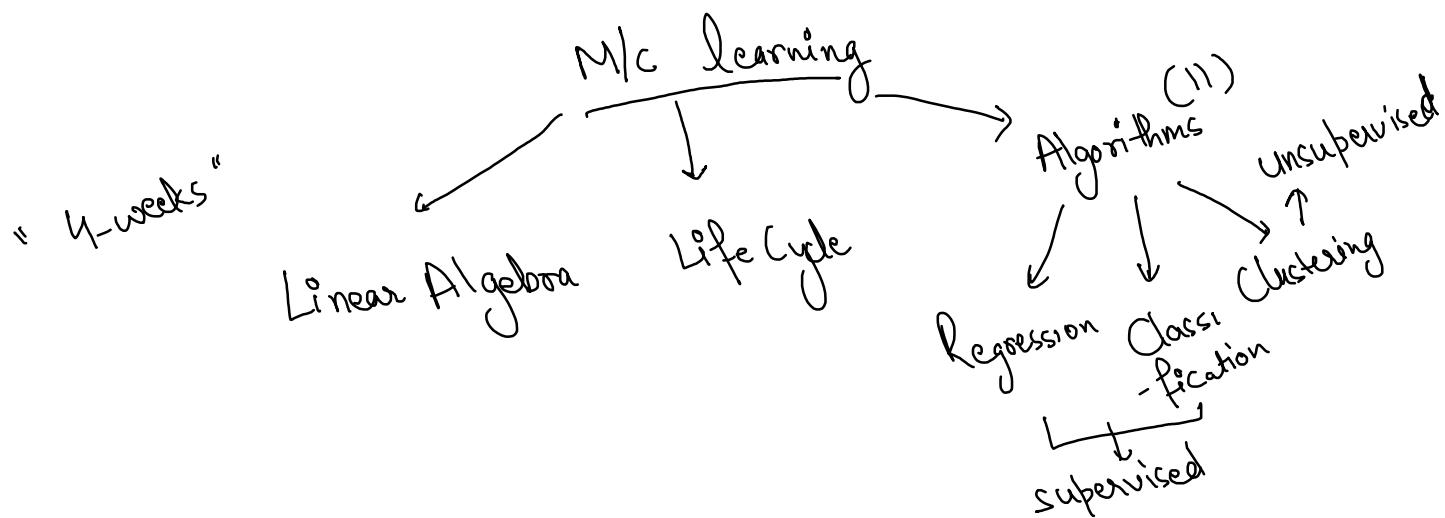
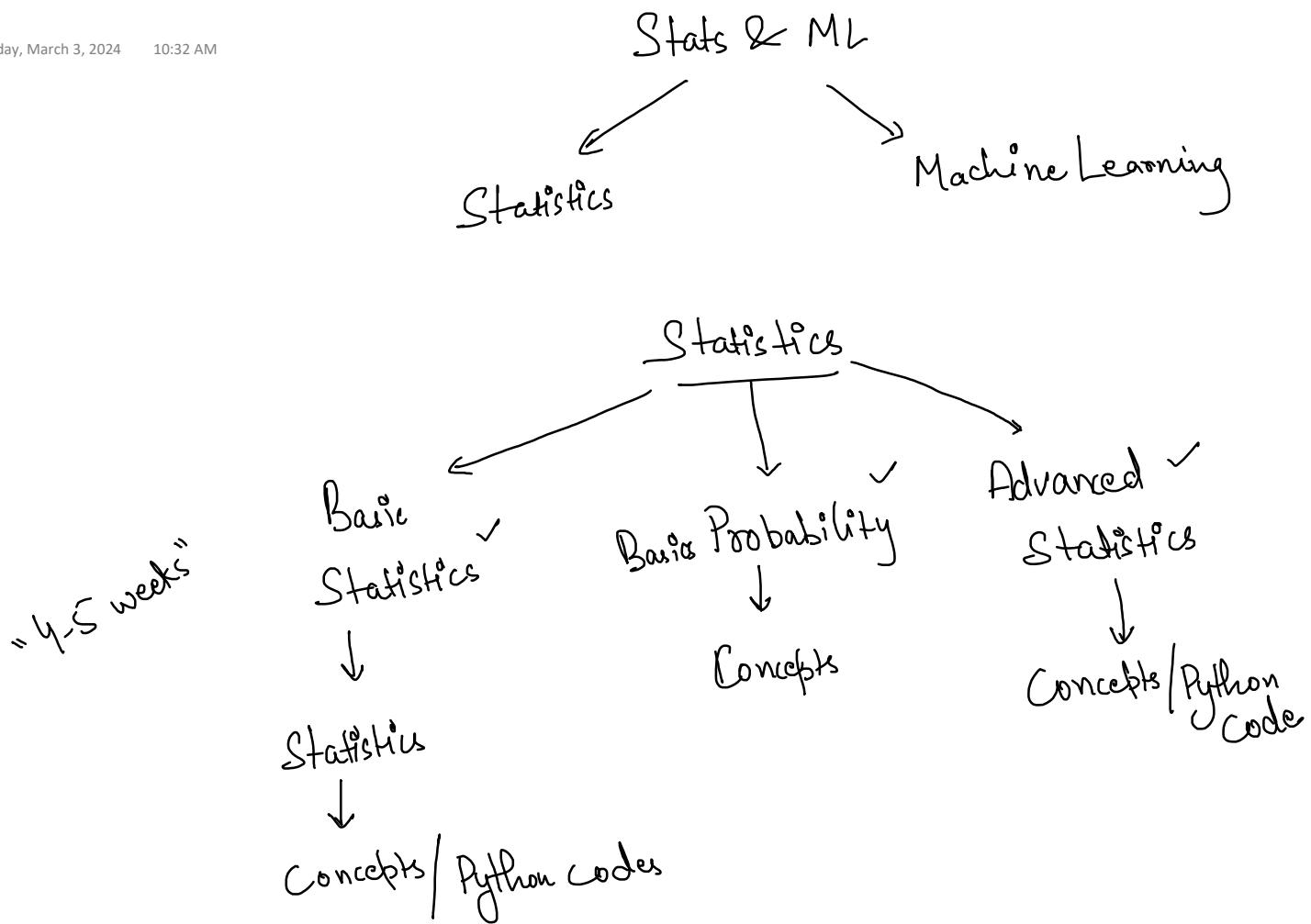
# Project

- Data Transformation
- Feature Engineering
- EDA
  - ↳ univariate
  - bivariate
  - multivariate
- Modelling
- Evaluation



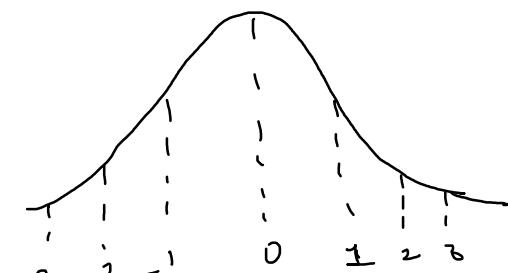
Air Ticket Price  
Prediction → Regression

10:25 am



## Z-score & probability value calculation

$$Z = \frac{x - \mu}{\frac{\sigma}{\sqrt{n}}}$$



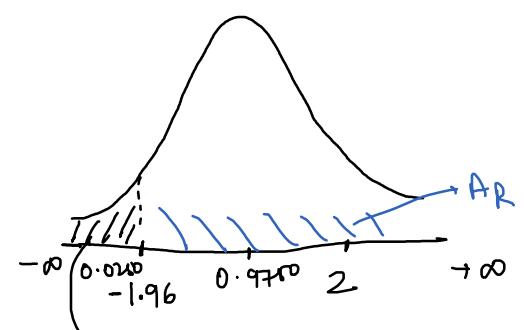
area under curve = probability

Q Z-score, can we relate it with probability or can we get prob. value from z-score?

Sol.

$$\int_{-\infty}^{\infty} \text{Pdf} = \text{Prob.} = 1$$

$$\int_{-\infty}^{-1.96} \text{Pdf} = \int_{-\infty}^{-1.96} \frac{1}{\sqrt{2\pi}} e^{-x^2/2}$$

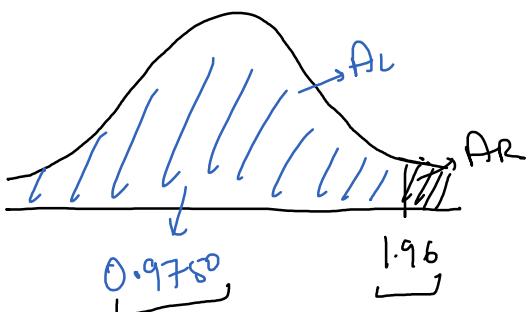


$A_L$  = area to the left  
 $A_R$  = area to the right

\* easier way  $\Rightarrow$  use z-table  $\Rightarrow \text{Prob} = 0.0250$

$$I = A_L + A_R$$

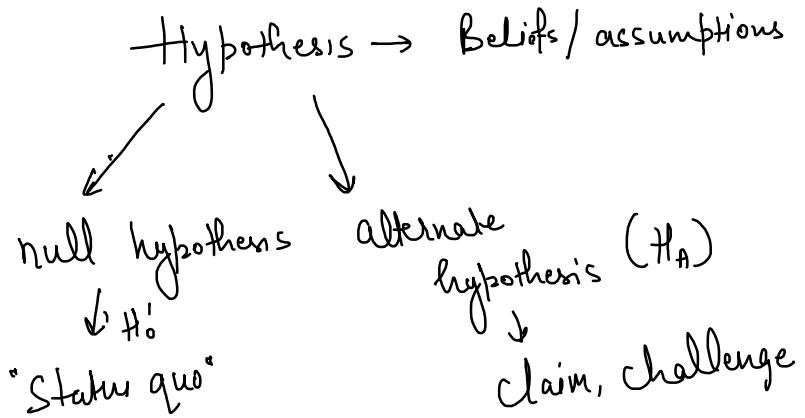
$$A_R = I - A_L$$



$$A_R = 1 - A_L$$

$$= 1 - 0.9750 = 0.0250$$

Hypothesis  
Testing



Q Police claims that a person is criminal?

Sol.  $H_0$ : innocent

$H_A$ : criminal

Q Bride claims groom has taken dowry?

Sol.  $H_0$ : Not guilty

$H_A$ : guilty  
hypothetical

In real dowry,

$H_0$ : Guilty

$H_A$ : Not guilty

Q India will win the world cup.

Sol.  $H_0$ : Any country can win except India

$H_A$ : India wins

Q I claim, that avg salary of teacher changed from 6 lpa?

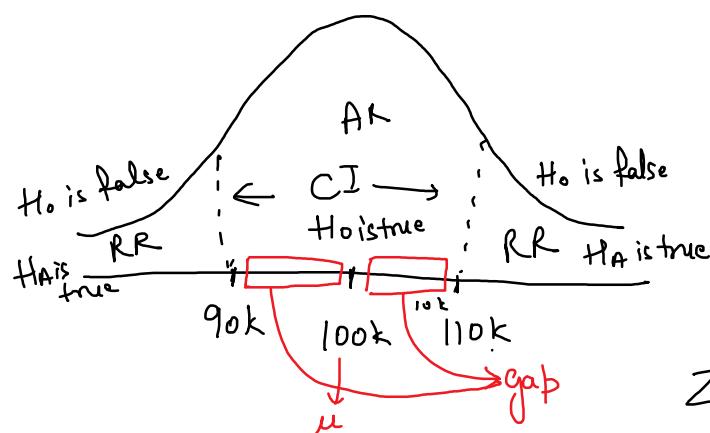
Sol.  $H_0$ :  $\mu = 6$  lpa

$$H_0: \mu = 6 \text{ lpa}$$

Build the criteria to test hypothesis:

1 → Acceptance Region Method:

- Q Data Scientists earn \$100,000 salary on avg  
g claim <sup>avg</sup> salary has changed from \$100,000.



$$H_0: \mu = \$100,000$$

$$H_A: \mu \neq \$100,000$$

$$Z = \frac{x - \mu}{\sigma / \sqrt{n}} \Rightarrow x - \mu = \left( Z \frac{\sigma}{\sqrt{n}} \right)$$

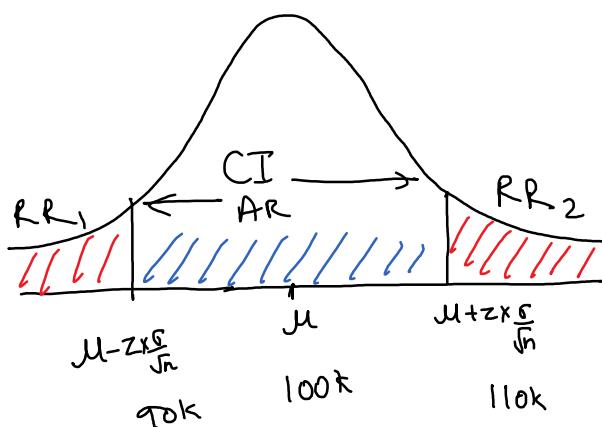
$$UL \Rightarrow \mu + \text{gap} \Rightarrow \mu + z \frac{\sigma}{\sqrt{n}}$$

$$LL \Rightarrow \mu - \text{gap} \Rightarrow \mu - z \frac{\sigma}{\sqrt{n}} \rightarrow \text{Margin of error}$$

CI = Confidence Interval

AR = acceptance region

RR = Rejection Region



$$RR_1 + I(AR) + RR_2 = 1$$

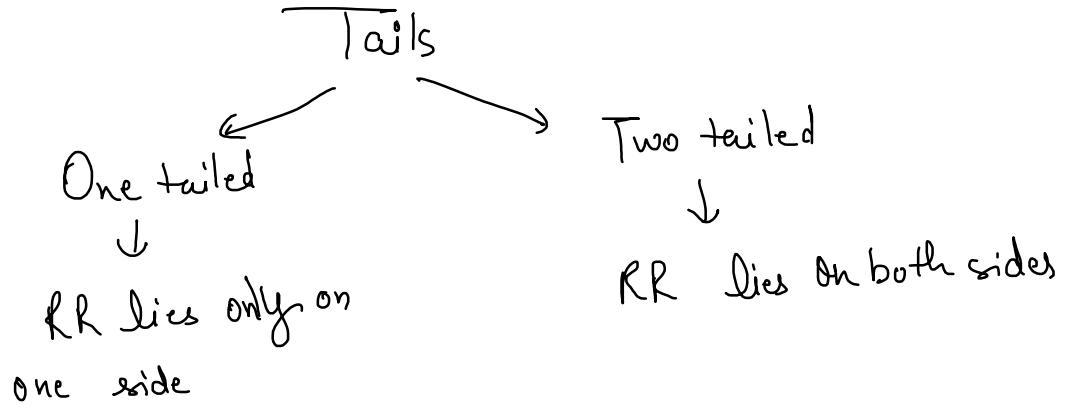
$$\boxed{r\tau + \delta \theta - 1}$$

$$\boxed{RR = RR_1 + RR_2}$$

$$CI + RR = 1$$

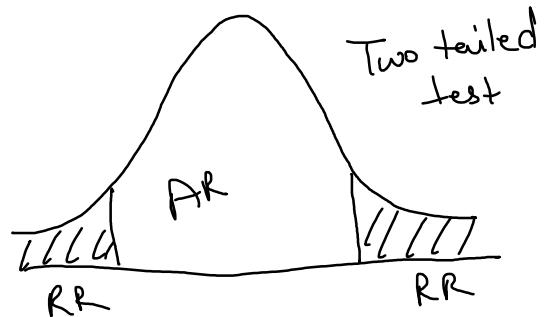
$$RR = RR_1 + RR_2$$

$$CI = 1 - RR$$



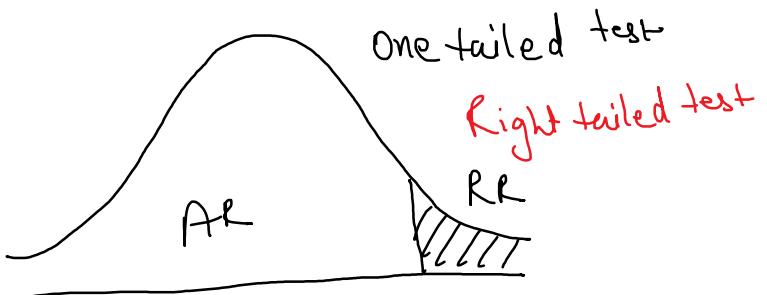
1)  $H_0: \mu = \$100,000$

$H_A: \mu \neq \$100,000$



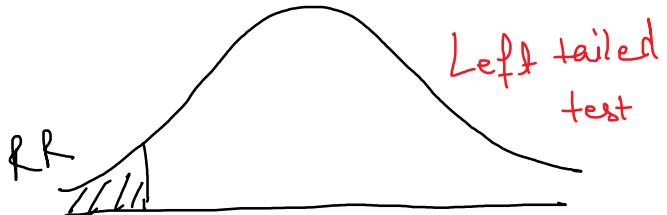
2)  $H_0: \mu \leq \$100,000$

$H_A: \mu > \$100,000$



3)  $H_0: \mu \geq \$100,000$

$H_A: \mu < \$100,000$



## 7 - Critical Value Method

## 2- Critical Value Method

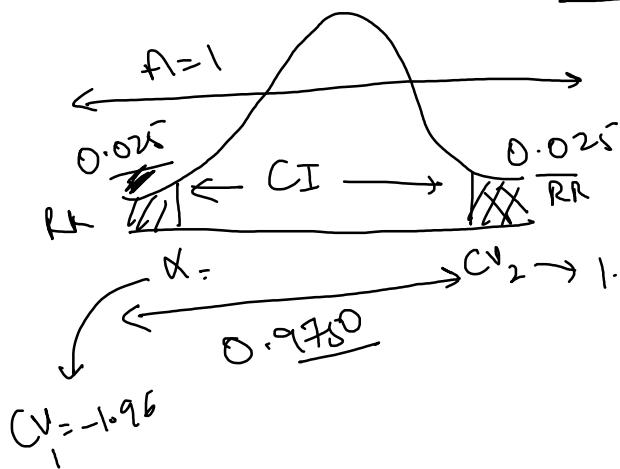


↓  
significance  
level  
( $\alpha$ )

Two-tailed

$\alpha = 0.05$

$$CI + SL = 1$$



$$Z_{cal} = \frac{x - \mu}{\sigma/\sqrt{n}}$$

$$CV_1 < Z_{cal} < CV_2 \Rightarrow H_0 \text{ is Right}$$

$$\Rightarrow Z_{cal} < CV_1 \text{ & } Z_{cal} > CV_2 \\ H_A \text{ is right}$$

3 P-value Method: prob. of null hypothesis to be true

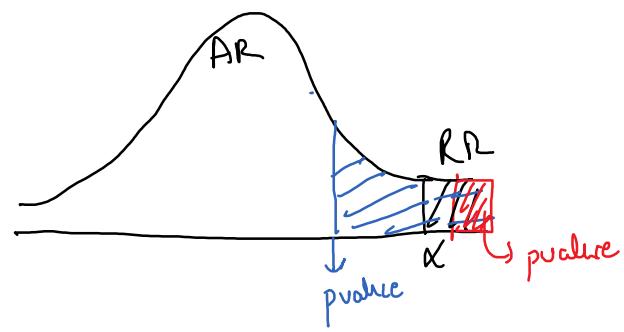
$\alpha$  (significance)

$p\text{value} < \alpha$

$p\text{value} > \alpha$

Accept  $H_0$

Reject  $H_0$



HT

Q A principal of school claims that students have above average IQ. A random sample of  $(30 \text{ students})$  is taken with a mean of 112.5. The mean & std dev of population is 100 & 15. Test your hypothesis.

Sol. ①  $H_0: \mu \leq 100$

$$H_A: \mu > 100$$

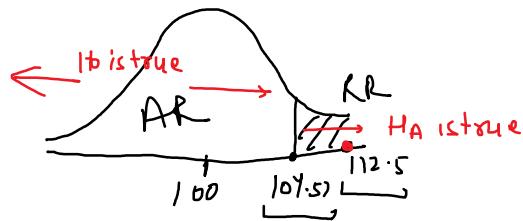


②  $\mu = 100, \sigma = 15, \bar{x} = 112.5$

$$UL = \mu + z \times \frac{\sigma}{\sqrt{n}}$$

$$= 100 + 1.65 \times \frac{15}{\sqrt{30}}$$

$$= 104.51$$



$$CI = 0.95$$

$$\alpha = 0.05$$

$$LL = 100 - 1.65 \times \frac{15}{\sqrt{30}}$$

$$Z = 1.65$$

$$= 95.15$$

$H_A$  is true

$$|104.51 - 112.5| < 1.65$$

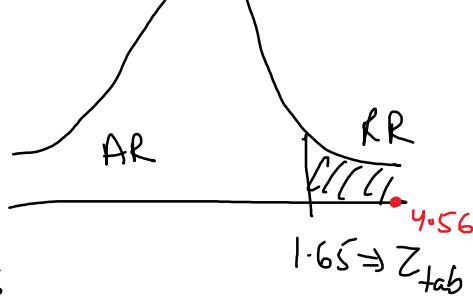
Reject  $H_0$ .

Q2 Critical value

$$Z_{cal} = \frac{\bar{x} - \mu}{\sigma / \sqrt{n}}$$

$$= \frac{112.5 - 100}{15 / \sqrt{30}} = 4.56$$

$$\alpha = 0.05$$



$$15/\sqrt{30} \quad 15/\sqrt{30}$$

Reject  $H_0$

3) P-value  $X = 0.05, z_{\text{cal}} = 4.56$

$$P(z_{\text{cal}} = 4.56) = 0.9999966 \quad ( \text{Graph of } AR )$$

$$\text{P-value} = AR = 1 - 0.9999966 = 0.0000034$$

$$\text{P-value} < X$$

Reject  $H_0$ .

Homework  $\Rightarrow$  same but principal claim is students have below avg IQ.  $\bar{x} = 90$

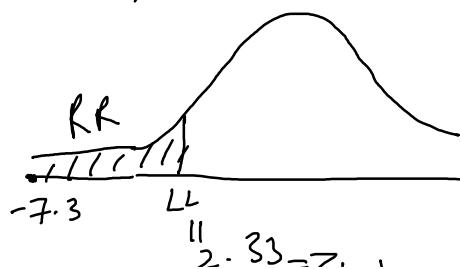
Q A company manufactures car batteries with avg life span of 2 years or more. An engineer believes this value to be less. Using 30 batteries, he measured the life span & found it to be 1.8 years with a std dev of 0.15. At 99%. CI, is there enough evidence to reject  $H_0$ ?

Sol.

$$H_A: \mu < 2$$

$$H_0: \mu \geq 2$$

$$X = 0.01 \Rightarrow Z_{\text{tab}}$$



$$\begin{aligned} CI &= 99 \\ X &= 1 - 0.99 \\ &= 0.01 \\ Z &= -2.29 \end{aligned}$$

$$\bar{X} = 0.01 \Rightarrow Z_{tab}$$

$$-7.3 - \frac{11}{2.33} = Z_{tab} = Z = -2.33 = 0.6)$$

①

$$Z_{cal} = \frac{1.8 - 2}{\frac{0.15}{\sqrt{30}}} = \frac{-0.2}{0.15/\sqrt{30}} = -7.30$$

(0.0099)

Reject  $H_0$ .

$$b) \underline{n=10} \rightarrow t\text{-distribution} \quad df = n-1 = 10-1 = 9$$

$$t = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}} \text{ or } \frac{s}{\sqrt{n}}}$$

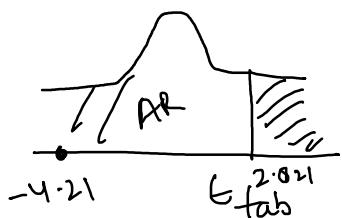
$df \rightarrow$  degrees of freedom  
 $df = n-1$   
 $\hookrightarrow$  logically independent values

$$t_{tab} = 2.821 \quad (0.01)$$

$$t_{cal} = \frac{1.8 - 2}{\frac{0.15}{\sqrt{30}}} = -4.21$$

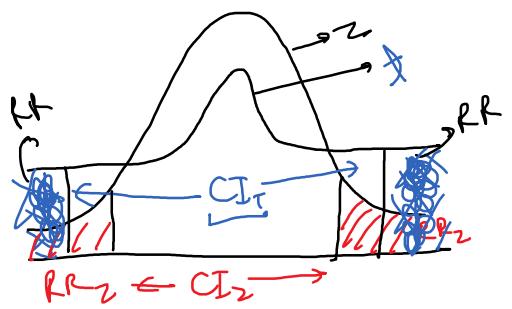
$$n=5 \quad \left\{ \begin{array}{c} 2 \\ 3 \\ 5 \\ 8 \\ x \\ 17 \end{array} \right. \quad \text{avg} = 7$$

$$df = n-1 = 4$$



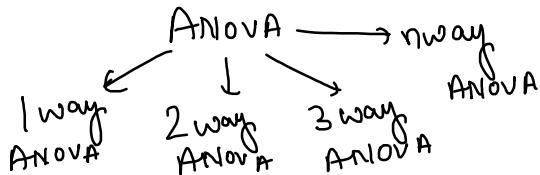
$$t_{tab} > t_{cal}$$

accept  $H_0$



$$O.O.S = R_{h_2} + t R_2$$

ANOVA:



One Way ANOVA

Q1: To assess the significance of possible variation in performance in a certain test between the convent schools of a city, a common test was given to a number of students taken at random from the 5<sup>th</sup> class of the 3 schools concerned. The result is given as follows:

| A  | B  | C  |
|----|----|----|
| 9  | 13 | 14 |
| 11 | 12 | 13 |
| 13 | 10 | 17 |
| 9  | 15 | 7  |
| 8  | 5  | 9  |

Make the Analysis of Variance of the given data. (Null Hypo: No Significance Variation in the schools).

Solution:

Null Hypothesis = No variation between schools  
Alt. Hypothesis = There is variation between schools

| Source of Variation | Sum of Square | Degrees of freedom                   | Mean Square                            | F           |
|---------------------|---------------|--------------------------------------|--|-------------|
| Between the Sample  | SSC           | (c-1) = Df <sub>f</sub> <sub>1</sub> | MSC = SSC/df <sub>f</sub> <sub>1</sub> | F = MSC/MSE |
| Within the sample   | SSE           | (n-c) = Df <sub>e</sub>              | MSE = SSE/df <sub>e</sub>              |             |

ANOVA  
+ Analysis of Variance

→ it is a extension of Z-test & t-test

→ F-statistics

→ It is used to compare variance among groups

$$\rightarrow f = \frac{SD_1^2}{SD_2^2} \Rightarrow SD_1 > SD_2$$

$$\bar{x}_A = 10 \\ \bar{x}_B = 11 \\ \bar{x}_C = 12$$

$$\bar{x} = \frac{\bar{x}_A + \bar{x}_B + \bar{x}_C}{3} = 11$$

SSC

$$\begin{array}{cccccc}
 \bar{x}_A - \bar{x} & (\bar{x}_A - \bar{x})^2 & (\bar{x}_B - \bar{x}) & (\bar{x}_B - \bar{x})^2 & (\bar{x}_C - \bar{x}) & (\bar{x}_C - \bar{x})^2 \\
 |0 - 11| = 1 & 1 & |11 - 11| = 0 & 0 & |12 - 11| = 1 & 1 \\
 |0 - 11| = 1 & 1 & |11 - 11| = 0 & 0 & |12 - 11| = 1 & 1 \\
 |0 - 11| = 1 & 1 & |11 - 11| = 0 & 0 & |12 - 11| = 1 & 1 \\
 |0 - 11| = 1 & 1 & |11 - 11| = 0 & 0 & |12 - 11| = 1 & 1 \\
 |0 - 11| = 1 & 1 & |11 - 11| = 0 & 0 & |12 - 11| = 1 & 1 \\
 \hline
 & \sum & & \sum & & \sum = 10
 \end{array}$$

$$SSC = 10 + 0 + 0 = 10 \quad | \quad MSC = \frac{10}{Df(c-1)} = \frac{10}{2} = 5$$

SSE

SSE

$$\begin{array}{cccccc}
 & (A - \bar{x}_A)^2 & (B - \bar{x}_B)^2 & (C - \bar{x}_C)^2 \\
 (A - \bar{x}_A) & 1 & 13 - 11 = 2 & 4 & 14 - 12 = 2 & 4 \\
 9 - 10 = -1 & & 12 - 11 = 1 & 1 & 13 - 12 = 1 & 1 \\
 11 - 10 = 1 & & & 1 & 17 - 12 = 5 & 25 \\
 13 - 10 = 3 & 9 & 10 - 11 = -1 & 16 & 7 - 12 = -5 & 25 \\
 9 - 10 = -1 & 1 & 15 - 11 = 4 & 36 & 9 - 12 = -3 & 9 \\
 8 - 10 = -2 & 4 & 5 - 11 = -6 & \hline & & 64 \\
 & \hline 16 & & 58 & & 
 \end{array}$$

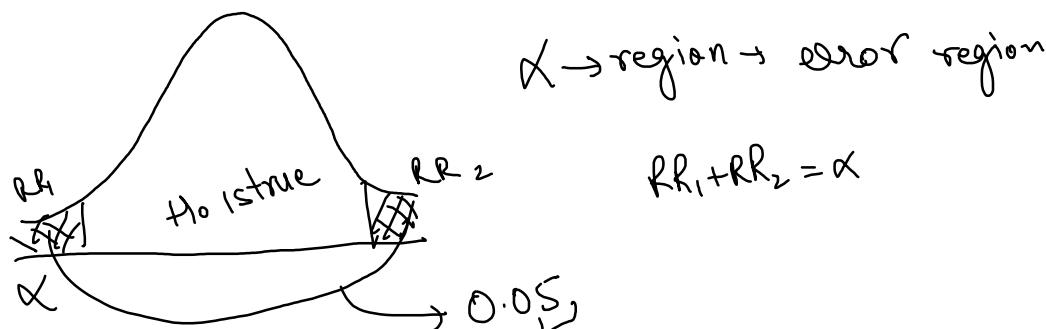
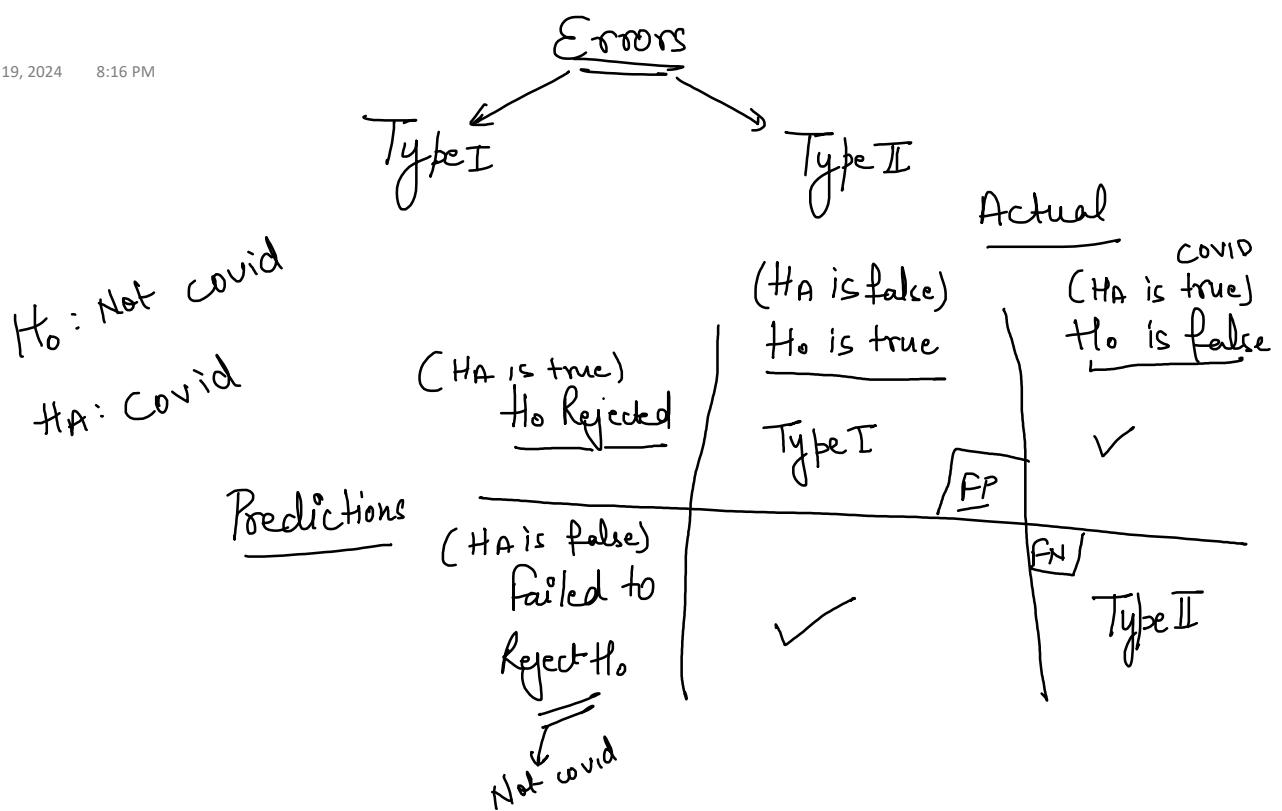
$$SSE = 16 + 58 + 64 = 138$$

$$\begin{aligned}
 df_2 &= n - c \\
 &= 15 - 3 = 12
 \end{aligned}$$

$$MSE = \frac{138}{12} = 11.5$$

$$\left. \begin{array}{l}
 f_{cal} = \frac{MSE}{MSCE} = \frac{5}{11.5} = 0.43 \\
 DF_1 = 2 = j_1 \quad \Rightarrow f_{tab} = 3.89 \\
 DF_2 = 12 = j_2
 \end{array} \right\}$$

$\therefore f_{tab} > f_{cal} \therefore f_{cal}$  failed to reject  $H_0$



Quantify  $\rightarrow$  Type I  $\rightarrow \alpha \rightarrow$  Above ex: 5% chance that  $H_0$  is rejected even though it was true.

Quantify  $\rightarrow$  Type II  $\rightarrow \beta \rightarrow \beta = 1 - \text{Power}$

$\downarrow$

ability of your test to take/make right decisions.

$\downarrow$

$n \dots \dots \dots$  # of sample

Relationship b/w Type I & Type II.

$$\frac{\text{Type I}}{\text{Type II}} \propto \frac{1}{\text{explore ct!}}$$

### CHI-SQUARE TEST

non-parametric  $\rightarrow$  no distribution

char-char situation  
(categorical)

$$df = (R-1)(C-1) \quad R \Rightarrow \text{rows} \\ C \Rightarrow \text{columns}$$

Q What is the relationship between Gender & Result?

|  |  | Result | Pass | fail |
|--|--|--------|------|------|
|  |  | Gender |      |      |
|  |  | M      | 60   | 40   |
|  |  | F      | 24   | 32   |

Sol.  $H_0$ : There is no relationship b/w Gender & result

$H_A$ : There is relationship b/w Gender & result

D NL D<sub>ad</sub> F<sub>ad</sub>

| Gender | M | 60 | 40 | 100 |
|--------|---|----|----|-----|
| F      |   | 24 | 32 | 56  |
|        |   | 84 | 72 | 156 |

$$\begin{array}{l} \text{Total males} = 100 \quad \text{Total females} = 56 \quad \text{Total pass} = 84 \\ \text{Total fail} = 72 \quad \text{Total people} = 156 \end{array}$$

Expected values:

$$\text{expected value}_{(\text{total males who passed})} = \frac{\text{Total males} \times \text{total pass}}{\text{total no of people}} = \frac{100 \times 84}{156} = 53.84$$

$$\text{expected value}_{(\text{total females who passed})} = \frac{\text{total females} \times \text{total pass}}{\text{total no of people}} = \frac{56 \times 84}{156} = 30.15$$

$$\text{expected value}_{(\text{total males who failed})} = \frac{\text{total males} \times \text{total fail}}{\text{total no of people}} = \frac{100 \times 72}{156} = 46.15$$

$$\text{expected value}_{(\text{total females who failed})} = \frac{\text{Total females} \times \text{total fail}}{\text{total no of people}} = \frac{56 \times 72}{156} = 25.84$$

$$ev_1 = 53.84, ev_2 = 30.15, ev_3 = 46.15, ev_4 = 25.84$$

Result Pass Fail

Gender

$$M \quad 53.84 \quad 46.15 = 100$$

$$F \quad \frac{30.15}{84} \quad \frac{25.84}{72} = \frac{56}{156}$$

Chi-square calculation

$$\chi^2 = \frac{(Actual - Expected)^2}{Expected}$$

$$(I) \quad \frac{(60 - 53.84)^2}{53.84} = 0.70$$

$$(II) \quad \frac{(40 - 46.15)^2}{46.15} = 0.81$$

$$(III) \quad \frac{(24 - 30.15)^2}{30.15} = 1.25$$

$$(IV) \quad \frac{(32 - 25.84)^2}{25.84} = 1.46$$

$$\chi^2_{cal} = 0.70 + 0.81 + 1.25 + 1.46 = 4.22$$

$$\chi^2_{tab} \Rightarrow \chi = 0.05, df = (R-1)(C-1)$$

$$= (2-1)(2-1) = 1 \times 1 = 1$$

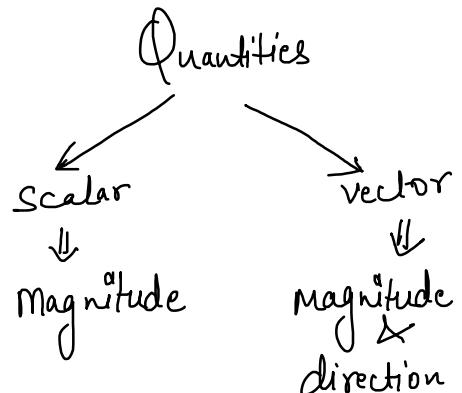
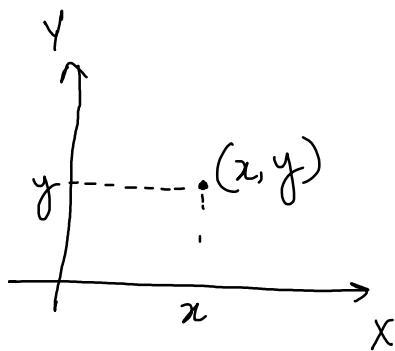
$$\chi^2_{tab} = 3.841$$

lets compare  $\chi^2_{\text{cal}}$  with  $\chi^2_{\text{tab}}$

$$\chi^2_{\text{cal}} > \chi^2_{\text{tab}}$$

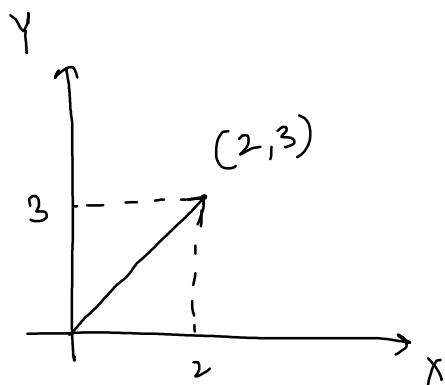
Reject  $H_0$

# Linear Algebra

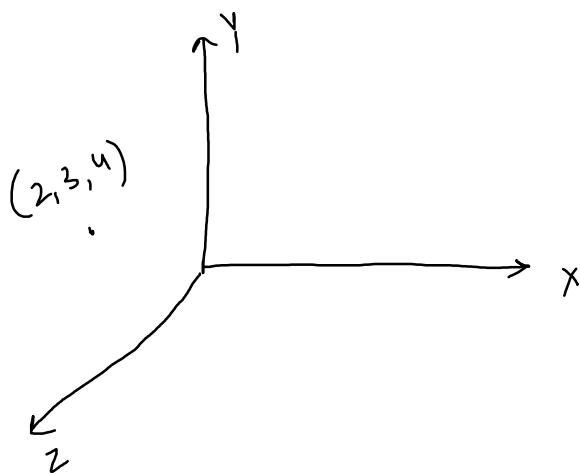


distance = 2 km (scalar)

(vector)  
displacement =  $\begin{pmatrix} 1 \\ -1 \end{pmatrix}$



$$\text{vector} = \begin{bmatrix} 2 \\ 3 \end{bmatrix} \Rightarrow 2d$$



$$\text{vector} = \begin{bmatrix} 2 \\ 3 \\ 4 \end{bmatrix} \Rightarrow 3d$$

$$\text{vector in } 6d = \begin{bmatrix} 1 & 2 & 3 & 4 & 5 & 6 \end{bmatrix} \Rightarrow 6d$$

Vectors in  $\text{nd} = [1 \ 2 \ 3 \ 4 \ 5 \ \dots \ n] \Rightarrow \text{nd vector}$

MATRIX  $\Rightarrow$  table of numbers

Rows

$$\begin{bmatrix} & & & & & \text{columns} \\ 1 & 1 & 1 & 1 & 1 & 1 \end{bmatrix}$$

Shape of matrix  $\Rightarrow$  Rows  $\times$  columns

$$g/p \xrightarrow{\text{Transformation (Linear)}} o/p$$

Addition:

$$\begin{bmatrix} a & b \\ c & d \end{bmatrix} + \begin{bmatrix} e & g \\ f & h \end{bmatrix} = \begin{bmatrix} a+e & b+g \\ c+f & d+h \end{bmatrix}$$

$$\begin{bmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \end{bmatrix} + \begin{bmatrix} 2 & 4 & 8 \\ 9 & 10 & 11 \end{bmatrix} \Rightarrow \begin{bmatrix} 3 & 6 & 11 \\ 13 & 15 & 17 \end{bmatrix}$$

Multiplication:

$$r \rightarrow r \ r \downarrow r \quad -$$

## Multiplication:

$$\begin{bmatrix} 1 & 2 \\ \cancel{1 \times 2} & \\ & 2 \\ & \cancel{2 \times 1} \end{bmatrix} \begin{bmatrix} 2 \\ 1 \end{bmatrix} = \begin{bmatrix} 1 \times 2 + 2 \times 1 \\ \cancel{1 \times 1} \end{bmatrix} = \begin{bmatrix} 4 \\ \cancel{1 \times 1} \end{bmatrix}$$

$$\begin{bmatrix} a & b \\ \cancel{c} & d \\ \cancel{2 \times 2} & \end{bmatrix} \times \begin{bmatrix} e & f \\ g & h \\ \cancel{2 \times 2} & \end{bmatrix} = \begin{bmatrix} ae+bg & af+bh \\ ce+dg & cf+dh \\ \cancel{2 \times 2} & \end{bmatrix}$$

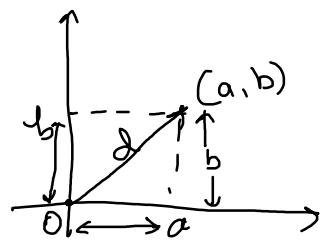
$$\begin{bmatrix} 1 & 2 \\ \cancel{3} & \cancel{4} \\ \cancel{2 \times 2} & \end{bmatrix} \times \begin{bmatrix} 1 & 2 \\ \cancel{3} & 4 \\ \cancel{2 \times 2} & \end{bmatrix} = \begin{bmatrix} 1 \times 1 + 2 \times 3 & 1 \times 2 + 2 \times 4 \\ 3 \times 1 + 4 \times 3 & 3 \times 2 + 4 \times 4 \\ \cancel{2 \times 2} & \end{bmatrix} = \begin{bmatrix} 1 \times 1 + 2 \times 3 & 1 \times 2 + 2 \times 4 \\ 3 \times 1 + 4 \times 3 & 3 \times 2 + 4 \times 4 \\ \cancel{2 \times 2} & \end{bmatrix}$$

a)  $a_{m \times n} \times b_{p \times q}$  &  $n \neq p$  : Multiplication not possible

b)  $a_{m \times n} \times c_{n \times q}$  : Multiplication possible : shape :  $m \times q$

Distance

Distance of a point from origin:



By Pythagoras Theorem,

$$d^2 = a^2 + b^2$$

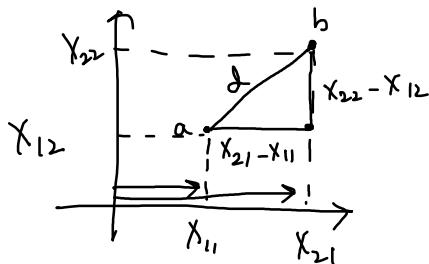
$$d^2 = b^2 + a^2$$

$$d = \sqrt{a^2 + b^2}$$

for nd, (extending idea to n dim)

$$d = \sqrt{a^2 + b^2 + c^2 + \dots + n^2}$$

Distance b/w two points:



$$a = [x_{11} \ x_{12}]$$

$$b = [x_{21} \ x_{22}]$$

By Pythagoras Theorem

$$d = \sqrt{(x_{21} - x_{11})^2 + (x_{22} - x_{12})^2}$$

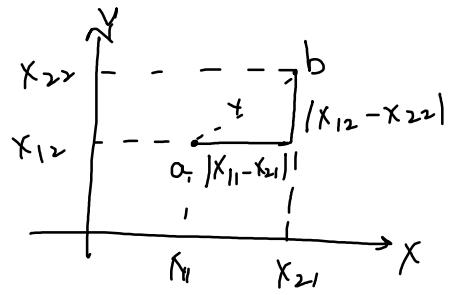
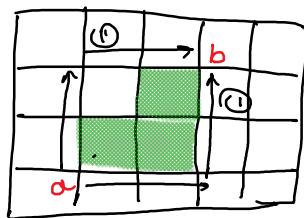
$$4^2 = 16$$

$$(-4)^2 = 16$$

euclidean distance  $\Leftrightarrow d = \sqrt{(x_{11} - x_{21})^2 + (x_{12} - x_{22})^2}$

euclidean distance =  $\left[ \sum_{i=1}^n |x_{1i} - x_{2i}|^2 \right]^{1/2} \rightarrow L2 \text{ Norm}$

### Manhattan distance



$$d = |x_{11} - x_{21}| + |x_{12} - x_{22}|$$

$$d = \sum_{i=1}^n |x_{1i} - x_{2i}| \rightarrow L1\text{-Norm}$$

### Minkowski distance

$$d = \left[ \sum_{i=1}^n |x_{1i} - x_{2i}|^p \right]^{1/p} \rightarrow Lp \text{ Norm}$$

$p = 1, 2, 3, \dots$

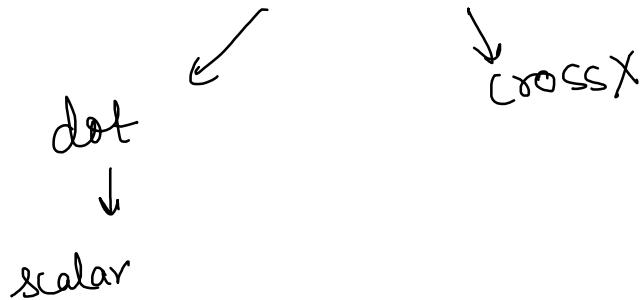
Let  $p=1$

$$d = \left[ \sum_{i=1}^n |x_{1i} - x_{2i}| \right] \Rightarrow \text{Manhattan distance}$$

Let  $p=2$

$$d = \left[ \sum_{i=1}^n |x_{1i} - x_{2i}|^2 \right]^{1/2} \Rightarrow \text{Euclidean distance}$$

Vector Multiplication



dot product in linear algebra  $\vec{a} = [a_1 \ a_2 \ a_3 \ \dots \ a_n]_{1 \times n}$

$$\vec{b} = [b_1 \ b_2 \ b_3 \ \dots \ b_n]_{1 \times n}$$

$$\vec{a} \cdot \vec{b} = [a_1 b_1 + a_2 b_2 + a_3 b_3 + \dots + a_n b_n]$$

Vector Representation

default  $\rightarrow$  column vector

$$\begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_n \end{bmatrix}$$

Row vector

$$\begin{bmatrix} a_1 & a_2 & \dots & a_n \end{bmatrix}$$

Vector

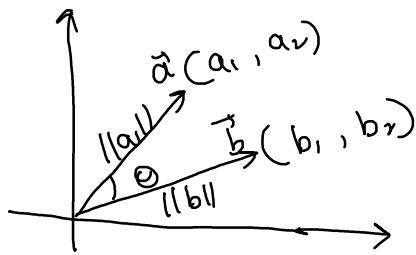
-

Vector

$$a = \begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_n \end{bmatrix}_{n \times 1} \quad b = \begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_n \end{bmatrix}_{n \times 1}$$

$$\begin{aligned} a \cdot b \Rightarrow a^T b &= \begin{bmatrix} a_1 & a_2 & \dots & a_n \end{bmatrix}_{1 \times n} \times \begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_n \end{bmatrix}_{n \times 1} \\ &= [a_1 b_1 + a_2 b_2 + \dots + a_n b_n]_{1 \times 1} \end{aligned}$$

Angle b/w two vectors:



(Geometric dot product)

$$\vec{a} \cdot \vec{b} = \|a\| \|b\| \cos \theta$$

$$a^T \cdot b = a \cdot b = [a_1 b_1 + a_2 b_2]$$

$$a_1 b_1 + a_2 b_2 = \|a\| \|b\| \cos \theta$$

$$\cos \theta = \frac{a_1 b_1 + a_2 b_2}{\|a\| + \|b\|}.$$

$$\|a\| = \sqrt{a_1^2 + a_2^2}$$

$$\|\vec{b}\| = \sqrt{b_1^2 + b_2^2}$$

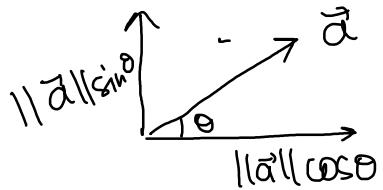
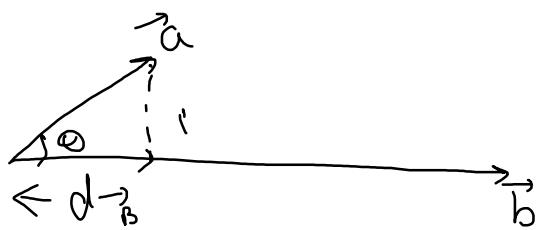
or

$$\theta = \cos^{-1} \left[ \frac{a_1 b_1 + a_2 b_2}{\|a\| + \|b\|} \right]$$

Projection:

$\uparrow - \rightarrow \vec{a}$

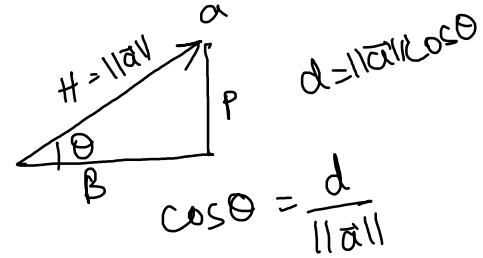
## Projection:



$$\textcircled{I} \quad d = \|\vec{a}\| \cos \theta$$

$$\textcircled{II} \quad \vec{a} \cdot \vec{b} = \underline{\|\vec{a}\|} \underline{\|\vec{b}\|} \underline{\cos \theta}$$

$$\textcircled{III} \quad \vec{a} \cdot \vec{b} = \underline{d} \|\vec{b}\|$$

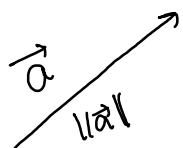


$$\boxed{d = \frac{\vec{a} \cdot \vec{b}}{\|\vec{b}\|}}$$

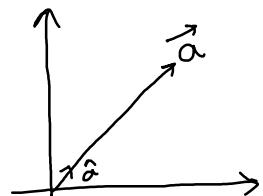
Projection of vector a on vector b

Unit vectors: → vector with a magnitude of 1.  
→ gives info regarding direction

unit vector → magnitude × direction  
magnitude

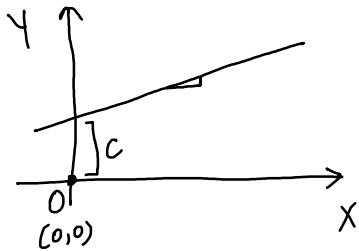


$$\hat{a} = \frac{\vec{a}}{\|\vec{a}\|}$$



## Lines and Planes

### Line



$c \rightarrow$  y-intercept  $\rightarrow$  value of  $y$  when  $x=0$   
 $m \rightarrow$  slope

$$y = mx + c$$

dependent variable    independent variable

$$m = \frac{y_2 - y_1}{x_2 - x_1} = \tan \theta = \frac{dy}{dx}$$

General Equation of line:  $Ax + By + C = 0$

$$\begin{cases} By = -C - Ax \\ y = -\frac{C}{B} - \frac{A}{B}x \end{cases}$$

$\downarrow$   $m$   
 $C$   
 (Intercept)

### Plane:

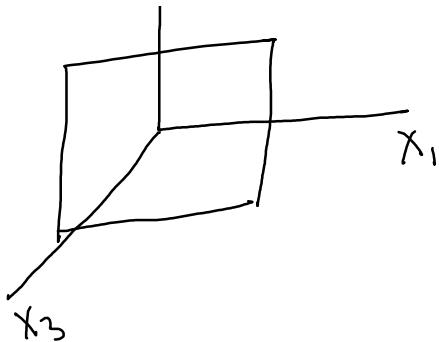
$$\omega_1 = 1, \omega_2 = 10, \omega_3 = 0$$



General Equation:

$$Ax_1 + By_1 + Cz_1 + D = 0$$

$$w_1=1, w_2=10, w_3=2$$



$$Ax + By + Cz + D = 0$$

↓ changing coeff

$$w_0 + w_1x_1 + w_2x_2 + w_3x_3 + w_4 = 0$$

↓

$$w_0 + w_1x_1 + w_2x_2 = 0 \text{ ?}$$

$$w_0 + w_1x_1 + w_2x_2 + \underline{w_3x_3} = 0$$

Above 3d: Hyperplane  $\Rightarrow w_0 + [w_1x_1 + w_2x_2 + w_3x_3 + \dots + w_nx_n] = 0$



$$w_0 + \underline{w^T \cdot x} = 0 \text{ [linear algebra way]}$$

hyperplane passing through origin  
 $w_0 = 0$

$$w^T x = 0$$

Eigen vector & Eigen values:

$$A\vec{x} = \lambda \vec{x}$$

eigen value (scalar)

eigen vector

Matrix      vector

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

$\vec{x}$   
 $\vec{x}$

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

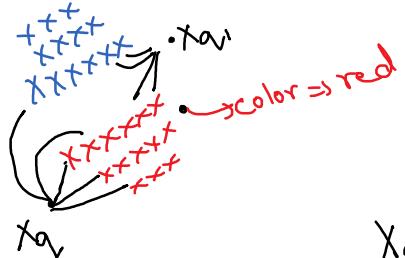
=  $\begin{matrix} 2 \\ \downarrow \\ \lambda \end{matrix}$   $\begin{bmatrix} 1 \\ 1 \end{bmatrix}$

## K-Nearest Neighbours

↳ you are like your neighbours

$K = 5 = \# \text{ neighbours}$

number of

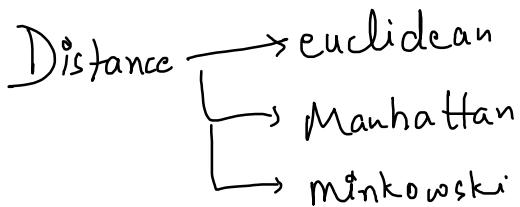


$$x_q(4R, 1B) = \text{Red}$$

$K = 4 = \# \text{ neighbours}$

$x_q(2R, 2B) = \text{Can't decide}$   
 ↓  
 hence, never  
 take  $K$  as even

hyperparameters:  $K = \# \text{ neighbours} \rightarrow K \uparrow / K \downarrow \rightarrow \text{better model}$

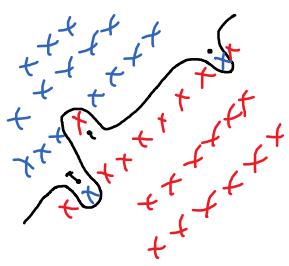


## Effect on $K$ :

$K=1$

⇒ No error in training ⇒ accuracy  $\geq 95\%$

⇒ large " in testing ⇒ accuracy  $\approx 60\%$



Overfitting

accuracy ↑

Reality "blue"

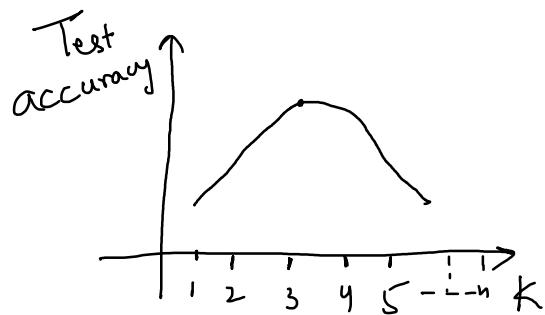
Overfitting  
 ↴ Training accuracy ↑  
 ↴ Testing " " ↓

$K=n$

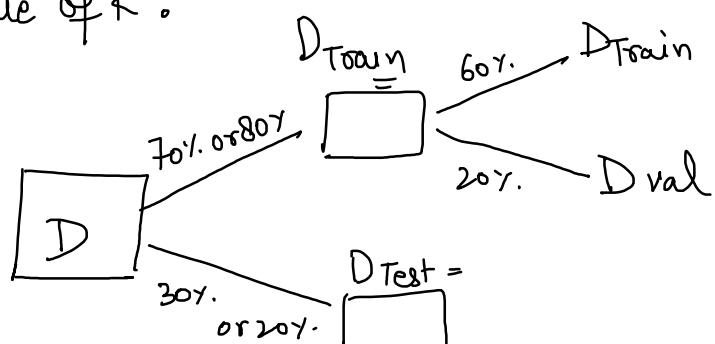
$x_2 = \text{Blue} \quad (\text{count(blue}) > \text{count(red)})$

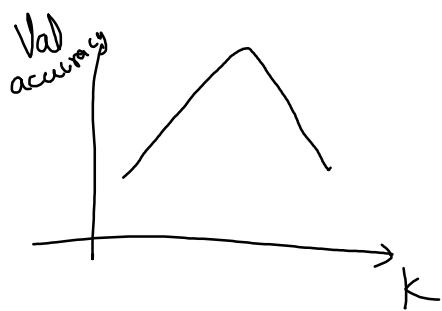
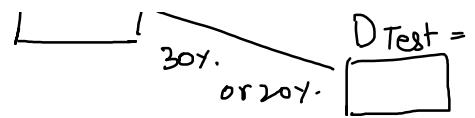
- Training error ↑
  - Testing error ↓
- Underfitting

Curve of  $K$  with accuracy :

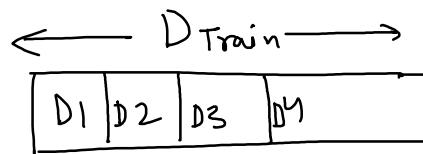


Choosing right value of  $K$ :





## $k'$ Fold cross validation



$$k' = 4$$

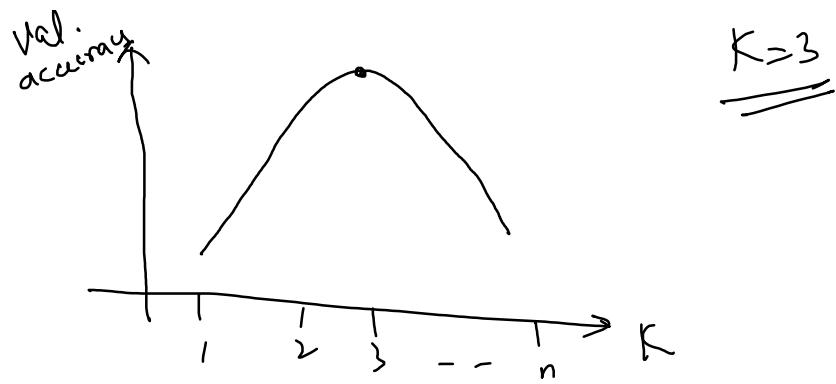
$$k' = 10 \text{ (Thumb Rule)}$$

Accuracy

$$\left. \begin{array}{c} a'_1 \\ a'_2 \\ a'_3 \\ a'_4 \end{array} \right\} \rightarrow a'^{\text{avg}}$$

| $k$<br>(#neighbours) | Training | Validation | Accuracy |
|----------------------|----------|------------|----------|
| 1                    | D1 D2 D3 | D4         | $a'_1$   |
| 1                    | D2 D3 D4 | D1         | $a'_2$   |
| 1                    | D1 D3 D4 | D2         | $a'_3$   |
| 1                    | D1 D2 D4 | D3         | $a'_4$   |
| <hr/>                |          |            |          |
| 2                    | D1 D2 D3 | D4         | $a^2_1$  |
| 2                    | D2 D3 D4 | D1         | $a^2_2$  |
| 2                    | D1 D2 D4 | D3         | $a^2_3$  |
| 2                    | D1 D3 D4 | D2         | $a^2_4$  |

$$\text{list} = [a^{\text{avg}}, a^{2 \text{ avg}}, a^{3 \text{ avg}}, \dots, a^{n \text{ avg}}]$$



$\Rightarrow$  Create KNN for me, by keeping  $\Rightarrow K = 3$ .

$$KNNC = 3$$

### Advantages:

- very easy to understand
- No assumption

### Disadvantage

- lazy learner
- space issue
- time complexity ↑



### Evaluation Metrics

CONFUSION MATRIX  $\Rightarrow$

|            |        | Actual |        | $T_P$ |
|------------|--------|--------|--------|-------|
|            |        | $O(-)$ | $I(+)$ |       |
| Prediction | $O(-)$ | $T_N$  | $F_N$  |       |
|            | $I(+)$ | $F_P$  | $T_P$  |       |

$$\text{Accuracy} = \frac{TN + TP}{TP + TN + FP + FN} \Rightarrow P=0.5 \rightarrow \text{Accuracy} \times \text{Class imbalance}^{\uparrow}$$

$\downarrow \text{Precision} \Rightarrow \frac{TP}{TP+FP \uparrow}$

$\downarrow \text{Recall} = \frac{TP}{TP+FN \uparrow}$

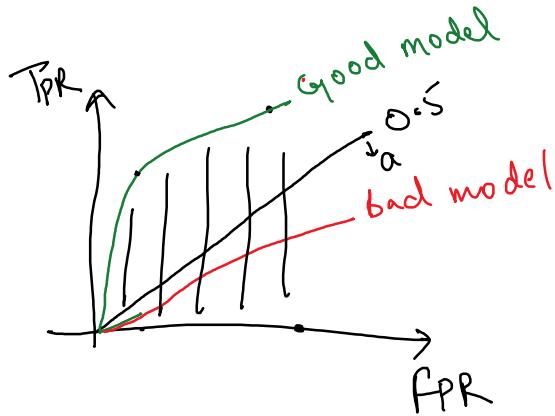
(Sensitivity)  
(TPR)

$$F1\text{-score} \Rightarrow \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

$$\text{SPECIFICITY} = 1 - \text{FPR}$$

(Opp. of  
Recall)

Roc Auc



$$TPR = \frac{TP}{TP+FN}$$

$$FPR = \frac{FP}{TN+FP}$$

$AUC > 0.5$  (For good model)  
 $AUC > 0.7$

$$\frac{P(A|B)}{P(B) \rightarrow \text{evidence}} = \frac{\text{likelihood}}{\text{prior}}$$

Bayes Theorem

posterior

What is the difference b/w likelihood & probability?

height  $\Rightarrow$  dataset

$$(h=170\text{cm} \mid \mu=150\text{cm}, \sigma=10\text{cm}) \Rightarrow \text{Prob.}$$

$$(\mu=150\text{cm}, \sigma=10\text{cm} \mid h=170\text{cm}) \Rightarrow \text{likelihood}$$

Naive Bayes  $\rightarrow$  Bayes Theorem  
 ignorant + innocent

$\rightarrow$  all features are independent of each other.

$$P(C_x/x_i) = \frac{P(x_i/C_x) P(C_x)}{P(x_i)}$$

$C_x \rightarrow$  class labels  
 $x_i \rightarrow$  input features

$$P(\text{Yes}/\text{outlook}) = P(\text{outlook}/\text{Yes}) P(\text{Yes})$$

| Outlook  | Temperature | Humidity | Windy | PlayTennis |
|----------|-------------|----------|-------|------------|
| Sunny    | Hot         | High     | False | No         |
| Sunny    | Hot         | High     | True  | No         |
| Overcast | Hot         | High     | False | Yes        |

$$P(\text{Yes}/\text{outlook}) = P(\text{outlook}/\text{Yes}) P(\text{Yes})$$

$$\frac{P(\text{outlook}) \times}{}$$

$$P(\text{No}/\text{outlook}) = P(\text{outlook}/\text{No}) P(\text{No})$$

$$\frac{P(\text{outlook}) \times}{}$$

$$P(\text{Yes}) = \frac{9}{14}$$

$$P(\text{No}) = \frac{5}{14}$$

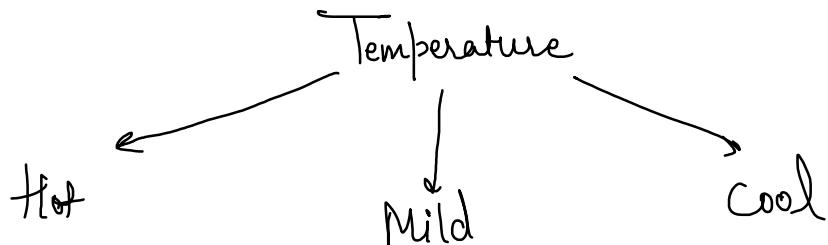
| Sunny    | Hot  | High   | False | No  |
|----------|------|--------|-------|-----|
| Sunny    | Hot  | High   | True  | No  |
| Overcast | Hot  | High   | False | Yes |
| Rainy ✓  | Mild | High   | False | Yes |
| Rainy ✗  | Cool | Normal | False | Yes |
| Rainy —  | Cool | Normal | True  | No  |
| Overcast | Cool | Normal | True  | Yes |
| Sunny    | Mild | High   | False | No  |
| Sunny    | Cool | Normal | False | Yes |
| Rainy ✓  | Mild | Normal | False | Yes |
| Sunny    | Mild | Normal | True  | Yes |
| Overcast | Mild | High   | True  | Yes |
| Overcast | Hot  | Normal | False | Yes |
| Rainy —  | Mild | High   | True  | No  |

Working :



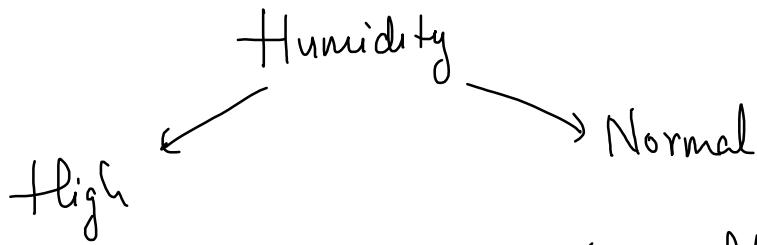
$$P(\text{sunny}/\text{Yes}) = \frac{2}{9} \quad P(\text{overcast}/\text{Yes}) = \frac{4}{9} \quad P(\text{Rainy}/\text{Yes}) = \frac{3}{9}$$

$$P(\text{sunny}/\text{No}) = \frac{3}{5} \quad P(\text{overcast}/\text{No}) = 0 \quad P(\text{Rainy}/\text{No}) = \frac{2}{5}$$



$$P(\text{Hot}/\text{Yes}) = \frac{2}{9} \quad P(\text{Mild}/\text{Yes}) = \frac{4}{9} \quad P(\text{Cold}/\text{Yes}) = \frac{3}{9}$$

$$P(\text{Hot}/\text{No}) = \frac{2}{5} \quad P(\text{Mild}/\text{No}) = \frac{2}{5} \quad P(\text{Cold}/\text{No}) = \frac{1}{5}$$

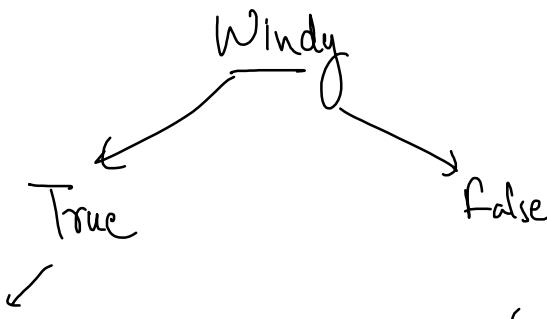


$$P(\text{High} | \text{Yes}) = \frac{3}{9}$$

$$P(\text{Normal} | \text{Yes}) = \frac{6}{9}$$

$$P(\text{High} | \text{No}) = \frac{4}{5}$$

$$P(\text{Normal} | \text{No}) = \frac{1}{5}$$



$$P(\text{True} | \text{Yes}) = \frac{3}{9}$$

$$P(\text{false} | \text{Yes}) = \frac{6}{9}$$

$$P(\text{True} | \text{No}) = \frac{3}{5}$$

$$P(\text{false} | \text{No}) = \frac{2}{5}$$

Q outlook  $\rightarrow$  overcast, temp  $\rightarrow$  cool, humidity  $\rightarrow$  high, wind  $\rightarrow$  true, tell me will I play?

$$P(\text{Yes} | \text{overcast, cool, high, true}) = P(\text{Yes}) \times P(\text{overcast} | \text{Yes}) \times P(\text{cool} | \text{Yes}) \\ \times P(\text{high} | \text{Yes}) \times P(\text{true} | \text{Yes})$$

$$= \frac{1}{3} \times \frac{2}{9} \times \frac{2}{9} \times \frac{3}{9} \times \frac{2}{9}$$

$$= \frac{2}{243} = \underline{\underline{0.0105}}$$

$$\begin{aligned}
 P(\text{No} \mid \text{overcast, cool, high, true}) &= P(\text{No}) \times P(\text{overcast} \mid \text{No}) \times P(\text{cool} \mid \text{No}) \times \\
 &\quad P(\text{high} \mid \text{No}) \times P(\text{true} \mid \text{No}) \\
 &= \frac{5}{14} \times 0 \times \times \times = 0
 \end{aligned}$$

$$P(\text{Yes} \mid \text{overcast, cool, high, true}) > P(\text{No} \mid \text{overcast, cool, high, true})$$

Yes, g will play!

$\underline{\Omega}$  Outlook  $\rightarrow$  sunny, temp  $\rightarrow$  cool, humidity  $\rightarrow$  high, windy  $\rightarrow$  true. Will g play?

$$P(\text{Yes} \mid \text{sunny, cool, high, true}) = \frac{2}{9} \times \frac{3}{9} \times \frac{3}{9} \times \frac{3}{9} \times \frac{9}{14} = 0.0052$$

$$P(\text{No} \mid \text{sunny, cool, high, true}) = \frac{3}{5} \times \frac{1}{5} \times \frac{4}{5} \times \frac{3}{5} \times \frac{5}{14} = 0.02$$

$$0.02 > 0.005$$

hence, No g will not play

Problem of zero probability:

$$\begin{aligned}
 \underline{\Omega} \quad P(\text{Yes} \mid \text{cloudy, cool, high, true}) &= P(\text{Yes}) \times P(\text{cloudy} \mid \text{Yes}) \times \dots \\
 &= 0
 \end{aligned}$$

$$P(\text{No} \mid \text{cloudy, cool, high, true}) = P(\text{No}) \times P(\text{cloudy} \mid \text{No}) \times \dots \\ = 0$$

cloudy isn't present in outlook!

(1) Ignore cloudy

$$2 \times 1 - 2 \Rightarrow P(\text{cloudy} \mid \text{Yes})$$

$$P(\text{Yes} \mid \text{cloudy, cool, high, true}) = P(\text{Yes}) \times P(\text{cool} \mid \text{Yes}) \times P(\text{high} \mid \text{Yes}) \times \\ P(\text{true} \mid \text{Yes}) \times 1$$

$$P(\text{No} \mid \text{cloudy, cool, high, true}) = P(\text{No}) \times P(\text{cool} \mid \text{No}) \times P(\text{high} \mid \text{No}) \\ \times P(\text{true} \mid \text{No}) \times 1$$

$P(\text{cloudy} \mid \text{No}) = 1$

(1) Laplace smoothing (var\_smoothing)

$$P(\text{cloudy} \mid \text{Yes}) = \frac{1 + \alpha}{n + \alpha} \rightarrow \begin{array}{l} \text{smoothing parameter} \\ n + \alpha \rightarrow \# \text{ distinct values} \end{array}$$

$\# \text{ data points}$   
 $(y=\text{yes})$

$\# \text{ columns}$   
 $\text{can take}$

effect of  $\alpha$ : 1000  $\xrightarrow{y=1}$  words,  $\rightarrow$  2 rare words

(1)  $\alpha = 0$

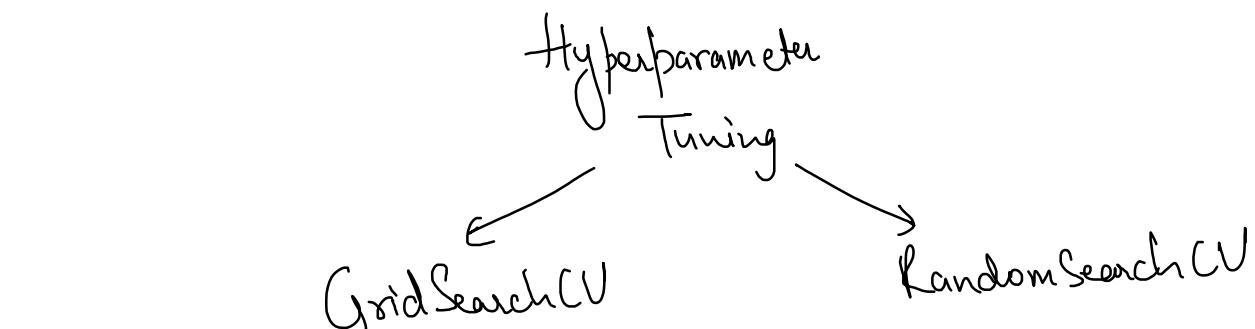
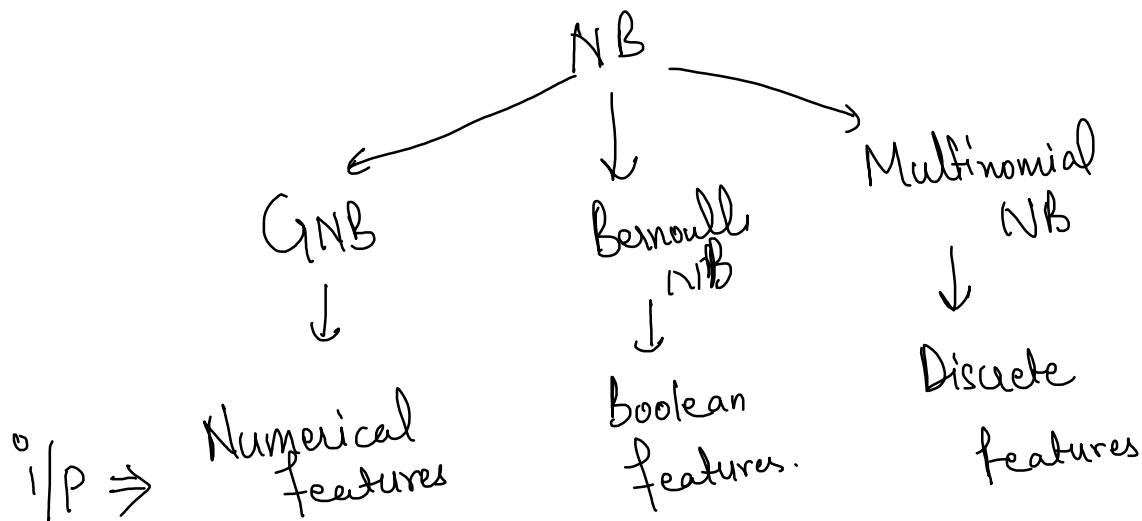
$$P(RW | Y=1) = \frac{2}{1000}$$

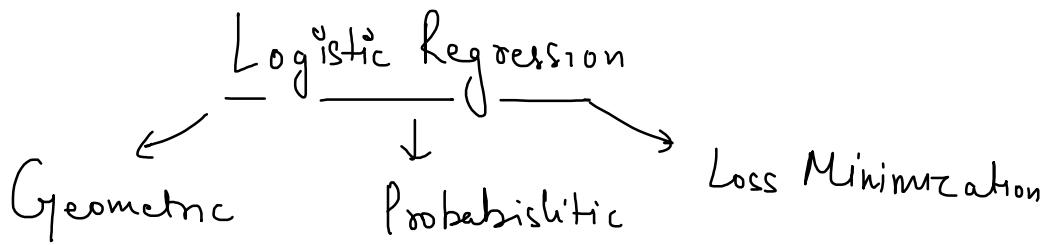
→ Overfitting  
↳ Variance  
(high)

⑪  $\alpha = 10000 \rightarrow$  underfitting → bias

$$P(W | Y=1) = \frac{2+10000}{1000 + 2 \times 10000} = \frac{10002}{21000} = 0.5$$

default value for  $\alpha = 1$ , → thumb rule





→  $g_t$  is used for binary classification

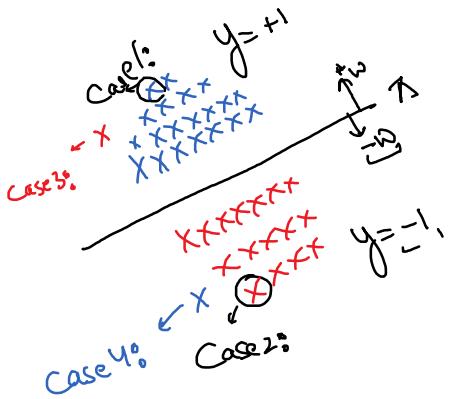


→ Data should be linearly separable

$$\text{Equation of plane: } w_0 + \sum_{i=1}^n w_i x_i = 0$$

plane is passing through origin

$$\sum_{i=1}^n w_i x_i = 0$$



①  $w^T x_i > 0 \rightarrow \text{for +ve class}$

②  $w^T x_i < 0 \rightarrow \text{for -ve class}$

lets multiply  $y_i$  with  $w^T x_i$

$$y_i \cdot w^T x_i$$

Case 1:  $y_i = +ve, w^T x_i > 0$

Case 2:  $y_i = -ve, w^T x_i < 0$

$y_i \cdot w^T x_i > 0 \rightarrow$    
 correct classification

$$y_i \cdot w^T x_i > 0$$

Case 3:  $y_i = -ve, \omega^T x_i > 0$

$$y_i \omega^T x_i < 0$$

Case 4:  $y_i = +ve, \omega^T x_i < 0$

$$y_i \omega^T x_i < 0$$

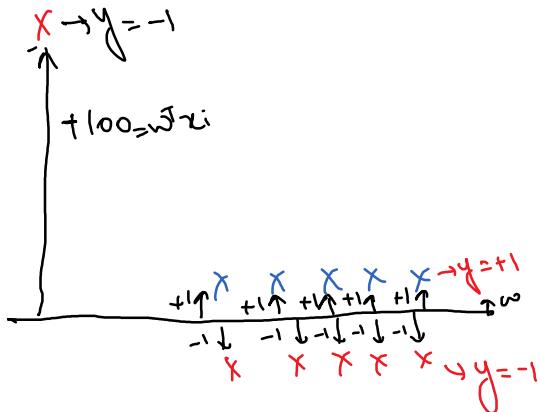
incorrect classification

if  $\begin{cases} y_i \omega^T x_i > 0 & \text{correct} \\ y_i \omega^T x_i < 0 & \text{incorrect} \end{cases}$

$$y_i \omega^T x_i >> 0$$

$\hookrightarrow$  i will be sure that  
classification is correct

Mathematical Objective Function  $\Rightarrow \underset{\omega}{\operatorname{argmax}} \left( \sum_i y_i \omega^T x_i \right) = \omega^*$

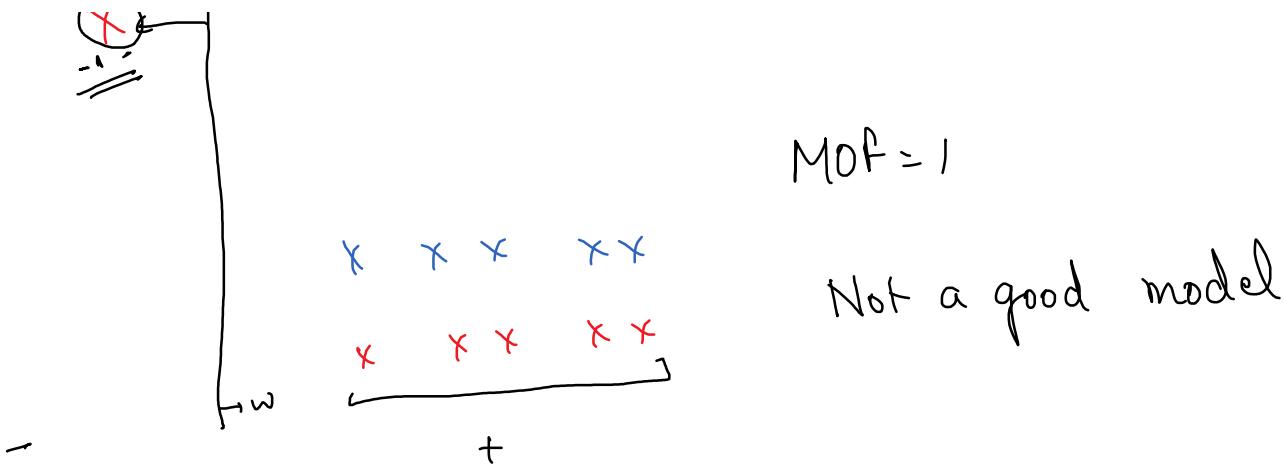


$$MOF = +1 + 1 + 1 + 1 + 1 + 1 + 1 + 1 + 1 + 1 - 100$$

$$= -90 \Rightarrow y_i \omega^T x_i$$

Not a good model





$$Mof = \underset{(w)}{\operatorname{argmax}} \left( \underbrace{y_i w^\top x_i}_{\rightarrow} \right) \rightarrow \text{outlier sensitive}$$

↙ Sigmoid function ↘  
 ↙ ↘  
 Probabilistic Squashing  
 interpretation values b/w  
 0 & 1

Robust to outliers

$$\tau(x) = \frac{1}{1+e^{-x}} \Rightarrow \text{sigmoid } f^n$$

$$\tau(y_i w^\top x_i) = \frac{1}{1 + e^{-y_i w^\top x_i}}$$

$$\text{MOF} \Rightarrow \underset{\omega}{\operatorname{argmax}} \left[ \sigma(y_i \omega^T x_i) \right]$$

$$\underset{\omega}{\operatorname{argmax}} \left[ \frac{1}{1 + e^{-y_i \omega^T x_i}} \right]$$

$\star \log \frac{1}{a} = -\log a$

$$\underset{\omega}{\operatorname{argmax}} \left[ \log \left( \frac{1}{1 + e^{-y_i \omega^T x_i}} \right) \right]$$

$$\underset{\omega}{\operatorname{argmax}} \left[ -\log \left( \frac{1}{1 + e^{-y_i \omega^T x_i}} \right) \right] \Rightarrow$$

$$\underset{\omega}{\operatorname{argmin}} \left[ \log \left( \frac{1}{1 + e^{-y_i \omega^T x_i}} \right) \right] \Rightarrow \text{loss function}$$

↓

*logistic loss*

Squashes value b/w 0 & 1

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

$x = -\infty$

$$\frac{1}{1 + e^{-\infty}} \rightarrow 0$$

$x = \infty$

$$\frac{1}{1 + e^{-\infty}} \rightarrow 1$$

$$e^{-\infty} = 0$$

## Probabilistic Interpretation

$$(0|p=1,0)$$

$$P = \frac{1}{1 + e^{-y}}$$

$$(1 + e^{-y})P = 1$$

$$P + P e^{-y} = 1$$

$$P e^{-y} = 1 - P$$

$$e^{-y} = \frac{1-P}{P}$$

$$e^y = \frac{P}{1-P} \rightarrow \text{odd's ratio}$$

Take  $\ln$  on both sides

$$\ln(e^y) = \ln\left(\frac{P}{1-P}\right)$$

$$y = \ln\left(\frac{P}{1-P}\right) \rightarrow \text{logit function}$$

$$y = \ln\left(\frac{p}{1-p}\right) \rightarrow \text{logit function}$$

$$\text{log loss} = - [y_i \log p(y_i) + (1-y_i) \log p(1-y_i)] \quad y_i \rightarrow 0 \text{ or } 1$$

Case 1:  $y_i \rightarrow$

|                      |            |
|----------------------|------------|
| prob $\rightarrow 1$ | $[0, 1]$   |
| geo $\rightarrow +1$ | $[-1, +1]$ |

geo:  $\log(1 + e^{-y_i w^T x_i}) \Rightarrow y=1 \Rightarrow \log(1 + e^{w^T x_i})$

prob:  $- [y_i \log p(y_i) + (1-y_i) \log p(1-y_i)]$

$$-\log p(y_i) \Rightarrow -\log \frac{1}{1+e^{-y_i}}$$

$$\Rightarrow -\log \frac{1}{1+e^{-w^T x_i}}$$

$$\Rightarrow -(-\log(1+e^{-w^T x_i}))$$

$$\Rightarrow \log(1+e^{-w^T x_i})$$

Derive Sigmoid from MOF:

$$\ln\left(\frac{P}{1-P}\right) = y_i w^T x_i$$

Take exp on both sides

$$e \left[ \ln \left( \frac{P}{1-P} \right) \right] = e^{y_i w^T x_i}$$

$$\frac{P}{1-P} = e^{y_i w^T x_i}$$

$$P = (1-P) e^{-y_i w^T x_i}$$

$$P = e^{y_i w^T x_i} - P e^{y_i w^T x_i}$$

$$P + P(e^{y_i w^T x_i}) = e^{y_i w^T x_i}$$

$$P(1+e^{y_i w^T x_i}) = e^{y_i w^T x_i}$$

$$P = \frac{e^{y_i w^T x_i}}{1+e^{y_i w^T x_i}}$$

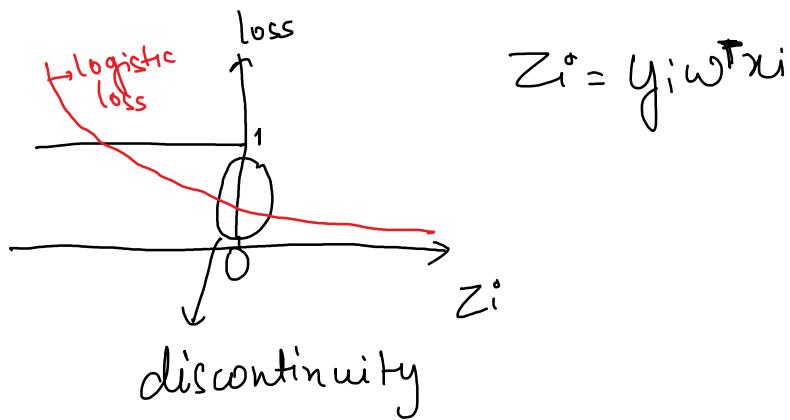
$$P = \frac{\cancel{e^{y_i w^T x_i}} / \cancel{e^{y_i w^T x_i}}}{1 / \cancel{e^{y_i w^T x_i}}} = \frac{1}{1 + e^{-y_i w^T x_i}}$$

$$P = \frac{e^{y_i w^\top x_i}}{1 + e^{y_i w^\top x_i}} = \frac{1}{1 + e^{-y_i w^\top x_i}}$$

Sigmoid function

$$P = \frac{1}{1 + e^{-y_i w^\top x_i}} = \sigma(y_i w^\top x_i)$$

### Loss Minimization : 0-1 loss



### Overfitting & Underfitting

$$w^* = \underset{w}{\operatorname{argmin}}$$

$$\sum_{i=1}^n \log(1 + e^{-y_i w^\top x_i})$$

$$y_i w^\top x_i \rightarrow \infty$$

### Regularization :

## ① RIDGE REGULARIZATION

$w^T \rightarrow$  very large

$$\text{Loss } f^n \Rightarrow \left[ \underset{w}{\operatorname{argmin}} \sum_{i=1}^n \underbrace{\log(1+e^{y_i w^T x_i})}_{\text{hyperparameter}} + \lambda \underbrace{w^T w}_{\text{hyperparameter}} \right]$$

## ② LASSO REGULARIZATION → feature selection

$$\text{Loss } f^n \Rightarrow \underset{w}{\operatorname{argmin}} \sum_{i=1}^n \log(1+e^{-y_i w^T x_i}) + \lambda \|w\|_1$$

Lasso creates sparsity  $\Rightarrow [1, 0, 0, 0, 1, 0, 0, 1, 1, 0, 0]$

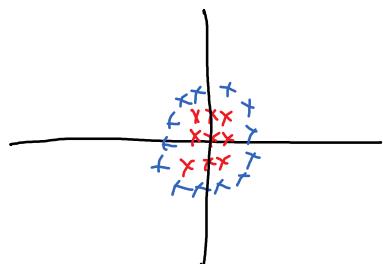
$\lambda \Rightarrow$  hyperparameter

$\lambda = 0$  overfitting

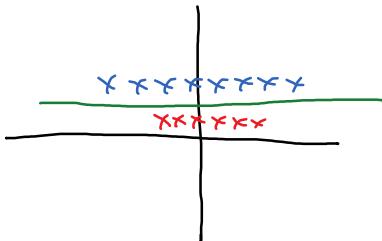
$\lambda = 1$  underfitting

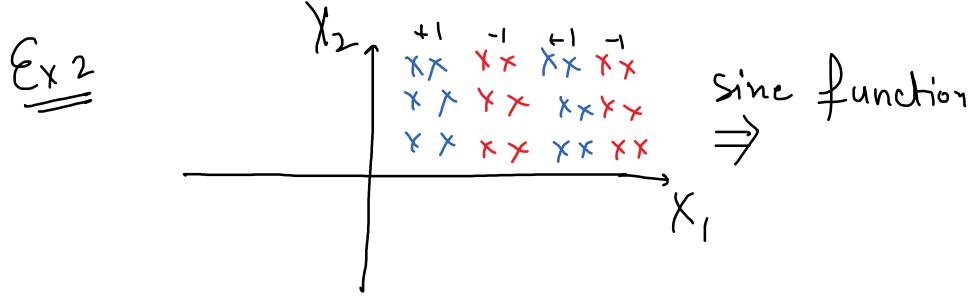
Feature

$x_1$



Square  
⇒

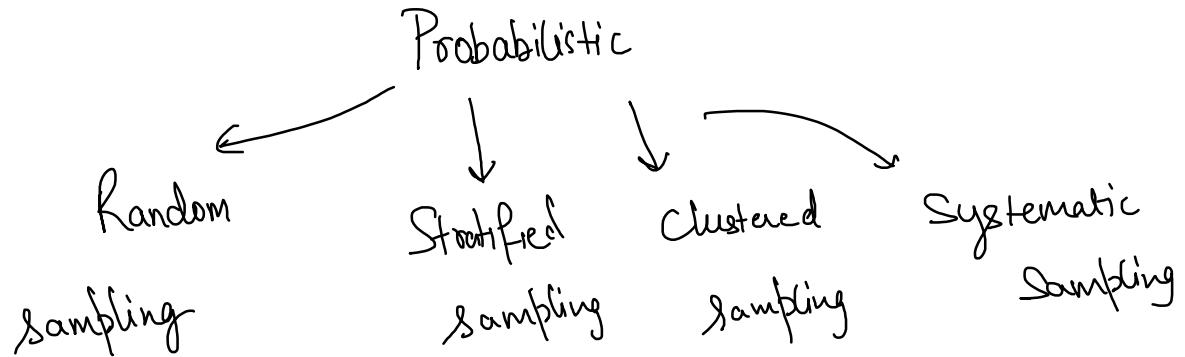




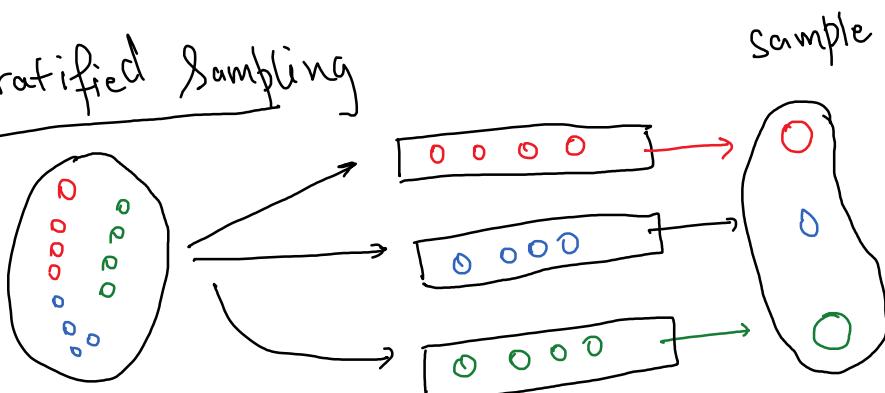
# Sampling Techniques

Probabilistic

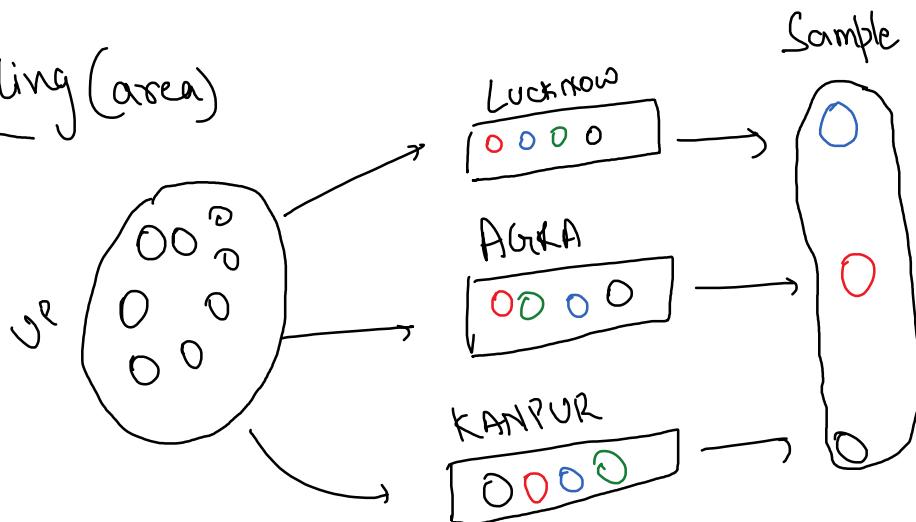
Non-Probabilistic



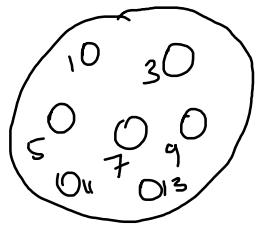
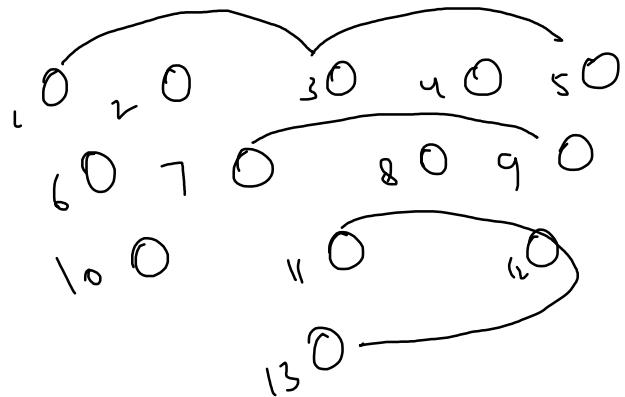
## Stratified Sampling



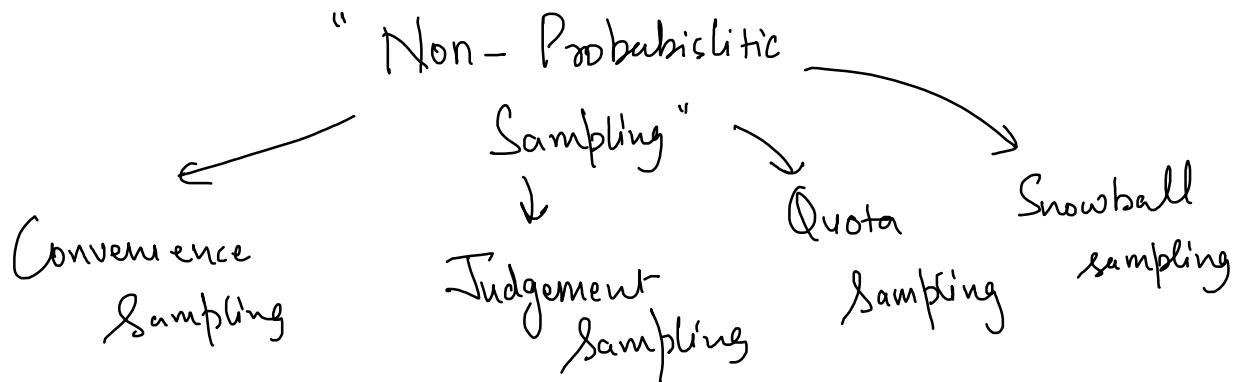
## Cluster Sampling (area)



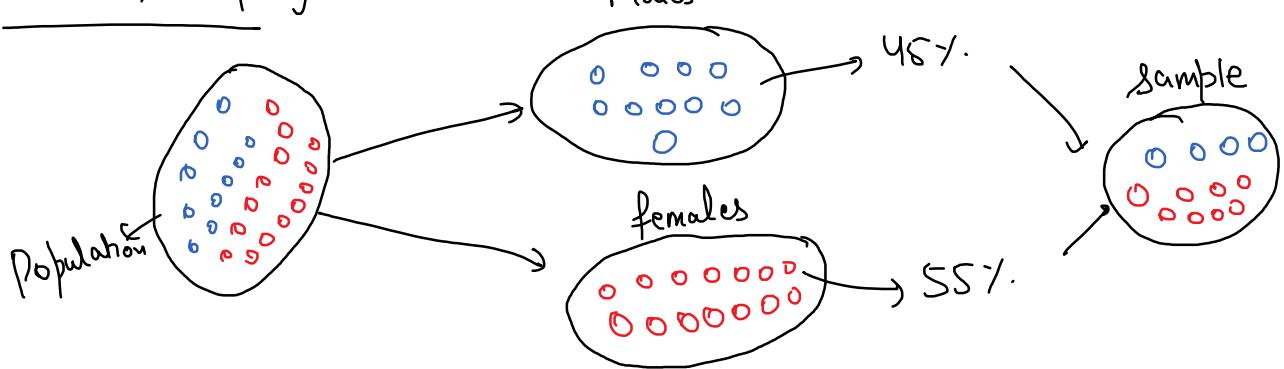
## Systematic Sampling



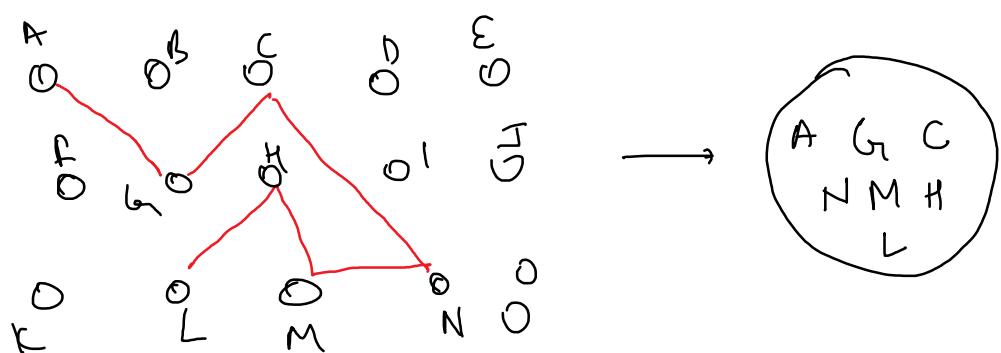
$$\text{System} \Rightarrow \frac{N}{n} \quad N \Rightarrow \frac{\text{Count of population}}{\text{Sample size}}$$



## Quota Sampling



## Snowball Sampling (Marketing)

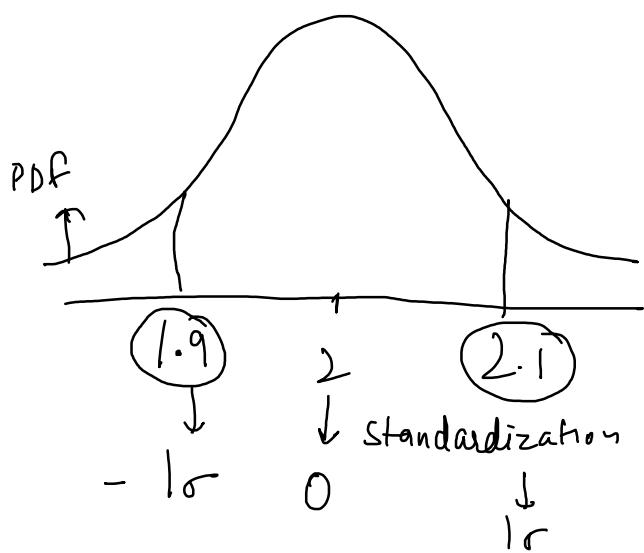


Z-score & probability values

$$\frac{Z = \bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}}$$

$\text{area} = \text{prob.}$

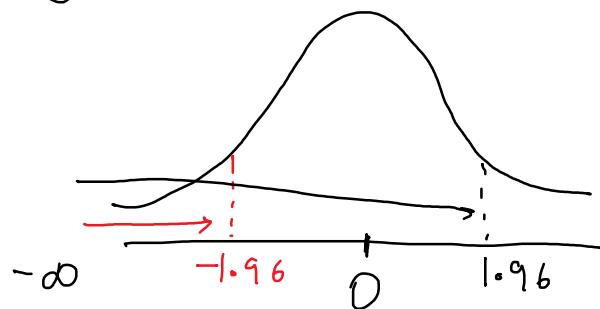
$$\int_{-\infty}^{\infty} \text{Pdf} = 1$$



Calculate Prob. from Z-score

$$Z_u = +1.96$$

$$Z_l = -1.96$$



$$\text{Prob} = \int_{-\infty}^{1.96} \text{Pdf} = ?$$

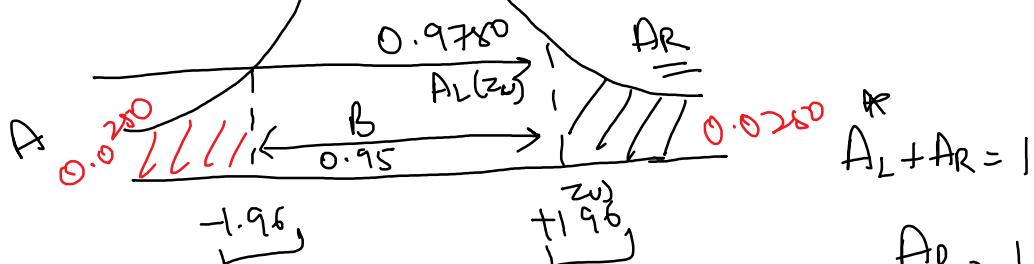
$$\int_{-\infty}^{-1.96} \text{Pdf} = ?$$

Instead of using integration, we will use z-table

- \* In z-table, area is always calculated from extreme left to the observation.

$$Z \text{ score} = \frac{x - \mu}{\frac{\sigma}{\sqrt{n}}} \rightarrow \text{std error}$$

Ex- AUC for entire curve = 1



$$A_L + A_R = 1$$

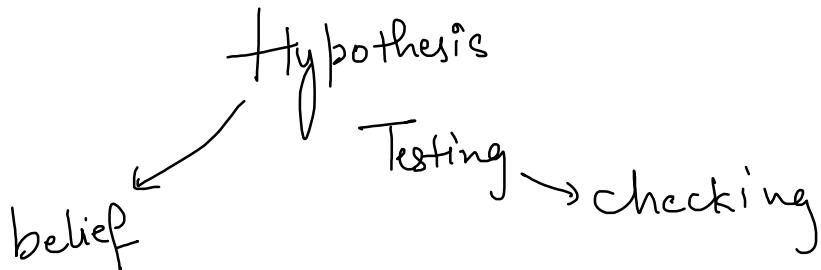
$$\begin{aligned} A_R &= 1 - 0.9750 \\ &= 0.0250 \end{aligned}$$

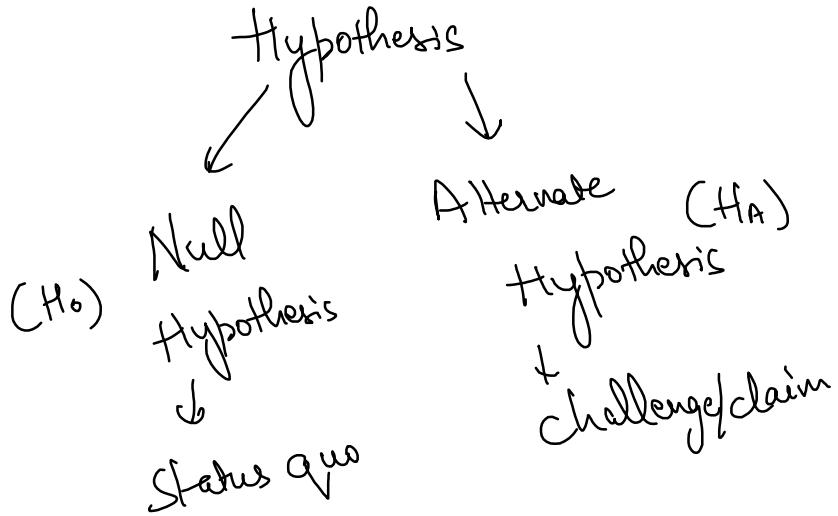
$$A + B + A_R = 1$$

$$0.0250 + B + 0.0250 = 1$$

$$B + 0.05 = 1$$

$$B = 1 - 0.05 = 0.95$$





Q Police claims, that a person is criminal ?

$H_0$  : Innocent

Alternate : Criminal

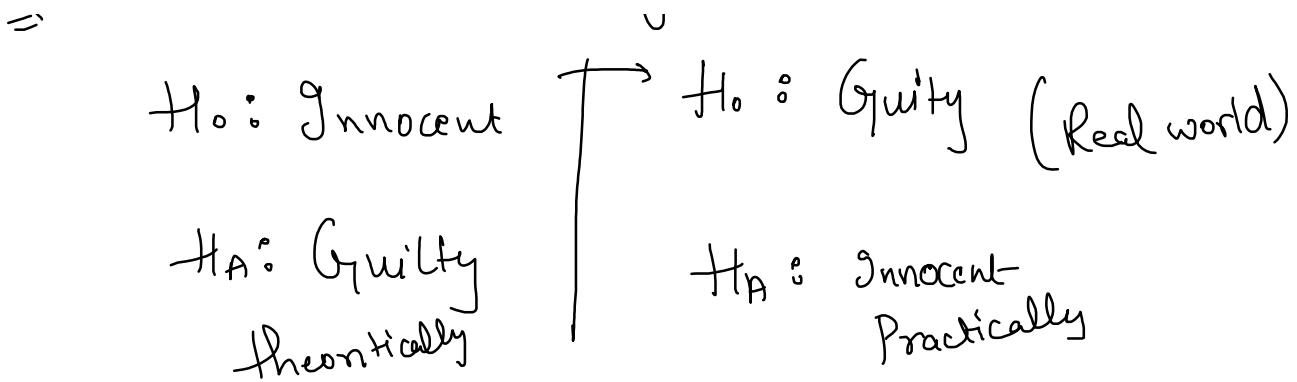
Q g claim, India will win the world cup?

$H_0$  : Any team can win the world cup

$H_A$  : India will world cup

Q Bride claims, that groom has taken dowry?

U . q . .  $\rightarrow H_0$  : Gravity  $r \propto \frac{1}{r^2}$



$\exists$  I claim, that  $\text{avg}$  salary of an engineer changed from \$100,000?

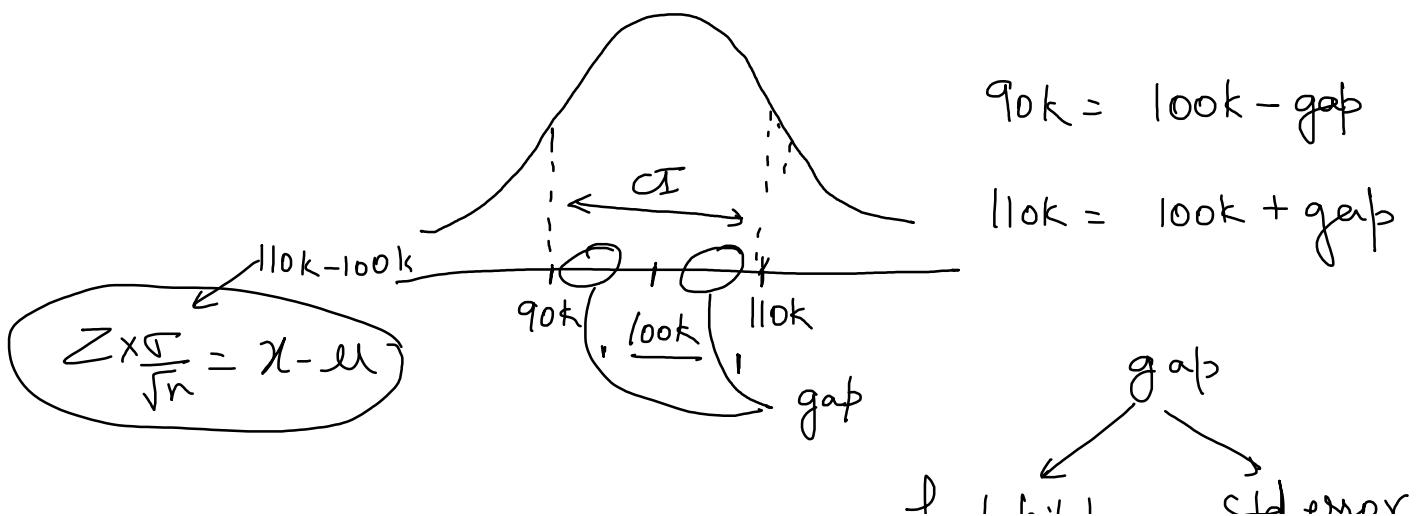
$$H_0: \mu = \$100,000$$

$$H_A: \mu \neq \$100,000$$

Building Criteria to test hypothesis:

i) a) Acceptance Region Method

$\exists$  Data Scientists earn \$100,000 salary on avg?



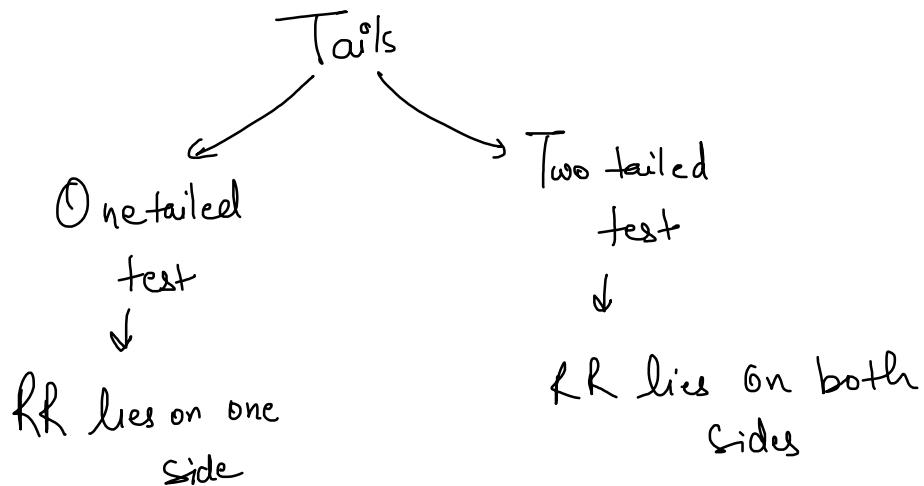
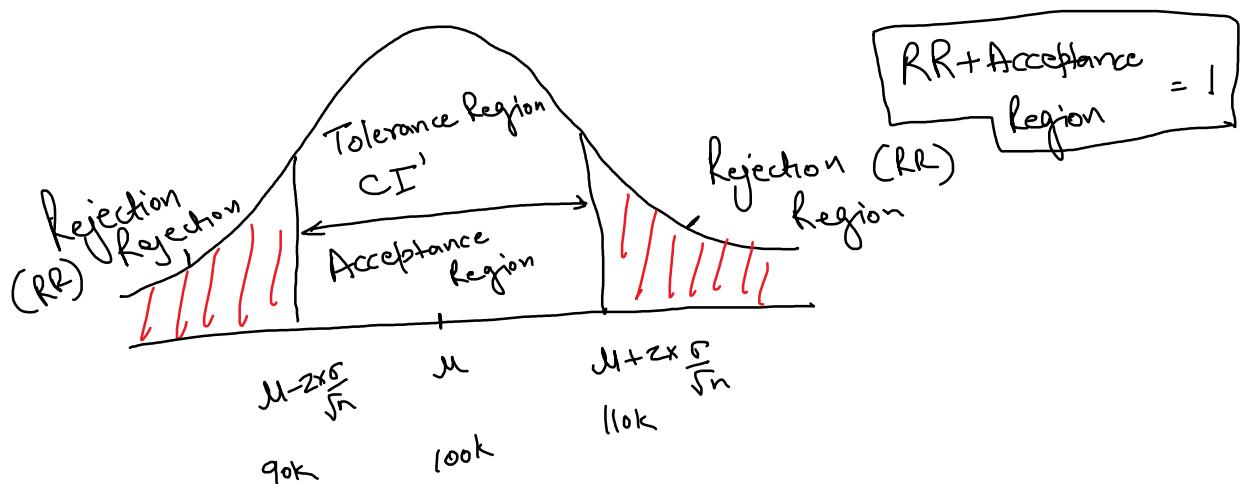
$$UL = \mu + z \times \frac{\sigma}{\sqrt{n}}$$

$$LL = \mu - z \times \frac{\sigma}{\sqrt{n}}$$

Margin  
of error

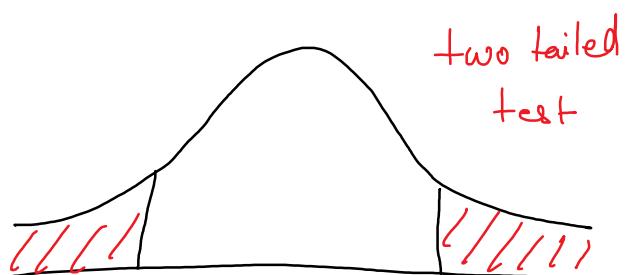
probability ↓  
z-score

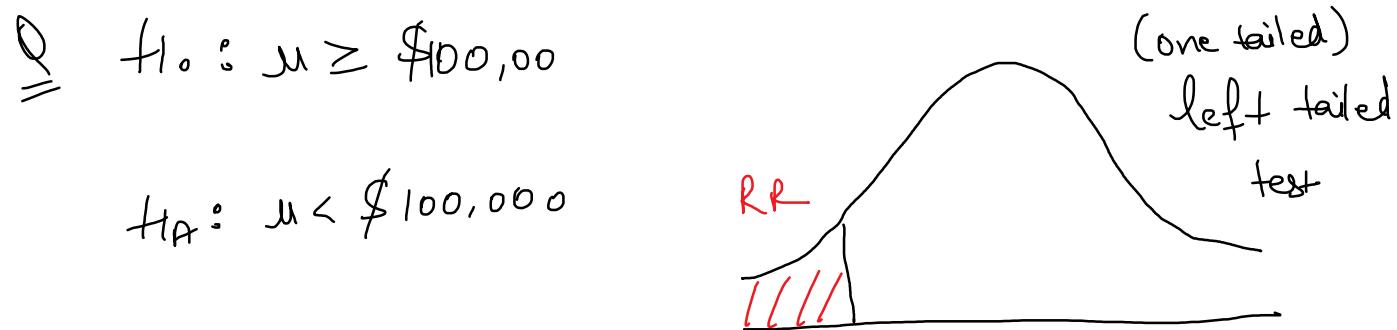
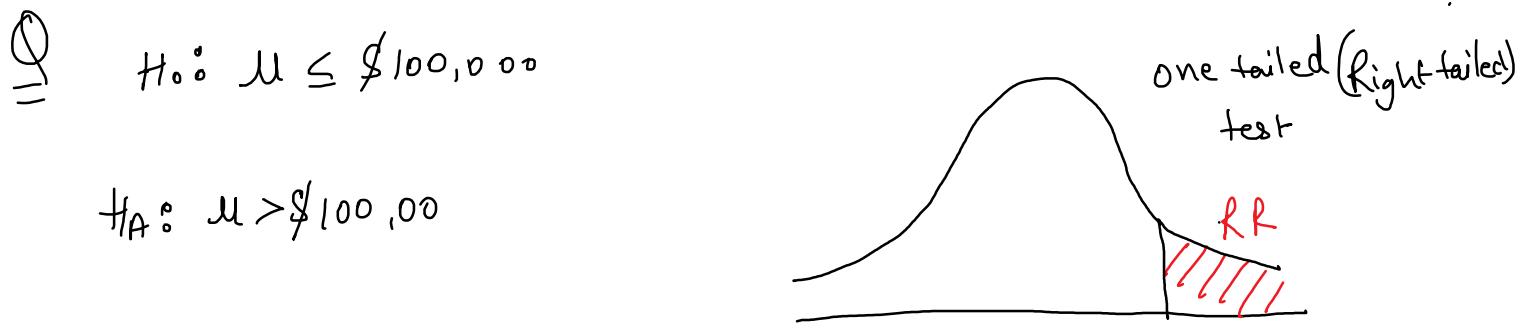
std error ↓  
 $\frac{\sigma}{\sqrt{n}}$



$\geq H_0: \mu = \$100,000$

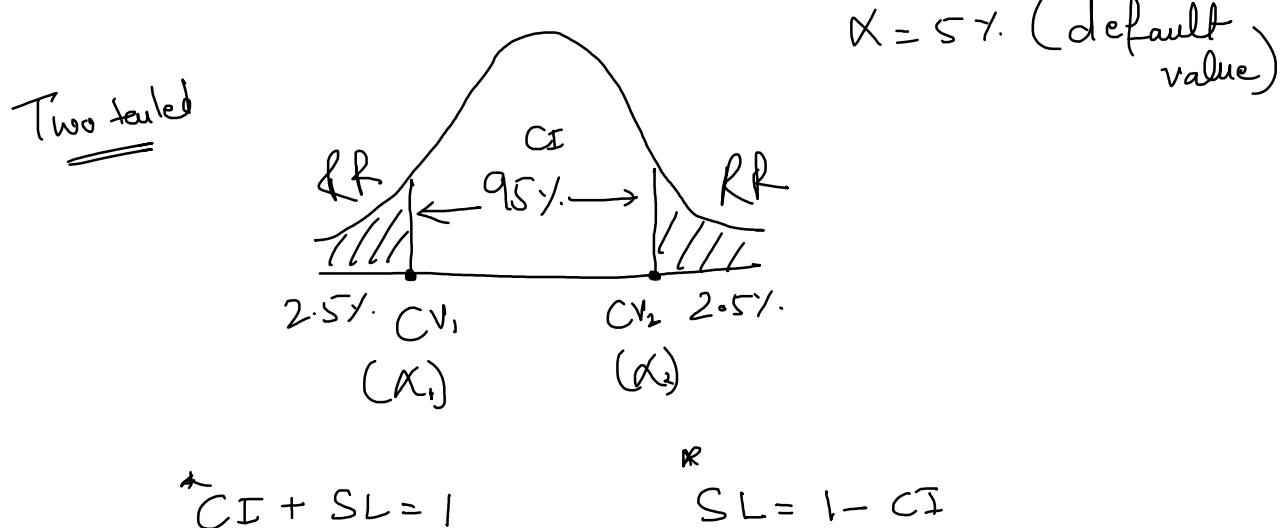
$H_A: \mu \neq \$100,000$





## 2 Critical Value Method:

$\alpha$  (significance level)  $\Rightarrow$  Marks your rejection region



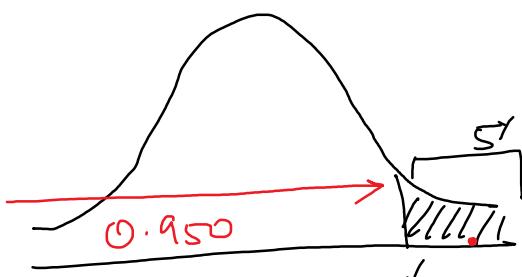
One tail

$$\alpha = 5\%$$

$CV \rightarrow Z\text{-table}$

One tail

$$\chi = 5\gamma$$



CV  $\rightarrow$  Z-table



Z-scores.

$$1 - 0.05$$

(II)  $Z_{\text{cal}}$  calculated from ztable  
 $Z_{\text{tab}}$

$$Z_{\text{cal}} = \frac{\chi - \mu}{\sigma / \sqrt{n}}$$

(III) Compare  $Z_{\text{tab}}$  with  $Z_{\text{cal}}$

(I)  $\stackrel{LT}{=} Z_{\text{cal}} > Z_{\text{tab}} \Rightarrow \text{Reject } H_0$

LT

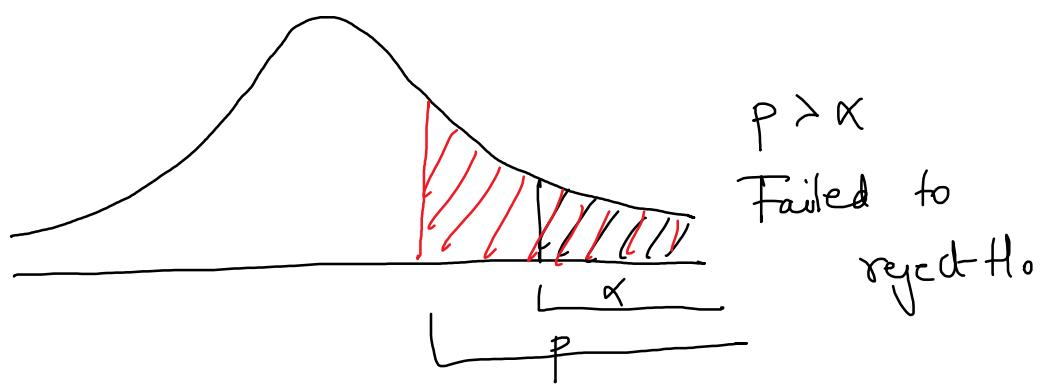
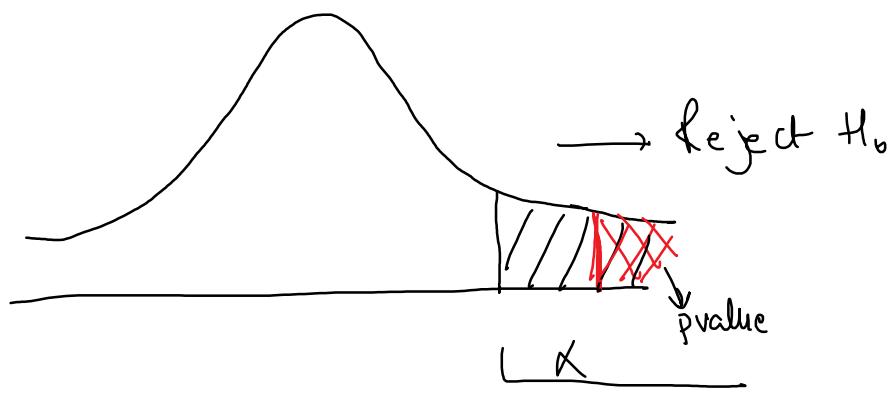
(II)  $Z_{\text{cal}} < Z_{\text{tab}} \Rightarrow \text{Reject } H_0$

3 P-value  $\Rightarrow$  The prob. of your null hypothesis to be true.

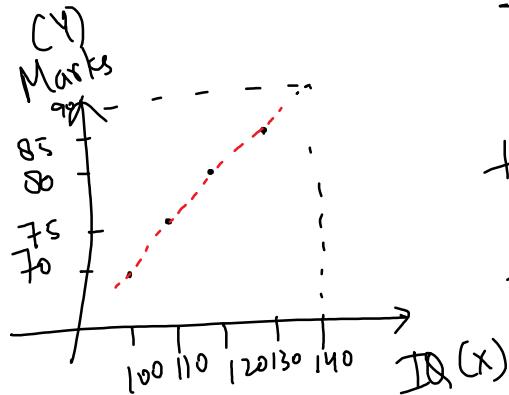


X

$P < \chi \Rightarrow \text{Reject } H_0$



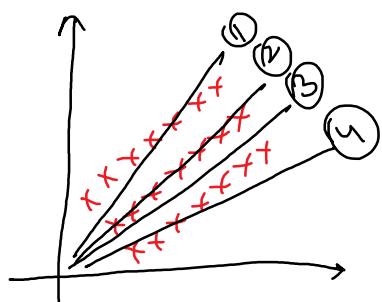
## Linear regression



How much marks can a person having IQ 140 score?

$$y = mx + c$$

Reality:



$L$  is best fit line

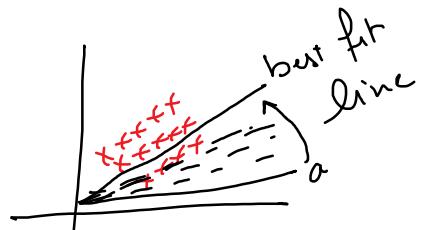
↳ line that fits your data best.

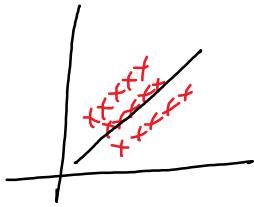
Linear regression

Ordinary Least  
Square

directly create  
best fit line

Gradient  
Descent  
↓  
min error



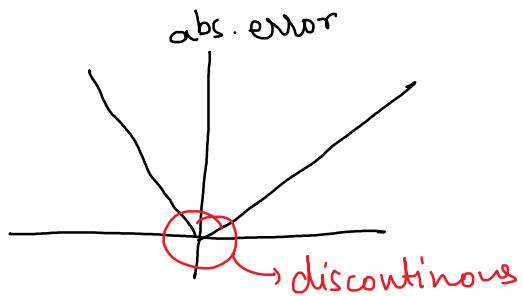
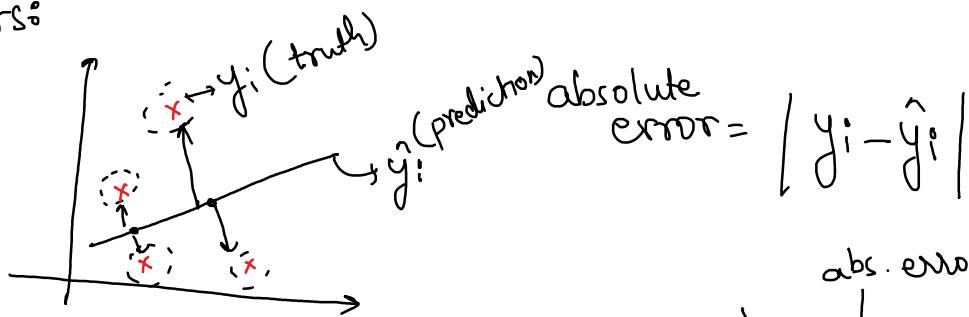


OLS

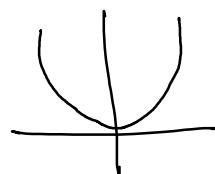
$$y = \omega_0 + \omega_1 x_1$$

$$y = \omega_0 + \omega_1 x_1 + \omega_2 x_2 + \dots + \omega_n x_n \Rightarrow \text{eqn of hyperplane}$$

Errors:



Squared error =  $\sum (y - \hat{y})^2$



$$E(m, b) = (y_i - \hat{y}_i)^2 = 0$$

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

$b$  intercept

$$E = \left[ y_i^* - (mx + b) \right]^2 = 0$$

$\frac{d}{db}$

$$\frac{dE}{db} = \frac{d \sum (y_i^* - (mx + b))^2}{db} = 0$$

$$\frac{d}{dx} x^n$$

$$\boxed{\frac{d}{dx} x^n = nx^{n-1}}$$

$$\Rightarrow \cancel{2} \sum (y_i^* - (mx + b)) (-1) = 0$$

$$\Rightarrow -2 \sum (y_i^* - mx - b) = 0$$

$$\sum (y_i^* - mx - b) = 0$$

Divide by  $n$  on both sides

$$\frac{\sum (y_i^* - mx - b)}{n} = \frac{0}{n} = 0$$

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

$$\bar{y}_i^* - m\bar{x}_i^* - \frac{n}{n} b = 0$$

$$\boxed{\bar{y}_i^* - m\bar{x}_i^* = \underline{b}}$$

→ intercept of  
best fit line

$m$

$\dots^2$

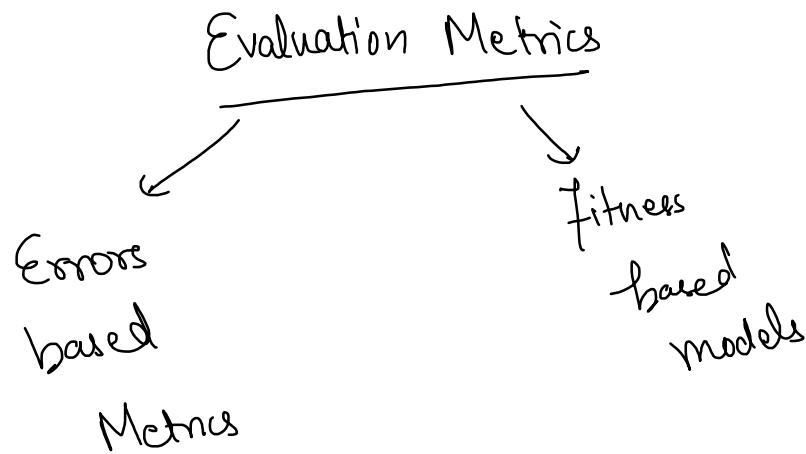
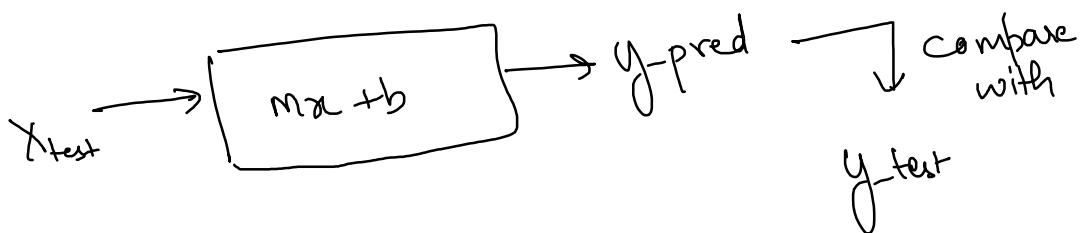
m

$$\frac{d E}{d m} = \frac{d \sum (y_i - mx_i - b)^2}{dm} = 0$$

↓ assignment

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

slope of  
best-fit line



## Error based:

1 → Mean absolute error  $\Rightarrow \frac{1}{n} \sum |y_i - \hat{y}_i|$

### Advantages:

- Same scale as that of data
- less sensitive to outliers

### disadvantages:

- can't be differentiated

2 → MSE (Mean Squared Error)  $\Rightarrow \frac{1}{n} \sum (y_i - \hat{y}_i)^2$

### Advantages:

- can be optimized

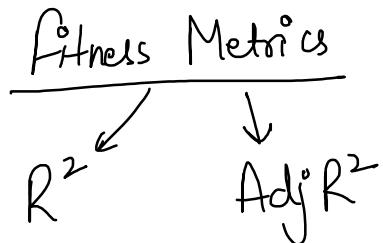
### disadvantages:

- hard to convey

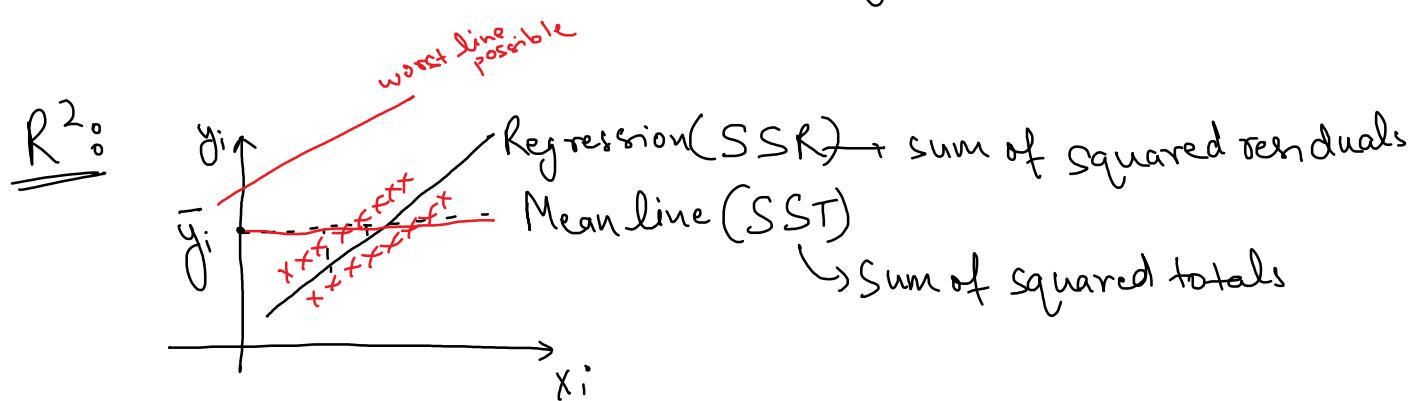
- sensitive to outliers.

3 → RMSE  $\Rightarrow \sqrt{MSE}$  ⇒ Root Mean Squared Error.

- can be conveyed easily
- less sensitive to outliers.



$$R^2 \xrightarrow{\quad} \downarrow \text{Adj } R^2$$



$$R^2 = 1 - \frac{SSR}{SST}$$

how well your model fits the data.

Case 1:  $SSR = 0$ ,  $SST = SST$

$$R^2 = 1 - \frac{SSR}{SST} = 1 - \frac{0}{SST} = 1 - 0 = 1 \rightarrow \text{overfitting}$$

Case 2:  $SSR = SST$  (when regression line coincide with mean line)

$$R^2 = 1 - \frac{SSR}{SST} = 1 - \frac{SST}{SST} = 1 - 1 = 0 \text{ (underfitting)}$$

Case 3:  $SSR > SST$

$$\left( \frac{SSR}{SST} > 1 \right)$$

$$R^2 = 1 - \frac{SSR}{SST} = -ve$$

Problem with  $R^2$ ?  $\rightarrow$  As your dimensions  $\uparrow$ ,  $R^2 \uparrow$

Adjusted  $R^2 \Rightarrow 1 - \left[ \frac{(1-R^2) \downarrow (n-1)}{(n-p-1) \downarrow} \right]$

# datapoints  
 $\frac{(1-R^2) \downarrow (n-1)}{(n-p-1) \downarrow}$

$$R^2 = 0.7 \\ = 0.75$$

# columns

$$p \Rightarrow 1 \Rightarrow M \\ \textcircled{3}$$

### Multicollinearity:

|     | $f_1$ | $f_2$ | $f_3$ |
|-----|-------|-------|-------|
| $w$ | 1     | 2     | 3     |
| $w$ | 0     | 3.5   | 3     |

$$f_1 = 1.5 f_2$$

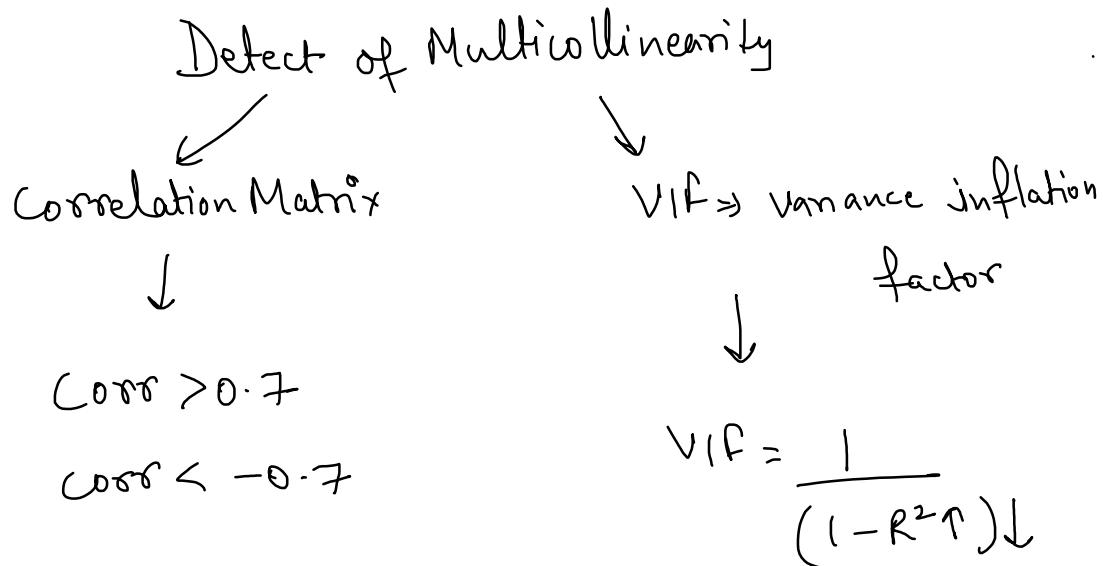
columns

$$y = 1 f_1 + 2 f_2 + 3 f_3 \quad \textcircled{1}$$

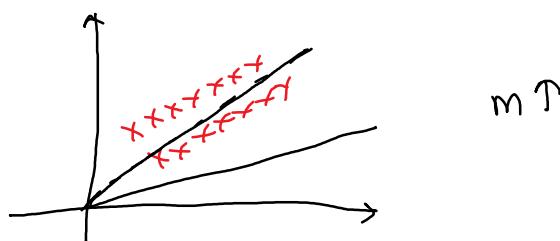
weights (coeffs)

$$y = 1.5 f_2 + 2 f_2 + 3 f_3$$

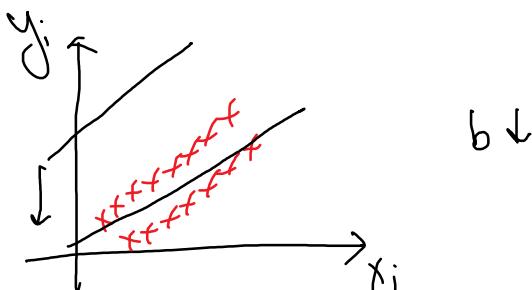
$$y = 0f_1 + 3.5f_2 + 3f_3$$



### Gradient Descent

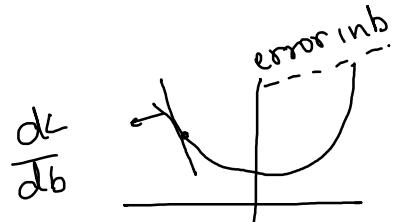


$y = mx + b$  (line rotation is done by slope)



$y = mx + b$  (intercepts can move the line up & down)

Steps :  $m = \text{constant}$ ,  $b = \text{variable}$



1) choose any random value of  $b$ .

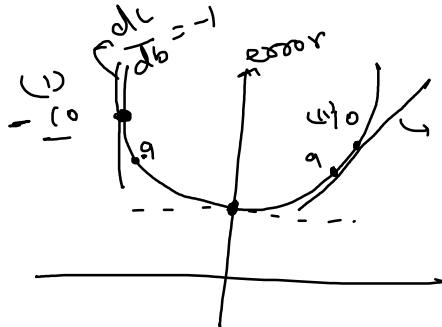
$$2) \frac{dL}{db} = \frac{d(y_i - \hat{y}_i)^2}{db} = \frac{d(y_i - mx_i - b)^2}{db}$$

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

$$3) b_{\text{next}} = b_{\text{old}} - \text{slope} \Rightarrow b_{\text{old}} - \left( \frac{dL}{db} \right)$$

$$b_{\text{old}} = 10$$

$$\text{(i)} \quad b_{\text{next}} = 10 - 1 \\ = 9$$



$$\frac{dL}{db} = -1$$

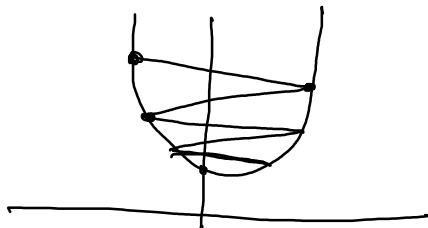
$$\text{(i)} \quad b_{\text{old}} = -10$$

$$\text{(i)} \quad b_{\text{next}} = -10 - (-1) \\ = -10 + 1 = -9 \quad \square$$

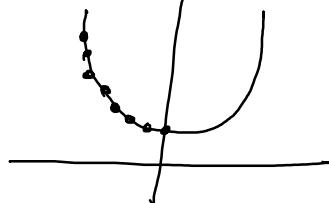
$$b_{\text{next}} = b_{\text{old}} - \eta \times \text{slope}$$

↓ learning rate (step size)

high value ( $\eta$ )



low value ( $\eta$ )

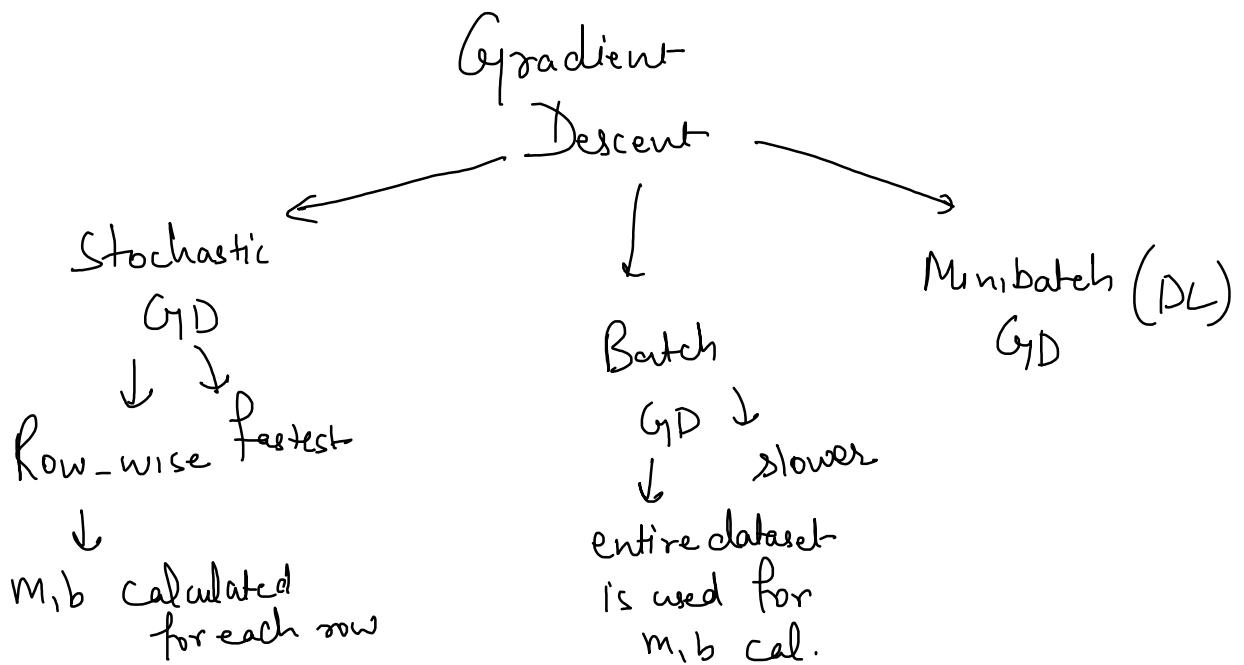


## Actual steps:

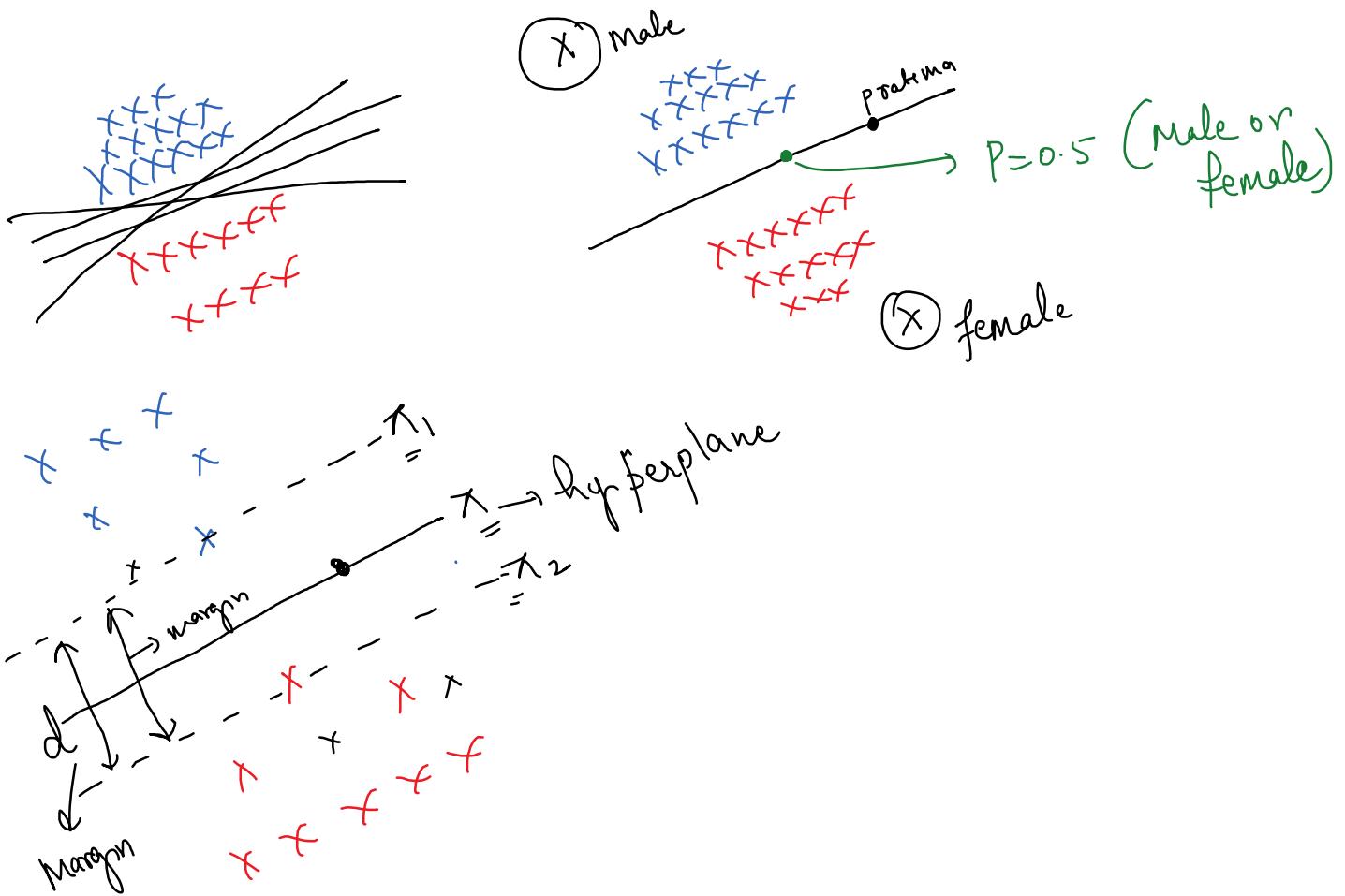
1) take random values of  $m, b$

2)  $\frac{\partial L}{\partial m}$  &  $\frac{\partial L}{\partial b}$   $\eta \Rightarrow$  hyperparameter

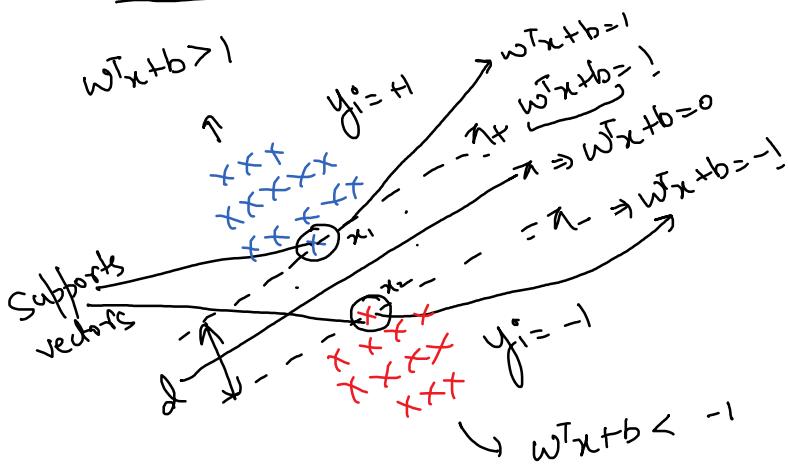
3)  $m_{\text{next}} = m_{\text{old}} - \eta \frac{\partial L}{\partial m}$  &  $b_{\text{next}} = b_{\text{old}} - \eta \frac{\partial L}{\partial b}$



# Support Vector M/c



## Mathematical Formulation:



$$w^T x > 0 \quad w^T x < 0$$

$$\begin{aligned} x_+ \Rightarrow w^T x_1 + b &= 1 \\ x_- \Rightarrow w^T x_2 + b &= -1 \end{aligned}$$

$$w^T(x_1 - x_2) = 2$$

Normalizing,

$$\left( \frac{w^T(x_1 - x_2)}{\|w\|} \right) = \frac{2}{\|w\|}$$

$$\text{Margin} = \frac{2}{\|w\|}$$

$$w^*, b^* = \underset{w}{\operatorname{argmax}} \left( \frac{2}{\|w\|} \right) \rightarrow \text{Hard margin SVM}$$

Building Constraints:

$$(I) \quad y_i(w^T x + b) = +1$$

$$(II) \quad y_i(w^T x + b) = -1$$

$$(III) \quad y_i(w^T x + b) > 1$$

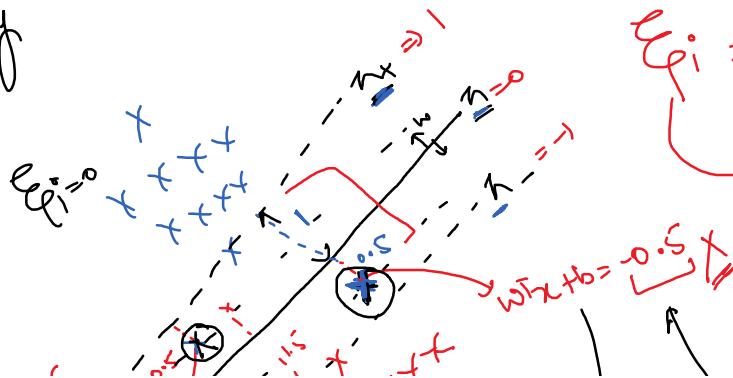
$$(IV) \quad y_i(w^T x + b) \geq 1$$

↓  
-ve

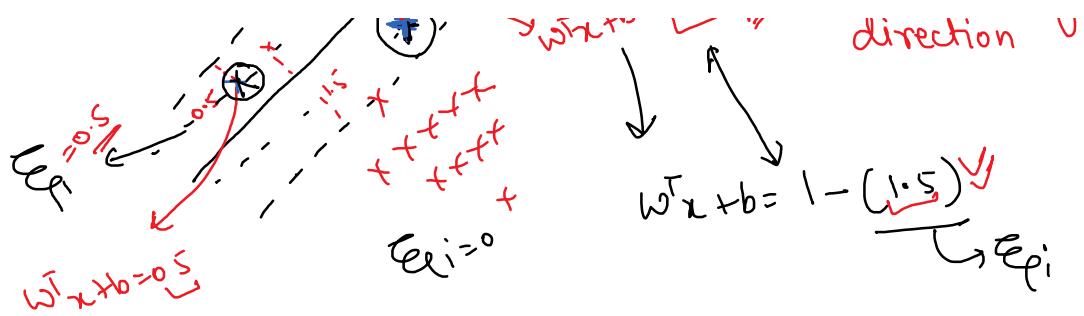
MOP  $w^*, b^* = \underset{w,b}{\operatorname{argmax}} \frac{2}{\|w\|}$  such that

$$y_i(w^T x_i + b) \geq 1$$

Reality



$e_{i+} \rightarrow$  represents misclassification  
distance of points from its  
correct hyperplane in obj.  
direction



$$w^*, b^* = \underset{w^*, b}{\operatorname{argmax}} \left( \frac{2}{\|w\|} \right) + C \sum_{i=1}^n \epsilon_i \text{ such that}$$

$$y_i(w^T x_i + b) \geq 1 - \epsilon_i$$

$$w^*, b^* = \underset{w, b}{\operatorname{argmin}} \frac{\text{regularizer}}{2} + C \sum_{i=1}^n \epsilon_i \rightarrow \text{soft margin}$$

loss

hyperparameter

1)  $C \uparrow \rightarrow$  less error  $\rightarrow$  overfitting

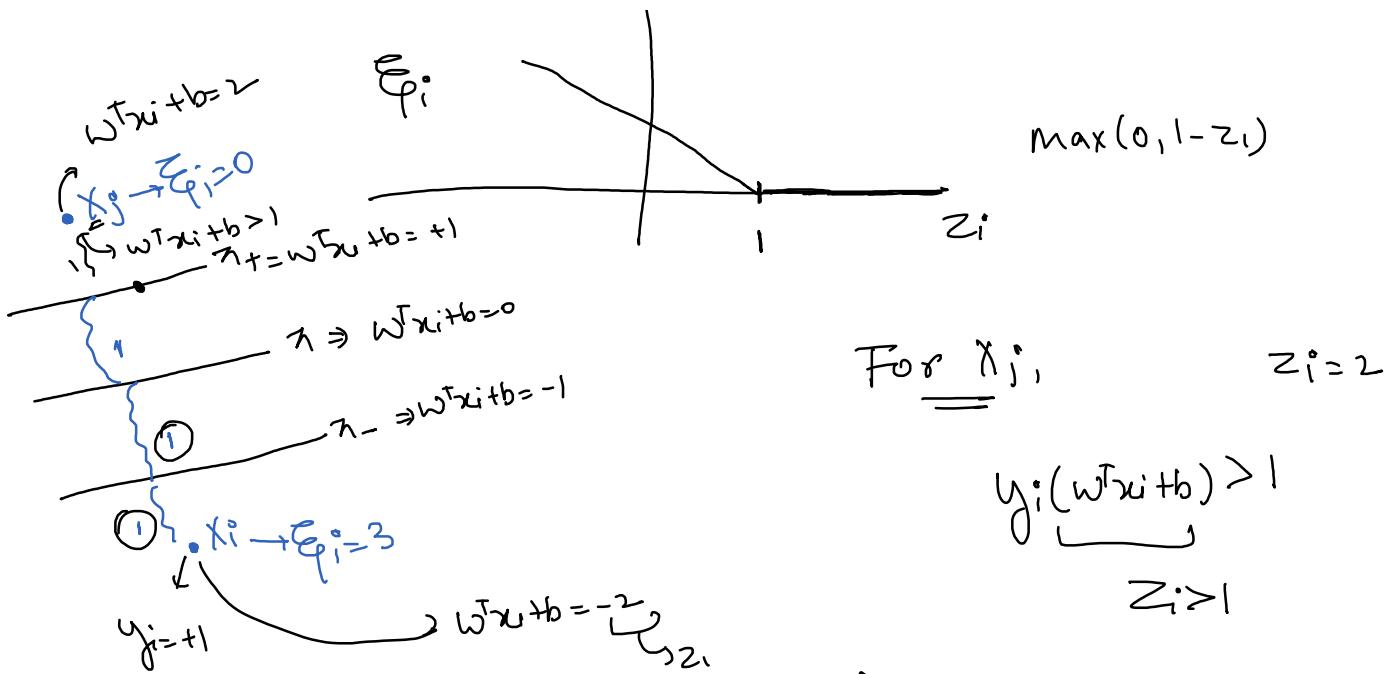
$$C \propto \frac{1}{\lambda}$$

2)  $C \downarrow \rightarrow$  more error  $\rightarrow$  underfitting

Loss Minimization : Hinge loss  $\Rightarrow \max(0, 1 - z_i)$

$$z_i = y_i(w^T x_i + b)$$





$$\text{For } \underline{x}_i, \quad y_i(w^T x_i + b) = 1 - (-2) = 3$$

$$\text{hinge loss} = \max(0, 1 - z_i)$$

$$\text{hinge loss} = \max(0, 3) = 3$$

$$\text{hinge loss} = \max(0, 1 - z_i)$$

$$= \max(0, 1 - 2)$$

$$= \max(0, -1) = 0$$

$$\boxed{\epsilon_i = 1 - z_i}$$

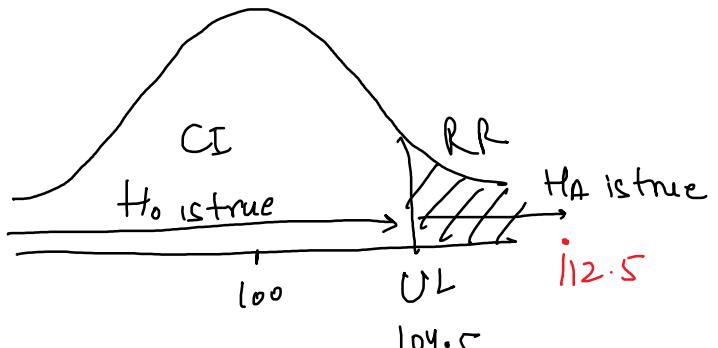
Q A principal claims that students have above average IQ. A random sample of 30 students is taken, mean = 112.5. The mean & std dev of population is 100 & 15. Test your hypothesis.

### Acceptance Region Method

- Sol.
- ①  $H_0: \mu \leq 100$
  - ②  $H_A: \mu > 100$
  - ③ Check for one-tailed / two-tailed test  
Right tailed test

$$\mu = 100, \sigma = 15, \bar{x} = 112.5, (\alpha = 0.05) \xrightarrow{\text{from z-table}} \text{z-score} = 1.65$$

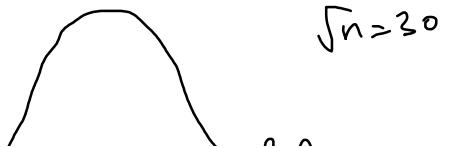
$$\begin{aligned} UL &= \mu + z \times \frac{\sigma}{\sqrt{n}} \\ &= 100 + 1.65 \times \frac{15}{\sqrt{30}} \\ &= 104.5 \end{aligned}$$



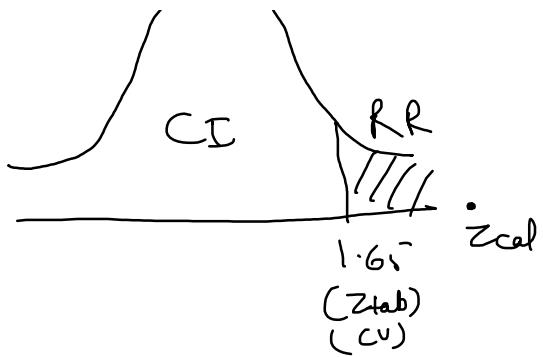
Reject  $H_0$

Q2 CV  $\alpha = 0.05 \rightarrow z = 1.65 \quad \bar{x} = 112.5 \quad \mu = 100 \quad \sigma = 15 \quad \sqrt{n} = 30$

$$Z_{cal} = \frac{\bar{x} - \mu}{\sigma / \sqrt{n}} = \frac{112.5 - 100}{15 / \sqrt{30}}$$



$$Z_{\text{cal}} = \frac{\bar{x} - \mu}{\sigma / \sqrt{n}} = \frac{112.5 - 100}{15 / \sqrt{30}} = 4.56$$

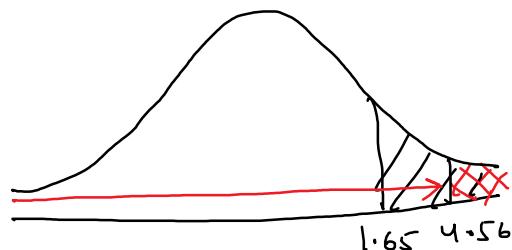


$Z_{\text{cal}} > Z_{\text{tab}} \Rightarrow \text{Reject } H_0$

P-value       $\alpha = 0.05$        $Z_{\text{tab}} = 1.65$

$$Z_{\text{cal}} = 4.56$$

$$P(Z_{\text{cal}} = 4.56) = 0.9999966$$



$$\begin{aligned} \text{P-value} &= 1 - P(Z_{\text{cal}} = 4.56) = 1 - 0.9999966 \\ &= 0.0000034 \end{aligned}$$

$$\alpha = 0.05$$

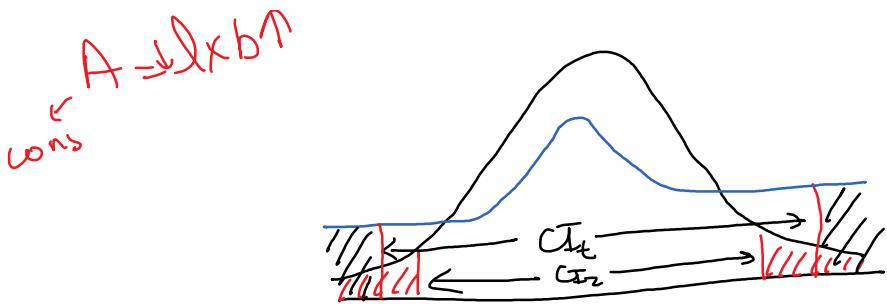
Compare p-value with  $\alpha$

$p < \alpha \rightarrow \text{Reject } H_0$

If sample size ( $n < 30$ )  $\rightarrow$  we don't use z-test  
we will t-test.

$A \rightarrow \Delta x b^{\uparrow}$





$\epsilon$ -dist

$$t = \frac{x - \mu}{\sigma/\sqrt{n}}$$

or

$$t = \frac{x - \mu}{s/\sqrt{n}} \rightsquigarrow S$$

Degrees of Freedom: logically independent values

$$\chi = \underline{\underline{7}}$$

$4 = \text{degrees of freedom}$

$2 \quad \quad \quad 4 \text{ independent values}$

$3 \quad \quad \quad \quad \quad$

$5 \quad \quad \quad \quad \quad$

$\emptyset \quad \quad \quad \quad \quad$

$x -$

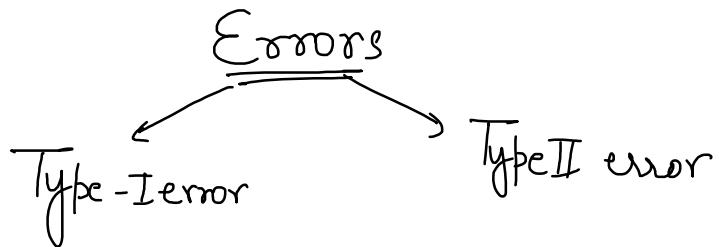
$$S = \frac{2 + 3 + 5 + 8 + n}{5}$$

$$\text{avg} = 5$$

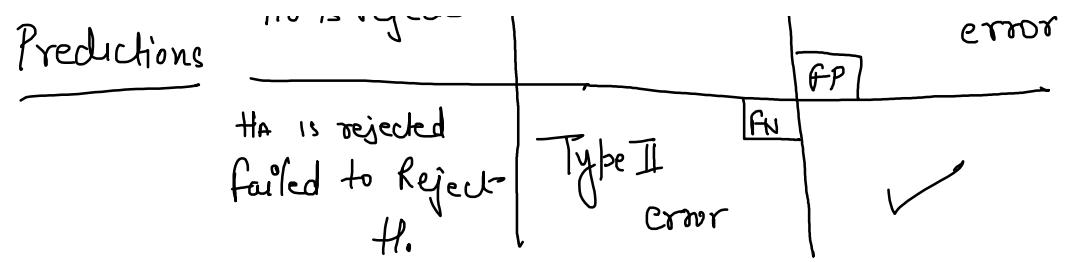
$$2S = 18 + n$$

$$\boxed{x = 7}$$

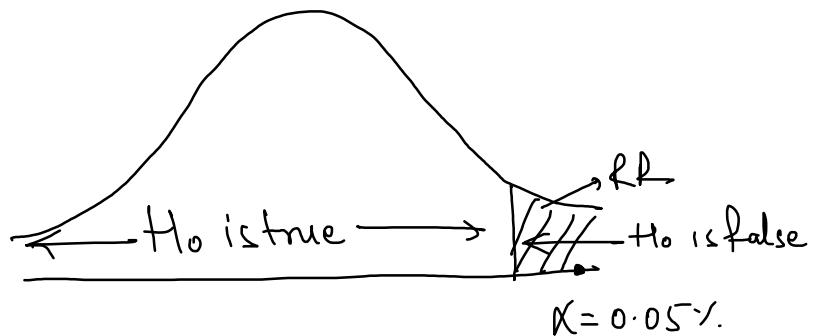
If you have  $n$  values,  $df = n - 1$



| <u>Predictions</u> | <u>Actual</u>  |               | $H_A$ is false<br>$H_0$ is true<br>Type-I error |
|--------------------|----------------|---------------|---|
|                    | $H_0$ is false | $H_0$ is true |   |
| $H_0$ is accepted  | ✓              |               |   |
| $H_0$ is rejected  |                |               | FP  |



Quantification of Type I error:



$$\text{Type I error} = \alpha \Rightarrow 5\%.$$

Quantification of Type II:

Power of test

↳ ability of test to make right decisions

Power of test  $\propto$  large no of sample

$$\text{Power} = 1 - \beta$$

$$\boxed{\beta = 1 - \text{Power}} \rightarrow \text{Type II error}$$

Relationship b/w Type I & Type II.

$$\text{Type I} \propto \frac{1}{\text{Type II}} \quad \text{explore} \left. \begin{array}{l} \\ \text{assignment} \end{array} \right\}$$

## CHI-SQUARE TEST

↗ Non-parametric (No distributions)  
 ↗ Character-character Situations  
 (categorical)

$$Df = (R-1) \times (C-1)$$

Q Is there a relationship b/w Gender & result?

| Result |        | Pass |      |
|--------|--------|------|------|
| Gender |        | Pass | Fail |
|        | Male   | 60   | 40   |
|        | Female | 24   | 32   |

Sol.  $H_0$ : There is no relationship b/w gender & result

$H_A$ : There is relationship b/w gender & result

| Result-Gender | Pass      | Fail      | Total      |
|---------------|-----------|-----------|------------|
|               | 60        | 40        | $= 100$    |
| Female        | 24        | 32        | $= 56$     |
|               | <u>84</u> | <u>72</u> | <u>156</u> |

Total males = 100

Total females = 56

Total passed = 84

Total failed = 72

Total = 156

Expected values:

$$EV_1 = \frac{\text{expected value}}{(\text{males who passed})} = \frac{\text{total males} \times \text{total passed}}{\text{total}}$$
$$= \frac{100 \times 84}{156} = 53.85$$

$$EV_2 = \frac{\text{expected value}}{(\text{females who passed})} = \frac{\text{total females} \times \text{total passed}}{\text{total}}$$
$$= \frac{56 \times 84}{156} = 30.15$$

$$EV_3 = \frac{\text{expected value}}{(\text{males who failed})} = \frac{\text{total males} \times \text{total failed}}{\text{total}}$$
$$= \frac{100 \times 72}{156} = 46.15$$

$$EV_4 = \frac{\text{expected value}}{(\text{females who failed})} = \frac{\text{total females} \times \text{total failed}}{\text{total}}$$
$$= \frac{56 \times 72}{156} = 25.84$$

expected values:

$$EV_1 = 53.85 \quad EV_2 = 30.15 \quad EV_3 = 46.15 \quad EV_4 = 25.84$$

| Gender  | Result | Pass               | Fail                                |
|---------|--------|--------------------|-------------------------------------|
| Males   |        | 53.85              | 46.15                               |
| Females |        | $\frac{30.15}{84}$ | $\frac{25.84}{72} = \frac{56}{156}$ |

Calculate  $\chi^2$ :  $\chi^2 = \frac{(Actual - expected)^2}{expected}$

$$\textcircled{I} \quad \frac{(60 - 53.85)^2}{53.85} = 0.7$$

$$\textcircled{III} \quad \frac{(40 - 46.15)^2}{46.15} = 0.81$$

$$\textcircled{II} \quad \frac{(24 - 30.15)^2}{30.15} = 1.25$$

$$\textcircled{IV} \quad \frac{(32 - 25.84)^2}{25.84} = 1.46$$

$$\begin{aligned} \chi_{\text{cal}}^2 &= 0.7 + 0.81 + 1.25 + 1.46 \\ &= 4.22 \end{aligned}$$

$$\chi_{\text{tab}}^2 \Rightarrow df = (R-1)(C-1) = (2-1)(2-1) = 1 \times 1 = 1$$

$$\alpha = 0.05$$

$$\sqrt{2}$$

$H_1 - \text{vs vs}$

$$\chi^2_{\text{tab}} = 3.841$$

Compare  $\chi^2_{\text{tab}}$  with  $\chi^2_{\text{cal}}$

Reject  $H_0$

$$\chi^2_{\text{tab}} < \chi^2_{\text{cal}}$$

$$3.841 < 4.22$$