M Tu W Th F S Su

Standardization

original value

$$St. = \frac{x - \mu}{\sigma}$$

μ → mean

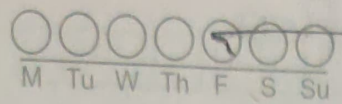
σ → standard deviation of original values

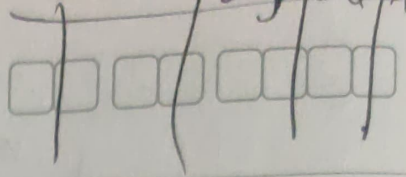
Day	$ fx rates		Daily trading vol
1	1.3	0.07	110000 (-0.25)
2	1.34	0.96	98700 (-0.85)
3	1.25	1.03	135000 1.1

Mean = 1.3

Std = 0.065

"Clustering"





① import library

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sn
from sklearn.cluster import KMeans
```

② Load the data → data = pd.read_csv('country_data.csv')

plot it → plt.scatter(data['Longitud'], data['Latitu'])
plt.show

③ data_with_features = data.iloc[:, 1:3]

④ kmeans = kmeans(2)
kmeans.fit(data_with_features)

⑤ identify_cluster = kmeans.fit_predict(data_with_features)

To print it → data_w_c = data.copy()
data_w_c['feature'] = identy_cluster

data_w_c → it will show with new feature

plt.scatter(data_w_c['Longitu'], dat['Latit'],
c = data_w_c['featur'], cmap = 'rainbow')

plt.show.

M Tu W Th F S Su

Cost function (Loss or error funct^{n})

→ measure of how well your algo models your dataset.

→ If your prediction too much deviate from actual results then cost is high

$$J(\beta) = \frac{1}{2m} \sum_{j=1}^{m} (y^{j} - \hat{y}^{j})^2$$

Gradient Decent

↳ Optimized algorithm based on continuous improvement toward reducing cost.

Regularization

M Tu W Th F S Su

is a way to reduce model overfitting and variance

1. Requires some addⁿ bias
2. Requires a search for optimal penality hyperparameter

Types of regularizatⁿ

1. L1 → Lasso regression
2. L2 → Ridge Regression
3. L1+L2 → Elastic Net

L1 regaliz² adds a penality eqⁿ to the absolute value of the magnitude of coefficient.

1. limit the size of coef.
2. can yeild sparse model

M Tu W Th F S Su

L2 regularizⁿ adds a penalty equal to square of the magnitude of coefficient.

(1) All coefficient shrink by the same factor

(2) Does not necessarily eliminate coefficient

feature scaling

Standardizⁿ : Rescale data to have a mean (μ) of 0 and standard deviatⁿ (σ) of 1.

$$X_{changed} = \frac{X - \mu}{\sigma}$$

Normalizatⁿ : Scales all values to be between 0 and 1

$$X_{changed} = \frac{X - X_{min}}{X_{max} - X_{min}}$$

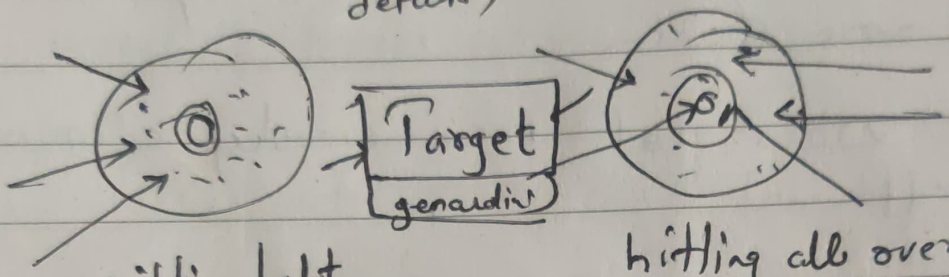
M Tu W Th F S Su

Cross Validation

Dartboard example-

Bias (too less details)

Variance (Too much details)



always hitting left of target (underfitting)

hitting all over dartboard some hit some mis (overfitting)

In machine learning we're trying to find a spot between bias and variance

predicting Movie Genre

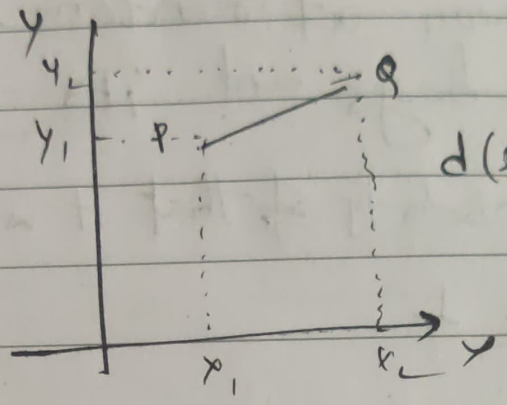
imdb Rating	Duration	Genre
8.0 (Mission Impos	160	Action
6.2 (Gadar 2)	170	Action
7.2 (Rocky & Rani	168	Comedy
5.2 (OMG 2)	155	Comedy

Calculate genre of Barbie with IMDB 7.4 & D=114

(1) K-nearest neighbors algo (K-NN)

M Tu W Th F S Su

step 1 - calculate Distance euclidean distance b/w the new movie & old



$$d(p,q) = \sqrt{(x_2-x_1)^2 + (y_2-y_1)^2}$$

predict barbie movie with is 7.4 and 114

Distance to (8.0, 160) =

MI to barbie

$$\sqrt{((7.4-8.0)^2 + (114-160)^2)} = \sqrt{0.36+2116} = 46.00$$

Gadar to barbie = 56.01

Rocky & Rani to Barbie = 54.00

OMG 2 to Barbie = 41.00 ← (41)

step2 selet nearest neighbor which is (41)

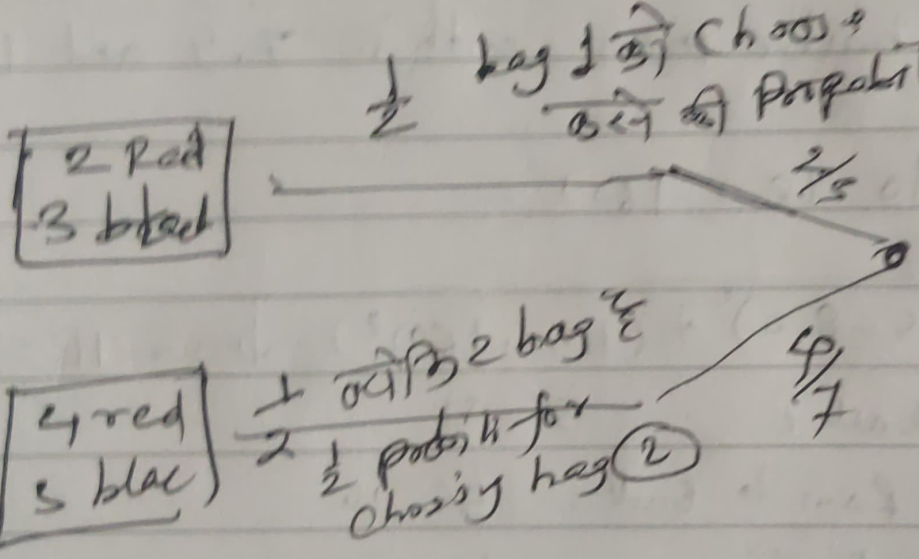
so barbie is **comedy** movie

K=1 तोने से

if we to **3** (K=) nearer neighbor for voting कर लो (classification)

Bayes' Theorem

Bag 1 have 2 Red 3 Black balls
Bag 2 have 2 Red 4 Black ball
One ball drawn at random from one of these bags and it's red
find the probability that it is drawn from bag 1



इसके Red होने की probability = $\frac{2}{2+3} = \frac{2}{5}$

Total probability of red event = $\frac{1}{2} \times \frac{2}{5} + \frac{1}{2} \times \frac{4}{7}$

$= \frac{2}{10} + \frac{4}{14} = \frac{1}{5} + \frac{2}{7} = \frac{7+10}{35}$

$\frac{17}{35}$

now come to Bayes theorem it is

M Tu W Th F S Su

a reverse problem

इसमे bag के बारे मे बुछ है कि कि bag से आई है ball

$$P\left(\frac{B_1}{Red}\right)$$

$$P\left(\frac{X}{X}\right) = \frac{P(X|Y) * P(Y)}{P(X)}$$

this is naive baise algo.

$$P(y|X_1, X_2 X_n) = \frac{P(X_1|Y) * P(P_2|Y) \cdots}{P(X_1) * (P_n(X_n)}$$

step 1 prior probability.

step 2 conditional probability.

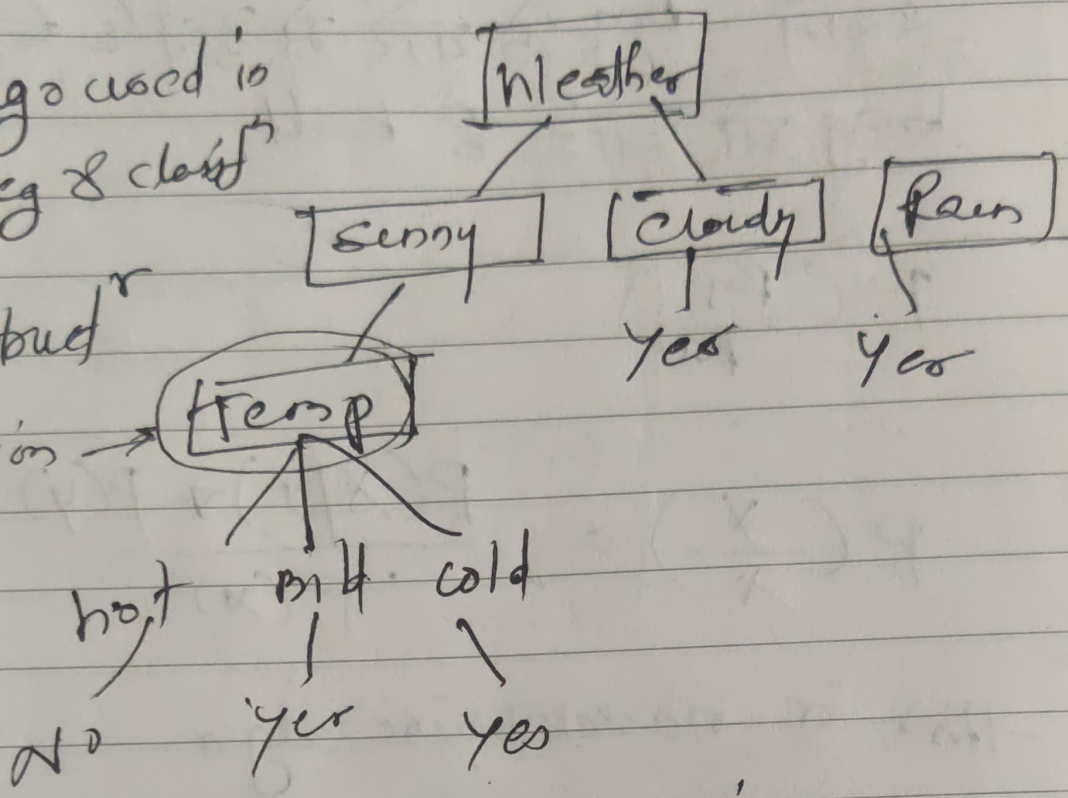
Decision Tree

M Tu W Th F S Su

↓

* ML algo used in both reg & classf"

* Tree structure
* Decision node



* entropy and information gain
* pruning

IDS & Random forest

Day	Weather	Temp	Humidity	Wind	Play?

Root node ?

for this you have to use entropy information gain on weather

step 1 – Entropy of full dataset

$$S[+9, -5] = \frac{-9}{14}\log_2\frac{9}{14} - \frac{5}{14}\log\frac{5}{14}$$

(yes, no)

$$= 0.94$$

step 2 → calculate attribute of Attributes

entropy of sunny $\{+2, -3\}$

$$= -\frac{2}{5}\log\frac{2}{5} - \frac{3}{5}\log\frac{3}{5} = 0.97$$

entropy of $\{-4, -0\} = 0$ (cloudy)

entropy of Rain $= \{-3, -2\} = 0.97$

IG = entropy of full ds $- \frac{5}{14}(\text{ent}(S)) - \ldots$

$= 0.296$

Root node ?

for this you have to use entropy

information gain on weather

step 1 – Entropy of full dataset

$$S[+9, -5] = \frac{-9}{14}\log_2\frac{9}{14} - \frac{5}{14}\log\frac{5}{14}$$

(yes, no)

$$= 0.94$$

step 2 → calculate attribute of Attributes

entropy of sunny $\{+2, -3\}$

$$= \frac{-2}{5}\log\frac{2}{5} - \frac{3}{5}\log\frac{3}{5} = 0.97$$

entropy of cloudy $\{-4, -0\} = 0$

entropy of Rain $= \{-3, -2\} = 0.97$

IG = entropy of full ds $- \frac{5}{14}(\text{ent}(S)) - \ldots$

$= 0.296$

M Tu W Th F S Su

conditional probability

$$P\left(\frac{E}{F}\right) = \frac{P(E \cap F)}{P(F)}$$

Tossing a coin 3 times (8)

S = {HHH HHT HTH HTT THH THT TTH TTT}

P(E) at least two tails = $\frac{4}{8}\left(\frac{1}{2}\right)$

P(F) ..

condition - find the probability of two tail and first coin should be head

M Tu W Th F S S

Date (1)

Supervised Learning

Classification
Logistic Reg
Naive Baise
Linear Descriminant Analysis (LDA)

Regression
Linear
Ridge
Lasso

Decision tree
Random forest
SVM (Support vector Machine)
k-nearest neighbor (k-NN
gradient Boosting algo
Neural network.

(2)

Unsupervised Learning

↓

PCA
K-means clus
Hirarical Cluster
DB Scan

M Tu W Th F S S

Supervised Learning (1)

Unsupervised Learning (2)

Supervised Learning

- Classification
  - Logistic Reg
  - Naive Bayes
  - Linear Discriminant Analysis (LDA)
  - Decision tree
  - Random forest
  - SVM (Support vector Machine)
  - k-nearest neighbor (k-NN
  - gradient Boosting algo
  - Neural network.
- Regression
  - Linear
  - Ridge
  - Lasso

Unsupervised Learning

- PCA
- k-means clus
- Hierarical Cluster
- DB Scan

price : fair.

mail : other

mean = 4.

price	Yes	No	P(y)	P(N)
low	2	2		
→ fair	2	2	2/6	
over	2	1		
	6	5		

$\frac{3}{6}$ $\frac{2}{6}$ $\frac{2}{6}$ $\frac{2}{5}$ $\frac{3}{5}$ $\frac{1}{5}$

mail	yes	No	P(y)	P(N)
Other	1	1	1/6	1/5
avg	3	1	3/6	1/5
ex	2	3	2/6	3/5

M Tu W Th F S Su

Safe	Ys	No	P(y)	
Yes	5	3	5/6	3/5
No	1	2	1/6	2/5

$$P(\text{pric}) = \frac{3}{6} \times \frac{1}{6} \times \frac{5}{6} \times \frac{8}{11}$$

$$\frac{1}{3} \times \frac{1}{6} \times \frac{5}{11} \qquad \frac{5}{88}$$

$$\frac{5}{18 \times 11} = \boxed{\frac{5}{198}} \qquad \boxed{.025}$$

$$1 - .025 = 0.97$$

$$\frac{6.0.25}{} + 0.975$$

M Tu W Th F S Su

Get the same for No

for normalized

① $\frac{P(\text{yes})}{P(N) + P(\text{yes})}$

② $\frac{P(\bar{N}O)}{P(Y) + P(N)}$

if ① is greater than ② then it is beneficial

M Tu W Th F S Su

KNN

IMDB Rating	Duration	Genre
8.0	160	Act
6.3	170	Act
7.2	168	comedy
8.2	155	comedy

predict the value of IMDB 7.4 and duration 114

(1) calculate