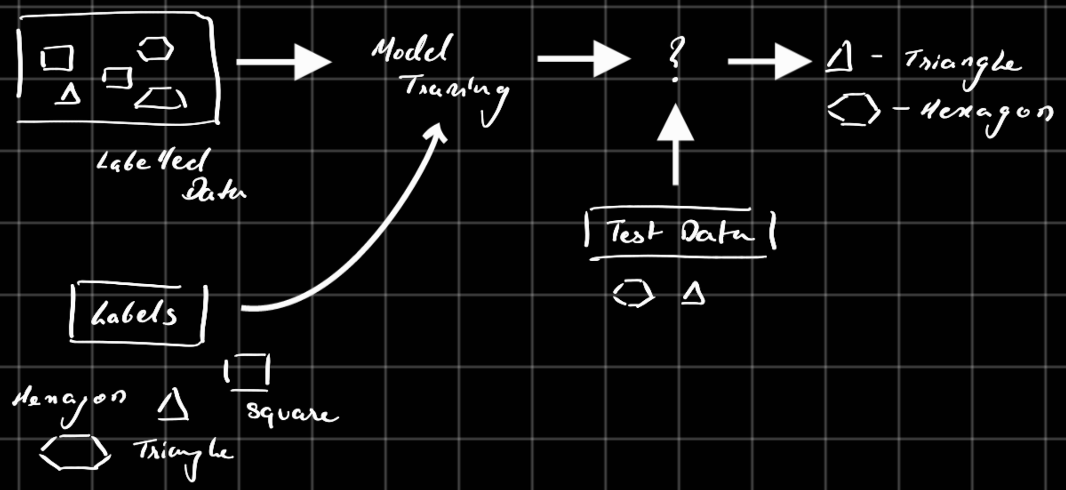
Machine Learning →

It is a branch of AI that enables the computer to self learn and improve over time.

It detects pattern's in data and learn from them.



Steps using ML →

1. Collecting Data
2. Preparing the Data
3. Traning the model
4. Evaluating the Data
5. Improving the Data

Applications of Machine Learning

1. Health Care
2. Fraud Detection
3. Recomendations
4. Online Search
5. Natural Language Processing
6. Smart Car's
7. Banking and Financial Services
8. Retail.

Machine Learning Life cycle →

1. Data gathering
2. Data Preparedness → Exploration, Pre processing
3. Data Wrangling
    - Repetation of data
    - Incorrect / Incomplete data
    - outliers
4. Analyse
5. Traning the Model
6. Testing the Model
7. Deployment

→ Gathering Data

Identify various data source.

Collect data

Integrate the data obtained from different sources.

By performing this we get a coherent set of data, Data Set.

→ Data preparation

Data exploration

To understand the nature, characteristics, format and quality of data.

We find correlation, general trend and outliners.

Data preprocessing

Processing of data to analyse it.

→ Data wrangling

Process of cleaning and converting raw data into useable format.

Missing values

Duplicate data

Invalid data

Noise

Corrupt / Meaningless data.

→ Data Analysis

Selection of analytical techniques

Building Model

Review the result

Classification, Regression, Cluster analysis, Association.

→ Train Model

→ Test Model

→ Deployment

Machine Learning

Supervised

Task Driven

1. Classification
2. Regression

1. linear regression for regression problem's.
2. support vector machine for classification problem.

Un-Supervised

Data Driven

1. Clustering
2. Association
3. Dimensionality Reduction

1. k-mean's for clustring problem's.
2. PCA algorithm is a feature Extraction approach.

Reinforcement

True AI

S/w component learn to react to an environment

Semisupervized →

Having a large sum of data out of which are few data are labled and most of it is unlabled.

Supervised learning → ex house price.

In supervised learning, the goal is to learn the mapping between a set of inputs and outputs.

Being able to adapt to new input's and make prediction is the crucial generalisation part of machine learning.

① Classification →

It is used to group the similar data points into different section's in order to classify them.

Decision boundaries →

Line drawn between classes.

Binary Classification →

Those classification tasks that have two class labels.

ex → Email spam detection (spam or not)

Decision Surface →

The entire area that is chosen to define a class

Multi-Class Classification →

Classification tasks having more than two class labels.

ex → Face classification.

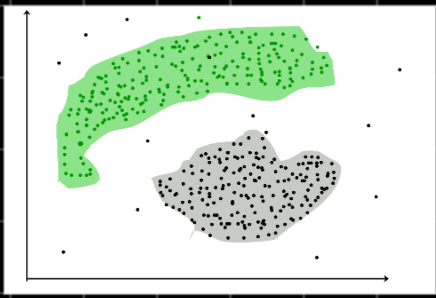
Multi-Label Classification →

Classification tasks that have two or more class labels

② Regression →

The difference between classion and regression is that regression is that regression outputs a number rather than a class.

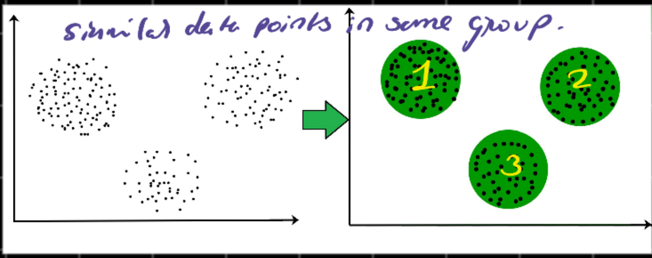
Unsupervised learning →



- Only input data is provided in the examples.
- There are no labelled example output to aim for.

① Clustering →

Is the task of dividing data points into a number of groups, such that data points in same group are more similar to data points in the same group.



② Association →

- In this you want to uncover the rules that describ your data.

Semi-Supervised Learning →

- Mix between survised and un-supervised approaches.
- The learning process isn't closely supervised with example outputs for every single input. also don't let algorith do its own thing and provide no form of feadback.

- Mix small amount of labelled data with large unlabeled dataset reduces the burden of having enough labelled data.

Reinforcement Learning →

Occasional positive and negative feedback is used to reinforce behaviours.

Reward-motivated behaviour is key in reinforce behaviour.

- Algorithm just aim's to maximise its rewards by playing the game over and over again.
- If you can frame a problem with a frequent 'score' as a reward, it is likely to be suited to Reinforcement learning.

Quantitative data →

Data which can be put into categories, measured or ranked.

ex :- Length, age, weight, cost.

Categorical data →

Data that has been placed into groups. An item cannot belong to more than one group at a time.

Chi-square test is used for analysis used for categorical data.

$$\chi^2 = \frac{\Sigma(\text{observed} - \text{expected})}{\text{expected}}$$

Continuous data →

Numerical data measured on a continous range or scale.

ex :- Person's height or weight and temprature

Qualitative data →

Data that uses words and descriptions.

- Qualitative data can be observed but is subjective and therefore difficult to use for making comparisons.

Nominal data →

Is used for naming or labelling variable's, without any quantitative value.

- Race have a number of categories, but there is no specific way to order from highest to lowest and vice-versa.

2. **Nominal**

Nominal data is made of discrete values with no numerical relationship between the different categories — mean and median are meaningless. Animal species is one example. For example, pig is not higher than bird and lower than fish.

Ordinary data →

Categorical data with an order

- The variables in ordinal data are listed in an ordered manner.
- The numbers are not mathematically measured or determined but are merely assigned as labels for opinion's.

Exploratory Data Analysis →

steps of Data exploration and Preparation

- Variable Identification
- Univariate Analysis
- Bivarite Analysis
- Missing values treatment
- Outliner treatment
- variable transformation
- Variable creation

Variable Identification →

Types of variable →

Predictor Variable : Gender, Prev_Exam_Marks, Height(cm), weight(kg)

Target variable : Play_Cricket

Data type →

Character : Student_id, Gender

Numeric : Prev_Exam_Marks, Height(cm), weight(kg), Play_Cricket.

Variable Category →

Categorical : Student_id, Gender

Continuous : Pre_Exam_Marks, Height(cm), weight(kg)

Uni-Varient →

Meathod to perform uni-varient analysis will depend on wheather the variable type is categorical or continuous.

Bi-Varient →

Finds out the relationship between two variables.

Missing Value Treatment

Why missing values treatment is required?

Missing data in the training set can lead to a biased mode because we have not analysed the behaviour and relationship with other variables correctly.

Name	Weight	Gender	Play Cricket/ Not
Mr. Amit	58	M	Y
Mr. Anil	61	M	Y
Miss Swati	58	F	N
Miss Richa	55		Y
Mr. Steve	55	M	N
Miss Reena	64	F	Y
Miss Rashmi	57		Y
Mr. Kunal	57	M	N

Gender	#Students	#Play Cricket	%Play Cricket
F	2	1	50%
M	4	2	50%
Missing	2	2	100%

Name	Weight	Gender	Play Cricket/ Not
Mr. Amit	58	M	Y
Mr. Anil	61	M	Y
Miss Swati	58	F	N
Miss Richa	55	F	Y
Mr. Steve	55	M	N
Miss Reena	64	F	Y
Miss Rashmi	57	F	Y
Mr. Kunal	57	M	N

Gender	#Students	#Play Cricket	%Play Cricket
F	4	3	75%
M	4	2	50%

Why data has missing values?

- Data extraction
- Data collection
- Missing completely at random
- Missing at random

How to handel missing or corrupted data in data set?

- Deletion →

List wise deletion and Pair wise deletion

deleted

List wise deletion

Gender	Manpower	Sales
M	25	343
~~F~~	~~.~~	~~280~~
M	33	332
~~M~~	~~.~~	~~272~~
~~F~~	~~25~~	~~.~~
M	29	326
	~~26~~	~~259~~
M	32	297

Pair wise deletion

Gender	Manpower	Sales
M	25	343
F	~~.~~	280
M	33	332
M	~~.~~	272
F	25	~~.~~
M	29	326
	26	259
M	32	297

- Mean / Median / Mode Imputation →

Filling the values with estimate one's.

i) Generalized Imputation →

we calculate the mean or median for all non missing values of that variable then replace missing value with mean or median.

avg. of all non missing values

List wise deletion

Gender	Manpower	Sales
M	25	343
F	z	280
M	33	332
M	z	272
F	25	.
M	29	326
	26	259
M	32	297

Pair wise deletion

Gender	Manpower	Sales
M	25	343
F	.	280
M	33	332
M	.	272
F	25	.
M	29	326
	26	259
M	32	297

avg = z

ii) Similar case Implementation →

we calculate avg for gender 'Male', 'Female' of non missing values.

avg →

avg →

List wise deletion

Gender	Manpower	Sales
M	25	343
F	.	280
M	33	332
M	.	272
F	25	.
M	29	326
	26	259
M	32	297

Pair wise deletion

Gender	Manpower	Sales
M	25	343
F	.	280
M	33	332
M	.	272
F	25	.
M	29	326
	26	259
M	32	297

- Prediction Model →
  - We divide our data set into (i) without missing data (ii) with missing data.
  - (i) data as training data set and the (ii) data set as test data set where the missing values are treated as target variable.

Draw backs →

1. The estimated values are usually more well-behaved than true values.
2. If no relationship with attribute in the data set and the attribute with missing values, then the model will not be precise for estimating missing values.

- KNN Imputation →

The missing values of an attribute are imputed using the given number of attributes that are most similar to the attribute whose values are missing.

- The similarity of two attributes is determined using distance function.

Advantages →

- K-nearest neighbour can predict both qualitative and quantitative attributes.
- Attributes with multiple missing values can be easily treated.

Disadvantages →

KNN algorithm is very time consuming in analyzing large dataset.

Regression Analysis →

Is a method to model a relationship between dependent and independent variables with one or more variable.

Regression helps us to understand how the value of dependent variable is changing corresponding to an independent variable.

It predicts temp, age, salary etc.

Dependent → Target

Independent → Predict

Dependent And Independent Variable →

$a + 9b = c$

Here 'a' and 'b' are independent can have any value, does not depend on any thing.

Here 'c' depends on 'a' and 'b' Dependent variable.

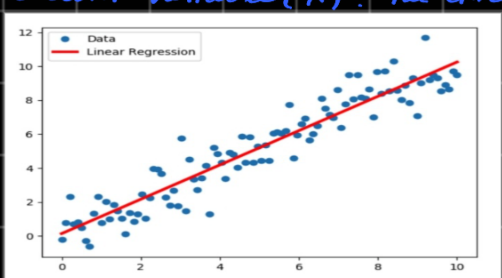
ex :- $a = 2$, $b = 10$ } Independent

$\therefore c = 92$ ~ Value depends on 'a' and 'b'.

Types of Regression Analysis →

Linear Regression →

Assumes a linear connection between a dependent variable (Y) and independent variable (X). The linear connection is defined as $Y = mx + c + e$.

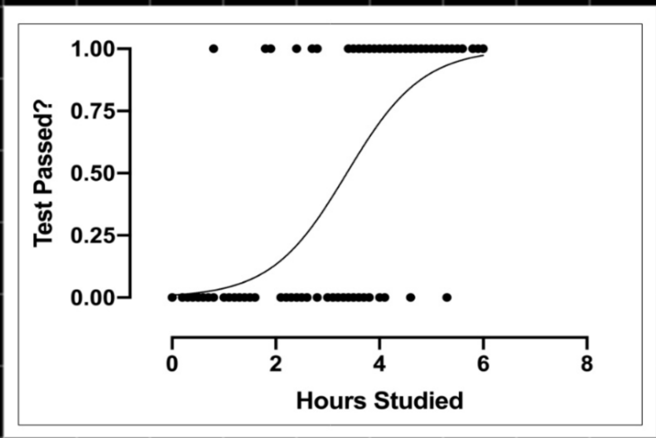


Mean Square Error $= \dfrac{\Sigma(\text{obsn} - \text{actual})^2}{n}$

Logistic Regression →

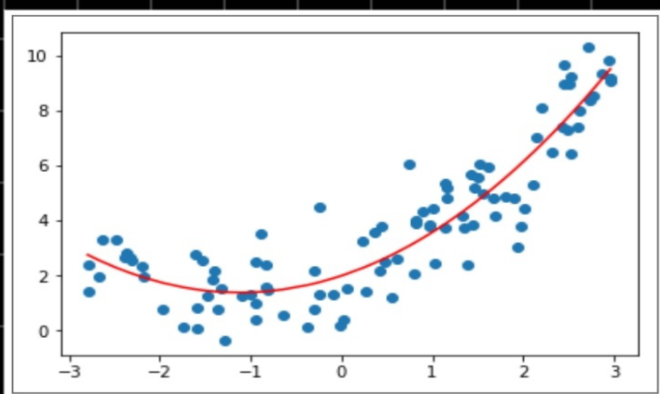
When the dependent variable is discrete. This technique is used to compute the probability of mutually exclusive occurances such as Pass/Fail. 0/1

The target target variable can take on only one of two values, and a sigmoid curve represents its connection to the independent variable, and probablity has a value between 0 and 1.



Polynomial Regression →

Is used to represent a non-linear relationship between dependent and Independent variables.

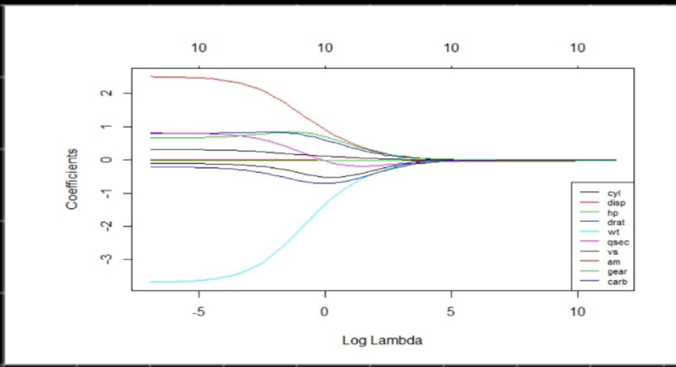


Ridge Regression →

When data exhibits multicollinearity, the ridge regression technique is applied when the independent variables are highly correlated.

The 'λ' variable in the ridge regression equation resolves the multicollinearity problem.

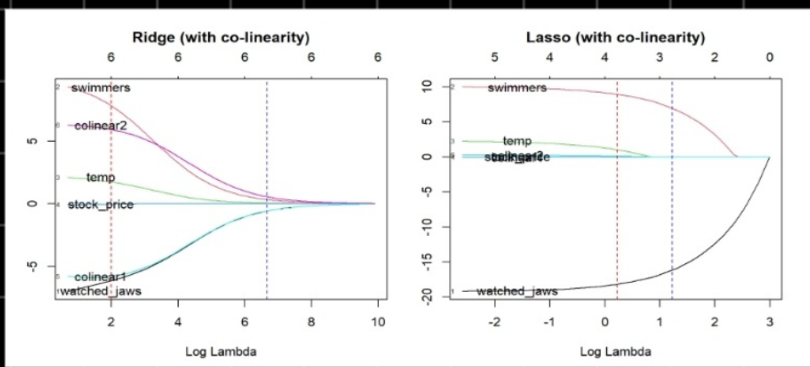
$$= \underset{\beta \in R^p}{\arg\min} \underbrace{\|y - X\beta\|_2^2}_{Loss} + \underbrace{\lambda\|\beta\|_2^2}_{Penalty}$$



Lasso Regression →

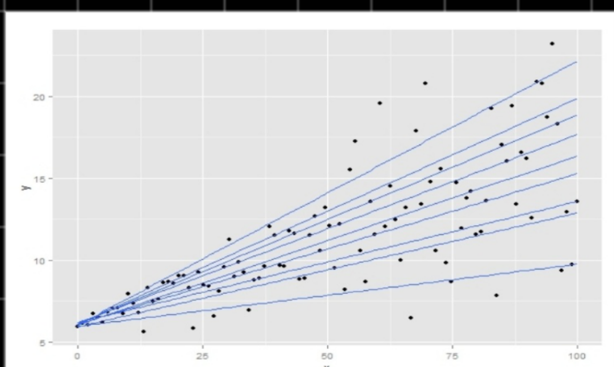
As ridge regression, Lasso technique penalizes the absolute magnitude of the regression coefficient.

Lasso regression technique employs variable selection, which leads to shrinkage of coefficient values to absolute zero.



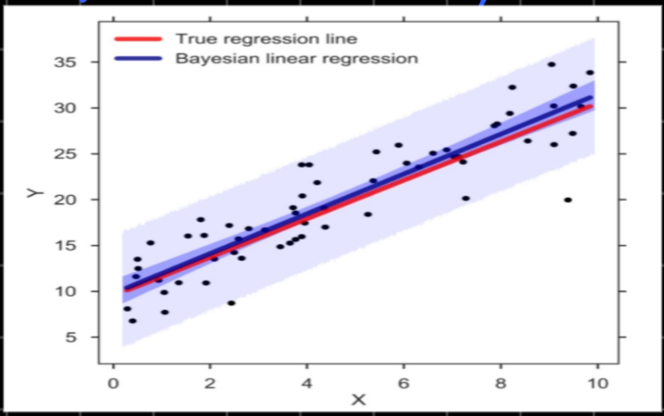
Quantile Regression →

- subset of linear regression technique.
- It is employed when the linear regression requirements are not met or when the data contains outliers.



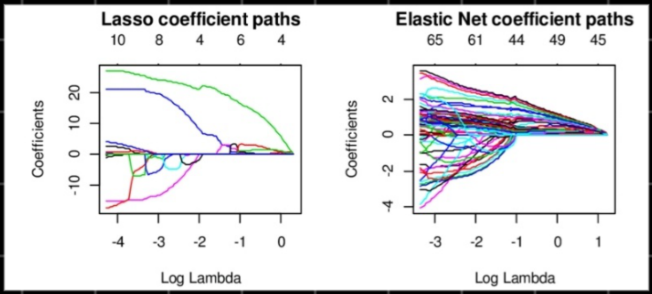
Bayesian Linear Regression →

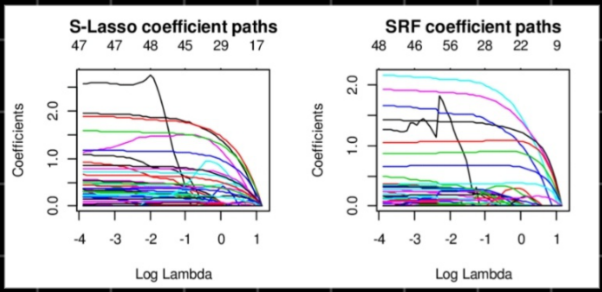
- Uses Bayes theoram to calculate the regression coefficients' values.
- This technique determines the features' posterior distribution. The approach outperform's ordinary linear geression in term's of stability.



Elastic Net Regression →

It combines ridges and lasso regression techniques that are partially useful when dealing with strongly corelated data. It regulizes regression models by utilizing the penalties associated with ridge and lasso regression method.

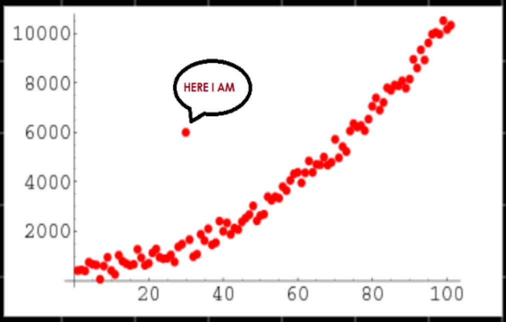




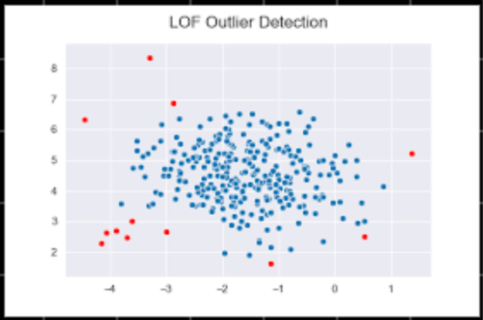
Outliers →

Values that look different from the other values in the data.

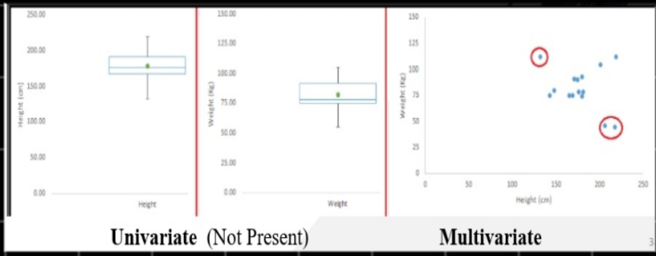
Univariate



Multivariate



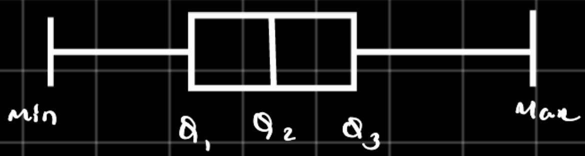
- Types of Outlier's →
  - Univariate
  - Multivariate



Univariate (Not Present) Multivariate

- **Outlier Detection** →

Box-plot, Histogram, Scatter Plot.



Box plot used to find the outliners.

$< Q_1 - IQR(1.5)$

$> Q_3 - IQR(1.5)$

$IQR - Q_3 - Q_1$

Binning →

It is used to categorize variables.

- **Outlier's Treatment** →
  - Deleting outlier values.
  - Using mean, median, mode imputation method.

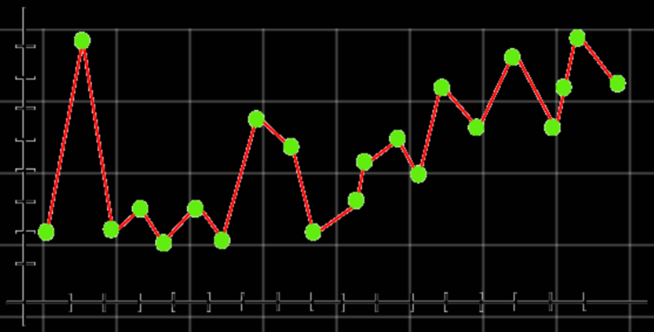
Multicollinearity →

When two or more independent variables are highly corelated with one another in a regression model.

Overfitting →

Occurs when our machine learning model tries to cover all the data points or more than the required data points present in the given data set.

Because of this, model starts catching noise and inaccurate values present in the data set, all these factors reduce the efficiency and accuracy of the model.
The overfitted model has low bias and high variance.



Mean → $\frac{\Sigma a_i}{n}$

Median → middle value

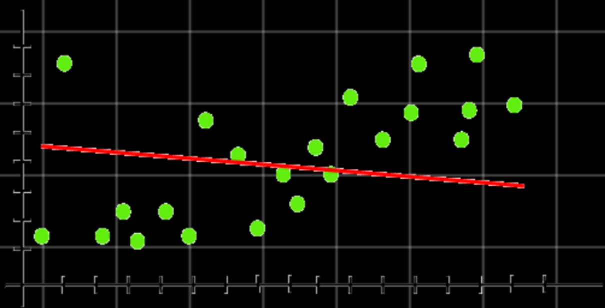
Mode → most occurance

How to avoid Overfitting →

- Cross Validation
- Traning with more data
- Removing features
- Early stopping training
- Regulization
- Ensembling

Underfitting →

When machine learning model is not able to capture the underlying trend of the data.



How to avoid underfitting →

- By increasing the training time of model.
- By increasing the number of features.

Signal →

Refers to the true underlying pattern of the data that helps the machine learning model to learn from the data.

Noise →

Unnecessary and irrelevant data that reduces the performance of the model.

Bias →

Prediction error that is introduced in the model due to oversimplifying the machine learning algorithms.

Or the difference between the predicted values and the actual values.

Variance →

If machine learning model performs well with the training dataset, but does not performs well with the test dataset, then varience occurs.

Residual sum of squares →

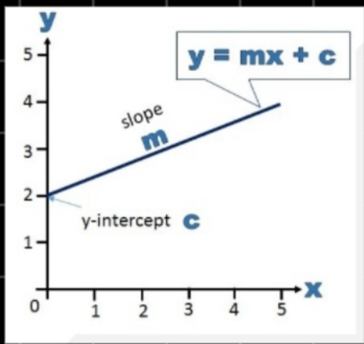
$$RSS = \Sigma(\text{actual output} - \text{predicted output})^2$$

$$RSS = \sum_{i=1}^{n} (y_i - (mx_i + c))^2$$

y = dependent variable
x = independent variable
m = slope of the line
c = y-intercept of the line
n = number of samples.

Formula for linear regression →

$$y = mx + c$$



y = dependent variable
x = independent variable
m = slope of the line
c = y-intercept of the line.

$$\underline{m\ (\text{slope})} = \frac{n\Sigma xy - (\Sigma x)(\Sigma y)}{n\Sigma x^2 - (\Sigma x)^2}$$

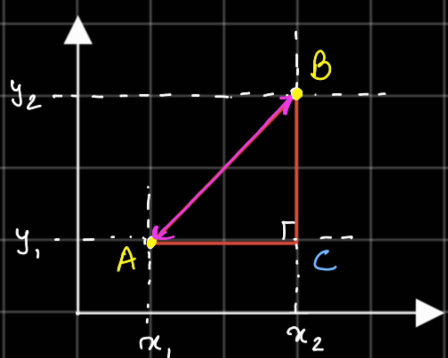
$$\underline{c\ (\text{intercept})} = \frac{\Sigma y \Sigma x^2 - \Sigma x \Sigma xy}{n\ \Sigma x^2 - (\Sigma x)^2}$$

y = dependent variable
x = independent variable
m = slope of the line
c = y-intercept of the line
n = number of samples

$$\frac{n\ \Sigma xy - \Sigma x \Sigma y}{n\ \Sigma x^2 - \Sigma x^2}$$

Distance Metrices →

- Eucledian →
  - It is the displacement between two points.
  - Calculated using "Pythagores Theoram".



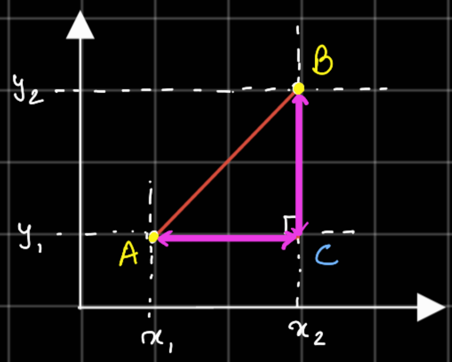
$$AB^2 = AC^2 + BC^2$$

$$= (x_2 - x_1)^2 + (y_2 - y_1)^2$$

$$AB = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2}$$ — for 2D plane.

$$AB = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2 + (z_2 - z_1)^2}$$ — for 3D plane

- Manhatton →
  - The distance between the two points measured along x-axis and y-axis.



$$\underbrace{AC + BC}_{\text{Manhattan}} = |(x_2 - x_1)| + |y_2 - y_1|$$

Types of Statistics →

- Descriptive Statistics →

  Is a way to organise, represent and describe a collection of data using tables, graphs and summary measures.

  ex :- The collection of people in a city using the internet or using Television.

Types of Descriptive Statistics →

- Measure of frequency

  frequency measurement displays the number of times a particular data occurs.

- Measure of dispersion

  range, variance, standard deviation are measure of dispersion.

- Measure of central tendency

  central tendency are mean, median, mode of the data.

- Measure of position

  measure of position describes the percentile and quartile ranks.

Mean →

$$\frac{\text{sum of no. of elements}}{\text{total no. of elements}} = \text{mean}$$

9, 3, 1, 8, 6, 3 ⇒ $\frac{30}{6} = \underline{\underline{5}}$

Median →

median is the middle number.

1, 3, 3, 6, 8, 9 $\qquad \frac{3+6}{2} = \underline{\underline{4.5}}$

Mode →

the most common number.

9, 3, 1, 8, 6, 3 ⇒ $\underline{\underline{3}}$ is the mode

Standard Daviation →

daviation of values / data from an average mean.

- Population Standard Deviation

$$\sigma = \sqrt{\frac{\Sigma(x-\mu)^2}{n}}$$

x - Value in data distribution

μ - population mean

$\bar{x}$ - Sample mean

n - total number of observation

- Sample Standard Deviation

$$s = \sqrt{\frac{(x-\bar{x})^2}{n-1}}$$

Variance →

the value of variance = (Standard Daviation)^2

$$= \sigma^2$$

$$= \frac{\Sigma(x-\mu)^2}{n}$$

Quartile →

its like median which divides the data in '2' parts, where as quartile divides the data into '4' parts.

1, 1, 2, 2, 2, 3, 3, 4, 5, 6, 6, 6, 6, 8, 8, 10, 10, 11, 11

Q_1 — Lower Quartile

Q_2 — Middle Quartile

Q_3 — Upper Quartile

Q_2 = Median

$$\text{Interquartile Range} = Q_3 - Q_1$$

$$\text{Semi Interquartile Range} = \frac{Q_3 - Q_1}{2}$$

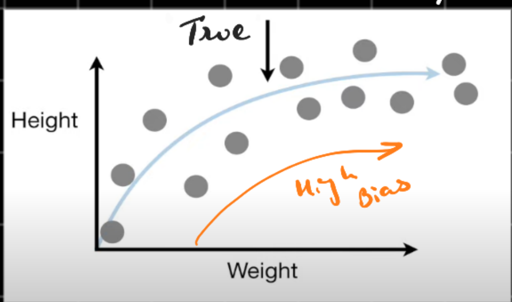
$$Q_1 = \frac{n+1}{4}$$

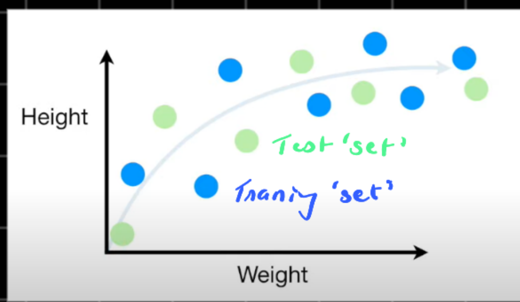
$$Q_3 = \frac{3(n+1)}{4}$$

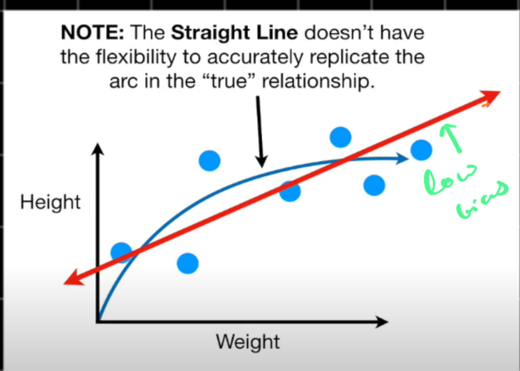
Bias →

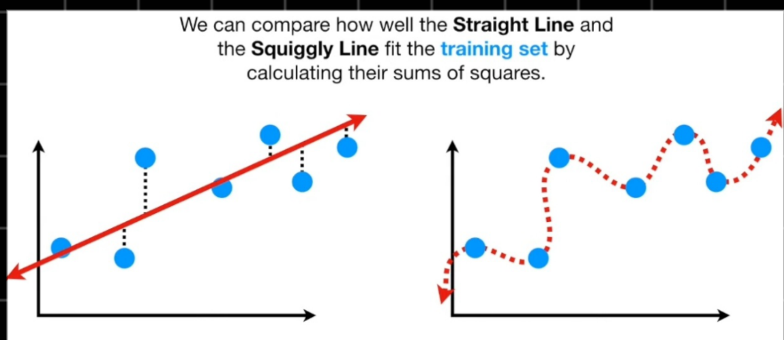
The inability for a machine learning method to capture the true relationship.

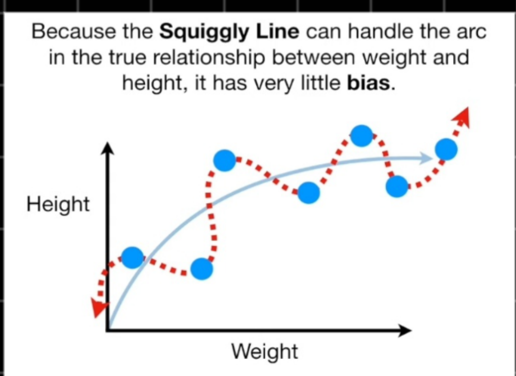
'Relationship'

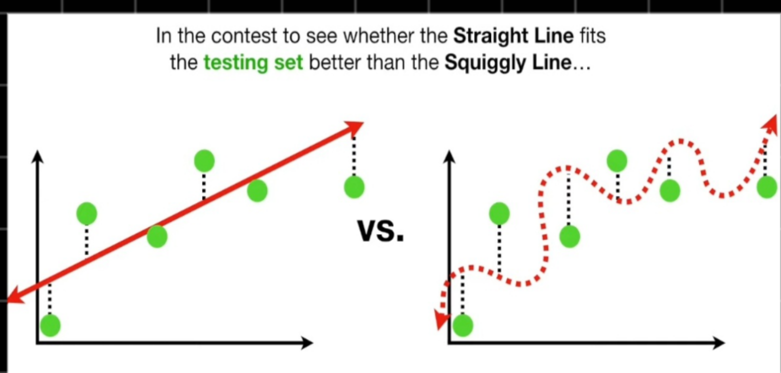








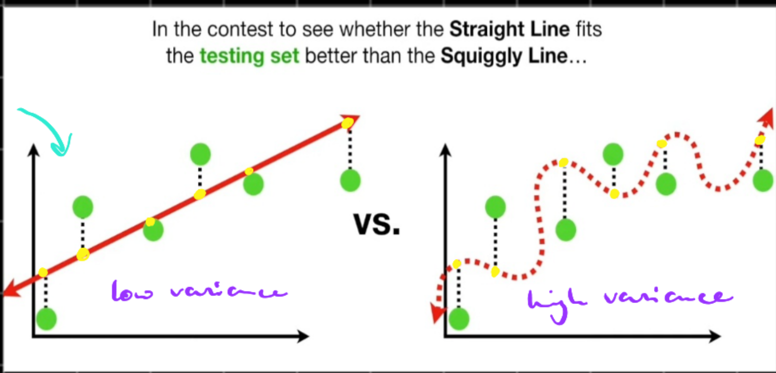




Variance →

Distance between the actual and predicted data points.

distance between • and • is var.



- Actual data points
- Predicted data points

Covariance → ✓ direction ✗ degree of proportion

shows the directional relationship between the two variables.

'+'ve	0	'−'ve
both in same direction		both in opp. direction

$$COV = \frac{\Sigma (x-\bar{x})(y-\bar{y})}{N-1}$$

$\bar{x}, \bar{y}$ = mean

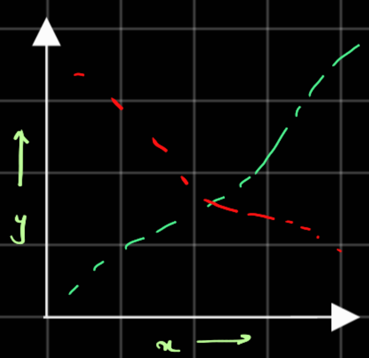
Correlation → ✓ direction ✓ strength / degree of proportion.

shows the direction of relation and in what proportion its changing.

Correlation coefficient → r +1 to −1

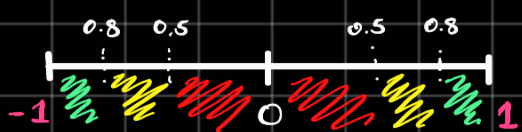
$$COR = \frac{cov(x,y)}{\sigma_x \times \sigma_y}$$ ← Standard deviation.

+ve, −ve Correlation →



x incr, y increases "+ve"

x incr, y decreases "−ve"



0 No correlation

→ low

→ moderate

→ strong

+, − 1 perfect +, −ve correlation.

Bayesian Inference (Intelligent Guess)

3 Techniques →

i. They do not rely on large data sets.

ii. Common sense, human reasoning

iii. Combining data

Bayes Theorem →

The probability of the occurance of an event A given that an event B has occured is called the conditional probability of A given B.

$P(A)$ = It will rain in our area

$P(B)$ = weather forecasting predicting that it will rain

$P(A/B) = 0.9$

$P(B/A) = \frac{P(A/B)\,P(B)}{P(A)} = \frac{0.9 \times 0.1}{0.27}$

$P(A/B) = \frac{P(B/A)\,P(A)}{P(B)}$

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

Same

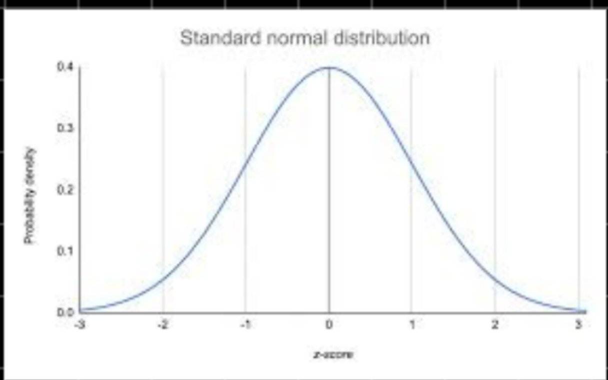
$\Rightarrow P(A/B) = \frac{P(B/A) \cdot P(A)}{P(B)}$

when there are more events

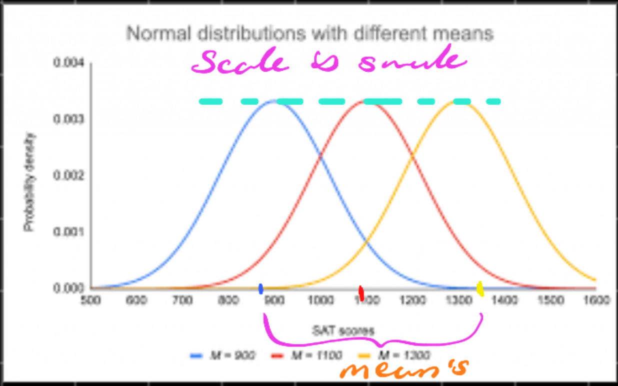
$$P(A) = \sum_{i=0}^{n} P(E_i)\,P(A/E_i)$$

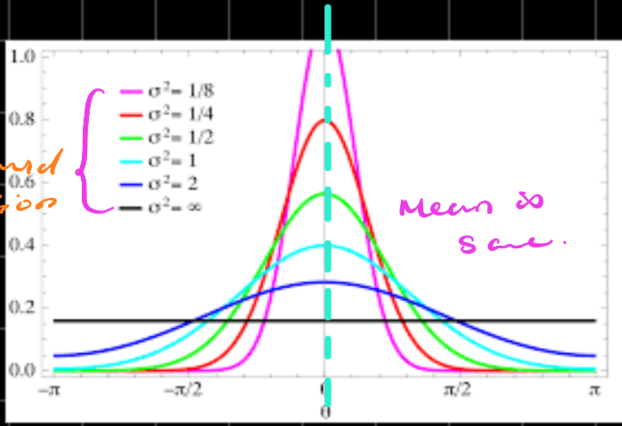
$$P(A/B) = \frac{P(B/A)\,P(A)}{P(E_1)\,P(A/E_1) + \ldots P(E_i)\,P(A/E_i)}$$

Normal Distribution →



- Mean = Median = Mode
- Mean → location of "bell curve"
- Standard deviation → scale of graph.





location of normal distribution changes as 'mean' changes

∘ scale of graph is same

'SD' shows the scale of graph, how broad or thin the area is.

∘ mean of graph is same

Statistical Hypothesis → (Hypothesis testing)

Population →

Sample →

Test (statistics) →

1) χ^2 test (chi)
2) t - student
3) fisher's z test

Hypothesis →

1) Null hypothesis (H_0)
2) Alternate hypothesis (H_1)

Level of significance (α) (5% 1%)

Level of confidence (c) (95% 99%)

n → random sample

$\bar{x}$ → mean

σ → variance

Null hypothesis $H_0 : \mu =$ "value"

Alternate hypothesis $H_1 : \mu \neq$ "value"

$$z = \left| \frac{\bar{x} - \mu}{\sigma / \sqrt{n}} \right|$$

$$\frac{\bar{x} - \mu}{\sigma / \sqrt{n}}$$

$$\frac{50}{\sqrt{50}} \quad \sqrt{} = 36$$

$$23.6 \longrightarrow 50 \longrightarrow 36$$

$$\% \ 1 = 1.6$$

$$\% = 1.9$$

Q There are two bag's e_1 an e_2 there are 2 white and 3 red in bag e_1, 4 white and 6 black, Find the probability of black ball drawn from e_2.

Ans $\Rightarrow \frac{1}{2}$

→ Statistical Hypothesis testing

$$y = mx + c$$

$$m = \frac{n\sum xy - (\sum x)(\sum y)}{n\sum x^2 - (\sum x)^2}$$

$$c = \frac{\Sigma y \Sigma x^2 - \Sigma x \Sigma xy}{n \Sigma x^2 - (\Sigma x)^2}$$

x	y
3	2
7	4
.	6
ω	8

$$m \Rightarrow \frac{4(144) - 20(144)}{4(120) - 400}$$

$$\Rightarrow \underline{0.95}$$

$$c = \frac{25(120) - 20(144)}{4(120) - 400}$$

$$y = mx + c$$

$$y = 0.95x + 1.5$$

$$\Rightarrow \underline{1.5}$$

To be followed in hypothesis testing →

1. Determine the tests is null hypothesis or alternative hypothesis.
2. Determine the significant value that be taken α as 0.05.
3. Calculate the probablity value
4. check wheather to reject or retain the hypothesis
5. Last you will draw conclussion through

Z - test

1. If we want to know about coperation mean and variance is given.
2. when the population is in
3.

$$z = \frac{X - \mu}{\sigma / \sqrt{n}}$$

$$COV(X,Y) = \frac{(\Sigma x - \bar{x}) \times (\Sigma y - \bar{y})}{n} \qquad COR(X,Y) = \frac{COV(X,Y)}{\sigma_X \sigma_Y}$$

correlation 'r' $\Rightarrow \dfrac{n\sum xy - \sum x \sum y}{\sqrt{n\sum x^2-(\sum x)^2}\sqrt{n\sum y^2-(\sum y)^2}}$

x	1	2	3	4	5
y	2	3	5	4	5

Scope Vector $\longrightarrow$

KNN $\longrightarrow$

Acid	Durability	Classify	(7, 3)		$\underline{\underline{K=3}}$
7	7	Bad	$\sqrt{0+16}$	$=4$	
7	4	Bad	$0 +$	$=1$	
3	4	Good	$\sqrt{17}$		
1	4	Good	$\sqrt{37}$		
3	4				

$$\frac{\dfrac{\sum(x-\bar{x})(y-\bar{y})}{n-1}}{\sigma_x \, \sigma_y}$$

Unit - 2

$y = mx + c$

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

$$y = \frac{1}{1+e^{-(mx+c)}}$$

* Binary logistic Regression. In book

	Outlook		
1.	S	H	N
2.	S	H	N
3.	O	H	Y
4.	R	M	Y
5.	R	C	Y
6.	R	C	N
7.	O	C	Y
8.	S	M	N
9.	S	C	Y
10.	R	M	Y
11.	S	M	Y
12.	O	M	Y
13.	O	C	Y
14.	R	M	N

Today (S, H)

$P(Y/\text{Today}) = P(S/Y) \times P(H/\text{Yes}) \times P(Y)$

$$P(9/14) = \frac{\cancel{2/9} \times 2/9 \times \cancel{9}}{14} \Rightarrow 2/63$$

	Yes	No	P(Y)	P(N)
Sunny	2	3	2/9	3/5
Overcast	4	0	4/9	0
Rainy	3	2	3/9	2/5
Total	9	5		

H	2	1	2/9	2/5
M	4	2	4/9	2/5
C	4	1	4/9	1/5
Total	9	5		

Confusion Matrix →

Recall = TP / Actual yes

Accuracy = $\frac{TP + TN}{Total}$

sensitivity = $\frac{\text{True Positive}}{\text{TP + False negative}}$

		Predicted No	Predicted Yes
Actual	No	TN	FP
Actual	Yes	FN	TP

TP → True Positive

TN → True Negative

FP → False Positive

FN → False Negative

Type 1 error

Type 2 error

specificity = $\frac{\text{True Negatives}}{\text{TN + false positive}}$

Error Rate = 1 - accuracy

$= \frac{FP + FN}{Total}$

precision = $\frac{TP}{FP + TP}$

Sensitivity → also known as recall or true positive range is the conditional property that the predicted class is positive given that actual class is positive.

specificity →

Piking the best splitting Attribute →

1. Information Gain
2. Gini Index

Instance	a_1	a_2	classification
1	T	T	+
2	T	T	+
3	T	F	−
4	F	F	+
5	F	T	−
6	F	T	−
7	F	F	−
8	T	F	+
9	F	T	−

$$\text{Entropy}(S) = -\sum_{i=1}^{n} \log_2 i(P_i)$$

$$= \frac{4}{9}\log_2\frac{4}{9} + \frac{5}{9}\log_2\frac{5}{9}$$

Information Gain →

T ← (3+, 1−)

F ← (1+, 4−)

$$IG = \text{Entropy}(S) - \sum_{v \in (T,F)} \frac{|S_v|}{S} \log_2 (S_v)$$

Q) 2 Red, 2 yellow, 4 black, to find out the entropy of this table.

$$-\left[\frac{2}{8}\log_2\frac{2}{8}+\frac{2}{8}\log_2\frac{2}{8}+\frac{4}{8}\log_2\frac{4}{8}\right]$$

$$\left[\log_2 1/4 = -2 \qquad \log_2 1/2 = \underline{-1}\right]$$

$$-\left[\frac{1}{\cancel{4}_2}(-\cancel{2})+\frac{1}{4}(-2)+\frac{1}{2}(-1)\right]$$

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

$$\Rightarrow \frac{3}{2} \Rightarrow \underline{1.5}$$

Q)

$$\frac{|S_v|}{|S_1|} = \frac{4}{9} \qquad \frac{|S_v|}{|S_T|} = \frac{5}{9}$$

$$\text{Entropy}(S) = -\left[\frac{4}{9}\log_2\frac{4}{9}+\frac{5}{9}\log_2\frac{5}{9}\right]$$

$$\text{Entropy}(S_T) = -\left[\frac{3}{4}\log_2\frac{3}{4}+\frac{1}{4}\log_2\frac{1}{4}\right]$$

$$\text{Entropy}(S_F) = -\left[\frac{1}{5}\log_2\frac{1}{5}+\frac{4}{5}\log_2\frac{4}{5}\right]$$

$$\text{Information Gain} = \text{Entropy}(S) - \sum_{v\in(T,F)} \frac{|S_v|}{S_T}\log_2(S_v)$$

$$\Rightarrow E(S) - \left[\frac{4}{9}\times\text{entropy}(S_T)+\frac{5}{9}\times\text{entropy}(S_F)\right)$$

Cross Validation →

K-fold cross validation

1	2	3	4	5	6	7	8	9

→ $K=3$

Recall =

divided into 3 sub sets of test set with no overlapping

1) $\underbrace{1\ 2\ 3}_{\text{Test set}} \mid \underbrace{4\ 5\ 6 \mid 7\ 8\ 9}_{\text{Train set}} \rightarrow E_1$

2) $\underbrace{1\ 2\ 3}_{\text{Test}} \mid \underbrace{4\ 5\ 6}_{\text{Test set}} \mid \underbrace{7\ 8\ 9}_{\text{set}} \rightarrow E_2$

3) $\underbrace{1\ 2\ 3 \mid 4\ 5\ 6}_{\text{Test set}} \mid \underbrace{7\ 8\ 9}_{\text{Test set}} \rightarrow E_3$

$$E = \frac{\text{Avg of } E_n}{K}$$

ASM →

① Information Gain

② Gini Index

Days	Outlook	Temprature	Play Game
1	S	H	NO
2	S	H	NO
3	O	H	Yes
4	R	M	Yes
5	R	C	Yes
6	R	C	NO
7	O	C	Yes
8	S	M	NO
9	S	C	Yes
10	R	M	Yes
11	S	M	Yes
12	O	M	Yes
13	O	H	Yes
14	R	M	NO

Random forest

Aggregation of 2 different model to give majority output. The attributes that we choose in aggregation is different and independent to each other. which gives our model more accurately. As compared to bootrapping. model.

Bootstrapping and Aggregation

the bootstrap.

$$E(S) = \frac{-5}{14}\log_2 \frac{5}{14} - \frac{9}{14}\log_2 \frac{9}{14}$$

Attribute → outlook

values (overcast, sunny, Rain)

$$E(S_{sunny}) = -\frac{2}{5}\log_2 \frac{2}{5} - \frac{3}{5}\log_2 \frac{3}{5} = 0.97$$

$(2^+, 3^-)$

$$E(S_{overcast}) = \frac{-4}{4}\log_2 \frac{4}{4} - \frac{0}{4}\log_2 \frac{0}{4} = 0$$

$(4^+, 0^-)$

$$E(S_{Rain}) = -\frac{3}{5}\log_2 \frac{3}{5} - \frac{2}{5}\log_2 \frac{2}{5} = 0.971$$

$$\text{Gain} = E(S) - \sum_{v(\text{overcast, Sunny, Rain})} \frac{|S_v|}{|S|}\,\text{Entropy}(S_v)$$

$$= 0.94 - \frac{5}{14}(0.97) - 0 - \frac{5}{14}(0.97)$$

$$= 0.24$$

Outlook

Sunny	Overcast	Rain
$(2+, 3-)$	$(4+, 0-)$	$(3+, 2-)$
$(D_1, D_2, D_8, D_9, D_{11})$	Yes	$(D4, D5, D6, D10, D14)$

Days	Temp	Humidity	Wind	Play
D_1	H	High	weak	No
D_2	H	High	strong	No
D_8	mild	High	weak	No
D_9	C	Low	weak	Yes
D_{11}	mild	Low	strong	Yes

H ⟶

C ——

M ⟶

High
low

+4, -0

+

$$\text{Gain} = E(S_{\text{sunny}}) - \frac{2}{5}(E_{S_H}) - \frac{2}{5}(E_{S\_}) - \frac{1}{5}(\;)$$

SVM Support vector Machine →

- Labeled data set.
- Comes under supervised learning.

labeled data → Machine learning Model

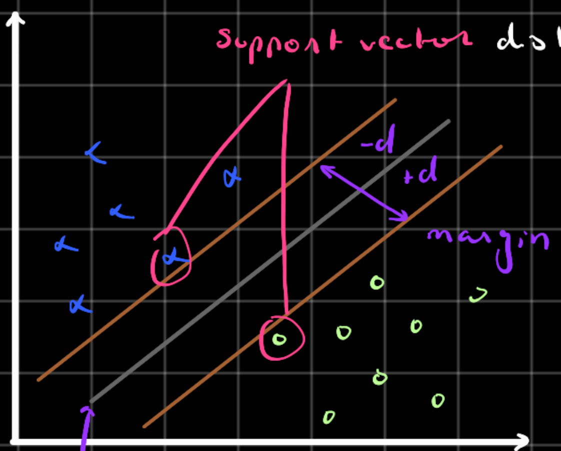
↓

Prediction ← new data

↓

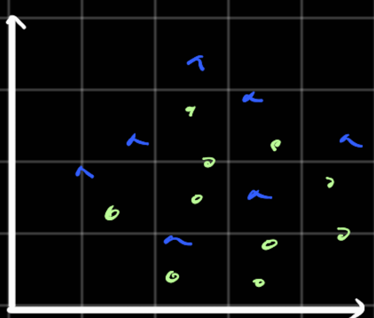
output

- To check in which class the new data will fall.



— linearly seperatable SVM

hyperplane should be formed such that margin is maximum.



separation under 3D space →

Low dimensional feature space

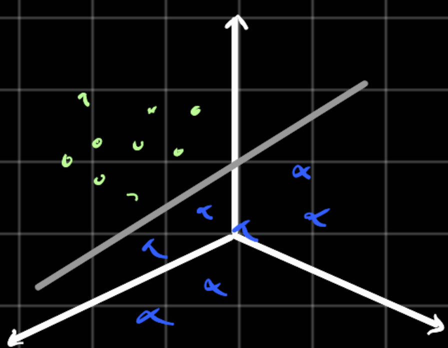
↓

Kernel

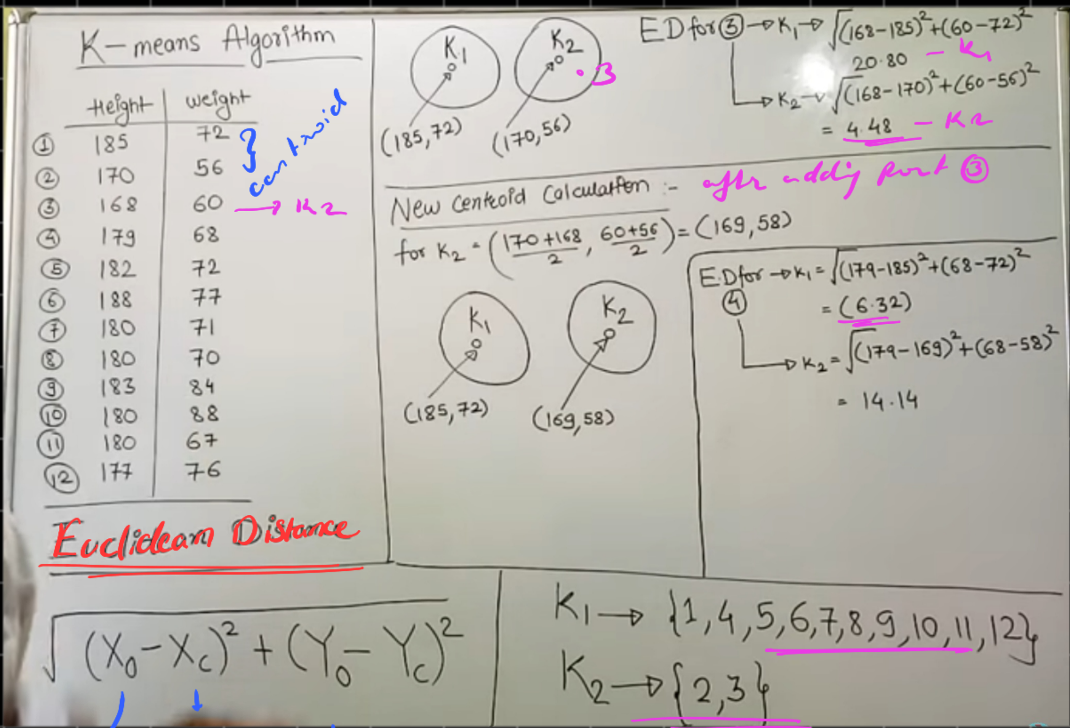
↓

high dimensional feature space

leave one out



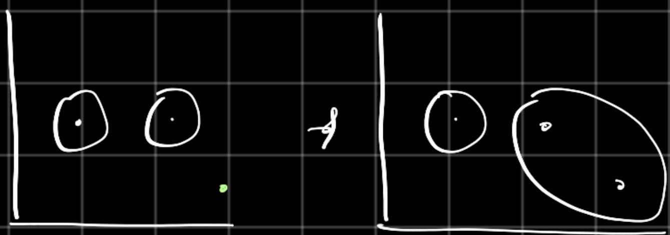
Kmean →



new point

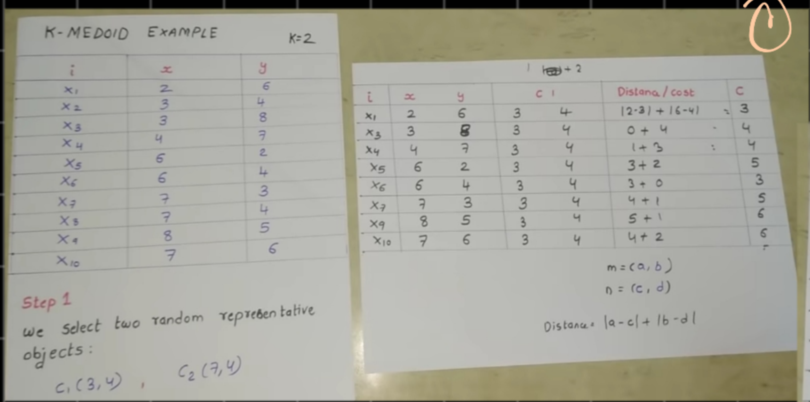
centroid location

- unsupervised learning.
- Centroid.
- After allocation of new data point. New to calculate mean.



overfitting

K - Medoid →



K-MEDOID EXAMPLE K=2

i	x	y
x_1	2	6
x_2	3	4
x_3	3	8
x_4	4	7
x_5	6	2
x_6	6	4
x_7	7	3
x_8	7	4
x_9	8	5
x_{10}	7	6

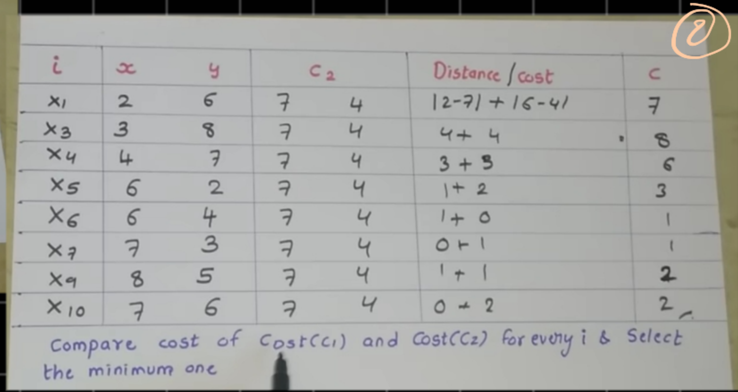
Step 1

we select two random representative objects:

$c_1 (3,4)$, $c_2 (7,4)$

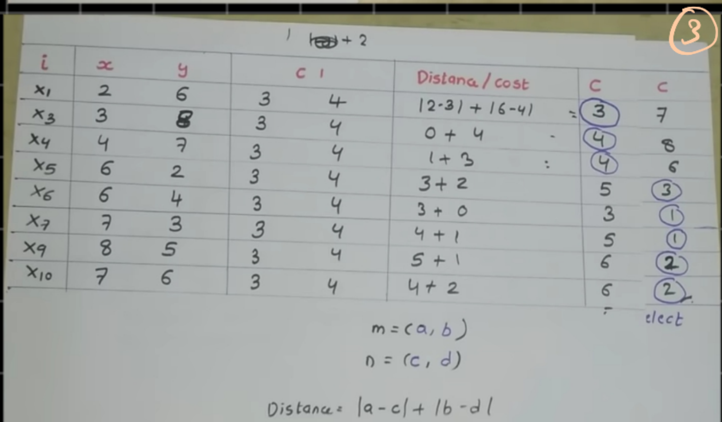
i	x	y	c1		Distance / cost	c
x_1	2	6	3	4	|2-3| + |6-4|	3
x_3	3	8	3	4	0 + 4	4
x_4	4	7	3	4	1 + 3	4
x_5	6	2	3	4	3 + 2	5
x_6	6	4	3	4	3 + 0	3
x_7	7	3	3	4	4 + 1	5
x_9	8	5	3	4	5 + 1	6
x_{10}	7	6	3	4	4 + 2	6

m = (a, b)
n = (c, d)
Distance = |a - c| + |b - d|



i	x	y	c_2		Distance / cost	c
x_1	2	6	7	4	|2-7| + |6-4|	7
x_3	3	8	7	4	4 + 4	8
x_4	4	7	7	4	3 + 3	6
x_5	6	2	7	4	1 + 2	3
x_6	6	4	7	4	1 + 0	1
x_7	7	3	7	4	0 + 1	1
x_9	8	5	7	4	1 + 1	2
x_{10}	7	6	7	4	0 + 2	2

Compare cost of Cost(c_1) and Cost(c_2) for every i & select the minimum one



i	x	y	c1		Distance / cost	c	c
x_1	2	6	3	4	|2-3| + |6-4|	3	7
x_3	3	8	3	4	0 + 4	4	8
x_4	4	7	3	4	1 + 3	4	6
x_5	6	2	3	4	3 + 2	5	3
x_6	6	4	3	4	3 + 0	3	1
x_7	7	3	3	4	4 + 1	5	1
x_9	8	5	3	4	5 + 1	6	2
x_{10}	7	6	3	4	4 + 2	6	2

select

m = (a, b)
n = (c, d)
Distance = |a - c| + |b - d|

(4)

Step II) then cluster are

cluster 1 : { (2,6), (3,8), (4,7), (3,4) }
cluster 2 : { (7,4), (6,2), (6,4), (7,3), (8,5), (7,6) }

Calculate total cost

$$T\ cost(x,c) = \sum_{i=1}^{d} |x_i - c_i|$$

Total cost = { cost((3,4), (2,6)), cost((3,4), (3,8)), cost((3,4), (4,7)), cost((7,4), (8,5)), cost((7,4), (6,2)), cost((7,4), (6,4)), cost((7,4), (7,3)), cost((7,4), (7,6)) },

= (3+4+4) + (3+1+1+2+2)

= 20

Step 3) Select one of non-Medoids O'
Let's O' = (7,3) i.e x_7

(5)

So now medoid's are C(3,4) & O'(7,3)

i	x	y	O'		Distance / Cost	= C
x_1	2	6	7	3	|2-7| + |6-3|	8
x_3	3	8	7	3	4+5	9
x_4	4	7	7	3	3+4	7
x_5	6	2	7	3	1+1	2
x_6	6	4	7	3	1+1	2
x_8	7	4	7	3	0+1	1
x_9	8	5	7	3	1+2	3
x_{10}	7	6	7	3	0+3	3

Compare the cost of cost(C_1) and cost(O') for every i & Select the minimum one

(6)

i	x	y	C_i		Distance / cost	C
x_1	2	6	3	4	|2-3| + |6-4|	= 3
x_3	3	8	3	4	0+4	= 4
x_4	4	7	3	4	1+3	= 4
x_5	6	2	3	4	3+2	= 5
x_6	6	4	3	4	3+0	= 3
x_8	7	4	3	4	4+0	= 4
x_9	8	5	3	4	5+1	= 6
x_{10}	7	6	3	4	4+2	= 6

(7)

C_i		Distance / cost	C	C
3	4	|2-3| + |6-4|	= 3	8
3	4	0+4	= 4	9
3	4	1+3	= 4	7
3	4	3+2	= 5	2
3	4	3+0	= 3	2
3	4	4+0	= 4	1
3	4	5+1	= 6	3
3	4	4+2	= 6	3

) for

(8)

Again create the cluster
cluster 1 : { (3,4), (2,6), (3,8), (4,7) }
cluster 2 : { (7,3), (6,2), (6,4), (7,4), (8,5), (7,6) }

current total cost = (3+4+4) + (2+2+1+3+3)
= 11 + 11
= 22

Step 4) So cost of swapping medoid from C_2 to O''s

S = current total cost − past total cost
= 22 − 20
= 2 > 0

so, moving O' would be a bad idea so previous choice was Good

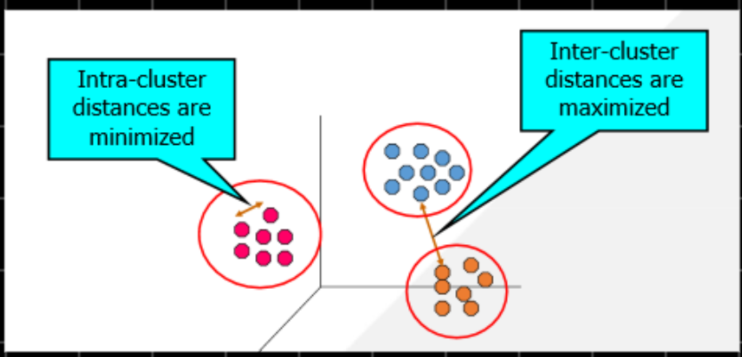
Cluster →

technique for finding similarity groups in data called clusters.

- It groups data instances that are similar to each other in one cluster and data instance that are very different from each other into different clusters.
- clustering is a mode of unsupervised learning
- Organize data into sensible groups.

- The data set has 3 natural groups of data points.

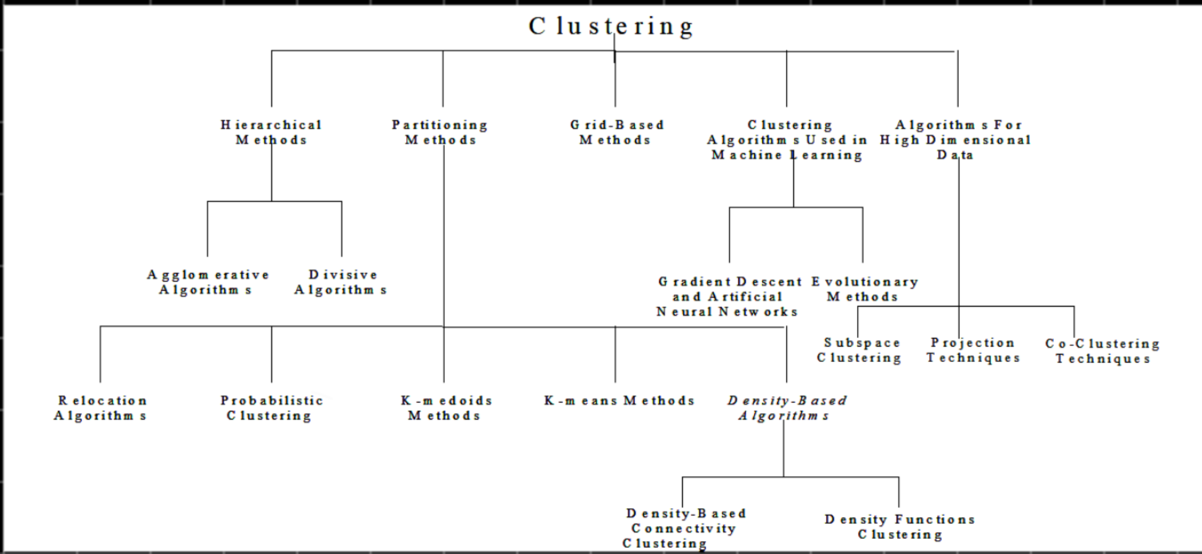
Inter , Intra - cluster



Types of clusters -->

- Well-separated clusters
- Center-based clusters
- Contiguous clusters
- Density-based clusters
- Property or Conceptual
- Described by an Objective Function

Types of clustering Algorithm -->



Agglomerative → Bottom-up | Divisive → Top-down

	P_1	P_2	P_3	P_4	P_5
P_1	0				
P_2	9	0			
P_3	3	7	0		
P_4	6	5	9	0	
P_5	11	10	2	8	0

Single Linkage →

	P_1	P_2	P_3	P_4	P_5
P_1	0				
P_2	9	0			
P_3	3	7	0		
P_4	6	5	9	0	
P_5	11	10	2	8	0

min

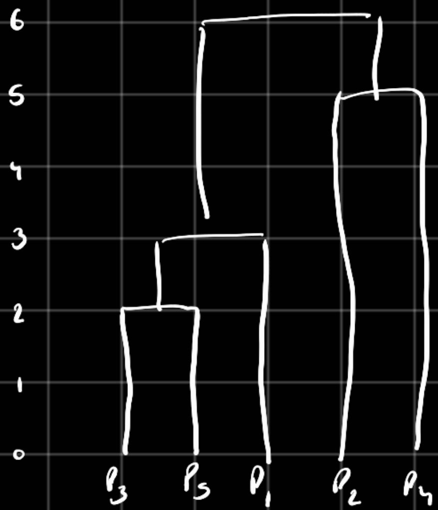
combine

min

$\min(d\, P_1, (P_3, P_5)) \rightarrow (3, 11) \Rightarrow 3$

$\min(d\, P_2, (P_3, P_5)) \rightarrow (7, 10) = 7$

$\min(d\, P_4, (P_3, P_5)) \Rightarrow (9, 8) \Rightarrow 8$



	P_1	P_2	P_3-P_5	P_4
P_1	0			
P_2	9	0		
P_3,P_5	3	7	0	
P_4	6	5	8	0

— next min

$\min(d\ P_2(P_1, P_3, P_5) \to (9, 7, 10) \Rightarrow 7$

$\min(d\ P_4(P_1, P_3, P_5) \to (6, 9, 8) \to 6$

	1,3,5	P_2	P_4
1,3,5	0		
P_2	7	0	
P_4	6	5	0

→ next min

$\min(d(P_2, P_4), (P_1, P_3, P_5)) \Rightarrow$

$(9, 7, 10), (6, 9, 8)$

$(7, 6) \Rightarrow 6$

	1,3,5	P_2, P_4
1,3,5	0	
P_2, P_4	6	0

Complete Linkage → distance between is max not min

	P_1	P_2	P_3	P_4	P_5
P_1	0				
P_2	9	0			
P_3	3	7	0		
P_4	6	5	9	0	
P_5	11	10	2	8	0

— min

$\max(d\ P_1(P_3, P_5)) \to (3, 11) \Rightarrow 11$

$P_2(P_3, P_5) \to (7, 10) = 10$

$P_4(P_3, P_5) \Rightarrow (9, 8) = 9$

	P_1	P_2	P_3, P_5	P_4
P_1	0			
P_2	9	0		
P_3, P_5	11	10	0	
P_4	6	5	9	0

— min

$\max(d\ P_1(P_2, P_4) \Rightarrow (6, 9) = 9$

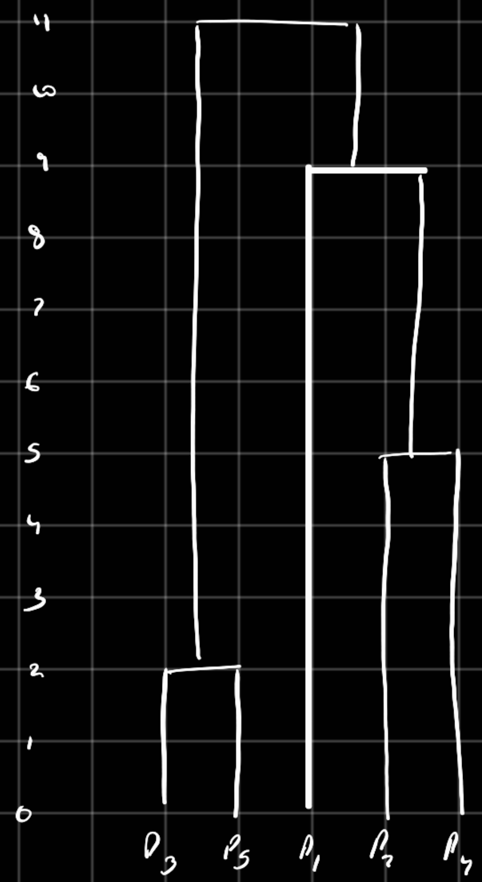
	P_1	2,4	3,5
P_1	0		
2,4	9	0	
3,5	11	10	0

— min

$\max(d((1,2,4),(3,5)))$

$((3,7,9),(11,10,8)) \Rightarrow 11$

	1,2,4	3,5
1,2,4	0	
3,5	11	0



Dendogram

Divisive →

	P_1	P_2	P_3	P_4	P_5	
P_1	0					
P_2	9	0				
P_3	3	7	0			— min_2
P_4	6	5	9	0		— min_3 — min_4
P_5	11	10	2	8	0	— min

P_5, P_3 — 2
P_3, P_1 — 3
P_4, P_2 — 5
P_4, P_1 — 6

$P_1\ P_2\ P_3\ P_4\ P_5$

$P_4\ P_2 \qquad P_1\ P_3\ P_5$

$P_4 \quad P_2 \quad P_1 \quad P_3\ P_5$

$P_3 \quad P_5$

→ Divisive Tree top-bottom.

Principle Component Analysis →

To extract important and usefull data from multivariate.

HD → LD

1) Point of view
2) Priority top down
3) No similarity between PC1 and PC2.

∗ To handel over fitting by dimensionality reduction.

1'nd point of view

2'st point of view

PC1 — highest priority
PC2
PC3
PC4
decrease.

Principle Component Analysis

X	Y
2.5	2.4
0.5	0.7
2.2	2.9
1.9	2.2
3.1	3.0
2.3	2.7
2	1.6
1	1.1
1.5	1.6
1.1	0.9

$\bar{X}$	$\bar{Y}$
1.81	1.91

mean

$$C = \begin{bmatrix} Cov(x,x) & Cov(x,y) \\ Cov(y,x) & Cov(y,y) \end{bmatrix} = \begin{bmatrix} 0.6165 & 0.6154 \\ 0.6154 & 0.7165 \end{bmatrix}$$

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

X	$X-\bar{X}$	$(X-\bar{X})(X-\bar{X})$
2.5	0.69	0.476
0.5	-1.31	1.7161

$\text{Sum} = \frac{5.5490}{9}$

Y	$Y-\bar{Y}$	$(Y-\bar{Y})(Y-\bar{Y})$
2.4	0.49	0.2401
0.7	-1.21	1.4641

$\text{Sum} = \frac{6.449}{9}$

calculate for all

X	Y	$X-\bar{X}$	$Y-\bar{Y}$	$(X-\bar{X})(Y-\bar{Y})$
2.5	2.4	0.69	0.49	0.3381
0.5	0.7	-1.31	-1.21	1.5851

$\text{Sum} = \frac{5.5390}{9}$

① Create the covariance matrix.

② Calculate eigen values → λ

③ Calculate eigen vector → v

$$C - \lambda I = 0$$

$$\begin{bmatrix} 0.6165 & 0.6154 \\ 0.6154 & 0.7165 \end{bmatrix} - \lambda \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} = 0$$

$$\begin{bmatrix} 0.6165-\lambda & 0.6154 \\ 0.6154 & 0.7165-\lambda \end{bmatrix}$$

$\rightarrow \lambda^2 - 1.333\lambda + 0.0630 = 0$

$\lambda_1 = 0.0490$
$\lambda_2 = 1.2840$ } eigen value

$$C v = \lambda v$$

$$\begin{bmatrix} 0.6165 & 0.6154 \\ 0.6154 & 0.7165 \end{bmatrix} \begin{bmatrix} X_1 \\ Y_1 \end{bmatrix} = 0.0490 \begin{bmatrix} X_1 \\ Y_1 \end{bmatrix}$$

① ②

$0.6165 X_1 + 0.6154 Y_1 = 0.0490 X_1$

$0.6154 X_1 + 0.7165 Y_1 = 0.0490 Y_1$

$\rightarrow 0.5674 X_1 = -0.6154 Y_1$

$\rightarrow 0.6154 X_1 = -0.6674 Y_1$

$X_1 = -1.0845 Y_1$

$$\begin{bmatrix} -1.0845^2 \\ 1^2 \end{bmatrix} \rightarrow 1.17614 + 1$$

square

$= \sqrt{2.17614}$

$= 1.47517$

$\rightarrow \begin{bmatrix} -0.7351 \\ 0.6778 \end{bmatrix}$ eigen vector for λ_1

calculating at $y = 1$, $x = ?$

$$\frac{-1.0845}{1.47517}$$

$X_2 = 0.92194 Y_2$

$\rightarrow \begin{bmatrix} 0.92194 \\ 1 \end{bmatrix}$ $= 0.8499 + 1$

$= \sqrt{1.8499}$

$= 1.3601$

$\Rightarrow \begin{bmatrix} 0.6778 \\ 0.7351 \end{bmatrix}$ eigen vector for λ_2

$\frac{1}{1.4}$ $\frac{2}{1.4}$

$14 \overline{)100}$ 0.701

98

20

Apriori Algorithm →

Apriori Algorithm

- Min Support = 50%
- Threshold Confidence = 70%

Eg:-

TID	Items
100	1 3 4
200	2 3 5
300	1 2 3 5
400	2 5

(I)

Itemset	Support
1	2/4 → 50%
2	3/4 → 75%
3	3/4 → 75%
4	1/4 → 25%
5	3/4 → 75%

(Itemset → 1,2,3,5)

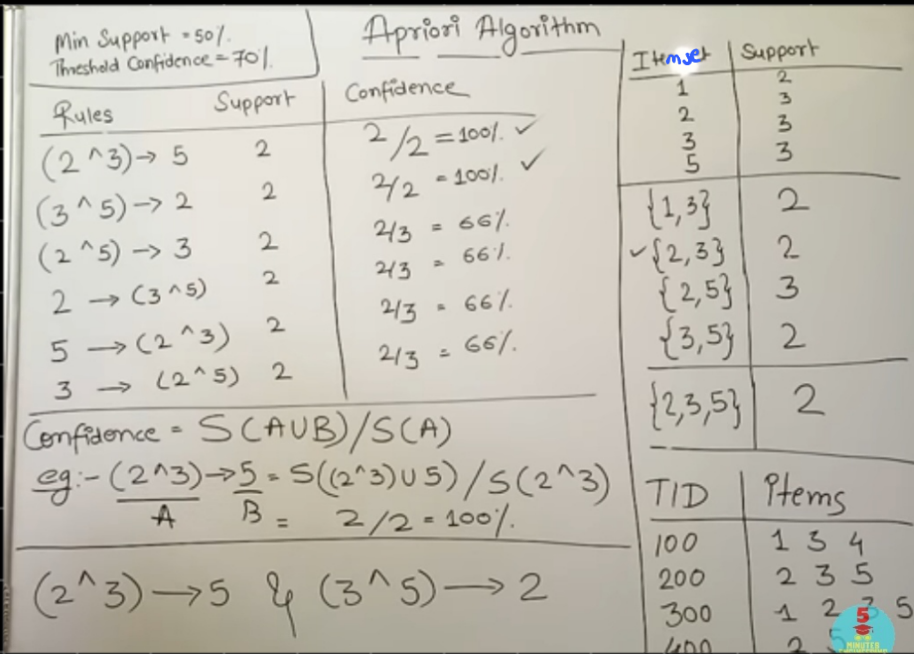
(II)

Itemset	Support
{1,2}	1/4 → 25% ✗
✓ {1,3}	2/4 → 50%
{1,5}	1/4 → 25% ✗
✓ {2,3}	2/4 → 50%
✓ {2,5}	3/4 → 75%
✓ {3,5}	2/4 → 50%

(III)

Itemset	Support
{1,3,5}	1/4 = 25%
{2,3,5}	2/4 = 50%
{1,2,3}	1/4 = 25%

- To find association between the two objects.
- Minimum support.



$$\text{confidence} = \frac{fr(A,B)}{fr(A)}$$

$$\text{support} = \frac{f(A,B)}{n}$$

$$\text{Lift} = \frac{su}{sup(A) \times sup(B)}$$

Market Basket Analysis

Frequent Patterns
- Itemsets
- Sequential Pattern

Itemsets → "Or"

ex – When some one buys "Milk" he might buy "Bread" too
{Milk, Bread}

Sequential Pattern → "Also"

ex :– When some one buys "Computer" also buy's "Antivirus".
{Computer, Antivirus}

When X is observed, Y is also observed.

x	y
4	11
8	4
13	5
7	14
8	8.5

$$\begin{bmatrix} \mathrm{Cov}(x,x) & (x,y) \\ (y,x) & (y,y) \end{bmatrix}$$

8.5
4)32
−32

$$\Sigma \frac{(x-\bar{x})(y-\bar{y})}{n-1}$$

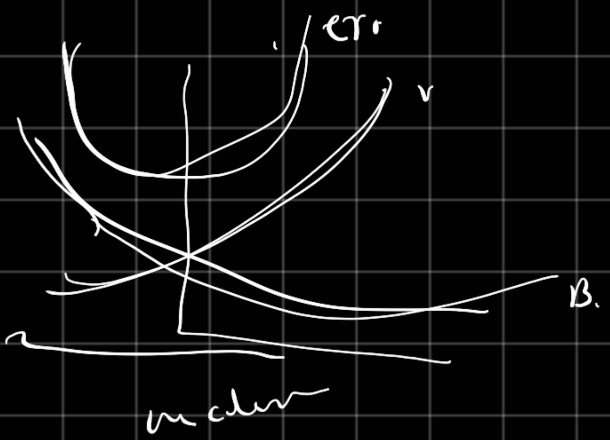
$(x-\bar{x})$	$(y-\bar{y})$	$(x-\bar{x})(x-\bar{x})$	$(y-\bar{y})(y-\bar{y})$	$(x \times y)$
−4	2.5	16	6.05	−10
0	−4.5	0	20.25	0
5	−3.5	25	12.25	−17.5
−1	5.5	1	30.35	−5.5
		42	68.90	−
		3	3	

$$\begin{bmatrix} \mathrm{cov}(x,x) & \mathrm{cov}(x,y) \\ \mathrm{cov}(\ ) & \mathrm{cov}(\ ) \end{bmatrix}$$

$C - \lambda I = 0 \quad \lambda_1 \quad \lambda_2$

$$C \begin{bmatrix} x_1 \\ y_1 \end{bmatrix} = \lambda \begin{bmatrix} x_1 \\ y_1 \end{bmatrix}$$

$$\begin{bmatrix} x_1 \\ y_1 \end{bmatrix}$$



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