Model Building – Unsupervised Learning

BY:- SHOBHIT TYAGI

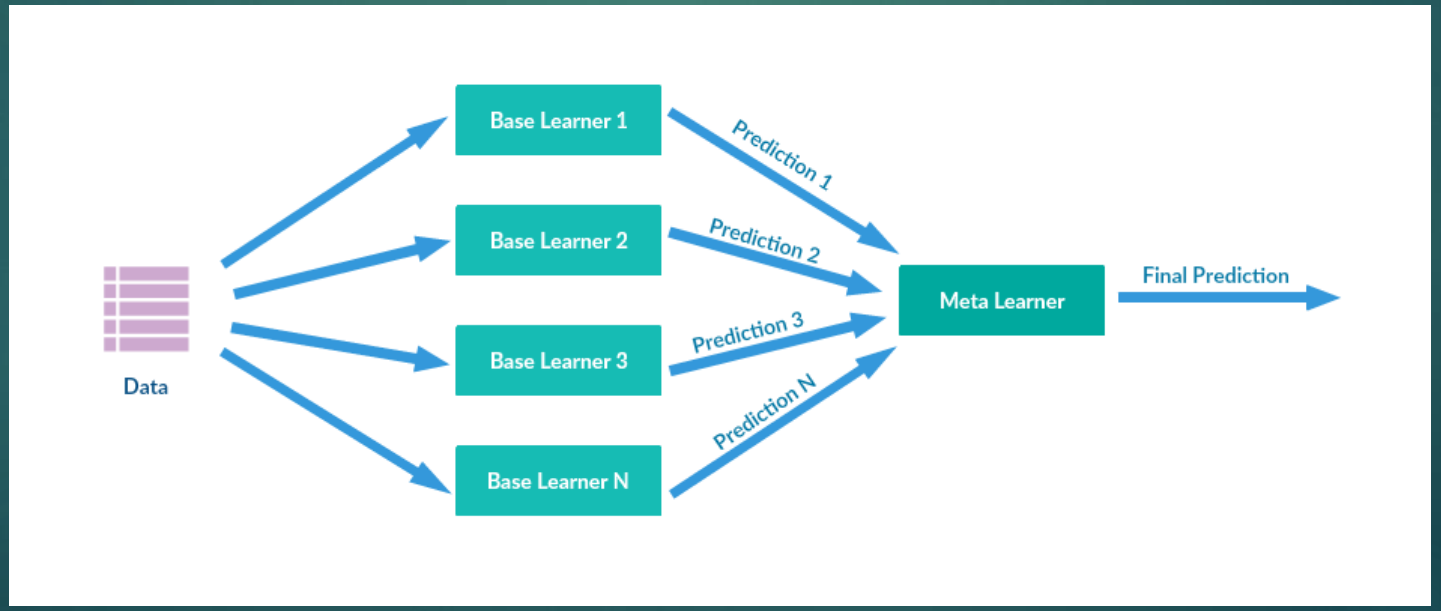
Contents

- Ensemble Methods
- Bayes Theorem
- Recap
- Clustering

Ensemble Methods

Ensemble Modelling

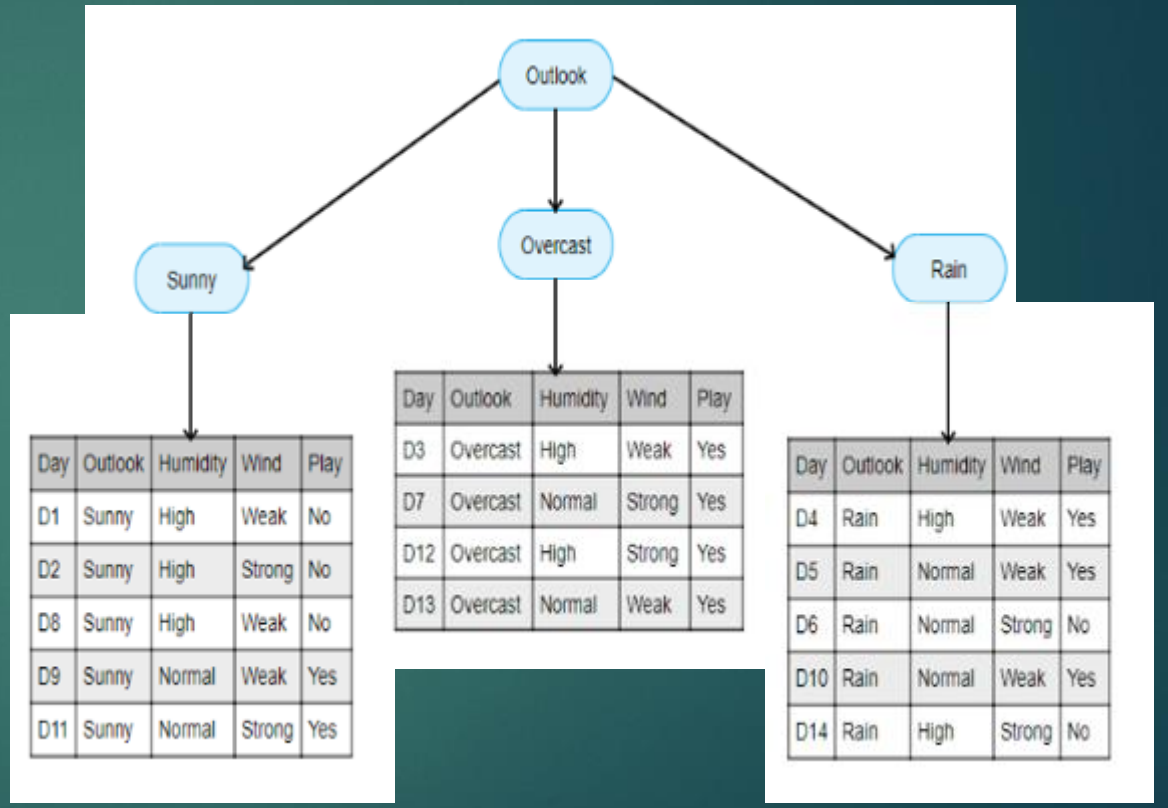
- It is a combination of multiple and diverse models.
- Each model in the ensemble makes a prediction.
- A final prediction is determined by a majority vote among the models.



Decision Tree

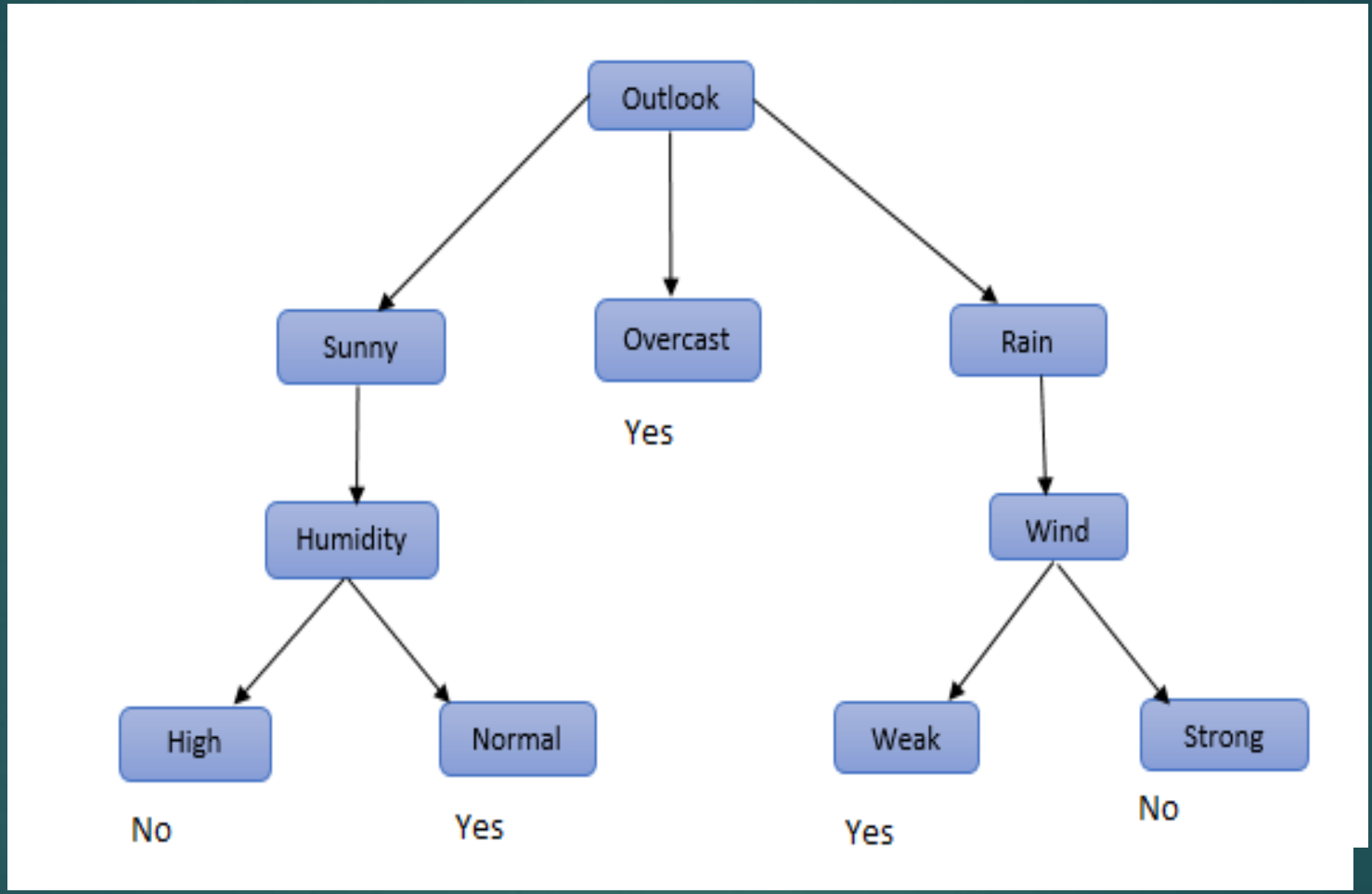
Day	Outlook	Humidity	Wind	Play
D1	Sunny	High	Weak	No
D2	Sunny	High	Strong	No
D3	Overcast	High	Weak	Yes
D4	Rain	High	Weak	Yes
D5	Rain	Normal	Weak	Yes
D6	Rain	Normal	Strong	No
D7	Overcast	Normal	Strong	Yes
D8	Sunny	High	Weak	No
D9	Sunny	Normal	Weak	Yes
D10	Rain	Normal	Weak	Yes
D11	Sunny	Normal	Strong	Yes
D12	Overcast	High	Strong	Yes
D13	Overcast	Normal	Weak	Yes
D14	Rain	High	Strong	No

Step - I



If Outlook is Rain and Wind is Strong .Will John Play Tennis?

Final Tree!!



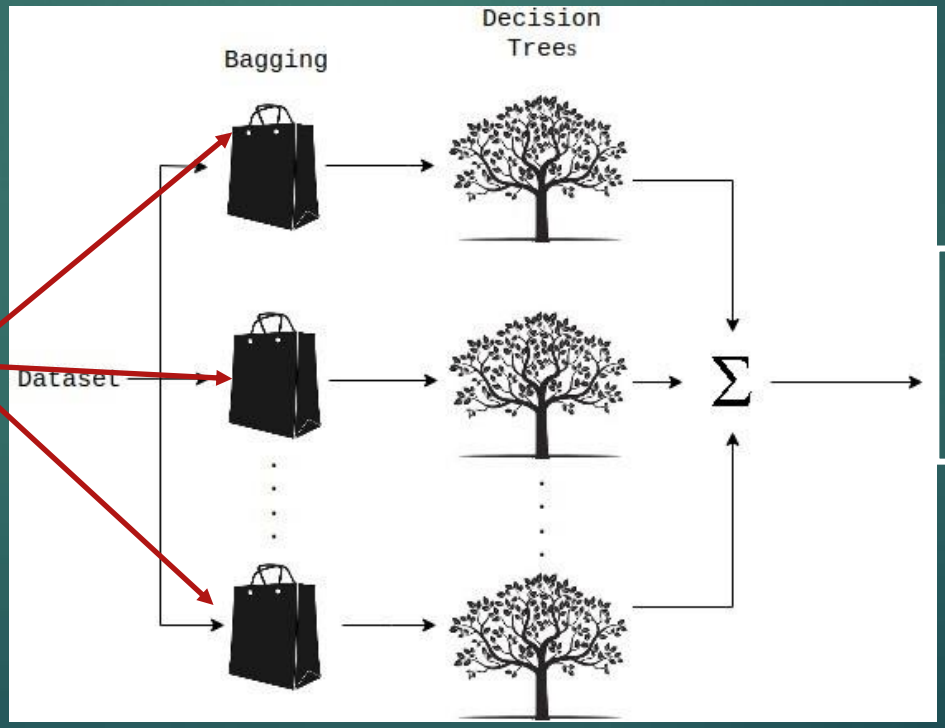
Step -2

Step -3

Bagging

Bootstrap Aggregation is used to reduce the variance for those algorithm that have high variance.

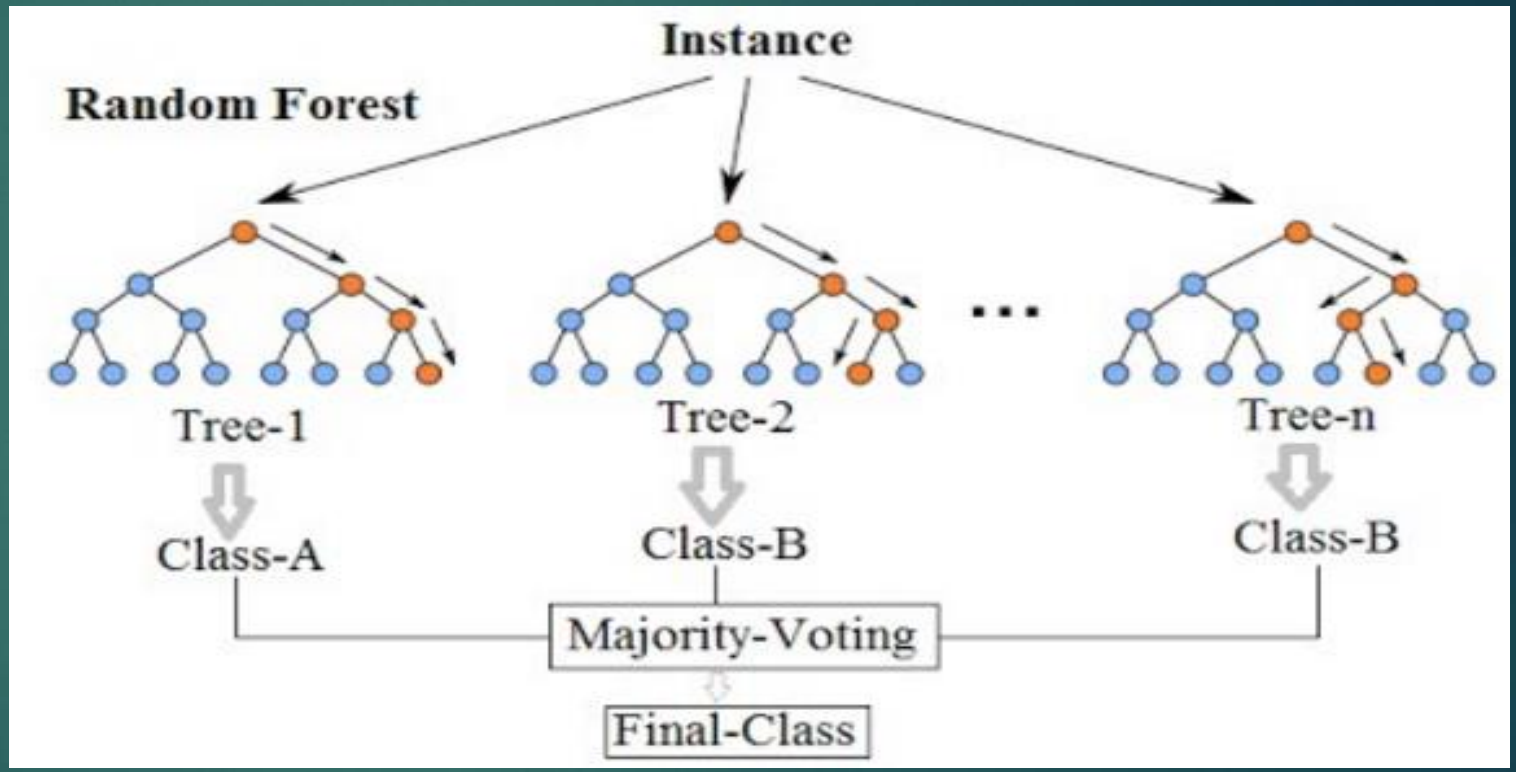
Subset of data from training sample chosen randomly with replacement



Average of all predictions from different trees are used which is more robust

Random Forest

- Extension over bagging
- Subset of data from training sample chosen randomly with replacement
- Random selection of features rather than using all features to grow trees.



Some Important Attributes

- **ntree** : Number of trees to grow
  - Larger number of trees produce more stable models
  - Require more memory and a longer run time.

- **mtry** : Number of variables available for splitting at each tree node.
  - For **classification** models, the default is the square root of the number of predictor variable.
  - For **regression** models, it is the number of predictor variables divided by 3 (rounded down).

What is a good split?

- Leaf should have homogeneous data.
- Minimizes the error/less impurity

- To split a particular columns or not is decided by below metrics:
  - Gini Index
  - Cross Entropy
  - Information Gain

} Measure Of Impurity

Entropy

- ***it is the measurement of the impurity or randomness in the data points.***
- Entropy is calculated between 0 and 1

Information Gain is applied to quantify which feature provides maximal information about the classification based on the notion of entropy

Gini Index, also known as Gini impurity, **calculates the amount of probability of a specific feature that is classified incorrectly when selected randomly.**

Gini index varies between values 0 and 1,

where 0 expresses the purity of classification, i.e. All the elements belong to a specified class or only one class exists there.

1 indicates the random distribution of elements across various classes. The value of 0.5 of the Gini Index shows an equal distribution of elements over some classes.

Information Gain

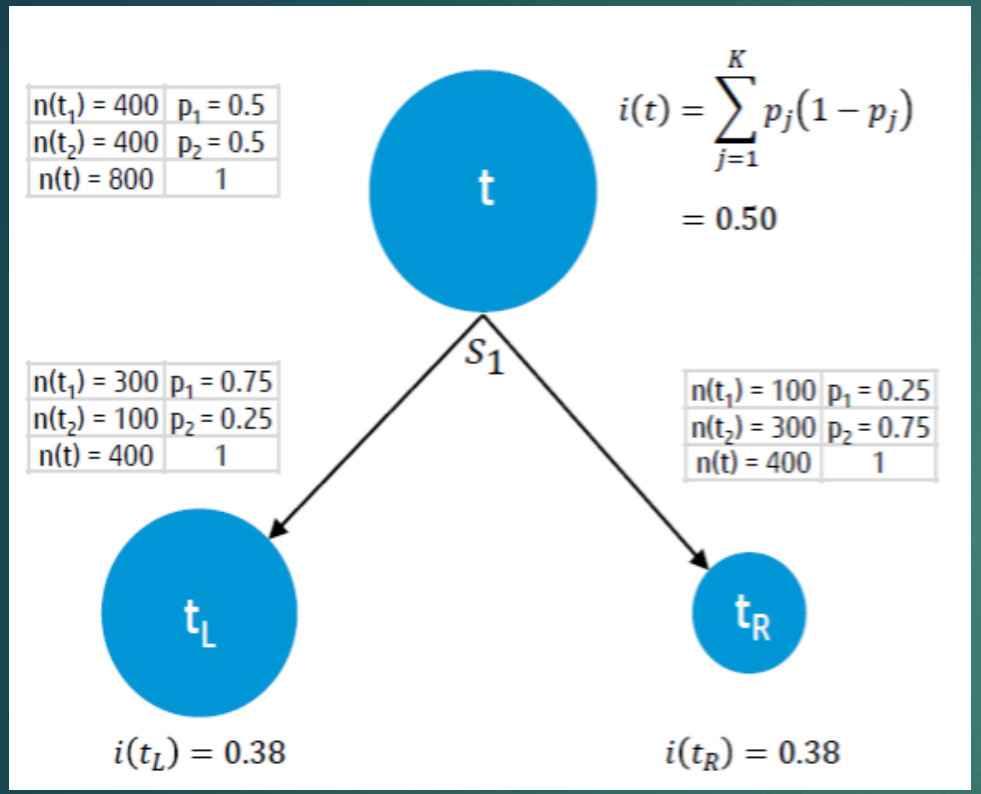
- It is the difference in impurity from the top node to the next node.
- The split which gives high Information Gain is considered for split.

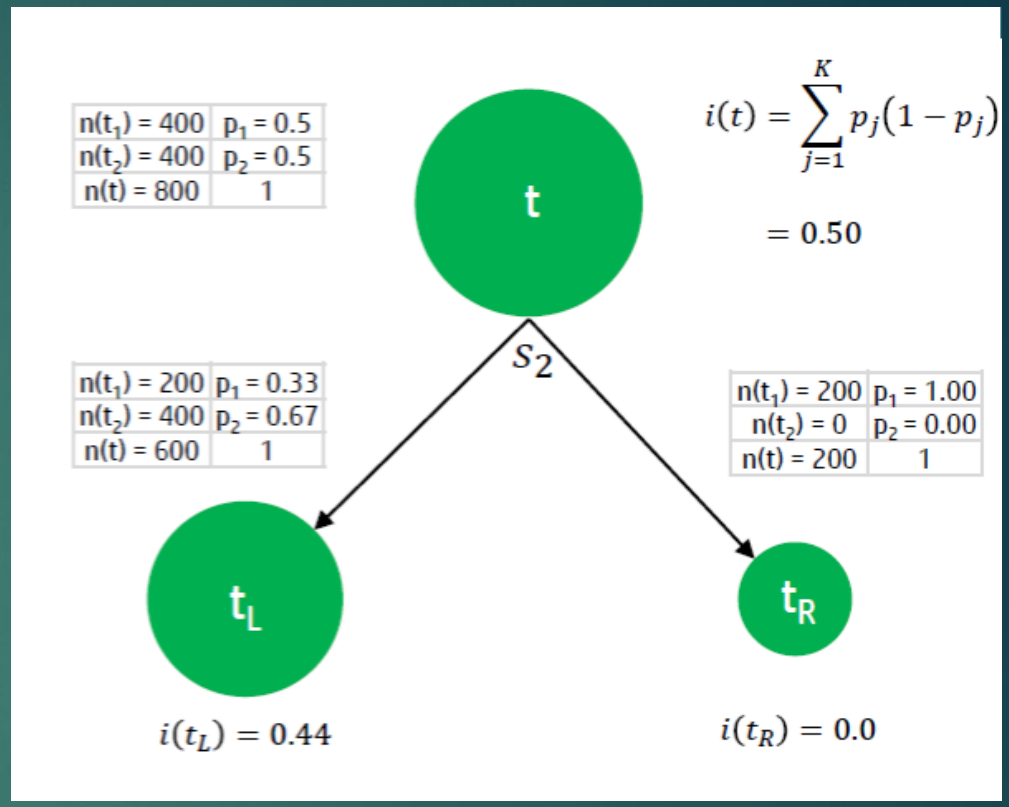
$$IG(D_p, f) = I(D_p) - \sum_{j=1}^{m} \frac{N_j}{N} I(D_j)$$

f: feature split on
D_p: dataset of the parent node
D_j: dataset of the jth child node
I: impurity criterion
N: total number of samples
N_j: number of samples at jth child node

maximize { Information Gain }

Gini Index



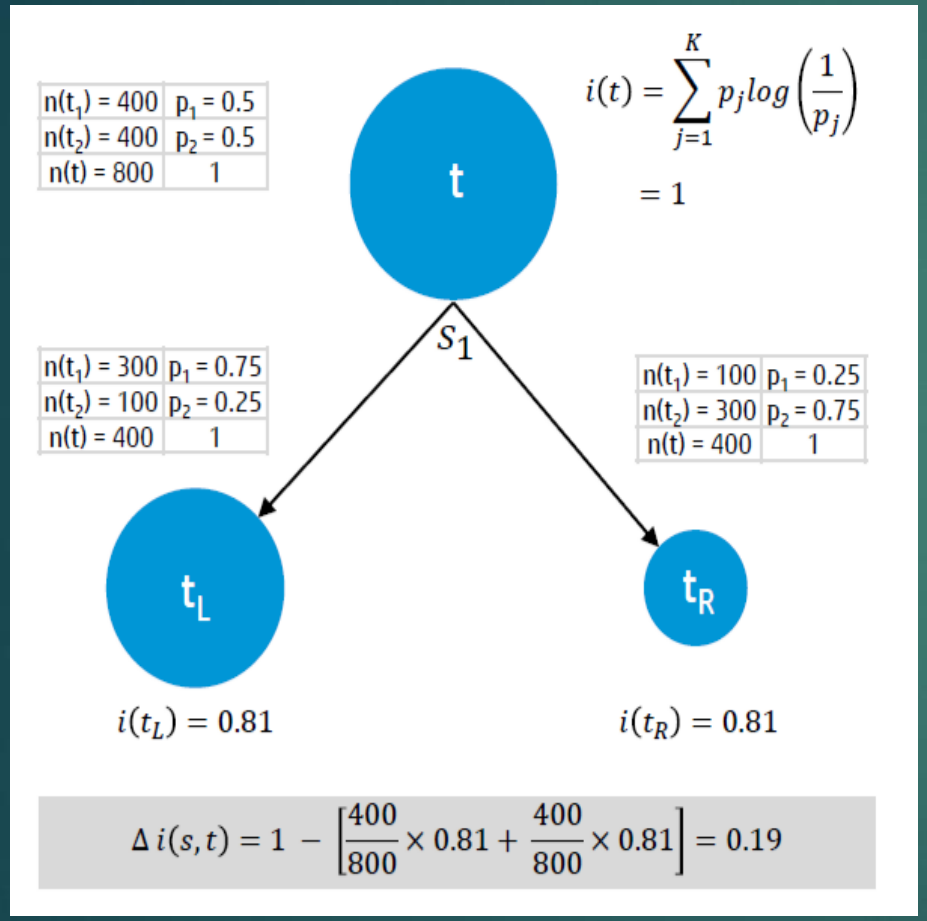


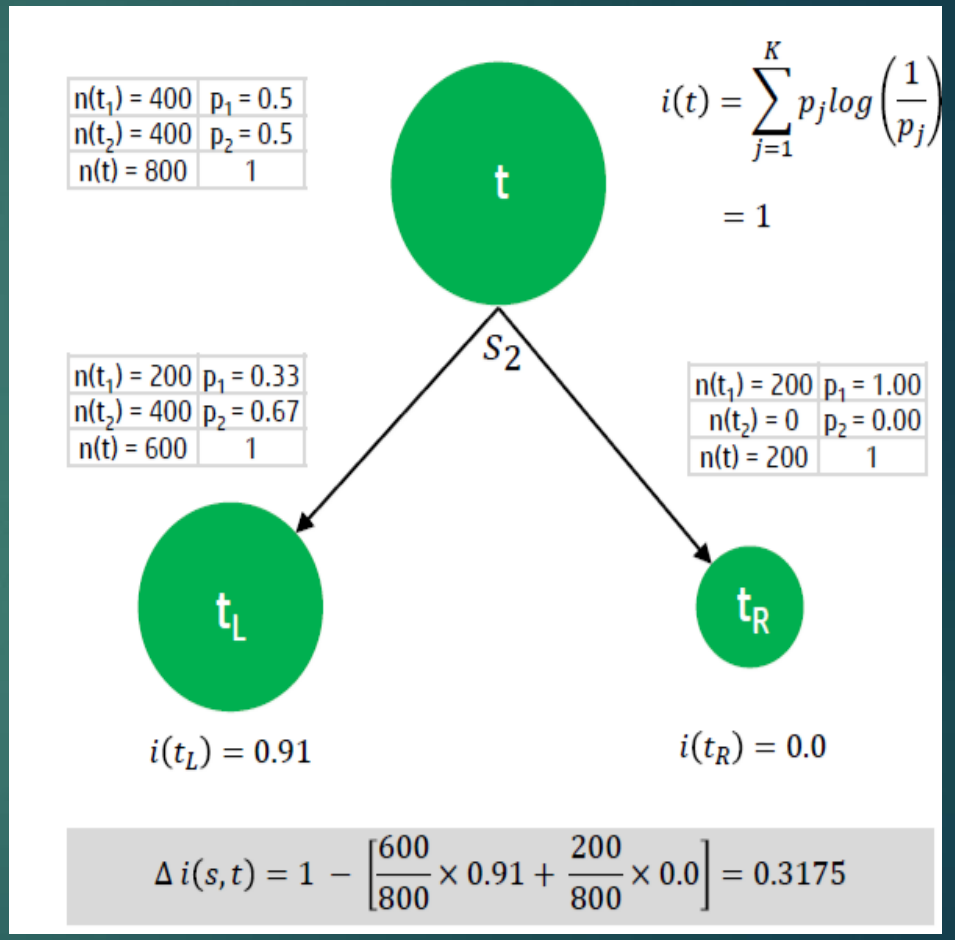
$$\Delta\, i(s,t) = 0.50 - \left[\frac{400}{800} \times 0.38 + \frac{400}{800} \times 0.38\right] = 0.12$$

$$\Delta\, i(s,t) = 0.50 - \left[\frac{600}{800} \times 0.44 + \frac{200}{800} \times 0.0\right] = 0.17$$

Information Gain

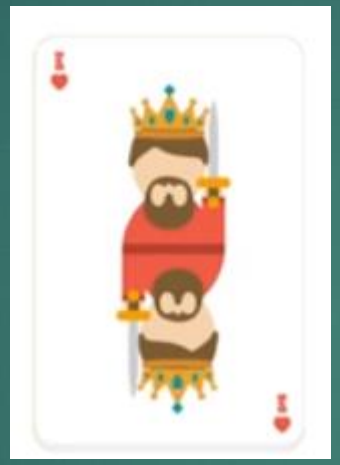
Cross Entropy





BAYES THEOREM

Bayes Theorem



$$P(King) = 4/52 = 1/13$$

$$P(Face) = 12/52 = 3/13$$

$$P(Face|King) = 1$$

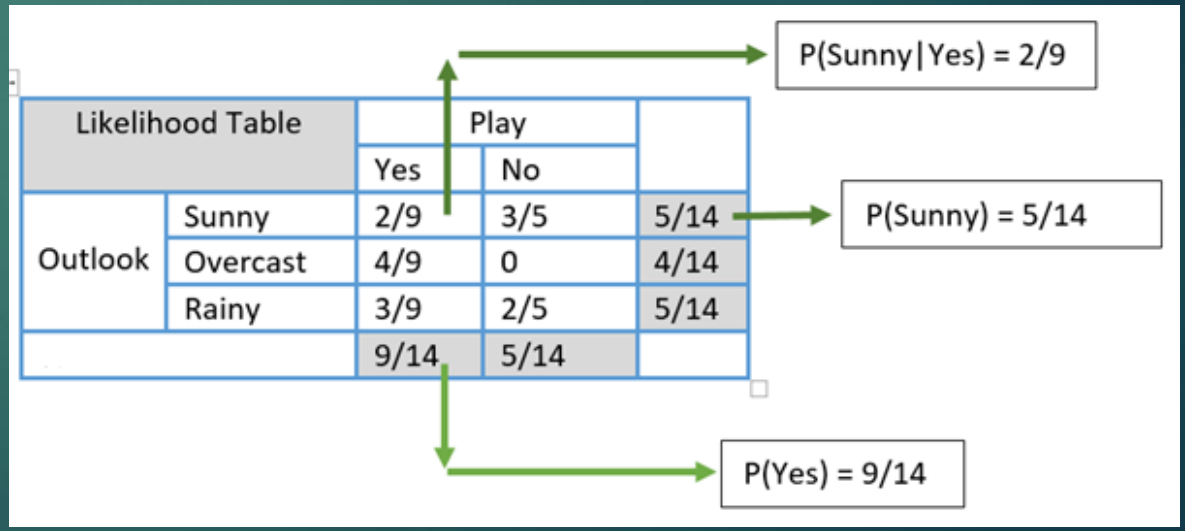
$$P(King|Face) = \frac{P(Face|King).P(King)}{P(Face)}$$

$$= \frac{1.(1/13)}{3/13} = 1/3$$

Naïve Bayes Algorithm

Day	Outlook	Humidity	Wind	Play
D1	Sunny	High	Weak	No
D2	Sunny	High	Strong	No
D3	Overcast	High	Weak	Yes
D4	Rain	High	Weak	Yes
D5	Rain	Normal	Weak	Yes
D6	Rain	Normal	Strong	No
D7	Overcast	Normal	Strong	Yes
D8	Sunny	High	Weak	No
D9	Sunny	Normal	Weak	Yes
D10	Rain	Normal	Weak	Yes
D11	Sunny	Normal	Strong	Yes
D12	Overcast	High	Strong	Yes
D13	Overcast	Normal	Weak	Yes
D14	Rain	High	Strong	No

Frequency Table		Play	
		Yes	No
Outlook	Sunny	2	3
	Overcast	4	0
	Rainy	3	2
Total		9	5



Naïve Bayes Algorithm

Likelihood Table		Play		
		Yes	No	
Humidity	High	3/9	4/5	7/14
	Normal	6/9	1/5	7/14
		9/14	5/14	

Likelihood Table		Play		
		Yes	No	
Wind	Weak	6/9	2/5	8/14
	Strong	3/9	3/5	6/14
		9/14	5/14	

Naïve Bayes Algorithm

Suppose we have a day with the following values

Outlook	=	Rain
Humidity	=	High
Wind	=	Weak
Play	=	?

Likelihood of 'Yes' on that day

= P(Outlook = Rain|Yes)* P(Humidity = High|Yes)* P(Wind = Weak|Yes)* P(Yes)

= 3/9 * 3/9 * 6/9 * 9/14 = 0.0476

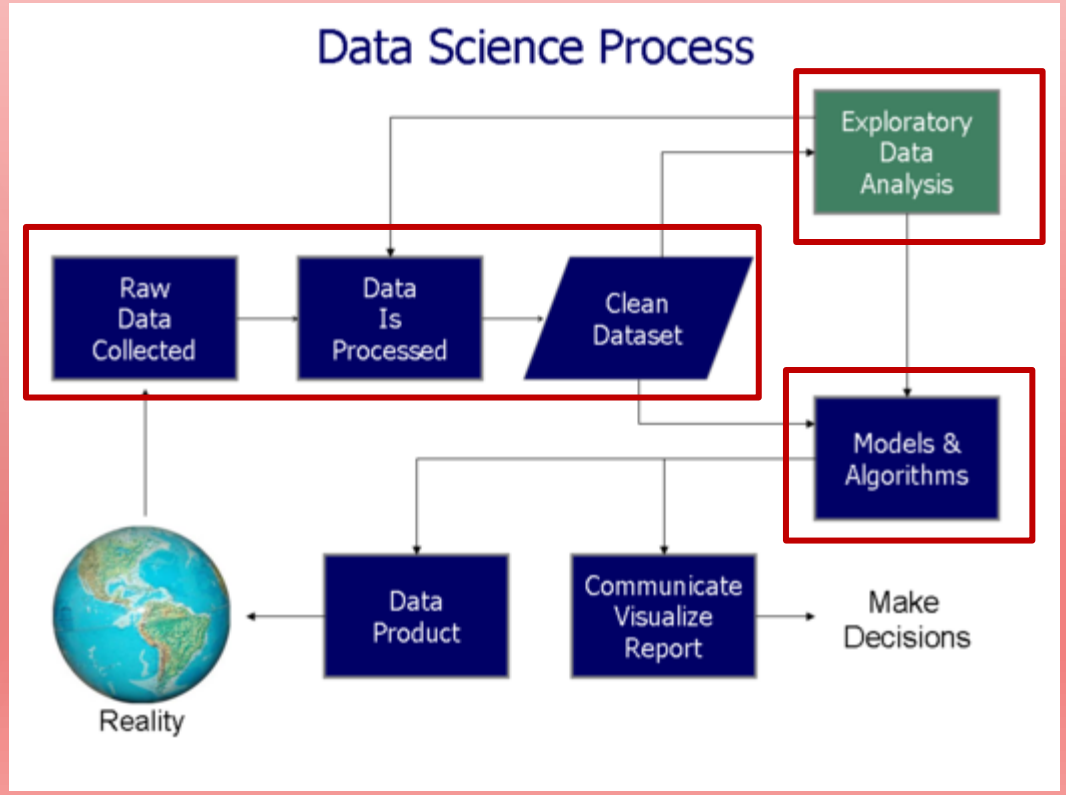
Likelihood of 'No' on that day

= P (Outlook = Rain|No) * P (Humidity = High| No) * P (Wind = Weak| No) * P(No)

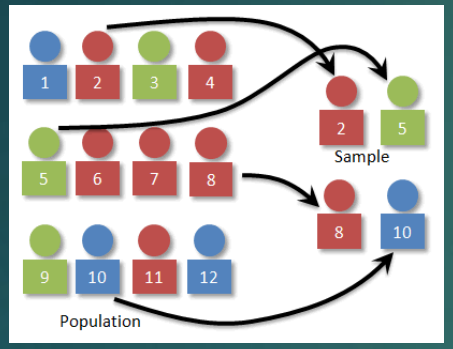
= 2/5 * 4/5 * 2/5 * 5/14 = 0 . 04571

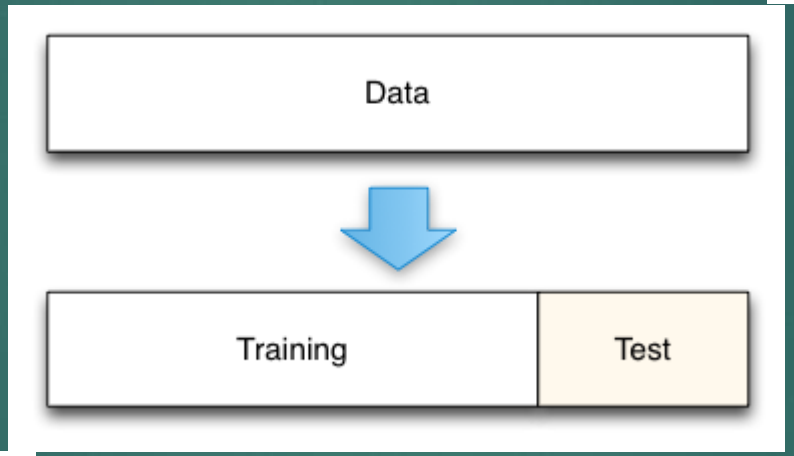
Recap

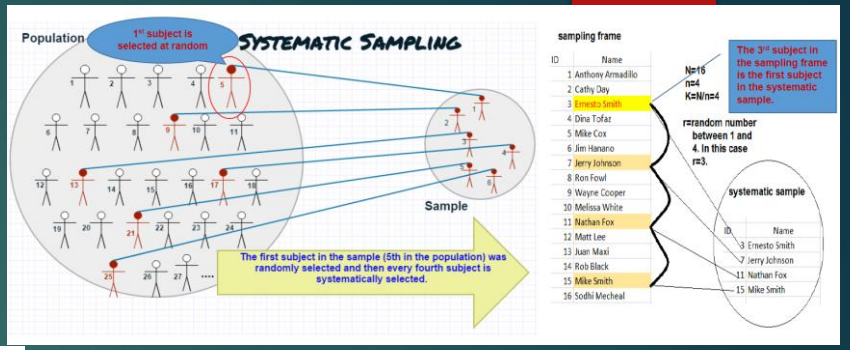
Data Science Process Flow



Split the data







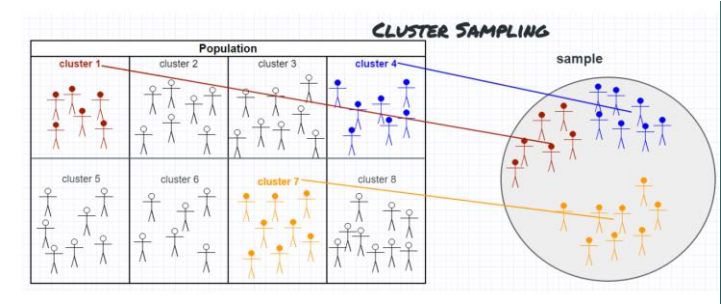
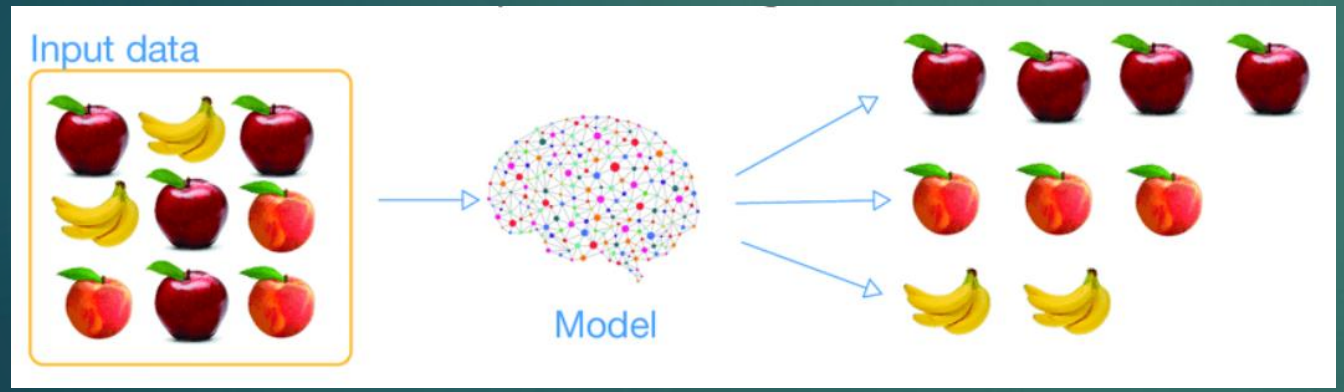
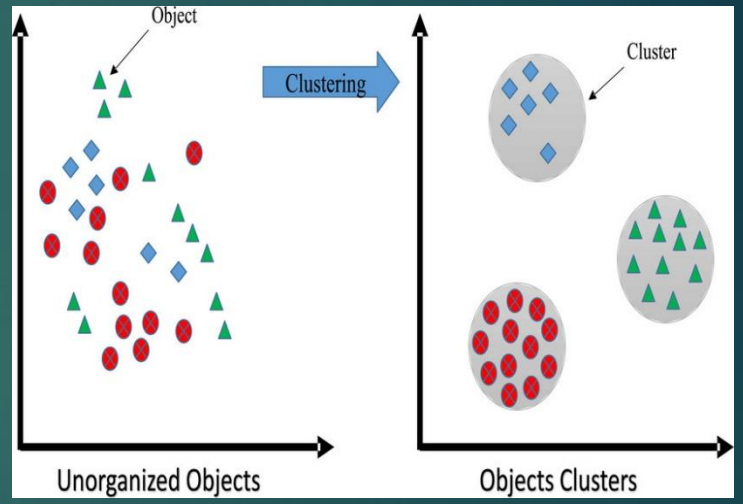


Fig. 3 Selecting datapoint from each subgroup

Modelling – Unsupervised Learning

Labels are not known, similar data is clustered together

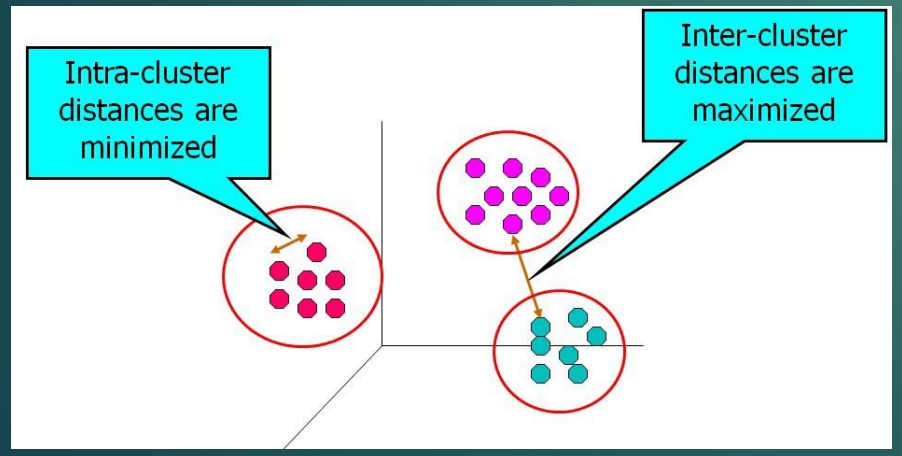


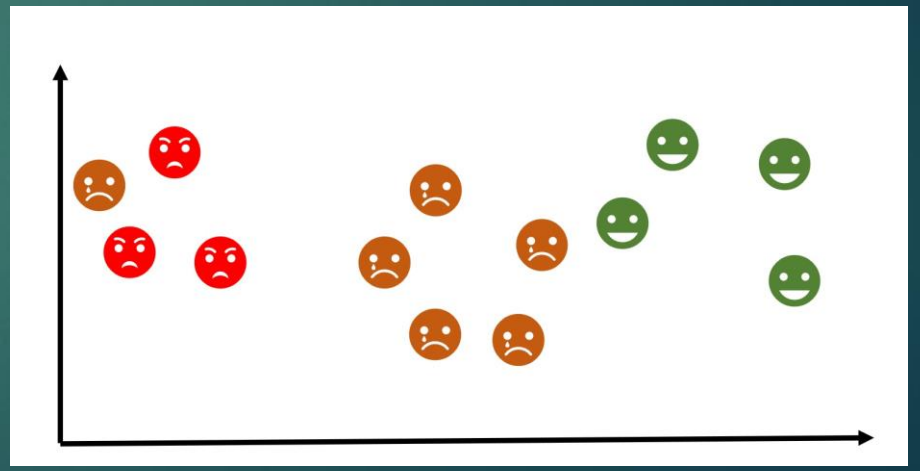


Unsupervised Learning - Clustering

What is Clustering

- Method of grouping data points that are close (or similar) to each other.
- Points within a cluster should be similar.
- Points from different clusters should be dissimilar.



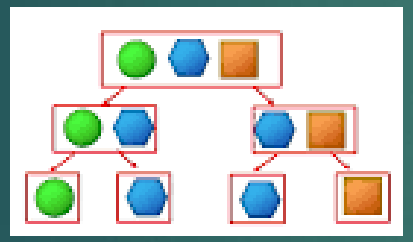


Types of Clustering

- **Connectivity Based**

- Data points closer in data space exhibit more similarity to each other than the data points lying farther away.

- E.g. Hierarchical Clustering

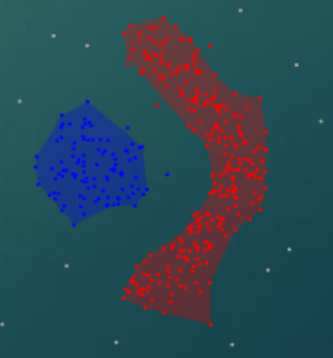


Centroid Based

- Similarity is derived by the closeness of a data point to the centroid of the clusters.
- E.g. K-Means algorithm

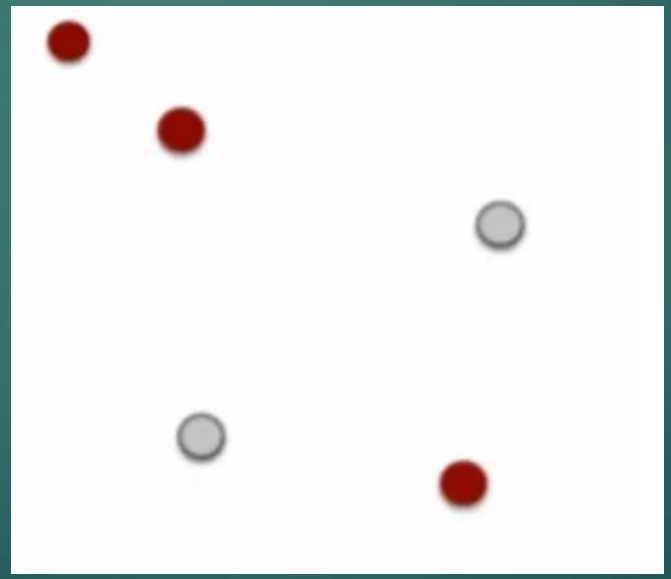
Density Based

- Point within different density regions are assigned in the same density region (cluster).
- E.g. DBSCAN



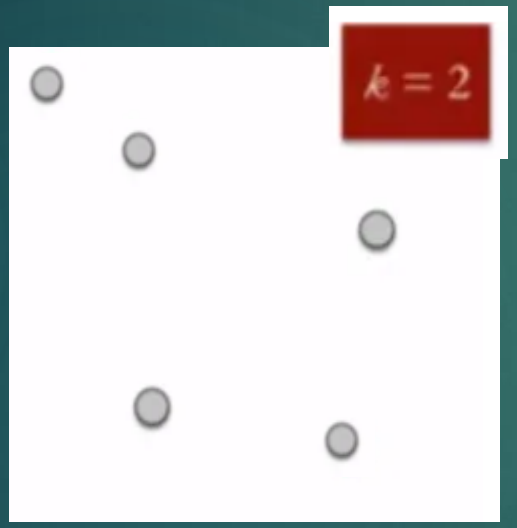
K-Means Clustering

- K means is an iterative clustering algorithm that aims to find local maxima in each iteration.

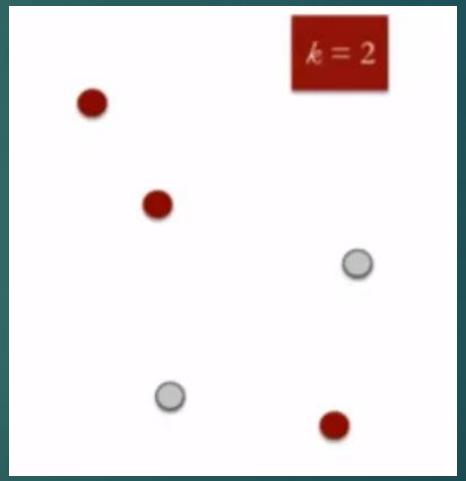


Algorithms Works

Step1:
Specify the desired number of clusters K

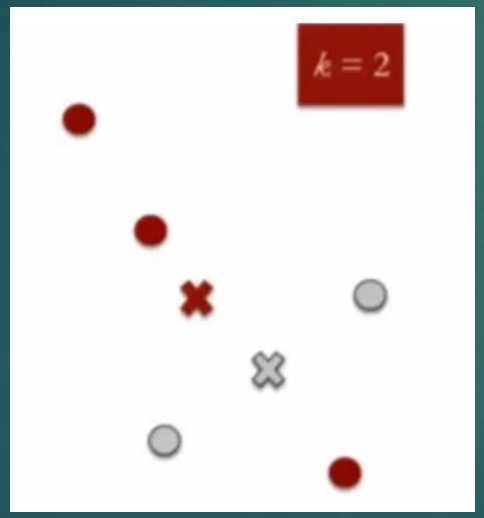


Step2:
Randomly assign each data point to a cluster

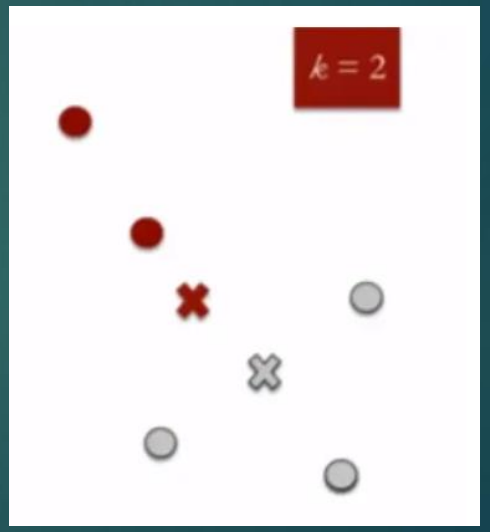


Algorithms Works

Step3:
Compute cluster centroids

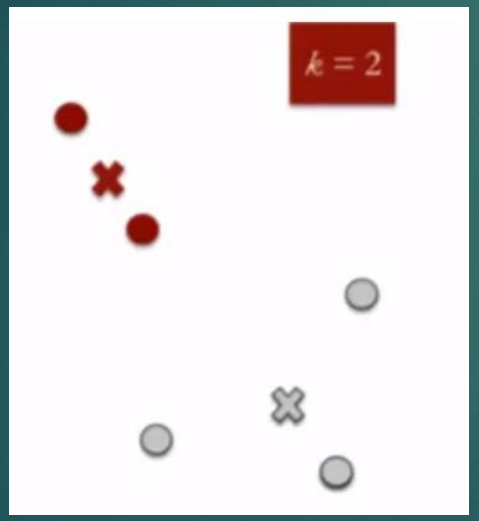


Step4:
Re-assign each point to the closest cluster centroid

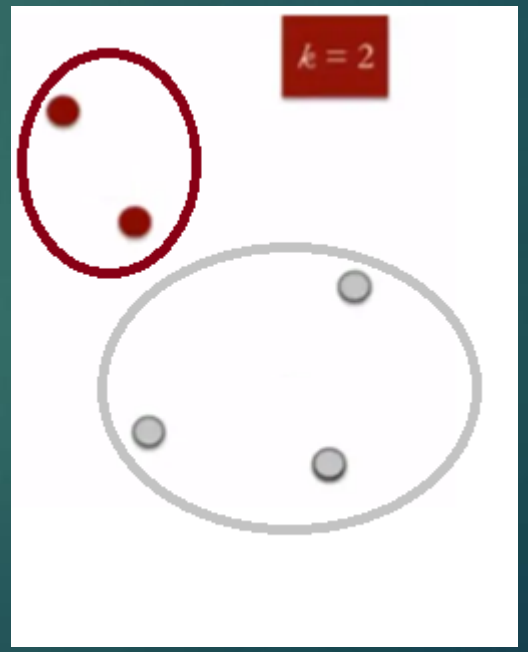


Algorithms Works

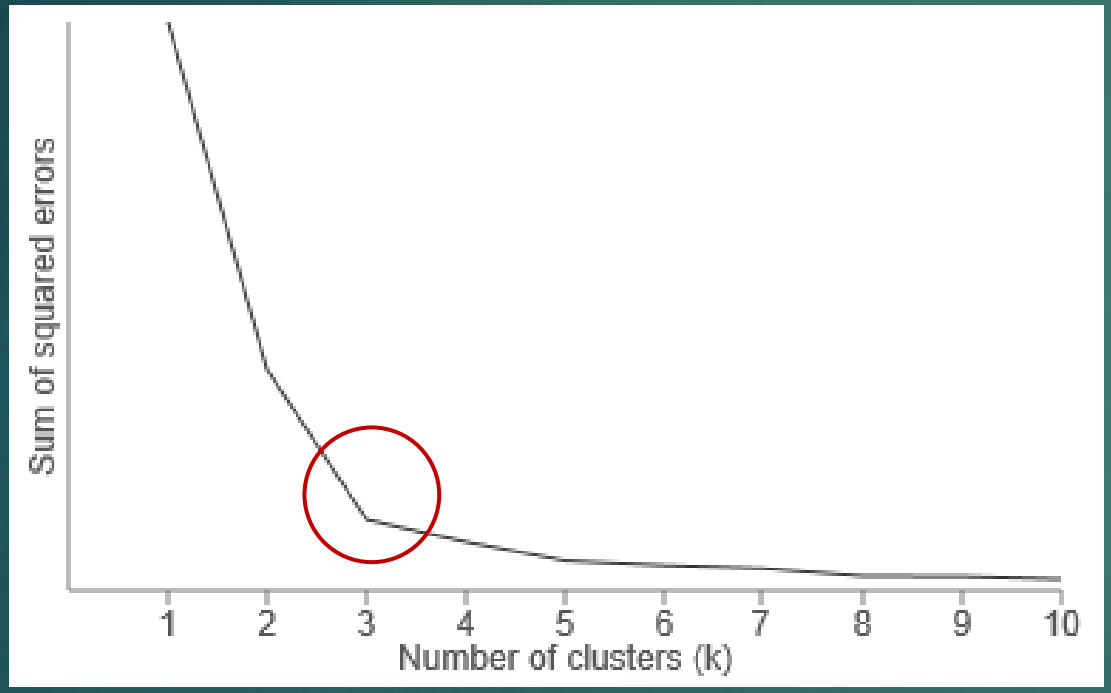
Step5:
Re-compute cluster centroids



Step6:
Repeat steps 4 and 5 until no improvements are possible



Finding Number of Clusters, k

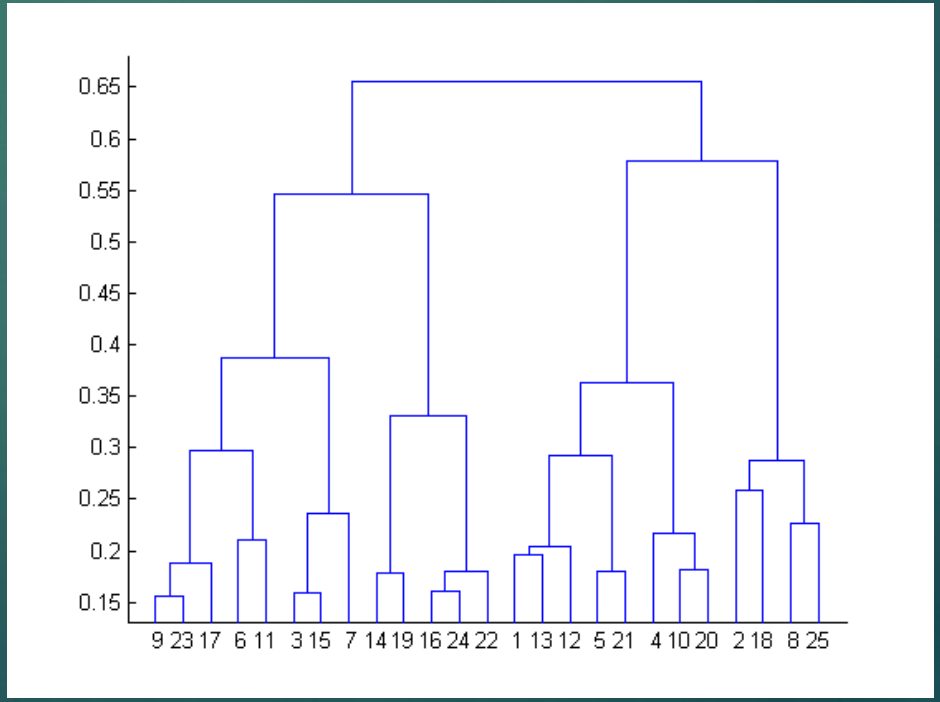


Because of the sudden drop we see an elbow in the graph. So the value to be considered for K is 3.

Elbow Method

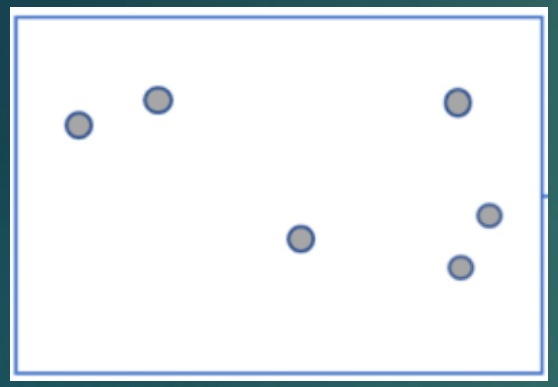
Hierarchical Clustering

- Algorithm that builds hierarchy of clusters.
- Initially each data point is considered as an individual cluster.
- At each iteration, the similar clusters merge with other clusters until one cluster or K clusters are formed

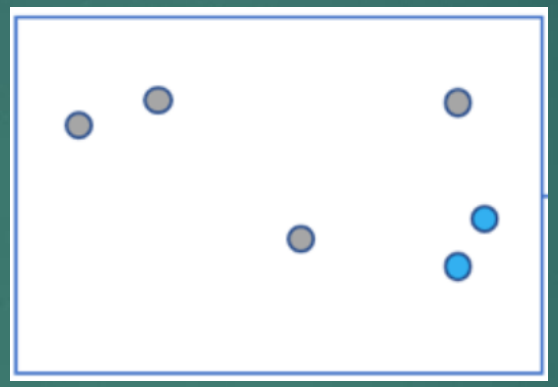


Algorithm Works

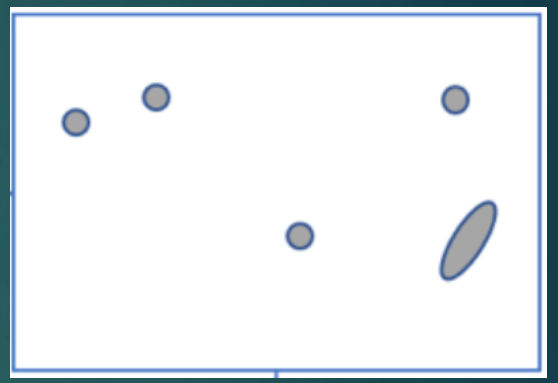
Data



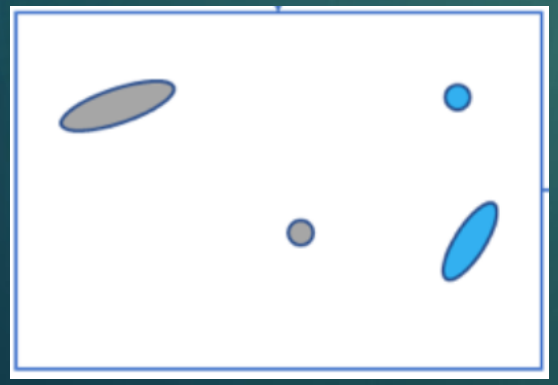
Step1



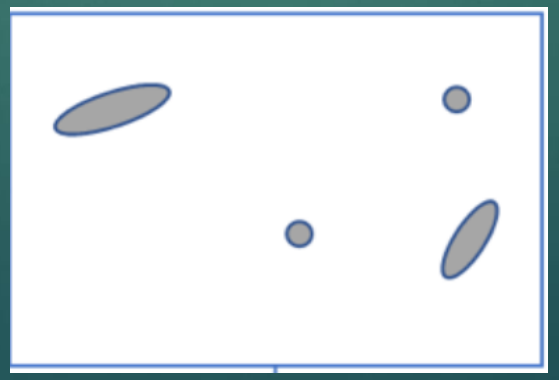
Step2



Step5



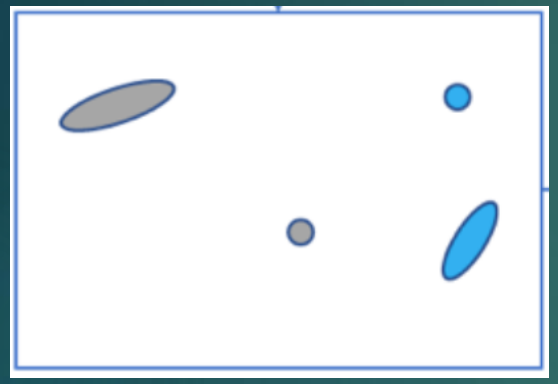
Step4



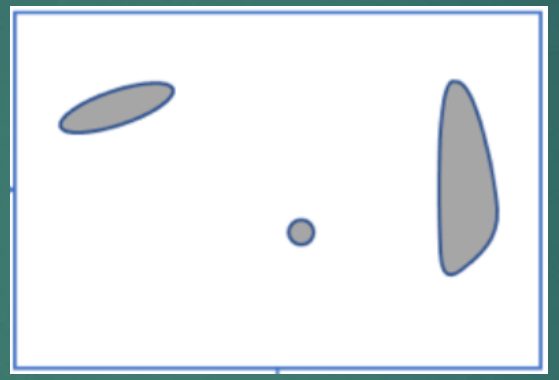
Step3

Algorithm Contd.

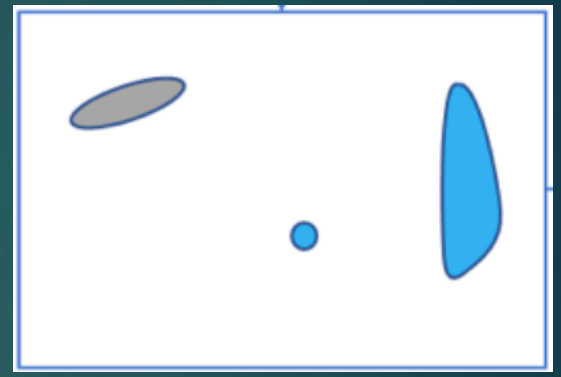
Step5



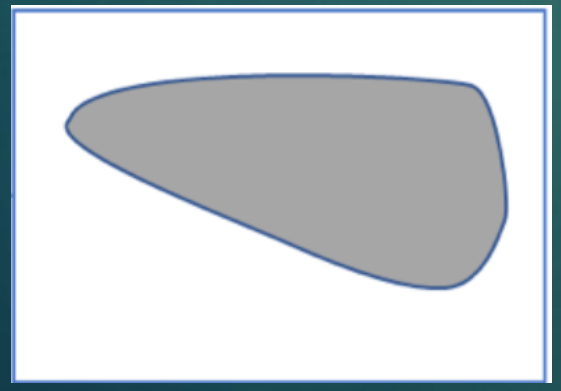
Step6



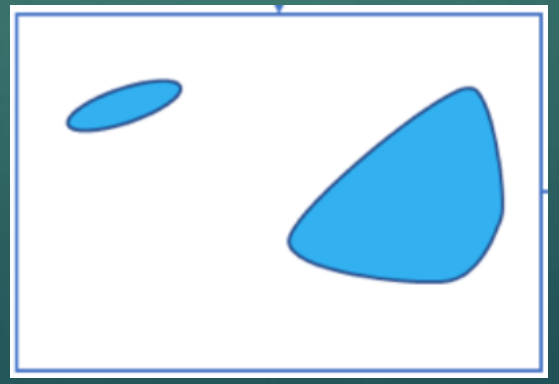
Step7



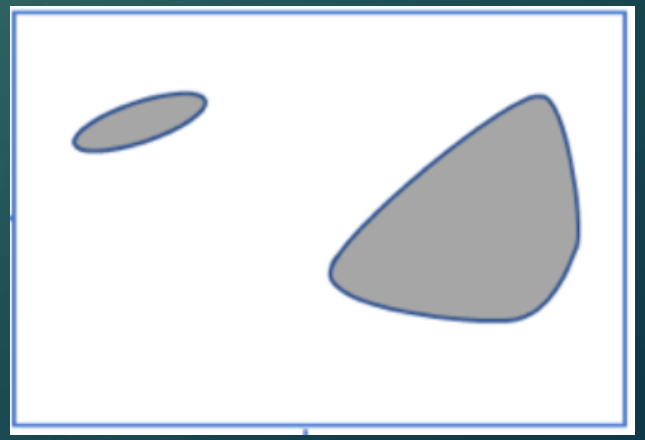
Step10



Step9

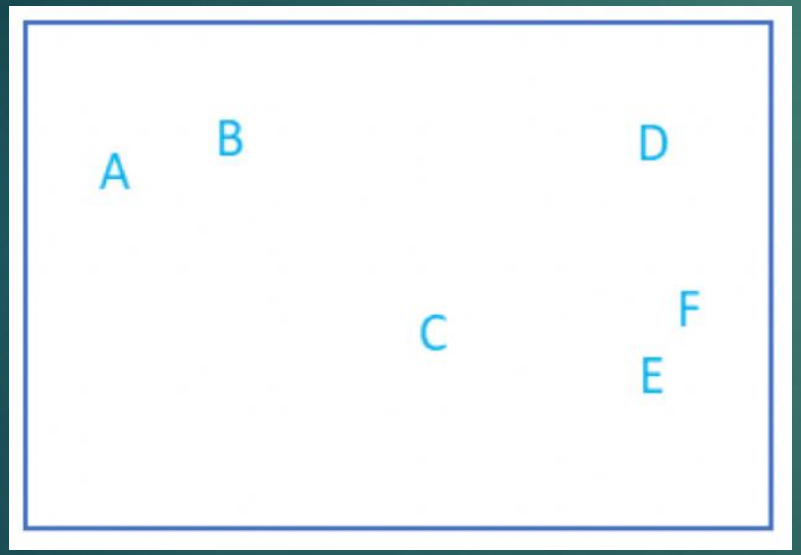


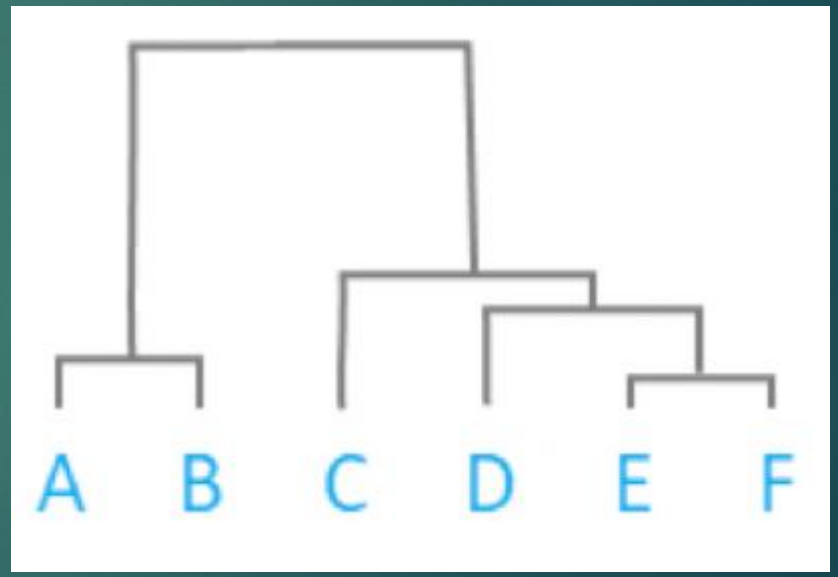
Step8



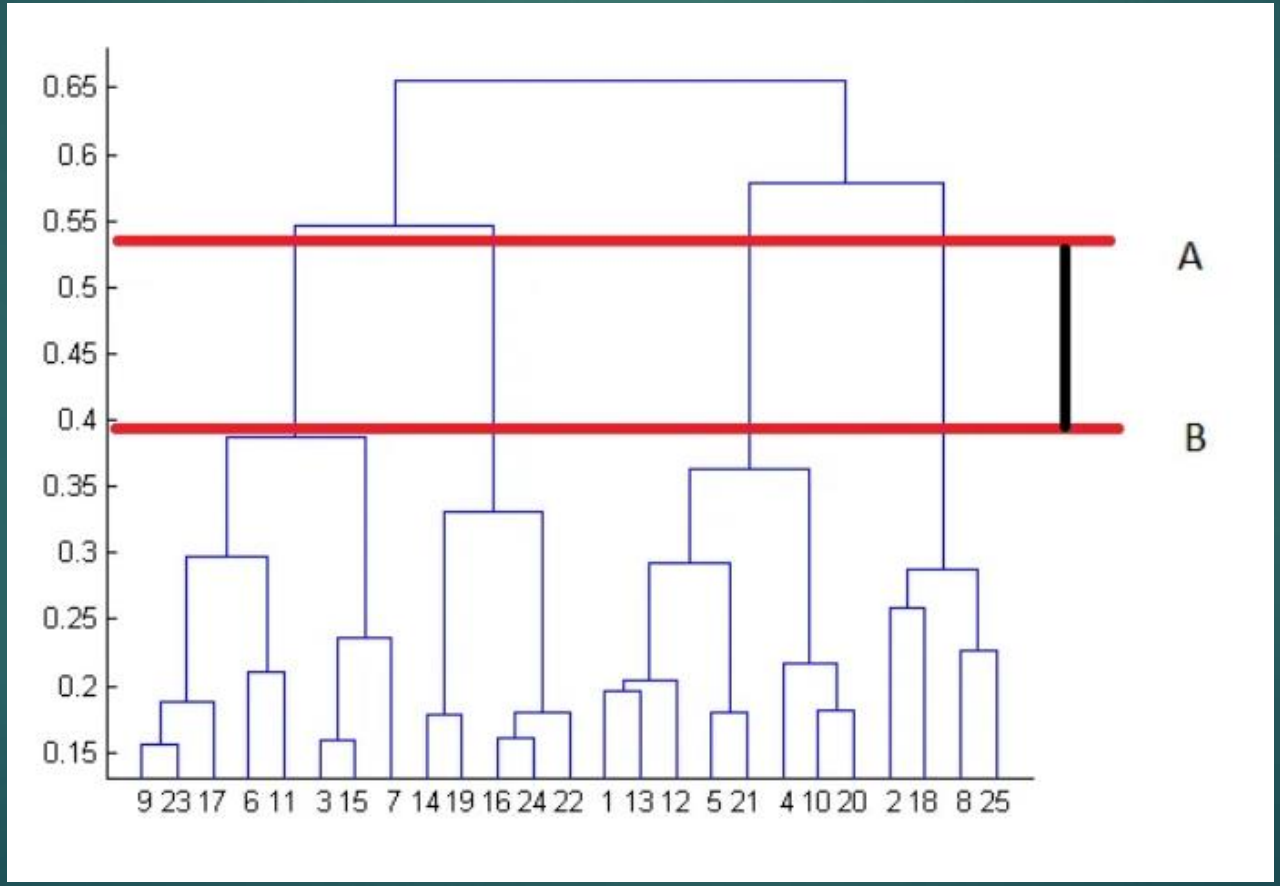
Dendrogram

It shows the hierarchical relationship between the clusters
A dendrogram is a tree-like diagram that records the sequences of merges or splits.





How to calculate number of clusters?

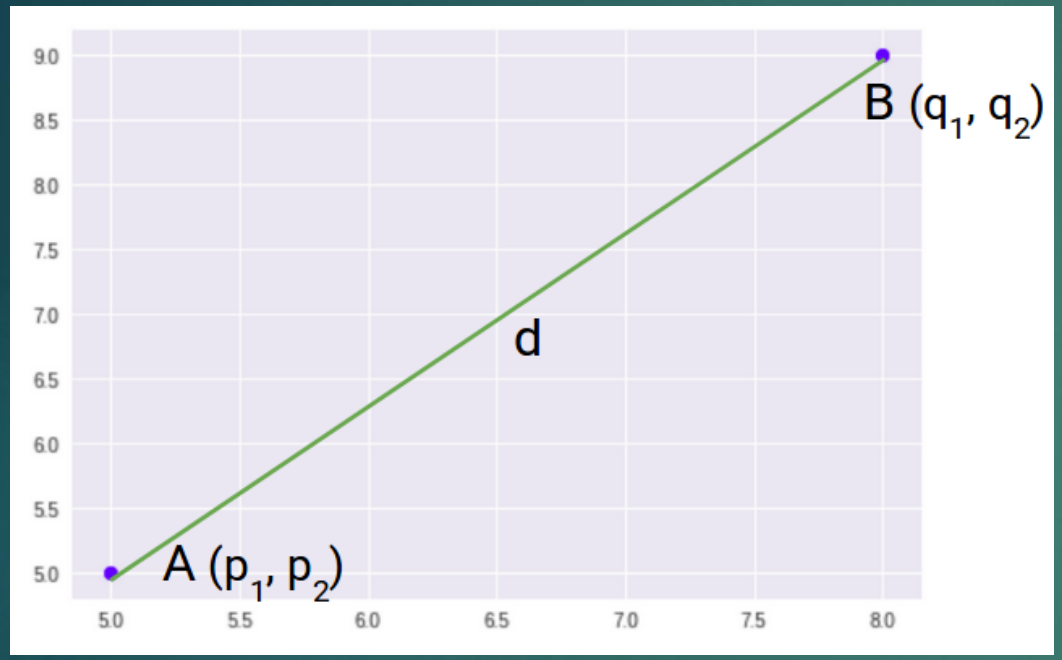


Types of Distance Metrics in Cluster Analysis

1. Euclidean Distance
2. Manhattan Distance
3. Hamming Distance
4. Cosine Similarity

EUCLIDEAN DISTANCE

Euclidean Distance represents the shortest distance between two points.

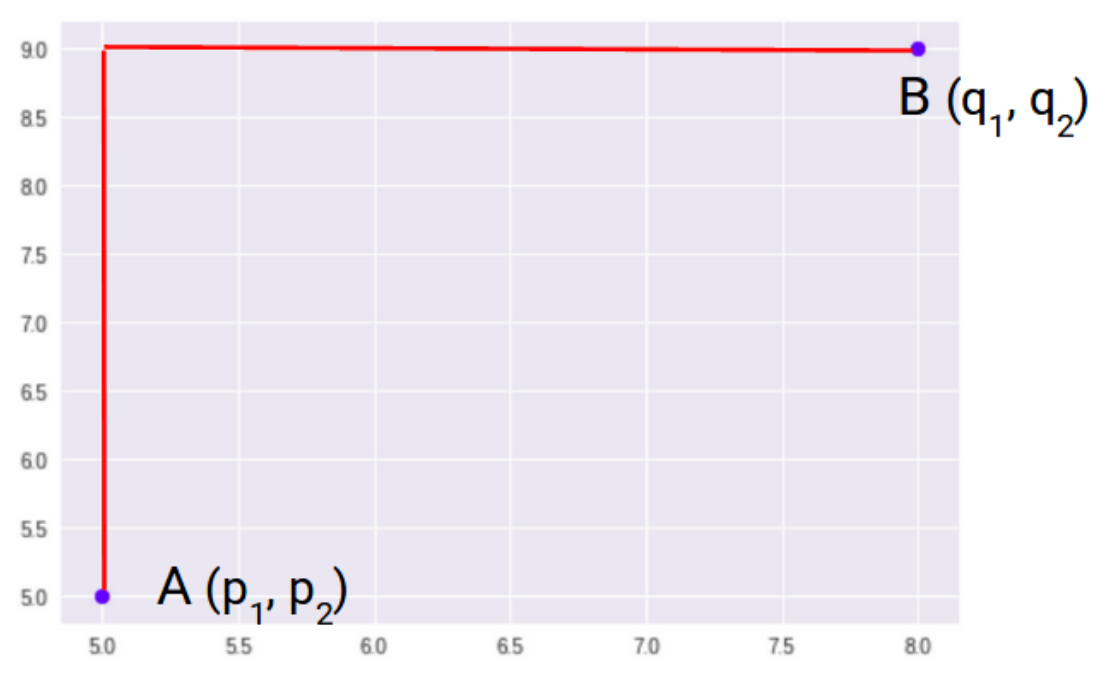


$$d = ((p_1 - q_1)^2 + (p_2 - q_2)^2)^{1/2}$$

$$D_e = \left(\sum_{i=1}^{n} (p_i - q_i)^2 \right)^{1/2}$$

Manhattan Distance

- Manhattan Distance is the sum of absolute differences between points across all the dimensions.
- We use Manhattan Distance if we need to calculate the distance between two data points in a grid like path.



Manhattan Distance is also known as city block distance.

$$d = |p_1 - q_1| + |p_2 - q_2|$$

$$D_m = \sum_{i=1}^{n} |p_i - q_i|$$

Example: x = [7, 3, 5] and y = [3, 2, 6].

The Manhattan distance is $|7 - 3| + |3 - 2| + |5 - 6| = 6$.

Hamming Distance

Hamming Distance measures the similarity between two strings of the same length.

“euclidean” and “manhattan”

Hamming Distance here will be 7

Cosine Similarity

- Cosine similarity is a metric used to measure how similar the documents are irrespective of their size.
- approach to match similar documents is based on counting the maximum number of common words between the documents.
- it measures the cosine of the angle between two vectors projected in a multi-dimensional space.

$$Cos\theta = \frac{\vec{a} \cdot \vec{b}}{\|\vec{a}\| \, \|\vec{b}\|} = \frac{\sum_1^n a_i b_i}{\sqrt{\sum_1^n a_i^2} \sqrt{\sum_1^n b_i^2}}$$

where, $\vec{a} \cdot \vec{b} = \sum_1^n a_i b_i = a_1 b_1 + a_2 b_2 + \cdots + a_n b_n$ is the dot product of the two vectors.

For example,

Let us consider that we have 2 vectors (both are 4 dimensional) -

X = [10, 4, 2, 5]

Y = [6, 4, 3, 2]

X.Y = (10 X 6) + (4 X 4) + (2 X 3) + (5 X 2) = (60 + 16 + 6 + 10) = 92

X.Y = 92

||X|| and ||Y|| are called P norm of the vectors.

Mathematically, the P norm of a given vector 'X' is represented as-

L2 norm of X,

||X|| = (|10|^2 + |4|^2 + |2|^2 + |5|^2) ^ ½ = (100 + 16 + 4 + 25) ^ 1/2 = 145 ^ 1/2

||X|| = 12.04

L2 norm of Y,

Y = [6, 4, 3, 2]

||Y|| = (|6|^2 + |4|^2 + |3|^2 + |2|^2) ^ ½ = (36 + 16 + 9 + 4) ^ 1/2 = 65 ^ 1/2

||Y|| = 8.06

cosine similarity of 2 vectors X and Y-

X = [10, 4, 2, 5] Y = [6, 4, 3, 2]

X.Y = 92

||X|| = 12.04 ||Y|| = 8.06

Cosine similarity (X, Y) = 92/ (12.04 X 8.06) = 92/97.06 = 0.94

L^p norm:

$$|X|_p = \left(\sum_{i=1}^{n} |X_i|^p\right)^{1/p}$$

Cosine similarity (X, Y) = 0.94

Thank you