MACHINE LEARNING WITH PYTHON

- ## CONFUSION MATRIX

Row represents an actual class

Column represents a predicted class

		Predicted	
		(N)	(Y)
Actual	(N)	TN	FP
	(Y)	FN	TP

{ Example }

Predicted Values.

		Positive(1)	Negative (0)
Actual Values	Positive(1)	TP (x)	FN (y)
	Negative(0)	FP (z)	TN (w)

Now,

1) Accuracy

$$A = \frac{TP + TN}{Total} = \frac{x + w}{x + y + z + w}$$

2) Precision

$$P = \frac{TP}{TP + FP} = \frac{x}{x + z}.$$

3) Recall or True Positive Rate

$$TPR = \frac{TP}{TP + FN} \Rightarrow \frac{x}{x + y}$$

4) False Positive Rate

$$FPR = \frac{FP}{TN + FP} = \frac{2}{w+2}.$$

5) F - Score

A way to combine both precision and recall into a single measure.

$$F_1 \text{ Score} = \frac{2 \cdot \times (\text{Recall} \times \text{Precision})}{\text{Recall} + \text{Precision}}$$

~~For multi~~

Generalised,

$$F_\beta \text{ Score} = \frac{(1+\beta)^{*} (\text{Recall} \times \text{Precision})}{\text{Recall} + \text{Precision}}$$

β signifies the importance of recall compared to precision

For eg. β = 2, means recall is twice as important as precision.

CLUSTER ANALYSIS

Clustering is a method of dividing the objects into clusters which are similar between them and are dissimilar to the objects belonging to another cluster.

It deals with finding a structure in collection of unlabelled data.

TYPES OF CLUSTERING

1) Hierarchical Clustering

→ It is separating data into different groups based on some measure of similarity.

→ Agglomerative - bottom up
Divisive - top down

→ In agglomerative clustering, we stop when we get a single cluster.

→ A tree like structure is formed in this clustering

2) Partitional Clustering

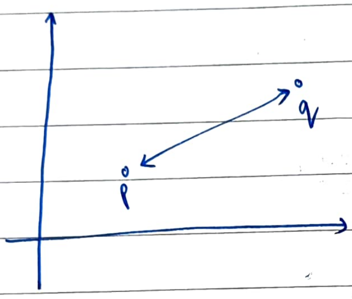
→ It is used to form clusters (k) for n objects where k <= n.

DISTANT MEASURES

1) **Euclidean Distance**

→ It is the distance between two points.

$$\boxed{d = \sqrt{\sum_{i=1}^{n} (q_i - p_i)^2}}$$

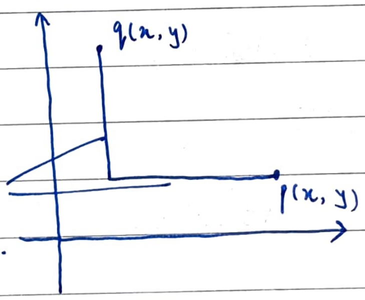


Disadvantages:

→ distance might be skewed

2). **Manhattan Distance**

It is the sum of the horizontal and vertical components or the distance between two points measured along axes at right angles.

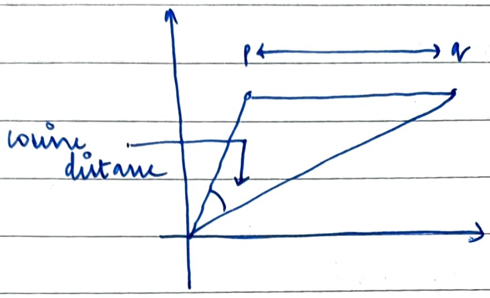


Mahattan Distance.

$$\boxed{d = \sum_{i=1}^{n} |q_x - p_x| + |q_y - p_y|}$$

3) Cosine Distance

It is used to measure the angle between the two vectors formed by joining the origin point.



Disadvantage:

→ Magnitude of vectors not taken into account.

$$\boxed{d = \frac{\sum_{i=0}^{n-1} q_i - p_i}{\sum_{i=0}^{n-1} (q_i)^2 + \sum_{i=0}^{n-1} (p_i)^2}}$$

4) Hamming Distance

It is the no. of values that are different between two vectors. It is used to compare two binary strings of equal length.

JACCARD COEFFICIENT

It measures the similarity of the two datasets items as the intersection of items divided by the union of the data items.

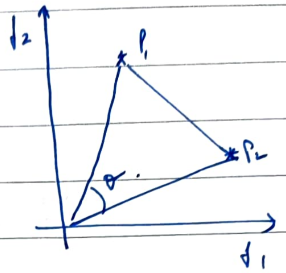
$$J(A, B) = \frac{|A \cap B|}{|A \cup B|}$$

COSINE SIMILARITY

It is used to find the similarity between two points.

$$\text{cos-simi} = \cos\theta .$$

$[-1, 1]$



$$\boxed{\text{cos-dist} = 1 - \text{cos-simi}}$$

PEARSON CORRELATION COEFFICIENT

$$r = \frac{n(\Sigma xy) - (\Sigma x)(\Sigma y)}{\sqrt{[n\Sigma x^2 - (\Sigma x)^2][n\Sigma y^2 - (\Sigma y)^2]}}$$

$r = 1$ - Positive correlation

$r = -1$ - Negative correlation.

SIMPLE MATCHING COEFFICIENT

$$SMC = \frac{\text{No. of Matching Attributes}}{\text{No. of Attributes.}}$$

$$= \frac{M_{00} + M_{11}}{M_{00} + M_{01} + M_{10} + M_{11}}$$

M_{00} - A : NO B : NO

M_{01} - A : NO B : YES

M_{10} - A : YES B : NO

M_{11} :- A : YES B : YES.

K - MEANS CLUSTERING

→ K - Means performs the division of objects into clusters that share similarities and are dissimilar to the objects in the other clusters.

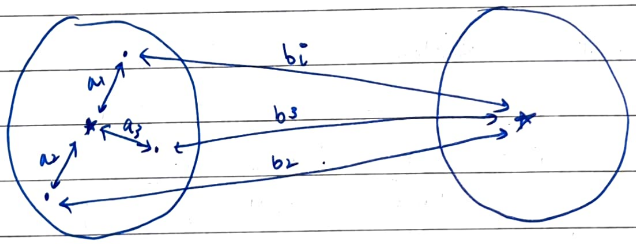
→ To determine the optimum value of k, we use hit & trial technique or the elbow technique.

→ K - means - Euclidean Distance
K - median - Manhattan Distance

Silhoutte Method

It is used to asses how good the cluster assignment is for that point

$$S_i = \frac{b_i - a_i}{\max(b_i, a_i)}$$



if $b_i > a_i$ then S_i = +ve meaning the data are not misclassified

Avg S_i ⇒ ≥ 0.5 - good evidence of reality of clusters
$0.25 - 0.5$ - Some evidence of reality of clusters
< 0.25 - Scant evidence of reality of clusters

/ /

Psuedo F - Statistic

Let, k = no. of clusters

$\Sigma n_j = N$: total sample size

x_{ij} = j^{th} data value in i^{th} cluster

m_i = centroid of i^{th} cluster

M = grand mean of all the data

$$D(a,b) = \sqrt{\Sigma (a_i - b_i)^2}$$

→ Sum of squares b/w clusters

$$SSB = \sum_{i=1}^{k} n_i \cdot Dist^2(m_i, M)$$

→ Sum of square within clusters | Sum of square of Error

$$SSE = \sum_{i=1}^{k} \sum_{j=1}^{n} Dist^2(x_{ij}, m_i)$$

→ Psuedo F - statistic

$$\boxed{F = \frac{MSB}{MSE} = \frac{SSB / k - 1}{SSE / N - k}}$$