UNIT IV – CLUSTERING AND UNSUPERVISED LEARNING

Similarity and distance measures – Cluster evaluation metrics – K-Means clustering – K-Medoids clustering – Hierarchical clustering (Agglomerative and Divisive) – Density-based clustering (DBSCAN) – Spectral clustering – Principal Component Analysis (PCA) – Dimensionality reduction using t-SNE and UMAP – Association rule mining – Anomaly detection – Federated Learning basics – Edge AI for distributed learning.

SIMILARITY AND DISTANCE MEASURES

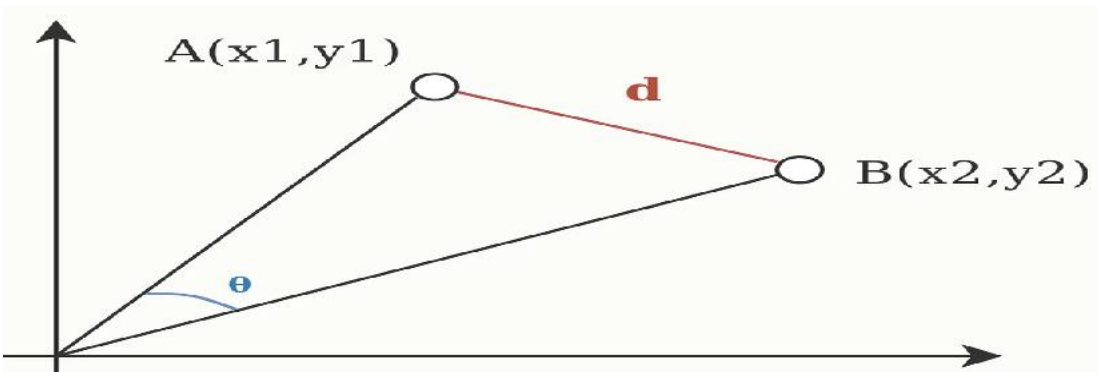
Similarity

- The state or fact of being similar or Similarity measures how much two objects are alike.
- Similarity measure in a machine learning context is a distance with dimensions representing features of the objects.
- If distance is small, two objects are very similar where as if distance is large we will observe low degree of similarity.

1) Cosine Similarity:

- Cosine similarity is a metric used to measure how similar the documents are irrespective of their size.
- Mathematically, it measures the cosine of the angle between two vectors projected in a multi-dimensional space.
- The cosine similarity is advantageous because even if the two similar documents are far apart by the Euclidean distance (due to the size of the document), chances are they may still be oriented closer together.
- The smaller the angle, higher the cosine similarity.

FORMULA



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

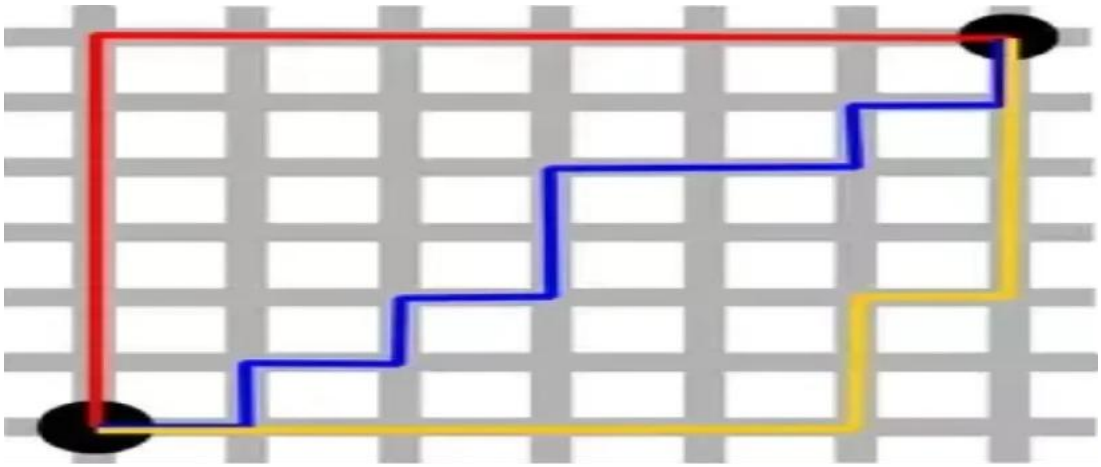
2) Manhattan distance:

- Manhattan distance is a metric in which the distance between two points is the sum of the absolute differences of their Cartesian coordinates.
- In a simple way of saying it is the total sum of the difference between the x-coordinates and y-coordinates.

FORMULA:

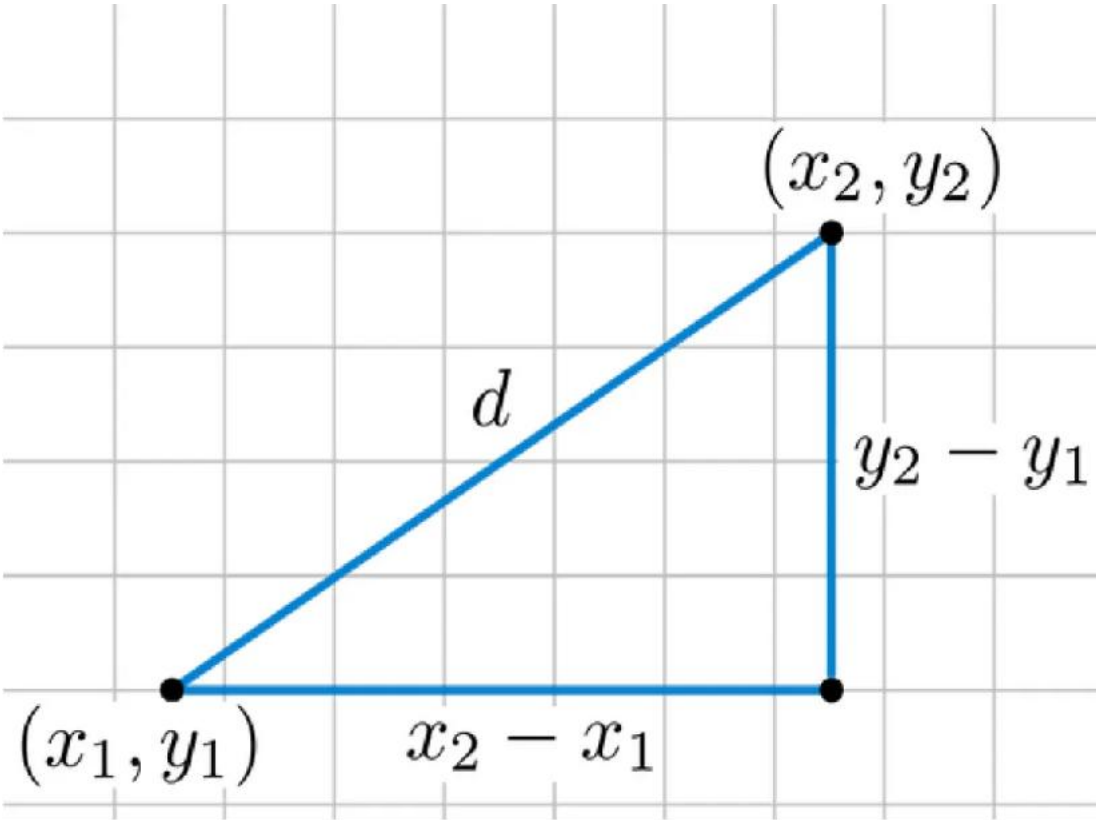
In a plane with p1 at (x1, y1) and p2 at (x2, y2)

$$|x_1 - x_2| + |y_1 - y_2|$$



3) Euclidean distance

- The Euclidean distance between two points in either the plane or 3-dimensional space measures the length of a segment connecting the two points.
- It is the most obvious way of representing distance between two points.
- The Pythagorean Theorem can be used to calculate the distance between two points, as shown in the figure below.



FORMULA:

- If the points (x1,y1)(x1,y1) and (x2,y2)(x2,y2) are in 2-dimensional space, then the Euclidean distance between them is

$$\sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2}.$$

4) Minkowski distance

- Minkowski distance is a generalisation of the Euclidean and Manhattan distances.

FORMULA

- The Minkowski distance of order p between two points is defined as

$$X = (x_1, x_2, \ldots, x_n) \text{ and } Y = (y_1, y_2, \ldots, y_n) \in \mathbb{R}^n$$

$$D(X,Y) = \left(\sum_{i=1}^{n} |x_i - y_i|^p \right)^{1/p}$$

5) Jaccard similarity:

- The Jaccard index, also known as Intersection over Union and the Jaccard similarity coefficient is a statistic used for gauging the similarity and diversity of sample sets.
- The Jaccard coefficient measures similarity between finite sample sets, and is defined as the size of the intersection divided by the size of the union of the sample sets.

FORMULA

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

