Everything on Clustering

Clustering is unsupervised learning. That means there is no training set, no true function and no classification

Clustering is about partitioning the data according to what we need to do. It groups together similar items. Similarity is domain/problem specific

To evaluate similarity distance functions are used. For numeric data we calculate :

Euclidean distance

Manhattan distance

For categorical data : 0s and 1s indicating presence or absence

Euclidean distance : $d(x,y) = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2 + ... + (x_n - y_n)^2}$

Manhattan distance : $d(x,y) = |x_1 - y_1| + |x_2 - y_2| + ... + |x_n - y_n|$

K-means Clustering Algorithm (Euclidean distance only)

Goal is to minimize the sum of square of distance from all data points to their means

Step-by-step process

1) Pick K different points from the data and assume they're cluster centres

2) Assign each point to the closest centre

3) Generate new cluster centres

4) Repeat until the cluster centres do not change

K-Means Disadvantages

1) Only measures the mean of each cluster, doesn't give any information about its shape → assume it's round

2) Requires to know K before start clustering

3) The distance measure, assumes that all ranges are equally important

Hierarchical Clustering Algorithm

1) Start with same number of clusters as data points. Every point is a cluster

2) Find the two clusters that are closest to each other and merge them into one

3) Calculate their centre

4) Repeat until you have the required number of clusters or everything is in one cluster, if no number is specified

A dendogram is used to represent the algorithm

- X-axis represents data points
- Y-axis represents number of iterations

Population size : Number of data points in cluster

Variance and range : Distance between data and centre