Clustering

Clustering is the classification of objects into different groups of data.

The types of clustering are

1. Hierarchical
    a. Agglomerative (Bottom-up) - Begins with each element as a separate cluster and merges them into successively larger clusters
    b. Divisive (Top-down) - Begins with one large cluster and divides it into successively smaller clusters.
    c. In this, data points can belong to more than one cluster.
2. Partitional
    a. Determines all clusters at once
    b. K-means or derivatives
    c. Points can belong only to one cluster

Agglomerative clustering

Algorithm

1. Compute proximity matrix
2. Let each data point be a cluster
3. Repeat
    a. Merge two closest clusters and update proximity matrix
4. End when only one cluster remains

Divisive Clustering

Divisive clustering is the exact inverse of agglomerative clustering, where the main cluster is separated based on the furthermost points by the proximity matrix.

Distance measures

1. Euclidean Distance
2. Manhattan Distance

K-Means Clustering

The K-Means algorithm is similar to the EM algorithm for Gaussian Mixtures, as they both try to find the centers of natural clusters of data. This algorithm separates data into k partitions, where k < n.

K-Means is done to minimise the mean squared error of data points to each cluster. This is done by calculating J as

$$J = \sum_{j=1}^{K} \sum_{n \in S_j} |x_n - \mu_j|^2$$

where x_n is the datapoint and μ_j is the centroid

The algorithm is

1. Begin by choosing a value of k
2. Put any initial partition that classifies the data into k clusters.
    a. Take the first k training examples as single-element clusters
    b. Assign each of the remaining N-k points to the cluster with the nearest centroid.
    c. After each assignment, recompute the centroid
3. Take each sample in the sequence and compute its distance from the centroid of each of the clusters. If the sample is not in the correct cluster, set it to the correct cluster and recalculate the centroid.
4. Repeat step 3 until convergence, where no reassignment occurs.

Bisecting K Means

In the bisecting K means method

1. Decide on a value K
2. Start with k=2 and split the data into 2 clusters
3. Choose one of the clusters with the greater mean intra-cluster distance and split it into 2
4. Repeat step 3 until k = K.

Disadvantages of K means clustering

1. When data is not a lot, initial grouping determines the cluster significantly
2. The number of clusters K must be determined beforehand, and does not yield the same result after each run
3. It is sensitive to the initial condition, and may cause local optima traps.