STA 3180 Statistical Modelling: Clustering

1. Given a set of data points, explain the differences between hierarchical and k-means clustering.

Solution: Hierarchical clustering is a type of clustering algorithm that builds a hierarchy of clusters. It starts by assigning each data point to its own cluster and then iteratively merges the most similar clusters until a single cluster is formed. K-means clustering is a type of clustering algorithm that partitions the data into k clusters based on the Euclidean distance between the data points and the cluster centroids.

2. Describe the steps involved in the k-means clustering algorithm.

Solution: The steps involved in the k-means clustering algorithm are as follows:

1. Select the number of clusters (k)

2. Randomly select k data points as the initial cluster centroids

3. Calculate the Euclidean distance between each data point and the cluster centroids

4. Assign each data point to the closest cluster centroid

5. Calculate the mean of all the data points in each cluster

6. Update the cluster centroids with the new means

7. Repeat steps 3-6 until the cluster centroids no longer change

3. Explain the concept of within-cluster sum of squares (WCSS).

Solution: Within-cluster sum of squares (WCSS) is a measure of the variability of the data points within a cluster. It is calculated by summing the squared Euclidean distances between each data point and the cluster centroid. The WCSS is used to evaluate the quality of the clusters and can be used to determine the optimal number of clusters.

4. What is the elbow method and how is it used to determine the optimal number of clusters?

Solution: The elbow method is a heuristic used to determine the optimal number of clusters for a given dataset. It plots the WCSS for different values of k and looks for an “elbow” in the plot, which indicates the optimal number of clusters. The elbow method is based on the assumption that the WCSS will decrease as the number of clusters increases, but at some point the decrease will become less significant. The point at which this occurs is the optimal number of clusters.