Project 3

Surya Prakash(20CS10038)

1 Introduction

The objective of this project is to segment customer data from an airline company using clustering techniques.

In this project, we use two clustering methods:

1. K-means clustering
2. Complete Linkage Divisive Clustering

2 K-means Clustering

This is a partition clustering technique in which the data points are partitioned into clusters based on their distance from the chosen centroids.

Initially, we start out with k random data points as the centroids, assign data points to each cluster depending on which centroid the data point is closest to, and then update the centroids for each cluster as the mean point of the cluster. We repeat this for 20 iterations.

3 Evaluating K-means Clustering

To evaluate the performance of K-means clustering we use silhouette score to measure the quality of clustering.

$$S = \frac{b - a}{max(a, b)}$$

3.1 K = 3

- Time taken for clustering = 2.91 seconds
- Time taken for score calculation = 102.2 seconds
- Silhouette Score = 0.627

3.2 K = 4

- Time taken for clustering = 2.65 seconds
- Time taken for score calculation = 99.4 seconds
- Silhouette Score = 0.633

3.3 K = 5

- Time taken for clustering = 3.23 seconds
- Time taken for score calculation = 98.15 seconds
- Silhouette Score = 0.559

3.4 K = 6

- Time taken for clustering = 3.74 seconds
- Time taken for score calculation = 100.06 seconds
- Silhouette Score = 0.523

3.5 Optimal K value

From the above results, it can be seen that for K = 4, we get the best Silhouette scores and the value of S = 0.633 indicates a clustering that is fairly separated.

4 Complete Linkage Divisive Clustering

This is a hierarchical clustering technique in which we start with all the data points as a single cluster and then split clusters until we have K clusters.

4.1 Complete Linkage

Distance between clusters is given by the complete linkage strategy. In this we take the distance as the distance between the furthest pair of points from the two clusters.

4.2 Split criterion

The split criterion chosen is that the larger of the two furthest clusters in the set of clusters will be split into two.

4.3 Split subroutine

For divisive clustering, we need to use some flat clustering technique like K-means to split the chosen cluster. Here we use K-means with k = 2.
Note: Choosing the split criterion as the cluster with maximum distance between farthest points gave poorer results where only one cluster was being repeatedly split into smaller clusters.

5 Similarity Evaluation

5.1 Jaccard Similarity

The Jaccard Similarity between two sets A, B is $\frac{|A \cup B|}{|A \cap B|}$
For the similarities between the clusters formed in the two cases with k = 4. We have the mappings and scores as,

- Cluster 0 in k-means clustering maps to cluster 0 in divisive clustering with a Jaccard similarity score of 0.485
- Cluster 1 in k-means clustering maps to cluster 3 in divisive clustering with a Jaccard similarity score of 0.725
- Cluster 2 in k-means clustering maps to cluster 1 in divisive clustering with a Jaccard similarity score of 0.276
- Cluster 3 in k-means clustering maps to cluster 2 in divisive clustering with a Jaccard similarity score of 0.278