HW #4

COMP 5130

1. (50 points) Both k-means and k-medoids algorithms can perform effective clustering. Illustrate the strengths and weaknesses of k-means in comparison with the k-medoids algorithm. Also, illustrate the strengths and weaknesses of these schemes in comparison with a hierarchical clustering scheme (such as AGNES).

1.1)

**K-Means Strengths:**

**Efficiency:** K-means is computationally more efficient than k-medoids, especially for large datasets.

**Scalability:** It can handle many data points and features.

**Cluster Shape:** Works well with clusters of spherical shapes.

**K- Means Weaknesses:**

**Sensitivity to Initial Seeds:** K-means is sensitive to the initial placement of centroids. Different initializations can lead to different results.

**Assumption of Equal Variance:** Assumes that clusters have equal variance, which may not always be the case.

3**Not Robust to Outliers:** Sensitive to outliers and noise in the data.

**K-Medoids Strengths:**

**Robust to Outliers:** More robust to outliers and noise due to its use of actual data points as representatives.

**No Assumption on Cluster Shape:** It can work with any arbitrary distance/similarity measure, making it more flexible.

**Stability:** More stable and less sensitive to initialization than k-means.

**K-Medoids Weaknesses:**

**Computationally Expensive:** K-medoids is more computationally expensive, especially for large datasets.

**Not Scalable:** Does not scale well with an increasing number of data points and features.

**Requires Distance Matrix:** Requires a precomputed distance matrix, making it impractical for some types of data.

1.2)

K-Means and K-Medoids Strengths:

Both k-means and k-medoids can be more scalable than hierarchical clustering, especially for large datasets.

Both methods allow for the specification of the number of clusters, providing more control over the result.

K-Means and K-Medoids Weaknesses:

Requires a priori knowledge of the number of clusters, which may not be known in advance.

Both methods are sensitive to the initial placement of centroids or medoids.

Hierarchical Clustering Strengths:

No Assumption on Number of Clusters: Does not require the number of clusters to be specified beforehand.

Hierarchy Exploration: Provides a hierarchical structure, allowing exploration of clusters at different granularity levels.

Hierarchical Clustering Weaknesses:

Computationally Expensive: Hierarchical clustering can be computationally expensive, especially for large datasets.

Lack of Scalability: May not scale well with an increasing number of data points.

Sensitive to Noise: Sensitive to noise and outliers, impacting the dendrogram structure.

2. (50 points) Clustering has been popularly recognized as an important data mining task with broad applications. Give one application example for each of the following cases:

(a) An application that takes clustering as a major data mining function.

Fraud Detection in Financial Transactions:

clustering is used for fraud detection in transactions as a major data mining function, it clusters similar transactions based on transaction amount, time, location, and user behavior, financial institutions can spot anomalous clusters that contain suspicious or fraudulent cases such as accessing from an unauthorized location, identity theft, or fraudulent transaction. clustering helps to group these transactions into clusters of typical spending, which helps in spotting unusual patterns that might require further investigation by the system of fraud detection or manual review.

(b) An application that takes clustering as a preprocessing tool for data preparation for other data mining tasks.

Market Segmentation for Product Recommendation:

clustering is used as a preprocessing tool to segment customers into distinct groups based on their preferences, demographics, and purchase history, when similar customers get clustered together, businesses can create targeted marketing campaigns and personalized product recommendations for each segment.