Data Science - Assignment 3 - Theoretical Questions

Mahla Entezari

1. Explain how Gaussian Mixture Models (GMM), K-means++, and Spectral Clustering work for clustering data. Highlight scenarios where each method is most suitable.

GMM (Gaussian Mixture Models): GMM assumes that data points come from a mix of several bell-shaped curves (Gaussians). Each point is assigned probabilities of belonging to each cluster. It uses a method called Expectation-Maximization to update these probabilities and cluster shapes.

Best for: When clusters overlap or have different shapes.

K-means++: This is an improved version of K-means. It chooses better starting points (centroids) before clustering, which helps avoid poor results. Then, each point is assigned to the nearest cluster center.

Best for: Round, well-separated clusters.

Spectral Clustering: This method builds a similarity graph from the data and uses math tools (like eigenvectors) to reduce the data to fewer dimensions. Then it applies K-means to those reduced dimensions.

Best for: Finding non-convex, oddly shaped clusters.

2. What strategies exist for clustering datasets containing both numerical and categorical variables?

- Convert categorical data using One-Hot Encoding or Label Encoding.
- Use distance measures like Gower distance that handle mixed types.
- Use algorithms designed for mixed data, like K-Prototypes.

3. Compare and contrast soft clustering (, GMM) with hard clustering (, K-means). When is soft clustering more appropriate?

Hard Clustering (, K-means): Each data point belongs to exactly one cluster.

Soft Clustering (, GMM): Each point is assigned a probability of belonging to each cluster.

Soft clustering is more appropriate when:

- Clusters overlap.
- There's uncertainty in the data.
- You want more flexible cluster shapes.

4. Can clustering be used for anomaly detection? Explain how DBSCAN or GMM can detect outliers.

DBSCAN: It finds dense regions as clusters. Points that don't fit into any dense area are labeled as outliers (or noise).

GMM: Each point gets a probability for each cluster. If none of these probabilities are high, the point is likely an outlier.

5. What challenges arise when clustering imbalanced datasets, and how can we ensure meaningful cluster formation?

Challenges:

- Large clusters can hide smaller ones.
- Algorithms may ignore or incorrectly merge small clusters.

Solutions:

- Use DBSCAN, which doesn't rely on cluster size.
- Use resampling techniques to balance the dataset.
- Use metrics like silhouette score to evaluate clustering quality.

6. You are given a dataset of customer transactions with the following features:

- **Annual income**
- **Age**
- **Total spending in the last year**

You apply hierarchical clustering using the Ward linkage method.

(a) How does hierarchical clustering build the cluster hierarchy? It starts with each point in its own cluster. Then, at each step, it joins the two clusters that increase the total variance the least. This process continues until all points are in one cluster. The result is a tree-like diagram called a dendrogram.

(b) How do you determine the optimal number of clusters from the dendrogram? Look for the largest vertical gap between horizontal lines in the dendrogram. Cutting the dendrogram at this point gives the most distinct grouping.

(c) Suppose you cut the dendrogram at 3 clusters. How can you interpret these clusters?

- Cluster 1: Young customers with low income and low spending — maybe new or budget shoppers.
- Cluster 2: Middle-aged customers with high income and high spending — loyal, valuable customers.
- Cluster 3: Older customers with medium income and medium spending — steady but less active shoppers.