Assignment 3 - Data Science
Theoretical Questions

Mohammad Hossein Basouli

May 15, 2025

Question 1

- **Gaussian Mixture Models**:

  1. It first initializes some gaussian distributions over the feature space.
  2. Then it estimates the probability of the data points based on the parameters.
  3. After that, it tries to maximize the likelihood of those parameters by changing them.
  4. At the end, it will apply the clustering on the dataset using a mixture of these gaussian distributions, by essentially assigning a probability of belonging to each of the distributions.

  It performs the best where*: the clusters are elliptic - we want to have a soft clustering - dimension is not too high - dataset is not too large*

- **K-means ++**:

  1. It first picks the first centroid.
  2. Then it will find the remaining centroid by assigning a probability, propotional to the distance from the first centroid.
  3. Then it will apply the **Ordinary K-means** to the dataset.

  It performs the best where*: the clusters are spherical and convex shaped - the dataset is large or the dimension is high*

- **Spectral Clustering**:

    1. It first obtains a **similarity matrix** from the dataset. (which IDK what it's exactly.)
    2. Then it finds the **graph Laplacian**. (which again, IDK what it's exactly.)
    3. After that, it tries to find the eigen vectors that correspond to the smallest eigen values of the **Laplacian matrix**, which captures the essential structure of the data.
    4. Each row of the matrix formed by the eigenvectors can be treated as a new data point in a lower-dimensional space.
    5. Then it will apply the **Ordinary K-means** to the dataset.

    It performs the best where*: the clusters are non-convex shaped - dataset is small*

Question 2

Following methods, used along with appropriate distance measures, could be used for mixed data type clustering:

1. **Hierarchical Clustering**
2. **DBSCAN**
3. **K-Prototypes**

Question 3

Soft Clustering is trying to somehow model the world under a uncertainty basis, whereas **Hard Clustering** tries to draw a rigid decision boundary and specify a single cluster that the data point could belong to. **Soft Clustering** might be a more appropriate approach when the clusters (or groups) are overlapping and we would rather want to have a uncertainty measure or a degree of confidence on how much it's possible for a data point to belong to a cluster.

Question 4

DBSCAN: In this algorithm, we have three types of data points; **Core**, **Reachable**, **Unreachable**, the latter case typically indicate outliers. **GMM**: When the algorithm, finally maximizes the expectation, we will have several different gaussian distributions over the features space. If we find out that a data point has a low probability in all of these distributions, then it's a outlier.

Question 5

We could utilize clustering methods that are robust to imbalance data; such as **DBSCAN**, **Hierarchical Clustering**. Also we might want to use metrics that take imbalanceness of the dataset into account as well, e.g. **Recall**, **F_1 Score**, **ROC Curve**.

Question 6

(a) Explaining **Agglomerative Hierarchical Clustering**:

- It first takes each datapoint as a single separate cluster.
- Then it begins by merging each pair of the clusters and calculate the centroid for that cluster.
- After that, it calculates the sum of squared distances of the data points in the merged cluster, from the obtained centroid, it then merges the pair of clusters that have the smallest sum of squared distances from their centroid.

(b) For obtaining the optimal number of clustering by just using the **dendrograms**, we could look at the obtained dendrogram and then cut it at the level above where the distance, between the clusters that join, is low, or at least, don't change with respect to the level above it.

(c) Cutting the dendrogram at three clusters, essentially mean that we have partitioned the customers into three different levels (or groups) in terms of their behavior. And if you ask what are the properties of each of these groups ? We should say that, it really depends on our dendrogram.