1. **(20 marks)** Explain the *k*-means clustering algorithm. Provide pseudo code of the

algorithm. It should be the version of the *k*-means clustering algorithm discussed in the

lectures. Implement the *k*-means clustering algorithm following your explanation and the

pseudo code. In the implementation, select initial cluster representatives randomly.

---------------------------------------------------------------------------------------------------------------------------------------------------

k-means clustering is a representative-based algorithm, which are a group of algorithms to find clusters in the given dataset by

* choosing some representatives for each cluster and
* some distance function to find closeness of the datapoints to these representatives.

The number of clusters is a hyperparameter and is generally denoted as ‘k’. Each cluster is represented by one representative, which may be initially chosen randomly or using some specific algorithm. The goal of the clustering is to determine the representatives which minimise the sum of distances of each datapoint with their closest representative. Thus, the following objective function is defined-

Following pseudo-code may be used to find the clusters in a given dataset, using k-means clustering algorithm:

Inputs required - number of clusters: ‘k’, Dataset: {}

1. Initialisation phase:
   1. Choose ‘k’ cluster representatives randomly from the dataset. These representatives may be denoted as CR: {}
2. Repeat until convergence (i.e., either run for fixed number of iterations OR till no object is assigned to a new cluster):
   1. Assignment phase:
      1. Assign all objects in the dataset to the closest representative () by comparing its Euclidean distance from all the representatives. This results in ‘k’ clusters.

It implies that for each , find a which minimises .

* 1. Optimisation phase:
     1. Compute the new representatives as the mean of each of the current clusters. Here, denotes the set of objects belonging to the cluster represented by .

denotes the number of objects in the set .

This step will update the set of cluster representatives, CR.

If the scale of input features are different, they are standardized by gauss normalisation before implementing this algorithm.

Following are the limitations of k-means clustering algorithm:

* The final clustering depends on initial cluster representatives. The algorithm might get stuck in local optima.
* A cluster representative is chosen by calculating the mean of all the objects in a cluster, which might result in the new representative that do not belong to the set of input objects. Such a cluster representative might be difficult to interpret, since it is not an actual instance in the cluster.
* The number of clusters is a hyperparameter, it might not represent the true clusters in the dataset.
* Euclidean distance is appropriate only for numerical features, not for categorical features.

2. (20 marks) Explain the k-means++ clustering algorithm. Provide pseudo code of the algorithm. It should be the version of the k-means++ clustering algorithm discussed in the lectures. Implement the k-means++ clustering algorithm following your explanation and the pseudo code.

k-means++ clustering algorithm is considered an improvement over k-means clustering algorithm in terms of choosing the initial cluster representatives. Instead of choosing all the representatives randomly, they are selected based on a probability value proportional to the squared distance from the closest cluster representative which is already chosen.

Following pseudo-code may be used to find the clusters in the given dataset, using k-means clustering algorithm:

Inputs required - number of clusters: ‘k’, Dataset D: {}

1. Initialisation phase:
   1. Choose one representative randomly from the dataset D.
   2. For every i = 2…, k:
      1. Select representative from the dataset with probability of selecting any object as a representative calculated as-

where, is the distance from to the closest cluster representative which is already chosen.

The set of these representatives may be denoted as CR: {}

1. Repeat until convergence (i.e., either run for fixed number of iterations OR till no object is assigned to a new cluster):
   1. Assignment phase:
      1. Assign all objects in the dataset to the closest representative () by comparing its Euclidean distance from all the representatives. This results in ‘k’ clusters.

It implies that for each , find a which minimises .

* 1. Optimisation phase:
     1. Compute the new representatives as the mean of each of the current clusters.

where, denotes the set of objects belonging to the cluster represented by , and

denotes the number of objects in the set .

This step will update the set of cluster representatives, CR.

Because of carefully chosen initial cluster representatives, k-means++ clustering performs better than randomly chosen initial representatives of k-means clustering. Randomly chosen representative may result in an unlucky selection leading to merging of different true clusters or splitting of a true cluster.