Name:	Roll No:	

Data Mining and Machine Learning

Quiz 3, II Semester, 2024–2025

25 March, 2025

Some questions below may have more than one correct answer. You get full credit if you select all correct options. You get partial credit if you select a non-empty, strict subset of correct options. You get zero credit if you select any incorrect option.

1. Which of the following are accurate statements about K-means clustering?

   (a) K-means clustering can only detect ellipsoid shaped clusters.

   (b) K-means clustering is not feasible for large datasets because of its computational complexity.

   (c) The silhouette score can help choose an optimum value of K.

   (d) The choice of initial centroids can influence the outcome of K means clustering.

2. Which of the following are accurate statements about hierarchical clustering?

   (a) Hierarchical clustering can only detect ellipsoid shaped clusters.

   (b) The definition of inter-cluster distance can influence the shape of the clusters.

   (c) Hierarchical clustering is not feasible for large datasets because of its computational complexity.

   (d) A dendrogram helps us evaluate the quality of the clustering.

3. Which of the following would qualify as semi-supervised learning?

   (a) Cluster the dataset, label the centroids, and extrapolate labels based on the centroids.

   (b) Cluster the dataset, redefine the points based on their distances from the centroids and perform classification on this new representation of the data.

   (c) Build a Naive Bayes model using a small subset of the data and use expectation-maximization to extend the labels to the entire dataset.

   (d) Use clustering on the feature values to merge data points.

4. Suppose we have a dataset in which the clusters are ellipsoidal. Which of the following methods would be effective to detect outliers?

   (a) Use K-means clustering and detect outliers based on their distance from the centroid.

   (b) Use hierarchical clustering and use complete link (maximum pairwise distance of points across clusters) to identify outliers.

   (c) Use density-based techniques and detect outliers based on local outlier factor.

   (d) Use expectation-maximization to model the dataset as a mixture of Gaussians and apply the standard definition of outliers for Gaussian distributions.