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. ✔

   Explanation:

   (b) K-means clustering is linear in the input size and hence works as well as any clustering algorithm could for large datasets.

   Note: (a) is poorly worded (the word "only" is subject to misinterpretation) so it will not be counted when scoring this question.

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.

   Explanation:

   (a) Since we can cluster at any level of the dendrogram, clusters can have arbitrarily complex shapes.

   (d) The dendrogram records the sequence in which clusters were combined. We can use the dendrogram to decide at which level of granularity to define clusters. It does not provide any information about the quality of the clusters.

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.

   Explanation:

   (b) Only the representation of each data item is changed. We still use the full set of labels, so it is not semi-supervised. In class, this was referred to as *preprocessing*.

   (d) This is just unsupervised learning. An example of this is the example where we clustered RGB values in the image with the ladybug on a flower, for object detection. No classification is involved, based on partial or total labelling.

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. ✔

   Explanation:

   (a) We cannot use this because outliers can distort K-means clustering, so we cannot cluster with outliers and then identify them based on the resulting clusters.

   (b) Complete link is one of the options to define inter-cluster distance to decide which pair of clusters to merge next. It has no connection to outlier detection.

   For (d), note that Gaussian clusters are elliptical in shape — recall the function `make_blobs()` in Scikit-Learn.
