Name:	Roll No:	Marks:

Data Mining and Machine Learning

Quiz 3, II Semester, 2025–2026

24 March, 2026

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 works well even for clusters that are not ellipsoidal.

   (b) K-means clustering can be used for large datasets because it is relatively efficient to compute. ✔

   (c) The silhouette score is useful to identify outliers.

   (d) To choose a set of well-separated initial centroids, we use a probabilistic strategy. ✔

   Explanation:

   (a) K-means clustering typically identifies ellipsoidal clusters and fails on clusters of different shapes.

   (c) The silhouette score is used to identify the quality of the clusters.

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

   (a) Hierarchical clustering works well even for clusters that are not ellipsoidal. ✔

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

   (c) Hierarchial clustering can be used for large datasets because it is relatively efficient to compute.

   (d) A dendrogram helps us choose the granularity of the clustering. ✔

   Explanation:

   (c) Computing intercluster distance is quadratic and hence not practical for large datasets.

3. Which of the following statements are true about dimensionality reduction?

   (a) One reason to reduce the dimensionality of the data is that points in high dimensions are very scattered. ✔

   (b) The singular vectors in SVD are listed in decreasing order of magnitude.

   (c) Locally linear embeddings (LLE) are not affected by rotation, reflection or scaling of the data. ✔

   (d) When PCA projects data to a lower number of dimensions, it can distort the local geometry around points. ✔

   Explanation:

   (b) All singular vectors are unit vectors, so they have the same magnitude. The associated singular *values* are in decreasing order of magnitude.

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.
