Sample Exam Questions 3050/4050, Dimensionality reduction and clustering!

Pooya Zakeri, November 2025

1- Which of the following methods is best suited for visualizing and preserving the local neighborhood relationships in data when reducing dimensions?

- A) Principal component analysis (PCA)
- B) Linear Discriminant Analysis (LDA)
- C) t-Distributed Stochastic Neighbor Embedding (t-SNE)
- D) Autoencoders
- E) K-means

2- In hierarchical clustering, suppose you choose Ward linkage to determine the distance between clusters. What characteristic does Ward linkage prioritize when merging clusters?

- A) Minimizing the maximum distance between samples in merged clusters
- B) Minimizing the increase in within-cluster variance
- C) Maximizing the number of clusters at each stage
- D) Maximizing the mean distance between cluster

3- In a k-means clustering application, the final clusters have high within-cluster variance. Which of the following techniques could help improve clustering by reducing within-cluster variance?

- A) Change the distance metric to Manhattan distance
- B) Decrease the number of clusters k
- C) Increase the number of clusters k
- D) Use average linkage

Solutions:

1) C) t-Distributed Stochastic Neighbor Embedding (t-SNE); t-SNE is specifically designed for visualizing high-dimensional data while preserving local neighborhood relationships.
- **A) PCA** – Linear, preserves global variance, not local neighborhoods.
- **B) LDA** – Supervised and focuses on class separation, not local structure.
- **D) Autoencoders** – Can preserve structure but not as explicitly focused on local neighborhoods as t-SNE.
- **E) K-means** – Clustering algorithm, not a dimensionality reduction or visualization method.

2) **B) Minimizing the increase in within-cluster variance**
Ward linkage merges clusters in a way that minimizes the increase in total within-cluster variance. It prefers cluster merges that keep clusters compact and homogeneous.
- **A** describes complete linkage.
- **C** is not related to linkage criteria.
- **D** does not describe any standard linkage method.

3) C) Increase the number of clusters k
When clusters are too spread out, their variance is high. You can fix this by creating more clusters, which results in smaller, tighter groupings.
- **A) Change to Manhattan distance** – This changes the geometry but does not reliably reduce variance.
- **B) Decrease k** – This would *increase* variance because clusters become larger.
- **D) Average linkage** – This is a hierarchical clustering method, not relevant to improving the result of k-means directly.