

Unsupervised Learning — Practice Questions

CM52054: Foundational Machine Learning
Practice set with fully worked answers

Part A - Concept & Short-Answer

- 1) **Supervised vs. unsupervised learning (notation).**
Define supervised learning and unsupervised learning using dataset notation.
- 2) **What is clustering?**
Give a concise definition of clustering in unsupervised learning.
- 3) **Motivations and applications.**
State three motivations or applications of clustering discussed in the lecture (any three).
- 4) **Good clustering criteria.**
State the two criteria for “good clustering” used in the lecture.
- 5) **Trade-off in clustering objectives.**
Briefly explain why “low intra-cluster distance” and “high inter-cluster distance” can be in tension.

Part B - Algorithms & Comparisons

- 6) **k-means: algorithm steps.**
State the two alternating steps of k-means and write the centroid update equation.
- 7) **k-means: stopping criterion.**
What stopping criterion is used for k-means in the lecture?
- 8) **k-means: non-determinism.**
Why is k-means considered non-deterministic?
- 9) **k-means: advantages and disadvantages.**
Give two advantages and two disadvantages of k-means.
- 10) **k-means: per-iteration complexity.**
Given N points, M features, and K clusters, state the per-iteration time complexity for (a) assignment and (b) centroid updates.
- 11) **k-means: one-iteration worked example (1D).**
Let $X = \{0, 2, 8, 10\}$ and initialise centroids $c_1 = 0$, $c_2 = 10$ using Euclidean distance. Perform one assignment + update iteration and report the new centroids.
- 12) **k-means: high-dimensional features.**
Explain briefly why k-means can be problematic in high-dimensional feature spaces.
- 13) **Hierarchical clustering: bottom-up vs top-down.**
Distinguish agglomerative (bottom-up) and divisive (top-down) hierarchical clustering.

14) **Single-linkage naming.**

Single-linkage clustering is also known as what?

15) **Hierarchical clustering: speed vs structure.**

According to the lecture, which approach is typically faster, and which tends to reflect global structure better?

Part C - More Challenging Questions

16) **k-means local optimum due to initialisation.**

Provide a concrete scenario showing that k-means can converge to a poor solution due to an unfortunate initialisation. Briefly explain the mechanism.

17) **Hierarchical clustering: linkage choice and cluster “chaining”.**

Single-linkage (nearest neighbour) can exhibit a “chaining” effect. Explain what this means and why it can be undesirable.