Unit III — Clustering (7 questions)

1. **Basics & Measures (Conceptual)**
   a) Define clustering and distinguish it from classification.
   b) Explain **similarity vs. dissimilarity** with two examples each. *(Easy)*
2. **Distance/Similarity Functions (Applied)**
   For the two vectors $x = (1, 2, 3)$, $y = (2, 4, 6)$, compute **Euclidean**, **Manhattan**, **Cosine similarity**, and briefly comment on how scale affects each measure. *(Medium)*
3. **Clustering Criteria (Theory + Short Derivation)**
   State the **minimum within-cluster distance (or WCSS)** criterion and show how it leads to the objective minimized by **k-means**. Why does minimizing WCSS not guarantee a global optimum? *(Medium)*
4. **K-means vs. K-medoids (Compare & Contrast)**
   Explain the **k-means** and **k-medoids** algorithms, comparing: initialization, objective, robustness to outliers, and time complexity. Give one dataset scenario where **k-medoids** clearly outperforms **k-means**. *(Medium)*
5. **Hierarchical Clustering (Linkage & MST)**
   a) Describe **single-link** and **complete-link** hierarchical agglomerative clustering; illustrate how chaining can occur.
   b) Explain how **Minimum Spanning Tree (MST)** can be used for clustering and when MST-based clustering is preferred. *(Medium)*
6. **Density-Based Clustering (DBSCAN)**
   Describe **DBSCAN**, including the roles of **ε (epsilon)** and **minPts**, the concepts of **core/edge/noise** points, and how DBSCAN handles clusters of arbitrary shape and noise. Provide one failure case for DBSCAN. *(Medium)*
7. **Data Realities (Visualization & Edge Cases)**
   a) List three visualization methods (e.g., t-SNE/UMAP, dendrograms, pair plots) to inspect cluster structure.
   b) Explain **unique clustering** vs. multiple valid partitions.
   c) Give two reasons why a dataset might show **no true clusters** (e.g., uniform or high-overlap distributions). *(Easy–Medium)*

Unit IV — Feature Selection (7 questions)

1. **Problem & Uses (Conceptual)**
   Define the **feature selection problem** and explain two practical benefits (e.g., improved generalization, reduced cost/latency). Contrast **feature selection** with **feature extraction**. *(Easy)*
2. **Sequential Methods (SFS/SBS/SFFS/SFBS)**
   Explain **Sequential Forward Selection (SFS)** and **Sequential Backward Selection (SBS)**. What is the **nesting effect** and how do **SFFS/SFBS** mitigate it? Give one scenario where SFS is preferable to SBS. *(Medium)*
3. **Branch and Bound (Exact Search)**
   Describe the **Branch-and-Bound** strategy for feature selection under a monotonic criterion. What does "monotonic" mean in this context? Discuss pros/cons vs. greedy methods. *(Medium)*
4. **(l, r) Algorithm (Heuristic Search)**
   Outline the **(l, r)** algorithm: how do the forward (add l features) and backward (remove r

features) steps alternate? When would you choose $(l, r) = (2, 1)$ over pure SFS? *(Medium)*

5. **Criterion Functions — Probabilistic Separability**
   Define a **probabilistic separability criterion** (e.g., **Bhattacharyya distance**, **Chernoff bound**) and explain how it guides feature subset ranking. Why can such criteria be preferable to raw classification accuracy during selection? *(Medium)*
6. **Criterion Functions — Interclass Distance / Scatter**
   Derive or state the **Fisher criterion** $J = \frac{(m_1 - m_2)^2}{s_1^2 + s_2^2}$ (1-D) and generalize its intuition to multi-class/multi-dimensional settings using **between-class** and **within-class** scatter matrices. Interpret high vs. low J. *(Medium)*
7. **Practical Pipeline & Overfitting Control (Applied)**
   Propose a **feature selection pipeline** for a high-dimensional dataset (e.g., gene expression): include train/validation splits, wrapper vs. filter choice, stability checks, and how to avoid **selection bias** (e.g., nested cross-validation). Give one metric you would track besides accuracy (e.g., calibration, F1, inference latency). *(Medium)*