Data Mining Viva – Questions with One-Line Answers

Experiment 1 – Distance Matrix

Q: What is Euclidean distance?

A: Euclidean distance: straight-line distance between two points.

Q: What is Manhattan distance?

A: Manhattan distance: sum of absolute differences.

Q: Why is the diagonal of a distance matrix always zero?

A: Diagonal is zero because distance(a,a)=0.

Q: What does a smaller distance indicate?

A: Smaller distance means higher similarity.

Q: Are distance matrices symmetric? Why?

A: It is symmetric because distance(a,b)=distance(b,a).

Experiment 2 – K-Means

Q: What is the purpose of K-means?

A: It partitions data into k clusters.

Q: What is a centroid?

A: A centroid is the mean of cluster points.

Q: Why does K-means sometimes give different results?

A: Different initial centroids give different results.

Q: Which distance measure is used in K-means?

A: Uses Euclidean distance commonly.

Q: When does K-means stop?

A: Stops when centroids stabilize.

Experiment 3 – Preprocessing

Q: What is data preprocessing?

A: Preprocessing cleans and prepares data.

Q: What is attribute selection?

A: Attribute selection picks relevant features.

Q: What does CFS evaluator do?

A: CFS selects features correlated to class.

Q: How are missing values handled?

A: Missing values replaced via mean/mode.

Q: Why is preprocessing important?

A: It improves accuracy and consistency.

Experiment 4 – Naive Bayes (Small)

Experiment 5 – Decision Tree

Q: What is entropy?

A: Entropy measures impurity.

Q: What is information gain?

A: Information gain is entropy reduction.

Q: Why was the root attribute chosen?

A: Root has highest gain.

Q: What algorithm does J48 implement?

A: J48 implements C4.5.

Q: How does a decision tree handle continuous attributes?

A: Continuous values are split by thresholds.

Experiment 6 – Apriori

Q: What is support?

A: Support is itemset frequency.

Q: What is confidence?

A: Confidence is rule strength.

Q: What does Apriori's downward closure property mean?

A: Downward closure prunes supersets.

Q: Why can Apriori be slow?

A: Slow due to candidate explosion.

Q: What is a frequent itemset?

A: Frequent itemset meets minimum support.

Experiment 7 – Hierarchical Clustering

Q: What is agglomerative clustering?

A: Agglomerative merges bottom-up.

Q: Difference between single-link and complete-link?

A: Single-link=min distance, complete-link=max distance.

Q: What is a dendrogram?

A: Dendrogram shows cluster merging.

Q: Does hierarchical clustering require k?

A: Does not require k initially.

Q: What is the difference between agglomerative and divisive?

A: Agglomerative merges, divisive splits.

Experiment 8 – FP-Growth

Q: What does FP-Growth avoid that Apriori does?

A: Avoids candidate generation.

Q: What is an FP-tree?

A: FP-tree is compressed prefix structure.

Q: Why is FP-Growth faster?

A: Faster due to fewer DB scans.

Q: What is a conditional FP-tree?

A: Conditional FP-tree mines patterns per item.

Q: Do FP-Growth and Apriori produce the same results?

A: Yes, both give same frequent itemsets.