Data Mining Viva Preparation Notes

(Unit-wise, Topic-wise Structured Questions and Answers)

Instructions

This document contains short viva questions and answers arranged unit-wise and topic-wise. Existing questions are retained and new questions are added after each unit.

Viva Questions and Answers – Data Mining (Unit-wise)

Unit 1: Introduction to Data Mining

1. **What is Data Mining?**
   Data mining is the process of discovering patterns and useful information from large datasets.

2. **What is Knowledge Discovery in Databases (KDD)?**
   KDD is the overall process of converting raw data into useful knowledge; data mining is one step in it.

3. **Give two motivations for data mining.**
   Huge data generation and the need for decision-making based on patterns.

4. **What are supervised techniques?**
   Techniques that use labelled data, e.g., classification.

5. **What are unsupervised techniques?**
   Techniques that use unlabelled data, e.g., clustering.

6. **What are the types of data mining tasks?**
   Classification, clustering, association rule mining, prediction, summarization.

7. **What are applications of data mining?**
   Fraud detection, market basket analysis, healthcare, recommendation systems.

8. **What is data quality?**
   The degree to which data is accurate, complete, consistent and reliable.

9. **What are data measurements?**
   Types of measurement scales: nominal, ordinal, interval and ratio.

10. **Difference between prediction and classification?**
    Prediction deals with continuous values; classification deals with categorical values.

11. **What is the goal of data mining?**
    To extract meaningful patterns and actionable knowledge from data.

12. **What is descriptive data mining?**
    It focuses on summarizing the general properties or patterns of data.

13. **What is predictive data mining?**
    It predicts unknown or future values using existing data patterns.

14. **What is meant by Big Data?**
    Extremely large datasets that are difficult to manage with traditional tools.

15. **State the 3 Vs of Big Data.**
    Volume, Velocity, and Variety.

16. **What is data heterogeneity?**
    When data comes from diverse sources and formats.

17. **What is noise in data?**
    Irrelevant or random errors that affect data accuracy.

18. **What is outlier?**
    A data point that differs significantly from other observations.

19. **What is metadata?**
    Data that describes other data.

20. **What is OLAP?**
    Online Analytical Processing; used for fast analysis of multidimensional data.

21. **Difference between OLAP and OLTP?**
    OLAP is for analysis; OLTP is for transaction processing.

22. **What is a data warehouse?**
    A centralized database used for reporting and analysis.

23. **What is data cleaning?**
    The process of detecting and correcting errors in data.

24. **Define pattern discovery.**
    Identifying trends, rules, or regularities in data.

Unit 2: Data Pre-Processing

26. **What is data preprocessing?**
    A set of techniques used to clean and prepare raw data for mining.

27. **What is data aggregation?**
    Combining multiple small objects into larger ones.

28. **What is sampling?**
    Selecting a subset of data to represent the entire dataset.

29. **What is dimensionality reduction?**
    Reducing the number of attributes while preserving important information.

30. **Name one dimensionality reduction method.**
    Principal Component Analysis (PCA).

31. **What is feature subset selection?**
    Selecting the best features that contribute most to prediction.

32. **What is feature creation?**
    Creating new features from existing ones.

33. **What is variable transformation?**
    Changing the scale or format of variables, e.g., normalization.

34. **What is normalization?**
    Scaling data to a small range like [0,1].

35. **Why is preprocessing important?**
    It improves the accuracy and performance of data mining models.

36. **What is data cleaning?**
    Removing noise, errors, and inconsistencies from data.

37. **What is data integration?**
    Combining data from multiple sources into a single dataset.

38. **What is data reduction?**
    Reducing data volume while maintaining the integrity of data.

39. **What is PCA used for?**
    To convert correlated variables into uncorrelated principal components.

40. **What is discretization?**
    Converting continuous data into categorical bins.

41. **What is binning?**
    Grouping a number of continuous values into smaller intervals.

42. **What is smoothing?**
    Removing noise by using techniques like bin means or medians.

43. **What is Z-score normalization?**
    Transforming data using mean and standard deviation.

44. **What is min–max normalization formula?**
    $v' = \frac{v - min}{max - min}$.

45. **What is data inconsistency?**
    When the same data shows different values across sources.

46. **What is missing data imputation?**
Filling missing values using mean, median, or model prediction.

47. **What is attribute construction?**
Creating new attributes from existing ones to improve model accuracy.

48. **What is curse of dimensionality?**
When high dimensional data becomes sparse and models perform poorly.

49. **Why is sampling used?**
To reduce computation time and cost while preserving data characteristics.

50. **What is random sampling?**
Selecting samples where every data item has equal chance of inclusion.

Unit 3: Cluster Analysis

51. **What is clustering?**
Grouping similar objects such that intra-cluster similarity is high.

52. **What is a similarity measure?**
A function that determines how close two data points are.

53. **What is Euclidean distance?**
A straight-line distance between two points used as a similarity measure.

54. **Name two types of clusters.**
Well-separated clusters and density-based clusters.

55. **What is K-means clustering?**
A partitional clustering algorithm that divides data into k clusters.

56. **What is a centroid in K-means?**
The mean point of all items in a cluster.

57. **How does K-means work?**
It iteratively assigns points to the nearest centroid and updates centroids.

58. **What is the main limitation of K-means?**
It requires the number of clusters k in advance.

59. **What is cluster validation?**
Methods used to evaluate the quality of clusters.

60. **How to determine the optimal number of clusters?**
Using the Elbow method or Silhouette score.

61. **What is hierarchical clustering?**
A method that builds clusters in a tree-like structure.

62. **Difference between agglomerative and divisive clustering?**
Agglomerative: bottom-up; Divisive: top-down.

63. **What is dendrogram?**
A tree diagram representing hierarchical clustering.

64. **What is Manhattan distance?**
Distance calculated as sum of absolute differences of coordinates.

65. **What is cosine similarity?**
It measures similarity based on angle between vectors.

66. **What is density-based clustering?**
Clustering based on dense regions, e.g., DBSCAN.

67. **What is DBSCAN?**
A clustering algorithm based on density and neighborhood parameters.

68. **What is minPts in DBSCAN?**
Minimum number of points to form a dense region.

69. **What is eps in DBSCAN?**
Maximum radius to search for neighboring points.

70. **What is a noise point in DBSCAN?**
A point that does not belong to any cluster.

71. **What is Silhouette score?**
A measure of cluster quality from -1 to 1.

72. **Define intra-cluster distance.**
Average distance between points in the same cluster.

73. **Define inter-cluster distance.**
Distance between cluster centroids.

Unit 4: Association Rule Mining

76. **What is association rule mining?**
Finding interesting relations among items in transaction data.

77. **What is a transaction dataset?**
Data where each record contains a list of purchased items.

78. **What is support?**
The proportion of transactions that contain an itemset.

79. **What is confidence?**
Probability that item Y occurs given X occurs.

80. **What is the Apriori algorithm?**
An algorithm to find frequent itemsets using the Apriori principle.

81. **What is the Apriori principle?**
If an itemset is frequent, all of its subsets must also be frequent.

82. **What are frequent itemsets?**
Itemsets whose support is above the minimum threshold.

83. **What is rule generation?**
Creating association rules from frequent itemsets.

84. **Give one application of association rule mining.**
Market basket analysis.

85. **What is lift?**
Lift measures how much more likely X and Y occur together than expected if independent.

86. **Give the formula for lift.**
Lift $= \frac{Support(XY)}{Support(X) \times Support(Y)}$

87. **What is leverage?**
Difference between actual and expected frequency of X and Y occurring together.

88. **What is conviction?**
Measures how strongly X implies Y.

89. **What is support count?**
The number of transactions containing an itemset.

90. **What is a candidate itemset?**
A possible frequent itemset generated during Apriori.

91. **What is pruning in Apriori?**
Removing candidates with infrequent subsets.

92. **What is FP-growth?**
A fast algorithm to find frequent itemsets using FP-tree.

93. **What is FP-tree?**
A compressed representation of transaction data.

94. **What is closed frequent itemset?**
A frequent itemset with no superset having the same support.

Unit 5: Classification

96. **What is classification?**
Assigning items to predefined categories.

97. **What is the Naive Bayes classifier?**
A probabilistic classifier based on Bayes' theorem assuming feature independence.

98. **What is the K-NN classifier?**
It assigns a label based on the majority class of k nearest points.

99. **What is a decision tree?**
A structure where nodes represent tests and leaves represent classes.

100. **What is overfitting?**
When a model performs well on training data but poorly on test data.

101. **What is a confusion matrix?**
A table showing actual vs. predicted classifications.

102. **What are evaluation metrics?**
Metrics like accuracy, precision, recall, F1-score.

103. **What is model evaluation?**
Assessing a classifier's performance.

104. **What is training data?**
Data used to teach the model.

105. **What is test data?**
Data used to evaluate model performance.

106. **What is entropy in decision trees?**
A measure of impurity in data.

107. **What is information gain?**
Reduction in entropy after a dataset split.

108. **Name one algorithm to build decision trees.**
ID3, C4.5, or CART.

109. **What is pruning in decision trees?**
Removing unnecessary branches to prevent overfitting.

110. **What is Bayes theorem?**
$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$

111. **What is prior probability?**
The probability of an event before evidence.

112. **What is posterior probability?**
Updated probability after considering evidence.

113. **What is K in K-NN?**
The number of nearest neighbors considered.

114. **What distance measures are used in K-NN?**
Euclidean, Manhattan, Minkowski.

115. **What is bias–variance tradeoff?**
Balance between underfitting and overfitting.

116. **Define accuracy.**
$\frac{TP+TN}{TP+TN+FP+FN}$

117. **What is precision?**
Proportion of predicted positives that are truly positive.

118. **What is recall?**
Proportion of actual positives correctly identified.