VISHNU INSTITUTE OF TECHNOLOGY (AUTONOMOUS)

Mid – II Examinations

Data Warehousing and Mining (CSE)

AIML and AIDS II-II

BIT BANK

Unit 3

1. What is model overfitting? [ ]
   A) When a model performs well on the training data but poorly on unseen data
   B) When a model performs well on unseen data but poorly on the training data
   C) When a model perfectly fits the training data and unseen data without any errors
   D) When a model fails to learn any patterns from the data

2. How does overfitting affect the generalization ability of a model? [ ]
   A) It improves the model's generalization ability
   B) It has no impact on the model's generalization ability
   C) It degrades the model's generalization ability
   D) It does not affect the model's generalization ability

3. Which technique helps to reduce overfitting in a model by introducing randomness during the training process? [ ]

   A) Regularization B) Feature scaling
   C) Principal component analysis (PCA) D) Bagging

4. What happens to the training error and the validation error as a model starts to overfit?
   A) Both errors increase [ ]
   B) Both errors decrease
   C) Training error decreases, while validation error increases
   D) Training error increases, while validation error decreases

5. Which evaluation metric is commonly used to assess model performance when dealing with imbalanced datasets ? [ ]
   A) Accuracy B) Precision C) Recall **D) F1 score**

6. How does increasing the complexity of a model affect the likelihood of overfitting? [ ]

   A) It decreases the likelihood of overfitting
   B) It has no impact on the likelihood of overfitting
   C) It increases the likelihood of overfitting
   D) It eliminates the possibility of overfitting

7. What is Naive Bayes classifier? [ ]

   A) A classification algorithm based on Bayes' theorem with the assumption of independence among features.

B) A regression algorithm based on the principle of maximum likelihood estimation.
C) A clustering algorithm that assigns instances to different groups based on their similarity.
D) An anomaly detection algorithm that identifies outliers in a dataset.

8. Which probability distribution is commonly used for continuous features in Naive Bayes classifier? [ ]
A) Gaussian distribution B) Bernoulli distribution
C) Poisson distribution D) Multinomial distribution

9. In Naive Bayes classifier, what is the effect of adding irrelevant features to the model?
A) It significantly improves the model's performance. [ ]
B) It has no impact on the model's performance.
C) It can degrade the model's performance.
D) It makes the model more robust to overfitting.

10. Which assumption of Naive Bayes classifier may not hold in real-world datasets? [ ]
A) Independence of features B) Dependence of features
C) Independence of class labels D) Dependence of class labels

11. What is a confusion matrix? [ ]
A) A matrix that displays the actual and predicted class labels of a classification model.
B) A matrix that represents the correlation between features in a dataset.
C) A matrix that stores the frequency of each unique value in a dataset.
D) A matrix that summarizes the performance of a regression model.

12. How many classes are typically represented in a confusion matrix? [ ]
A) 1 B) 2 C) 3 **D) Any number greater than or equal to 2**

13. What does the F1 score measure in a confusion matrix? [ ]
A) The overall correctness of the model's predictions
B) The balance between precision and recall
C) The proportion of true negatives in the dataset
D) The ability of the model to detect positive instances correctly

14. Which metric is calculated by dividing the true positives (TP) by the sum of true positives and false positives? [ ]
A) Accuracy
B) Precision
C) Recall
D) F1 score

15. What is the main purpose of using a confusion matrix in model evaluation? [ ]
A) To visualize the distribution of data in a dataset
B) To determine the most important features in a model
C) To evaluate the performance of a classification model
D) To identify the outliers in a dataset

Unit 4

1. What is the Apriori algorithm used for in data mining? [ ]

   A) Association rule mining
   B) Clustering
   C) Regression analysis
   D) Anomaly detection

2. What is the main objective of the Apriori algorithm? [ ]
   A) To find frequent itemsets in a transactional dataset
   B) To classify instances into predefined categories
   C) To predict continuous numerical values
   D) To detect outliers or anomalies in a dataset

3. What does the support measure represent in the Apriori algorithm? [ ]
   A) The frequency of an itemset in the dataset
   B) The confidence of an association rule
   C) The significance of a clustering result
   D) The error of a regression model

4. How does the Apriori algorithm generate candidate itemsets? [ ]
   A) By combining frequent itemsets from the previous level
   B) By selecting random itemsets from the dataset
   C) By using a clustering algorithm
   D) By applying feature selection techniques

5. What is the pruning step in the Apriori algorithm? [ ]
   A) Removing infrequent itemsets from consideration
   B) Removing outliers from the dataset
   C) Reducing the dimensionality of the dataset
   D) Combining similar itemsets into clusters

6. How does the Apriori algorithm handle the "apriori property"? [ ]
   A) By generating candidate itemsets that are subsets of frequent itemsets
   B) By randomly selecting itemsets from the dataset
   C) By using a decision tree to mine association rules
   D) By applying feature scaling to the dataset

7. How does the Apriori algorithm determine the association rules from frequent itemsets? [ ]

   A) By applying a minimum confidence threshold
   B) By selecting the itemsets with the highest support
   C) By performing feature selection on the itemsets
   D) By using a clustering algorithm

8. What is the key drawback of the Apriori algorithm in terms of computational efficiency? [ ]

   A) It requires a large amount of memory to store itemsets
   B) It is sensitive to the order of itemsets in the dataset
   C) It cannot handle continuous or numerical data
   D) It has a high time complexity for large datasets

9. What is the difference between frequent itemsets and association rules in the context of the Apriori algorithm? [ ]
   A) Frequent itemsets are itemsets that meet the minimum support threshold, while association rules are generated from frequent itemsets.
   B) Frequent itemsets are generated by applying the minimum confidence threshold, while association rules represent the patterns found in the dataset.
   C) Frequent itemsets are generated by pruning infrequent itemsets, while association rules are generated by combining frequent itemsets.
   D) Frequent itemsets represent the entire dataset, while association rules represent the subsets of the dataset.

10. What is the main objective of the FP-growth algorithm? [ ]
    A) To discover frequent itemsets in a transactional dataset
    B) To perform clustering on a dataset
    C) To classify instances into predefined categories
    D) To predict continuous numerical values

11. How does the FP-growth algorithm handle the generation of frequent itemsets? [ ]
    A) By recursively constructing conditional FP-trees
    B) By using a breadth-first search algorithm
    C) By applying feature selection techniques
    D) By performing dimensionality reduction

12. What is the role of the support count in the FP-growth algorithm? [ ]

    A) It represents the frequency of an itemset in the dataset
    B) It determines the number of branches in the FP-tree
    C) It is used to measure the confidence of association rules
    D) It represents the purity of clusters in the dataset

13. What is the advantage of the FP-growth algorithm over the Apriori algorithm? [ ]
    A) It does not require the generation of candidate itemsets
    B) It has a lower time complexity for large datasets
    C) It can handle datasets with missing values
    D) It can handle both categorical and continuous data

14. What is the minimum support threshold in association analysis? [ ]
    A) The minimum number of transactions an itemset must appear in to be considered frequent
    B) The minimum number of items in an itemset
    C) The maximum number of transactions in the dataset
    D) The maximum number of iterations allowed in the algorithm

15. What is the difference between frequent itemsets and infrequent itemsets in association analysis? [ ]

    A) Frequent itemsets have high support, while infrequent itemsets have low support.
    B) Frequent itemsets have high confidence, while infrequent itemsets have low confidence.
    C) Frequent itemsets contain more items than infrequent itemsets.
    D) Frequent itemsets are more significant than infrequent itemsets.

16. Which data format is commonly used for association analysis? [ ]
    A) Transactional format B) Text format
    C) Image format D) Graph format

17. What do you mean by support(A)? [ ]
    A) Total number of transactions containing A
    B) Total Number of transactions not containing A
    C) Number of transactions containing A / Total number of transactions
    D) Number of transactions not containing A / Total number of transactions

18. Which of the following is direct application of frequent itemset mining? [ ]

    A) Social Network Analysis
    B) Market Basket Analysis
    C) Outlier Detection
    D) Intrusion Detection

19. What is the relation between a candidate and frequent itemsets? [ ]
    (a)A candidate itemset is always a frequent itemset
    (b)A frequent itemset must be a candidate itemset
    (c)No relation between these two
    (d)Strong relation with transactions

20. Which algorithm requires fewer scans of data? [ ]
    (a)Apriori
    (b)FP Growth
    (c)Naive Bayes
    (d)Decision Trees

21. What will happen if support is reduced? [ ]

    (a) Number of frequent itemsets remains the same
    (b)Some itemsets will add to the current set of frequent itemsets
    (c) Some itemsets will become infrequent
    (d)Can not say

22. Frequency of occurrence of an itemset is called as _____ [ ]

    (a) Support
    (b) Confidence
    (c) Support Count
    (d)Rules

23. When do you consider an association rule interesting? [ ]
    (a) If it only satisfies min_support
    (b) If it only satisfies min_confidence
    (c) **If it satisfies both min_support and min_confidence**
    (d) There are other measures to check so

24. How is the support of an itemset calculated in association analysis? [ ]

A) By counting the number of transactions containing the itemset and dividing it by the total number of transactions
B) By dividing the support of the itemset by the support of the antecedent
C) By dividing the support of the itemset by the support of the consequent
D) By multiplying the support of the antecedent and consequent in an association rule

25. Which of the following statements is true regarding the relationship between support and confidence? [ ]

A) High support always guarantees high confidence.
B) High confidence always guarantees high support.
C) Support and confidence are independent measures in association analysis.
D) There is a positive relationship between support and confidence.

26. In e-commerce, association analysis can be applied to: [ ]
A) Personalize product recommendations
B) Detect credit card fraud
C) Monitor network security
D) Analyze weather patterns

27. Which of the following statements about closed frequent itemsets is true? [ ]
A) Closed frequent itemsets are itemsets that appear frequently in a dataset.
B) Closed frequent itemsets are itemsets that have no frequent supersets.
C) Closed frequent itemsets are itemsets that have no infrequent subsets.
D) Closed frequent itemsets are itemsets that have the highest support in the dataset.

28. How are minimal frequent itemsets different from closed frequent itemsets? [ ]
A) Minimal frequent itemsets have the highest support, while closed frequent itemsets have the lowest support.
B) Minimal frequent itemsets have no frequent subsets, while closed frequent itemsets have no frequent supersets.
C) Minimal frequent itemsets have the highest confidence, while closed frequent itemsets have the lowest confidence.
D) Minimal frequent itemsets and closed frequent itemsets are the same.

29. How does the FP-Tree technique handle the issue of memory usage compared to the Apriori algorithm? [ ]

A) It requires less memory due to its compressed representation of frequent itemsets.
B) It requires more memory due to the construction of a large tree structure.
C) It uses a different memory management technique that has no impact on memory usage.
D) It requires the same amount of memory as the Apriori algorithm.

30. What is the main drawback of the Apriori algorithm in terms of performance? [ ]

A) It requires multiple scans of the entire database.
B) It only works well with small datasets.
C) It cannot handle sparse datasets.
D) It is unable to discover frequent itemsets.

Unit 5

1. What is k-Means algorithm used for? [ ]
   a) Classification b) Regression **c) Clustering** d) Feature selection

2. What is the main goal of k-Means clustering? [ ]
   a) Minimizing within-cluster variance b) Maximizing between-cluster variance
   c) Minimizing misclassification rate d) Maximizing entropy

3. Which of the following is NOT a step in the k-Means algorithm?
   a) Initialization b) Assignment **c) Evaluation** d) Update

4. How does the k-Means algorithm initialize the cluster centroids? [ ]
   a) Randomly selecting k data points as centroids
   b) Calculating the mean of all data points as the initial centroids
   c) Determining centroids based on a pre-defined criterion
   d) Using the first k data points as centroids

5. What is the convergence criterion for the k-Means algorithm? [ ]
   a) Maximum number of iterations
   b) Minimum decrease in within-cluster variance
   c) Maximum increase in between-cluster variance
   d) Minimum number of misclassified instances

6. Which distance metric is commonly used in k-Means clustering? [ ]
   a) Euclidean distance b) Manhattan distance
   c) Hamming distance d) Cosine similarity

7. What is the time complexity of the k-Means algorithm? [ ]
a) O(n) b) O(n log n) **c) O(I*K*m*n)** d) O(kn^2)

8. Which of the following best describes Agglomerative Hierarchical Clustering? [ ]

   a) Top-down approach **b) Bottom-up approach**
   c) Greedy approach d) Divide-and-conquer approach

9. How does Agglomerative Hierarchical Clustering start the clustering process? [ ]
a) Each data point is assigned to its own cluster.
b) The entire dataset is considered as a single cluster.
c) Randomly selecting a subset of data points as initial clusters.
d) Pre-defining the number of clusters before clustering.

10. Which of the following is true about the dendrogram in Agglomerative Hierarchical Clustering? [ ]
a) It shows the pairwise distances between data points.
b) It displays the hierarchy of merged clusters.
c) It represents the within-cluster variance.
d) It indicates the number of clusters in the dataset.

11.Which of the following is the time complexity of Agglomerative Hierarchical Clustering? [ ]

a) O(n)
b) O(n log n)
c) O(n^2logn)
d) O(n^3)

12. What does DBSCAN stand for? [ ]
a) Density-Based Spatial Clustering of Applications with Noise
b) Distance-Based Spatial Clustering of Applications with Noise
c) Density-Based Sequential Clustering with Noise
d) Distance-Based Sequential Clustering with Noise

13. What is the main advantage of the DBSCAN algorithm? [ ]
a) It can handle clusters of arbitrary shape
b) It guarantees convergence to the global optima
c) It is computationally efficient for large datasets
d) It does not require the number of clusters to be predefined

14. What is the key concept used in DBSCAN for identifying clusters? [ ]
a) Density b) Distance c) Centroids d) Principal components

15. Which of the following points is classified as noise in DBSCAN? [ ]
a) Core point b) Border point **c) Outlier point** d) Cluster center point

16. How does DBSCAN determine the density of a region? [ ]
a) By counting the number of points within a fixed radius
b) By calculating the average distance between points in a region
c) By analyzing the distribution of distances between points
d) By considering the proximity to other dense regions

17. Which of the following is the output of DBSCAN? [ ]
a) Clusters and noise points b) Distance matrix
c) Centroids and labels d) Silhouette coefficients

18. Which of the following best describes clustering in data mining? [ ]
a) The process of labeling data points with predefined categories.
b) The process of identifying associations between data points.
c) The process of grouping similar data points together based on their attributes.
d) The process of predicting future values based on historical data.

19. Which of the following clustering algorithms is based on centroid-based clustering? [ ]
a) k-Means b) DBSCAN c) Agglomerative Clustering d) OPTICS

20. Which clustering algorithm constructs a hierarchy of clusters? [ ]
a) k-Means b) DBSCAN **c) Agglomerative Clustering** d) OPTICS

21. What is the objective function used in K-means clustering? [ ]
a) Sum of squared errors b) Maximum likelihood estimation
c) Entropy d) Jaccard similarity coefficient

22. What is Bisecting K-means clustering? [ ]
a) A hierarchical clustering algorithm b) A density-based clustering algorithm

c) A partition-based clustering algorithm d) A graph-based clustering algorithm

23. How does Bisecting K-means split a cluster to create two child clusters? [ ]
a) By selecting the cluster with the highest intra-cluster similarity
b) By selecting the cluster with the lowest intra-cluster similarity
c) By randomly selecting a data point and assigning it to a new cluster
d) By calculating the centroid of the cluster and splitting based on distance to the centroid

24. What is the main advantage of Bisecting K-means clustering compared to traditional K-means clustering? [ ]
a) It guarantees convergence to the global optimum
b) It is faster and more efficient for large datasets
c) It can handle arbitrary cluster shapes
d) It does not require specifying the number of clusters in advance

25. Which of the following is required by K-means clustering? [ ]

a) defined distance metric b) number of clusters
c) initial guess as to cluster centroids **d) All the mentioned**

26. Clustering is _____ [ ]

(a) Supervised learning **(b)Unsupervised learning**
(c)A & B Both (d)None of Above

27. A good clustering method will produce high quality clusters with [ ]
(a)High inter class similarity (b)Low intra class similarity
(c)High intra class similarity (d)No inter class similarity

28. The learning which is used to find the hidden pattern in unlabeled data is called? [ ]

(a)Unsupervised Learning (b)Supervised Learning
(c)Reinforcement Learning (d)None of above

29. Which of the following is not clustering method? [ ]
A) Density-Based B) Hierarchical Based
C) Partitioned-Based **D) Project Based**

30. Hierarchical Based Methods Consist of which category? [ ]

A) Divisive B) Agglomerative
C) both a and b D) None of these