By Mónica Atiaga
Batch - DSNB1222

MCQ …

1. Among the following identify the one in which dimensionality reduction reduces.
   a. Performance
   b. statistics
   c. Entropy
   d. Collinearity

   Answer:
   d. Collinearity. By reducing dimensionality and eliminating highly correlated variables, the issue of collinearity can be addressed, leading to improved accuracy and stability of models.

2. Which of the following machine learning algorithm is based upon the idea of bagging?
   a. Decision Tree
   b. Random Forest
   c. Classfication
   d. SVM

   Answer:
   b. Random Forest is a ensemble learning algorith that combines multiple decision trees to make predictions. It employs the technique of bagging, which stands for bootstrap aggregating. Bagging involves creating multiple subsets of the original dataset through random sampling with replacement.

3. Choose a disadvantage of decision trees among the following.
   a. Decision tree robust to outliers
   b. Factor analysis
   c. Decision Tree are prone to overfit
   d. all of the above

   Answer:
   c. Decision trees are prone to overfitting.
   While decision trees have several advantages, such as being able to handle both categorical and numerical data, and being interpretable, they are prone to overfitting. Overfitting occurs when a decision tree model becomes too complex and captures noise or irrelevant patterns in the training data, leading to poor generalization performance on unseen data.

4. What is the term known as on which the machine learning algorithms build a model based on sample data?
   a. Data Training
   b. Sample Data
   c. Training data
   d. None of the above

   Answer:
   a. Data Training.

5. Which of the following machine learning techniques helps in detecting the outliers in data?
   a. Clustering
   b. Classification
   c. Anamoly detection
   d. All of the above

   Answer:
   c. Anomaly detection, is specially designed to identify observations or instances that deviate significantly from the normal behavior or patterns in a dataset. Outliers can be considered as anomalies, as they represent data points that are significantly different from the majority of the data.

6. Identify the incorrect numerical functions in the various function representation of machine learning.
   a. Support Vector
   b. Regression
   c. Case based
   d. Classification

   Answer:
   c. Case based. The other options are all valid representations of different types of machine learning techniques.

7. Analysis of ML algorithm needs
   a. Statistical learning theory
   b. Computational learning theory
   c. None of the above
   d. Both a and b

   Answer:
   d. Both a and b. Both statistical learning theory and computational learning theory.

8. Identify the difficulties with the k-nearest neighbor algorithm.

a. Curse of dimensionality
b. Calculate the distance of test case for all training cases
c. Both a and b
d. None

Answer:
c. Both a and b.
Curse of dimensionality refers to the problem of increasing dimensionality leading to sparsity of data. Decreasing he accuracy and efficiency of the k-NN algorithm.

Calculating the distance of a test case for all training cases. It can lead to increased computational complexity and slower prediction time.

9. The total types of the layer in radial basis function neural networks is _____
a. 1
b. 2
c. 3
d. 4

Answer:
b. 2. 1) Input layer (receives the input features or variables). 2) Radial Basis Function Layer (Performs the nonlinear transformation of the input data using radial basis functions.).

10. Which of the following is not a supervised learning
a. PCA
b. Naïve bayes
c. Linear regression
d. KMeans

Answer:
d. KMeans, is an unsupervised learning algorithm.

11. What is unsupervised learning?
a. Number of groups may be known
b. Features of groups explicitly stated
c. Neither feature nor number of groups is known
d. None of the above

Answer:
c. Neither feature non number of groups is known. The objective of unsupervised learning is to discover hidden patterns or structures in the data without prior knowledge of the desired outcomes.

12. Which of the following is not a machine learning algorithm?
    a. SVM
    b. SVG
    c. Random Forest Algorithm
    d. None of the above

    Answer:
    b. SVG is not a recognized machine learning algorithm.
    SVM(Support Vector Machine) and Random Forest Algorithm are both widely used ML algorithms.

13. _____ is the scenario when the model fails to decipher the underlying trend in the input data
    a. Overfitting
    b. Underfitting
    c. Both a and b
    d. None of the above

    Answer:
    b. Underfitting occurs when a machine learning model is too simple or lacks the complexity to capture the underlying patterns or relationships in the data. The model may have high bias and is unable to fit the training data well or make accurate predictions.

14. Real-Time decisions, Game AI, Learning Tasks, Skill acquisition, and Robot Navigation are applications of …………
    a. Reinforcement learning
    b. Supervised learning
    c. Unsupervised Learning
    d. None of the above

    Answer:
    a. Reinforcement learning. It is a type of machine learning where an agent learns to make sequential decisions in an environment to maximize a reward signal.

15. What is called the average squared difference between classifier predicted output and actual output?
    a. Mean relative error
    b. Mean squared error
    c. Mean absolute error
    d. Root mean squared error
    Answer:
    b. Mean squared error. (MSE) is a common metric used to measure the average squared difference between predicted values and actual values.

16. Logistic regression is a ……….. regression technique that is used to model data having a ………… outcome.
    a. Linear, binary
    b. Linear, numeric
    c. Nonlinear, binary
    d. Nonlinear, numeric

    Answer:
    a Linear, binary. Logistic regression is a linear regression technique that predicts the probability of the occurrence of a certain event or class based on input features.

17. You are given reviews of few netflix series marked as positive, negative and neutral. Classifying reviews of a new netflix series is an example of
    A. supervised learning
    B. unsupervised learning
    C. semisupervised learning
    D. reinforcement learning

    Answer:
    A. supervised learning. There are the corresponding target labels(positive, negative or neutral sentiment).

18. Following is powerful distance metrics used by Geometric model
    A. euclidean distance
    B. manhattan distance
    C. both a and b
    D. square distance

    Answer:
    C. both a and b.
    Both are widely use in various geometric models, such as clustering algorithms(e.g., k-means) or nearest neighbor algorithms (e.g., k-nearest neighbors).

19. Which of the following techniques would perform better for reducing dimensions of a data set?
    A. removing columns which have too many missing values
    B. removing columns which have high variance in data
    C. removing columns with dissimilar data trends
    D. none of these

    Answer:

B. removing columns which have high variance in data. By removing columns with high variance, you can potentially reduce the number of features or dimensions while still retaining important information.

20. Supervised learning and unsupervised clustering both require which is correct according to the statement.
    A. output attribute.
    B. hidden attribute.
    C. input attribute.
    D. categorical attribute

    Answer:
    C. input attribute. The input attributes/features are essential for the analysis and modeling process.

21. What is the meaning of hard margin in SVM?
    (A) SVM allows very low error in classification
    (B) SVM allows high amount of error in classification
    (C) Underfitting
    (D) SVM is highly flexible

    Answer:
    (A) SVM allows very low error in classification. Hard margin SVM refers to the scenario where the SVM algorithm aims to find a decision boundary that maximally separates the classes while allowing very low or zero error in classification.

22. Increase in which of the following hyper parameter results into overfit in Random forest? (1). Number of Trees. (2). Depth of Tree, (3). Learning Rate
    (A) Only 1
    (B) Only 2
    (C) 2 and 3
    (D) 1,2 and 3

    Answer:
    (B). Only 2. Depth of Tree. Increasing the depth of the trees in a Random Forest model can lead to overfitting.

23. Below are the 8 actual values of target variable in the train file: [0,0,0, 0, 1, 1,1,1,1,1], What is the entropy of the target variable?
    (A) -(6/10 log(6/10) + 4/10 log(4/10))
    (B) 6/10 log(6/10) + 4/10 log(4/10)
    (C) 4/10 log(6/10) + 6/10 log(4/10)
    (D) 6/10 log(4/10) – 4/10 log(6/10)

Answer:
(A) -(6/10 log(6/10) + 4/10 log(4/10)).

The probability of class 0 occurring is 4/10, and the probability of class 1 occurring is 6/10.
Formula: -(p * log2(p) + q * log2(q)), where p and q are the probabilities of the two classes.

Entropy = -(4/10 * log2(4/10) + 6/10 * log2(6/10))

24. Lasso can be interpreted as least-squares linear regression where
    (A) weights are regularized with the l1 norm
    (B) weights are regularized with the l2 norm
    (C) the solution algorithm is simpler

    Answer:
    (A). weights are regularized with the l1 norm.
    Lasso regression, also known as L1 regularization, adds a penalty term to the least-squares linear regression objective function. This penalty term is the l1 norm (sum of absolute values) of the regression coefficients (weights).

25. Consider the problem of binary classification. Assume I trained a model on a linearly separable training set, and now I have a new labeled data point that the model properly categorized and is far away from the decision border. In which instances is the learnt decision boundary likely to change if I now add this additional point to my previous training set and re-train? When the training model is,
    (A) Perceptron and logistic regression
    (B) Logistic regression and Gaussian discriminant analysis
    (C) Support vector machine
    (D) Perceptron

    Answer:
    (D). Perceptron.
    Perceptron algorithm is a linear binary classification algorithm that iteratively updates the decision boundary based on misclassified instances. If a new labeled data point that is far away from the decision boundary is added to the training set and retraining is performed, the Percepton algorithm will likely update the decision boundary to account for this new point.

    Logistic regression and Gaussian discriminant analysis are also linear binary classification algorithms. However, they are not inherently designed to update the decision boundary based on individual misclassified instances.

SVM aims to find the maximum-margin decision boundary that separates the classes and is less influenced by individual data points.

26. Assume you've discovered multi-collinear features. Which of the following actions do you intend to take next?
(1). Both collinear variables should be removed.
(2). Instead of deleting both variables, we can simply delete one.
(3). Removing correlated variables may result in information loss.
We may utilize penalized regression models such as ridge or lasso regression to keep such variables.

(A) Only 1
(B) Only 2
(C) Either 1 or 3
(D) Either 2 or 3

Answer:
(D) Either 2 or 3.

Instead of deleting both variables (option 1), it is often preferable to remove just one of the collinear variables (option 2). Removing both variables can result in the loss of valuable information and potentially lead to a less informative model.

Alternatively, utilizing penalized regression models such as Ridge or Lasso regression (option 3) can help address the multicollinearity issue.

27. A least squares regression study of weight (y) and height (x) yielded the following least squares line: y = 120 + 5x. This means that if the height is increased by one inch, the weight should increase by what amount?
(A) increase by 1 pound
(B) increase by 5 pound
(C) increase by 125 pound
(D) None of the above

Answer:
(B) increase by 5 pound

28. The line described by the linear regression equation (OLS) attempts to ____?
(A) Pass through as many points as possible.
(B) Pass through as few points as possible
(C) Minimize the number of points it touches
(D) Minimize the squared distance from the points

Answer:

(D) Minimize the squared distance from the points.

The goal of linear regression using the Ordinary Least Squares (OLS) method is to find the line that minimizes the sum of the squared differences between the predicted values (on the line) and the actual values of the dependent variable (points).

29. For two real-valued attributes, the correlation coefficient is 0.85. What does this value indicate?
    (A) The attributes are not linearly related
    (B) As the value of one attribute increases the value of the second attribute also increases
    (C) As the value of one attribute decreases the value of the second attribute increases
    (D) The attributes show a curvilinear relationship

    Answer:

    (B) As the value of one attribute increases the value of the second attribute also increases

30. Which neural network architecture would be most suited to handle an image identification problem (recognizing a dog in a photo)?
    (A) Multi Layer Perceptron
    (B) Convolutional Neural Network
    (C) Recurrent Neural network
    (D) Perceptron

    Answer:

    (B) Convolutional Neural Network.

    Convolutional Neural Network (CNNs) are specifically designed for image-related tasks and have proven to be highly effective in image recognition and computer vision tasks.

    CNNs excel at capturing spatial relationships and patterns in images, which makes them well-suited for tasks like object recognition, including identifying dogs in photos.