ANSWERS

1. The correct option among the given choices is:

d) Collinearity

Dimensionality reduction techniques aim to reduce the number of features or variables in a dataset while preserving the relevant information. One common problem in statistical analysis is collinearity, which refers to a high degree of correlation between predictor variables. Collinearity can negatively affect statistical models by inflating standard errors and making it difficult to determine the individual effects of each predictor.

By applying dimensionality reduction techniques such as principal component analysis (PCA) or factor analysis, it is possible to combine correlated variables into a smaller set of uncorrelated variables known as principal components or factors. This reduces the collinearity among the variables, making it easier to interpret and analyze the data.

Therefore, dimensionality reduction can effectively reduce collinearity in a dataset.

2. The machine learning algorithm that is based upon the idea of bagging is:

b) Random Forest

Random Forest is an ensemble learning algorithm that combines multiple decision trees through a technique called bagging (bootstrap aggregating). Bagging involves training each decision tree on a different subset of the training data, randomly sampled with replacement. Each tree in the random forest is trained independently and makes predictions individually. The final prediction of the random forest is determined by aggregating the predictions of all the individual decision trees, such as through majority voting for classification problems or averaging for regression problems.

3. The correct option among the given choices is:

c) Decision Trees are prone to overfit

Decision trees have certain disadvantages, and one of them is their tendency to overfit the training data. Overfitting occurs when a decision tree captures the noise or irrelevant patterns in the training data, resulting in a complex and highly branched tree that fits the training data extremely well but fails to generalize well to unseen data.

Decision trees are capable of learning intricate and complex decision boundaries, which makes them prone to overfitting, especially when the tree depth increases or when there are a large number of features. This overfitting can lead to poor performance on new, unseen data as the decision tree becomes too specific to the training set.
To mitigate the overfitting issue, various techniques can be applied, such as setting a maximum tree depth, pruning the tree, or using ensemble methods like random forests.

These techniques help control the complexity of the decision tree and improve its generalization ability.

Therefore, the disadvantage of decision trees mentioned among the options is that they are prone to overfitting (option c).

4. The correct term for building a model based on sample data in machine learning is:

c) Training data

Training data refers to the labeled dataset used to train a machine learning model. It consists of a collection of input samples (features) and their corresponding output labels or target variables. The training data is used by machine learning algorithms to learn the patterns and relationships within the data and build a model that can make predictions or classify new, unseen data.

5. The correct option among the given choices is:

c) Anomaly detection

Anomaly detection is a machine learning technique specifically designed to detect outliers or anomalies in data. Outliers are data points that deviate significantly from the majority of the data points, either in terms of their values or patterns. Anomalies can represent important events or errors in the data that require special attention.
Clustering and classification, while useful in various machine learning tasks, are not specifically designed for outlier detection. Clustering algorithms group similar data points together, which can indirectly help in identifying outliers as points that do not belong to any cluster or form their own separate clusters. Classification algorithms focus on assigning labels to data points based on their features, but they are not primarily intended for outlier detection.

Therefore, the machine learning technique that specifically helps in detecting outliers in data is "Anomaly detection" (option c).

6. The correct option among the given choices is:

c) Case based

The term "Case based" is not typically used to represent a specific numerical function in machine learning. It is more commonly associated with a type of reasoning or problem-solving approach called case-based reasoning (CBR). CBR involves solving new problems by retrieving and adapting solutions from similar past cases.

7. The correct option among the given choices is:

d) Both a and b

The analysis of machine learning algorithms requires the application of both statistical learning theory and computational learning theory. Here's a brief explanation of each:

a) Statistical learning theory: Statistical learning theory provides a framework for understanding the statistical properties and generalization capabilities of machine learning algorithms. It focuses on studying the sample complexity, generalization bounds, and trade-offs between bias and variance. Statistical learning theory helps us understand how well a machine learning algorithm will perform on unseen data based on its performance on the training data.

b) Computational learning theory: Computational learning theory, also known as theoretical computer science, deals with the computational aspects of machine learning algorithms. It focuses on studying the computational complexity, efficiency, and scalability of learning algorithms. Computational learning theory helps us analyze the time and space complexity of algorithms, understand their convergence properties, and develop efficient learning algorithms.

8. The correct option among the given choices is:

c) both a and b

The k-nearest neighbor (KNN) algorithm has some difficulties, which include:

a) Curse of dimensionality: The curse of dimensionality refers to the challenge faced by algorithms when dealing with high-dimensional data. As the number of features or dimensions increases, the data becomes more sparse, and the distance between points becomes less meaningful. This can result in decreased accuracy and increased computational complexity for KNN, as the algorithm relies on distance calculations between data points.

b) Calculating the distance of a test case for all training cases: In the KNN algorithm, the distance between a test case and all training cases needs to be calculated to determine the k nearest neighbors. This process can become computationally expensive, especially when dealing with large datasets or high-dimensional data. As the dataset grows, the number of distance calculations required increases, which can impact the efficiency and scalability of the algorithm.

9. The correct answer is:

c) 3

Radial Basis Function (RBF) neural networks typically consist of three types of layers:

Input Layer: The input layer receives the input data, which is usually a vector of features or variables.

Hidden Layer: The hidden layer of an RBF neural network is responsible for computing the activation of each hidden neuron or unit. Each neuron in the hidden layer calculates the similarity or distance between the input data and a prototype vector associated with that neuron using a radial basis function. The activation of the hidden layer units determines the weights of the connections to the output layer.

Output Layer: The output layer of an RBF neural network computes the final output based on the weighted inputs from the hidden layer. The activation function used in the output layer depends on the type of problem being solved, such as regression or classification.

Therefore, the total types of layers in a radial basis function neural network is 3 (option c).

10. The correct answer is:

d) KMeans

KMeans is not a supervised learning algorithm. It is an unsupervised learning algorithm used for clustering. Unsupervised learning algorithms do not require labeled training data but instead aim to discover patterns, structures, or relationships within the data. KMeans clusters the data points into K distinct groups based on their similarity.

11. The correct option is:

c) Neither feature nor number of groups is known

Unsupervised learning is a type of machine learning where the algorithm learns patterns, structures, or relationships in data without the use of explicitly labeled or pre-classified examples. In unsupervised learning, the data provided to the algorithm is unlabeled, meaning that there are no predetermined target variables or class labels associated with the data.

12. The correct answer is:

b) SVG

SVG is not a commonly known machine learning algorithm. It does not correspond to any widely recognized abbreviation or acronym in the field of machine learning.

13. The correct answer is:

b) Underfitting

Underfitting occurs when a machine learning model is unable to capture the underlying trend or patterns in the input data. It happens when the model is too simple or lacks

complexity to adequately represent the data. In an underfitting scenario, the model may have high bias and low variance.

14. The correct answer is:

a) Reinforcement learning

Reinforcement learning is a machine learning approach that deals with an agent learning to make sequential decisions in an environment to maximize a reward signal. It is specifically designed for scenarios where an agent interacts with an environment, learns from feedback in the form of rewards or punishments, and makes decisions based on a trial-and-error learning process.

15. The correct answer is:

b) Mean squared error

The mean squared error (MSE) is a commonly used metric to measure the average squared difference between the predicted output of a classifier (or any other model) and the actual output. It is calculated by taking the average of the squared differences between the predicted and actual values.

The MSE is calculated using the following formula:

MSE = (1/n) * Σ(predicted - actual)^2

where n is the total number of samples.

16. The correct answer is:

a) Linear, binary

Logistic regression is a linear regression technique that is used to model data with a binary outcome. It is a popular algorithm for binary classification tasks, where the goal is to predict a binary or categorical variable that takes on two distinct classes or categories, such as "yes/no" or "true/false."

17. The correct answer is:

a) Supervised learning

Classifying reviews of a new Netflix series based on whether they are positive, negative, or neutral is an example of supervised learning. In supervised learning, a model is trained on labeled data, where each data point (review in this case) is associated with a known target variable (positive, negative, or neutral sentiment in this case).

18. The correct answer is:

C. both a and b

Euclidean distance and Manhattan distance are both powerful distance metrics used by geometric models.

Euclidean distance, also known as straight-line distance, calculates the shortest distance between two points in a straight line. It is computed as the square root of the sum of squared differences between corresponding coordinates.

Manhattan distance, also known as city block distance or L1 distance, measures the distance between two points by summing the absolute differences between their coordinates. It gets its name from the concept that it is the distance a taxi would have to travel along the streets of a city, moving only horizontally and vertically.

19. The correct answer is:

D. none of these

None of the techniques mentioned in options A, B, or C are specifically designed for reducing the dimensions of a dataset. Let's break down each option:

A. Removing columns with too many missing values: This technique, known as missing value imputation or deletion, focuses on handling missing data rather than dimensionality reduction. It involves removing columns with a high percentage of missing values or applying imputation methods to fill in missing values.

B. Removing columns with high variance: This technique, often used for feature selection or feature engineering, aims to identify and remove features that have low variance or do not contribute significantly to the target variable. While it can help in improving model performance and reducing overfitting, it is not specifically focused on dimensionality reduction.

C. Removing columns with dissimilar data trends: This option refers to identifying columns that exhibit different trends or patterns compared to the rest of the dataset. While it can be relevant in certain scenarios for data preprocessing or feature engineering, it is not a direct approach for dimensionality reduction.

For dimensionality reduction, techniques such as Principal Component Analysis (PCA), Linear Discriminant Analysis (LDA), or feature selection methods like Recursive Feature Elimination (RFE) or SelectKBest can be employed. These techniques specifically aim to reduce the number of input features while retaining the most relevant information or capturing the most significant variance in the data.
Therefore, the correct answer is "D. none of these" as none of the mentioned techniques are primarily focused on reducing the dimensions of a dataset.

20. The correct answer is:

C. input attribute.

Both supervised learning and unsupervised clustering techniques require input attributes or features to learn patterns and make predictions or cluster data points.

In supervised learning, the input attributes are used to train a model to learn the relationship between the input attributes and a target variable or output attribute. The model is then used to make predictions on new, unseen data based on the learned patterns.

In unsupervised clustering, the input attributes are used to group similar data points together based on their feature similarities or distances. The clustering algorithm analyzes the input attributes to identify patterns or structures in the data without the need for labeled target variables.

21. The correct answer is:

(A) SVM allows very low error in classification

In Support Vector Machines (SVM), the concept of margin refers to the separation between the decision boundary and the closest data points of different classes. The hard margin in SVM refers to the scenario where the SVM algorithm aims to find a decision boundary that maximizes this margin while allowing very low error in classification.

22. The correct answer is:

(D) 1, 2, and 3

In Random Forest, an ensemble learning algorithm, increasing any of the following hyperparameters can potentially lead to overfitting:

Number of Trees: Increasing the number of trees in the random forest can lead to overfitting if the number becomes too large. More trees allow the model to capture complex relationships in the training data, but it also increases the risk of overfitting by memorizing noise or specific patterns in the training set.

Depth of Tree: Increasing the depth of individual decision trees in the random forest can make them more complex and capable of fitting the training data very closely. However, overly deep trees can also memorize noise or outliers, leading to overfitting.

Learning Rate: In random forests, the learning rate is not a standard hyperparameter. It is typically associated with gradient boosting algorithms. Nevertheless, if we consider a boosting algorithm with a learning rate, increasing the learning rate can lead to overfitting. A high learning rate allows the model to learn faster and potentially fit the training data too closely, resulting in overfitting.

23. The correct answer is:

(A) -(6/10 log(6/10) + 4/10 log(4/10))

To calculate the entropy of the target variable, we need to determine the proportion of each class in the dataset and compute the entropy using the formula:

Entropy = - (p * log2(p) + q * log2(q) + ...)

where p and q represent the proportions of the two classes (0 and 1 in this case), respectively.

In the given dataset, there are 6 instances of class 0 and 4 instances of class 1. Thus, the proportions are p = 6/10 and q = 4/10.

Substituting these values into the entropy formula, we get:

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

Simplifying this expression gives:

Entropy = -(6/10 log(6/10) + 4/10 log(4/10))

Therefore, the correct answer is (A) -(6/10 log(6/10) + 4/10 log(4/10)).

24. The correct answer is:

(A) weights are regularized with the l1 norm

Lasso, short for Least Absolute Shrinkage and Selection Operator, is a regularization technique used in linear regression. It adds a penalty term to the ordinary least squares objective function, where the weights of the regression model are regularized with the l1 norm (also known as the Lasso norm or L1 regularization).

25. The correct answer is:

(D) Perceptron

When using the Perceptron algorithm for binary classification, the decision boundary is updated incrementally based on each misclassified point during the training process. Adding a new data point to the training set, even if it is far away from the decision boundary and properly categorized by the current model, can potentially change the decision boundary because the Perceptron algorithm continues to update and refine the boundary with each iteration.

26. The correct answer is:

(D) Either 2 or 3

When dealing with multi-collinear features, which are highly correlated with each other, there are multiple options to handle the situation. Let's analyze each option:

Both collinear variables should be removed: This option suggests removing both variables that exhibit high collinearity. By removing both variables, we eliminate the redundancy and reduce the risk of multicollinearity. However, this approach may result in the loss of potentially valuable information if the variables have unique contributions to the model.

Instead of deleting both variables, we can simply delete one: This option suggests removing only one of the collinear variables while keeping the other. By removing one variable, we can still retain some information while reducing the collinearity. This approach is a practical solution if one of the variables is deemed less important or less relevant in the context of the problem.

Removing correlated variables may result in information loss. We may utilize penalized regression models such as ridge or lasso regression to keep such variables: This option acknowledges the potential loss of information when removing correlated variables. Instead of directly removing the variables, penalized regression models such as ridge regression or lasso regression can be used. These models introduce regularization terms that penalize the coefficients of correlated variables, effectively shrinking their impact and reducing the multicollinearity.
Both options 2 and 3 address the issue of multicollinearity while considering the potential loss of information. Option 2 retains one of the collinear variables, while option 3 applies penalized regression to mitigate the effects of collinearity.

Therefore, the correct answer is (D) Either 2 or 3.

27. The correct answer is:

(B) Increase by 5 pounds

In the given least squares line equation: y = 120 + 5x, the coefficient of x is 5. This coefficient represents the slope of the regression line, indicating how much the dependent variable (weight, y) changes for a one-unit increase in the independent variable (height, x).

Therefore, if the height is increased by one inch (x increases by 1), the weight (y) is expected to increase by the amount of the coefficient, which is 5 pounds in this case. Thus, the correct answer is (B) Increase by 5 pounds.

28. The correct answer is:

(D) Minimize the squared distance from the points

The line described by the linear regression equation (OLS) aims to minimize the sum of the squared differences between the predicted values of the dependent variable (y) and the actual values of the dependent variable (y). This method is known as Ordinary Least Squares (OLS) regression.

By minimizing the sum of squared differences, the regression line tries to find the best-fitting line that closely matches the data points. It achieves this by reducing the overall distance between the predicted values and the actual values of the dependent variable.

Therefore, the correct answer is (D) Minimize the squared distance from the points.

29. The correct answer is:

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

A correlation coefficient of 0.85 indicates a strong positive linear relationship between the two real-valued attributes. A correlation coefficient ranges between -1 and 1, where a value close to 1 indicates a strong positive correlation.

In this case, a correlation coefficient of 0.85 suggests that as the value of one attribute increases, the value of the second attribute tends to increase as well. The relationship is linear, meaning that the change in one attribute is directly proportional to the change in the other attribute in a positive direction.

Therefore, the correct answer is (B) As the value of one attribute increases, the value of the second attribute also increases.

30. The correct answer is:

(B) Convolutional Neural Network

Convolutional Neural Networks (CNNs) are particularly well-suited for image identification problems, such as recognizing a dog in a photo. CNNs are specifically designed to process and analyze visual data, leveraging the spatial structure and local dependencies within images.