1) Among the following identify the one in which dimensionality reduction reduces.

 a) Performance

b) statistics

c) Entropy

d) Collinearity

Ans :- d) Collinearity

Dimensionality reduction techniques are used to reduce the number of features or variables in a dataset while preserving essential information. Collinearity refers to the situation when two or more features in a dataset are highly correlated. Dimensionality reduction methods can help address collinearity problems by transforming the original features into a lower-dimensional space, making the data more manageable and less prone to multicollinearity.

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

Ans:- b) Random Forest

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

Ans:- among the given choices is (c). Decision trees are prone to overfitting, especially when they are allowed to grow deep and complex.

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

Ans:- c) Training Data

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

Ans:- c) Anomaly detection

Anomaly detection is the machine learning technique specifically designed to identify and detect outliers in data. Anomalies, also known as outliers.

6) Identify the incorrect numerical functions in the various function representation of machine learning.

a) Support Vector

b) Regression

c) Case based

d) Classification

Ans:- c) Case based

7) Analysis of ML algorithm needs

a) Statistical learning theory

b) Computational learning theory

c) None of the above

d) Both a and b

Ans:- d) Both a and b

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

Ans:- c) Both a and b

9) The total types of the layer in radial basis function neural networks is _____

a) 1

b) 2

c) 3

d) 4

Ans:- b) 2

10) Which of the following is not a supervised learning

a) PCA

b) Naïve bayes

c) Linear regression

d) KMeans

Ans:- a) PCA

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

Ans:- c) Neither feature nor number of groups is known

12) Which of the following is not a machine learning algorithm?

a) SVM

b) SVG

c) Random Forest Algorithm

d) None of the above

Ans:- b) SVG

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

Ans:-c) Both a and b

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

Ans:- c) Unsupervised Learning

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

Ans:- b) Mean squared error

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

a) Linear, binary

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

Ans:- A. supervised learning

18) Following is powerful distance metrics used by Geometric model

A. euclidean distance

B. manhattan distance

C. both a and b

D. square distance

Ans:- Both a and b

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

Ans:- B. removing columns which have high variance in data

Removing columns with high variance in the data can perform better for reducing the dimensions of a dataset. High variance in a column indicates that the data points in that column are spread over a wide range of values. Such columns may not contribute significantly to the patterns or relationships in the data and might add noise to the analysis.

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

Ans:-c) input attribute.

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

(A) SVM allows very low 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

Ans:-D) 1,2, 3

1. Number of Trees: Increasing the number of trees in the forest can lead to overfitting, especially if the number of trees becomes too large. Having more trees can cause the model to memorize noise and outliers in the training data, leading to reduced generalization performance on unseen data.

2. Depth of Tree: Increasing the depth of the individual trees in the forest can also lead to overfitting. Deeper trees can capture intricate details of the training data, including noise, and may not generalize well to new data.

3. Learning Rate: Although Random Forest doesn't have a learning rate hyperparameter, boosting algorithms, like Gradient Boosting, often have a learning rate parameter. When using boosting with decision trees, increasing the learning rate can lead to overfitting. Higher learning rates may cause the model to overemphasize the influence of individual trees, making it more sensitive to noise and overfitting the training data.

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)

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

The entropy of the target variable can be calculated using the formula for entropy:

Entropy = - Σ (P(xi) * log2(P(xi)))

where P(xi) is the probability of each class label xi in the target variable.

In this case, there are 6 instances with the label 0 and 4 instances with the label 1.

P(0) = 6/10 = 0.6

P(1) = 4/10 = 0.4

Now, we can calculate the entropy:

Entropy = - (0.6 * log2(0.6) + 0.4 * log2(0.4))

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

ANS:-(A) weights are regularized with the l1 norm

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

Ans:- D) Perceptron

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

(C) Either 1 or 3. Both removing collinear variables and using penalized regression models are valid approaches to address multicollinearity, depending on the specific requirements of the analysis and the importance of the correlated features

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

(B) increase by 5 pounds

In the given least squares regression line, the coefficient of the independent variable (height, x) is 5. This coefficient represents the change in the dependent variable (weight, y) for a one-unit increase in the independent variable.

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

ANS:- (D) Minimize the squared distance from the 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

Ans:- (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

Ans:- (B) Convolutional Neural Network