ADEOLA ODUNEWU
INTERNSHIP NO: DS2311
ASSESSMENT: INTERNSHIP

21) When implementing linear regression of some dependent variable on the set of independent
variables B0 B1… Bx where r is the number of predictors, which of the following statements will
be true?

Solution

Linear regression is fundamentally concerned with finding the best weights (coefficients) for predicting the dependent variable, and it does so by minimizing the sum of squared errors using the method of ordinary least squares.

B Linear regression is about determining the best predicted weights by using the method of ordinary least squares.

22) What indicates that you have a perfect fit in linear regression?

Solution
D The value $R^2 = 1$, which corresponds to SSR = 0

R^2 (coefficient of determination) is a measure of how well the independent variables explain the variability of the dependent variable in a linear regression model. It ranges from 0 to 1, where 1 indicates a perfect fit.

23) In simple linear regression, the value of what shows the point where the estimated regression line crosses the y axis?

Solution
B B0

The estimated regression line is represented by the equation:

Y = B0 + B1X

Where B0 Is the intercept, which represents the point where the regression line crosses the y-axis (the value of X is zero).

24) Which one represents an underfitted model?

Solutions
D Underfitting occur when on the training data fails to learn the underlying patterns. Has display by the top-left plot.

25) There are five basic steps when you're implementing linear regression:

Solution
C d, e, c, b, a

d Import the packages and classes that you need.
e Create a regression model and fit it with existing data.
c Apply the model for predictions.
b Provide data to work with, and eventually do appropriate transformations.
a Check the results of model fitting to know whether the model is satisfactory.

26) Which of the following are optional parameters to LinearRegression in scikit-learn?

Solution
In scikit-learn, the LinearRegression class has several optional parameters. Among the options you provided:

b) fit_intercept
c) normalize
d) copy_X
e) n_jobs

These are optional parameters for the LinearRegression class in scikit-learn

27) While working with scikit-learn, in which type of regression do you need to transform the array of inputs to include nonlinear terms such as X^2?

Solution
C polynomial regression

In polynomial regression, transform the array of inputs by including nonlinear terms such as X^2 X3. This allows the model to capture nonlinear relationships between the independent variable(s) and the dependent variable.

28) You should choose statsmodels over scikit-learn when:

Solution
C You need more detailed results.

Statsmodels is often preferred over scikit-learn when you need more detailed statistical results from your regression analysis.

29) ____Numpy_____ is a fundamental package for scientific computing with Python. It offers
comprehensive mathematical functions, random number generators, linear algebra routines, Fourier transforms, and more. It provides a high-level syntax that makes it accessible and productive.

Solution
B NumPy

NumPy is a fundamental package for scientific computing with Python. It provides essential functionality for numerical operations, including comprehensive mathematical functions, random number generators, linear algebra routines, Fourier transforms, and more.

30) ___Seaborn_____ is a Python data visualization library based on Matplotlib. It provides a high-level interface for drawing attractive and informative statistical graphics that allow you to explore and understand your data. It integrates closely with pandas data structures.

Solution
B Seaborn

Seaborn is a Python data visualization library based on Matplotlib. It provides a high-level interface for creating informative and aesthetically pleasing statistical graphics. Seaborn is particularly well-suited for visualizing complex datasets and integrates closely with pandas data structures, making it easy to work with DataFrame objects.

PART TWO (2)

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

Solution
D Collinearity

Dimensionality reduction techniques, such as Principal Component Analysis (PCA), can help reduce collinearity in a dataset. Collinearity refers to the situation where two or more independent variables in a regression model are highly correlated, which can cause issues such as instability in the estimation of coefficients.

42) Which of the following machine learning algorithm is based upon the idea of bagging?

Solution
B Random Forest

Random Forest is a machine learning algorithm based on the idea of bagging (Bootstrap Aggregating). Bagging involves creating multiple subsets of the training dataset by randomly sampling with replacement and then training a separate model on each subset. In the case of Random Forest, the base models are decision trees.

43) Choose a disadvantage of decision trees among the following.

Solution
C Decision Tree are prone to overfit

One of the disadvantages of decision trees is that they can be prone to overfitting, especially if the tree is allowed to grow too deep and becomes too complex.

44) What is the term known as on which the machine learning algorithms build a model based on sample data?

Solution
C Training data

The process by which machine learning algorithms build a model based on sample data is known as training.

45) Which of the following machine learning techniques helps in detecting the outliers in data?

Solution
C Anomaly detection

Anomaly detection is a machine learning technique specifically designed for identifying outliers or anomalies in a dataset.

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

Solutions
C Case-based

Case-based: is not a numerical function typically associated with machine learning. Instead, it refers to a type of reasoning or problem-solving approach in which new problems are solved by adapting solutions that were used to solve similar past problems.

47) Analysis of ML algorithm needs

Solution
D Both a and b

Analysis of machine learning algorithms can involve both statistical learning theory and computational learning theory.

48) Identify the difficulties with the k-nearest neighbor algorithm.

Solution
C Both a and b

a) Curse of Dimensionality: As the number of features or dimensions increases, the distance between data points also increases, making it difficult for the k-nearest neighbor algorithm to accurately identify neighbors.

b) Calculate the distance of the test case for all training cases: In k-nearest neighbor algorithm, the distance between the test case and all training cases needs to be calculated, which can be computationally expensive and slow, especially as the size of the dataset grows.

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

Solution
C 3
Radial Basis Function (RBF) neural networks typically consist of three (3)

50) Which of the following is not a supervised learning

Solution
A) PCA (Principal Component Analysis)

PCA (Principal Component Analysis) is not a supervised learning algorithm. It is an unsupervised learning technique used for dimensionality reduction and data compression.