Name: Adeola Odunewu
Internship N0: DS2311
Assessment: PFA Worksheet Set 1

MACHINE LEARNING WORKSHEET-1

Solutions

1. D) Both A and B

In Linear Regression, we use the method of Least Squares Error (option A) and Maximum Likelihood (option B) to find the best fit line for the data. These methods aim to minimize the difference between the predicted values from the regression line and the actual observed values in the dataset.

2. A) Linear regression is sensitive to outliers

Outliers can significantly impact the performance of linear regression models, as they can disproportionately influence the fitting of the regression line. Linear regression aims to minimize the sum of squared differences between predicted and actual values, and outliers can have a substantial impact on model accurate.

3. B) Negative

A line falls from left to right if the slope is negative. The slope of a line represents the rate of change of the dependent variable with respect to the independent variable.

4. C) Both of them

Both regression and correlation can provide a symmetric relationship between the dependent variable and the independent variable.

In regression, the linear connection between dependent and independent variables is symmetric. Swapping the roles of the variables doesn't alter the regression line. Correlation, measuring the strength and direction of a linear relationship between two variables, is also symmetric. The correlation coefficient remains the same, irrespective of whether one variable is treated as the dependent or independent variable.

5. C) Low bias and high variance

Overfitting occurs when a model is too complex, capturing noise in the training data rather than the underlying pattern. This often happens when a model has low bias (it can fit the training data well) but high variance (it is sensitive to small fluctuations in the training data).

6. B) Predictive model

The correct term for a model that involves predicting labels or categories as Predictive model

7. D) Regularization

Lasso and Ridge regression are both regularization techniques used in linear regression to prevent overfitting.

8. D) SMOTE

To address the issue of an imbalanced dataset, one common technique is using Synthetic Minority Over-sampling Technique. SMOTE generates synthetic samples for the minority class to balance the class distribution.

9. A) TPR and FPR

The AUC-ROC (Area Under the Receiver Operating Characteristic) curve is constructed using True Positive Rate (Sensitivity) on the y-axis and False Positive Rate on the x-axis.

10. B) False

A higher area under the curve (AUC) indicates a better model. AUC values range from 0 to 1, where a value of 1 represents a perfect model that has perfect discrimination between the positive and negative classes

11. A) Construction bag of words from an email

Feature extraction involves transforming raw data into a format that can be used for machine learning. For example in natural language processing (NLP), constructing a bag of words from an email is a form of feature extraction where the words in the text are represented as a set of features for analysis.

12. D) It does not make use of a dependent variable.

The Normal Equation is a closed-form solution used to compute the coefficients of a linear regression model. It does not involve choosing a learning rate, and it does not require iteration.

13. Regularization:

Regularization is a technique aimed at preventing overfitting and improving a model's ability to generalize to new data. It involves adding a penalty term to the model's objective function, discouraging the use of overly complex models with large coefficients. Regularization helps strike a balance between fitting the training data well and maintaining a simpler model, and its effectiveness is controlled by a hyperparameter.

14. Algorithms used for Regularization

Logistic Regression
Support Vector Machines (SVM):
Neural Networks (ANN)
Decision Trees and Random Forests

15. Error present in linear regression equation?

In linear regression, "error" denotes the variance between predicted and actual values of the dependent variable in the dataset. It reflects the unexplained divergence that the model cannot account for. However, the regression equation aims to depict the link between independent and dependent variables, factors exist that hinder the model's ability to perfectly foresee observed outcomes.

The error term captures the unexplained variability in the dependent variable. The goal in linear regression is to minimize these errors by determining coefficients that result in the most precise line fit. The commonly used ordinary least squares (OLS) method is employed to estimate these coefficients, reducing the sum of squared errors. This sum quantifies the squared differences between predicted and actual values for each data point.

Reducing these error terms contributes to a model that best characterizes the linear association between independent and dependent variables. However, acknowledging that some level of error is inevitable, the aim remains to keep it minimal for proficient predictions and effective generalization to new data.

Linear Regression Equation:

Y = B0 + B1X1 + B2X2 + … + BnXn + E

Where:

Y: represent the dependent variable
X: represent the independent variable
E: represent the error term