Machine Learning- Assignment

Q.1- Which of the following methods do we use to find the best fit line for data in Linear Regression?

Answer:- A) Least Square Error

Q.2- Which of the following statement is true about outliers in linear regression?

Answer:- A) Linear regression is sensitive to outliers

Q.3- A line falls from left to right if a slope is _____?

Answer:- B) Negative

Q.4- Which of the following will have symmetric relation between dependent variable and independent variable?

Answer:- B) Correlation

Q.5- Which of the following is the reason for over fitting condition?

Answer:- C) Low bias and high variance

Q.6- If output involves label then that model is called as:

Answer:- B) Predictive modal

Q.7- Lasso and Ridge regression techniques belong to _____?

Answer:-D) Regularization

Q.8- To overcome with imbalance dataset which technique can be used?

Answer:- D) SMOTE

Q.9- The AUC Receiver Operator Characteristic (AUCROC) curve is an evaluation metric for binary classification problems. It uses _____ to make graph?

Answer:- A) TPR and FPR

Q.10- In AUC Receiver Operator Characteristic (AUCROC) curve for the better model area under the curve should be less.

Answer:-B) False

Q.11- Pick the feature extraction from below:

Answer:- B) Apply PCA to project high dimensional data

Q.12-Which of the following is true about Normal Equation used to compute the coefficient of the Linear Regression?

Answer:- A) We don't have to choose the learning rate. **B**) It becomes slow when number of features is very large.

Q.13- Explain the term regularization?

Answer:- Regularization is a technique to prevent overfitting in machine learning models.

Overfitting occurs when a model performs well on the training data but poorly on the test data or new data.

Regularization reduces the complexity of the model by adding a penalty term to the loss function, which shrinks the coefficients of the features towards zero. This way, the model can learn the general patterns in the data and avoid fitting the noise.

In other words, regularization methods typically lead to less accurate predictions on training data but more accurate predictions on test data.

There are different types of regularization techniques, such as lasso, Ridge and Elastic Net.

They differ in how they calculate the penalty term and how they affect the coefficients.

Lasso uses the absolute value of the coefficients, Ridge uses the squared value, and Elastic Net uses a combination of both.

*Bias-Variance tradeoff:-

This concession of increased training error for decreased testing error is known as bias-variance trade-off.

Bias-Variance tradeoff is a well-known problem in machine learning.

-Bias measures the average difference between predicted value and true values

As bias increases, a model predict less accurately on a training dataset. High bias refers to high error in training.

-Variance measures the difference between predictions across various realization of a given model.

As variance increases, a model predicts less accurately on unseen data. High variance refers to high error during testing and validation.

Bias and Variance thus inversely represent model accuracy on training and test sets respectively.

*Regression model fits:-

By increasing bias and decreasing variance, regularization resolves model overfitting.

Overfitting describes models with low bias and high variance. However, if regularization introduces to much bias, then a model will underfit.

Underfitting describes models characterized by high bias and high variance.

*Types of regularization with linear models:-

Linear regression and logistic regression are both predictive models underpinning machine learning.

Linear regression makes continuous quantitative predictions while logistic regression produces discrete categorical predictions.

There are three main forms of regularization for regression model:-

1) Lasso regression (or L1 regularization)
2) Ridge regression (or L2 regularization)
3) Elastic Net regularization.

In all three techniques, the strength of the penalty term is controlled by lambda, which can be calculated using various cross-validation techniques.

Q.14-Which particular algorithms are used for regularization?

Answer:- Regularization is a technique used to reduce overfitting and improve the generalization performance of a machine learning model.

It involves adding a penalty term to the loss function during training, which shrinks the models coefficients and prevents them from becoming too large or complex.

There are three commonly used regularization algorithms, which differ in the type and amount of penalty they apply to the coefficients. They are:-

1) L1 regularization or Lasso regression:-

This algorithm adds the absolute value of the coefficients as a penalty term to the loss function.

This results in some coefficients becoming exactly zero, which means that the corresponding features are eliminated from the model.

L1 regularization can be useful for feature selection and sparse models.

2) L2 regularization or Ridge regression:-

This algorithm adds the squared value of the coefficients as a penalty term to the loss function.

This results in the coefficients being reduced, but not eliminated completely.

L2 regularization can be useful for reducing multicollinearity and improving stability of the model.

3) Elastic net regularization:-

This algorithm combines both L1 and L2 regularization and adds a weighted sum of the absolute and squared values of the coefficients as penalty term to the loss function.

This results in a trade-off between features selection and stability, and can be useful for models with many correlated features.

In statistics, these methods are also dubbed “Coefficient shrinkage”, as they shrink predictor coefficient values in the predictive model.

In all three techniques, the strength of the penalty term is controlled by lambda, which can be calculated using various cross-validation techniques.

Q.15- Explain the term error present in linear regression equation?

Answer:- In linear regression, the error is the difference between the observed value and the predicted value of the output variable foe given input variable.

The error reflects how well the regression model fits the data.

A smaller error means a better fit, and a larger error means a worse fit.

The error can be calculated using various metrics, such as mean squared error (MSE), root mean squared error (RMSE) or mean absolute error (MAE).

These methods are based on squaring, taking the square root or taking the absolute value of the error, and then finding the average over all the observations.

The error is also known as the residual or the disturbance term.

The error term is often denoted by e in the equation.

Equation:- $Y=B0+B1X+e$

Mathematically, the error for the i-th observation can be written as :

Error for i-th term:- $ei= Yi - Y^$

Where Yi is the observed value and $Y^$ is the predicted value of the output variable for the i-th input variable.

The predicted value is obtained by plugging the input variable into the regression equation, which has the form:

$Y^= B0+B1X$

Where B0 is the intercept and B1 is the slope of the regression line.

The error can be visualized as the vertical distance between the observed value and the regression line.