Assignment 1 ML Fliprobo internship

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

A) Least Square Error B) Maximum Likelihood

C) Logarithmic Loss D) Both A and B

Answer : A

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

A) Linear regression is sensitive to outliers B) linear regression is not sensitive to outliers

C) Can't say D) none of these

Answer : A

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

A) Positive B) Negative

C) Zero D) Undefined

Answer : B

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

variable?

A) Regression B) Correlation

C) Both of them D) None of these

Answer : B

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

A) High bias and high variance B) Low bias and low variance

C) Low bias and high variance D) none of these

Answer : C

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

A) Descriptive model B) Predictive modal

C) Reinforcement learning D) All of the above

Answer : B

7. Lasso and Ridge regression techniques belong to _____?

A) Cross validation B) Removing outliers

C) SMOTE D) Regularization

Answer : D

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

A) Cross validation B) Regularization

C) Kernel D) SMOTE

Answer : D

9. The AUC Receiver Operator Characteristic (AUCROC) curve is an evaluation metric for binary

classification problems. It uses _____ to make graph?

A) TPR and FPR B) Sensitivity and precision

C) Sensitivity and Specificity D) Recall and precision

Answer : A

10. In AUC Receiver Operator Characteristic (AUCROC) curve for the better model area under the

curve should be less.

A) True B) False

Answer : B

11. Pick the feature extraction from below:

A) Construction bag of words from a email

B) Apply PCA to project high dimensional data

C) Removing stop words

D) Forward selection

Answer : A

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

Regression?

A) We don’t have to choose the learning rate.

B) It becomes slow when number of features is very large.

C) We need to iterate.

D) It does not make use of dependent variable.

Answer : A B D

13. Explain the term regularization?

Answer : It’s a technique to prevent overfitting and improve the generalisation of models. Regularization helps to address the problem of overfitting by reducing model variance, improving model interpretability, and making the model more robust to noisy data. The strength of regularization is controlled by a hyper-parameter (alpha) that needs to be tuned during model training.

There are two types of commonly used regulairzation techniques :

1. L1 Regularization (Lasso): Involves adding a penalty proportional to the absolute value of the coefficients (L1 norm) to the loss function. It encourages sparsity in the model by shrinking less important feature coefficients to zero, effectively performing feature selection.
2. L2 Regularization (Ridge): Involves adding a penalty proportional to the square of the coefficients (L2 norm) to the loss function. It penalizes large coefficients and encourages the model to find simpler and smoother models by spreading the error across all features.

14. Which particular algorithms are used for regularization?

Answer : Regularization techniques can be applied to various machine learning algorithms to improve their performance and generalization. Some of the commonly used algorithms that incorporate regularization include:

1. Linear Regression (Ridge Regression and Lasso Regression):

   Ridge Regression: Uses L2 regularization to penalize large coefficients.
   Lasso Regression: Uses L1 regularization to promote sparsity by shrinking coefficients to zero.

2. Logistic Regression:

   Regularization can be applied to logistic regression using either L1 (Lasso) or L2 (Ridge) penalties to control overfitting.

3. Support Vector Machines (SVM):

   SVMs can use regularization through the choice of the regularization parameter C, which controls the trade-off between maximizing the margin and minimizing the classification error.

4. Neural Networks:

   Regularization techniques such as Dropout, L1/L2 regularization on weights (weight decay), and batch normalization are commonly used to prevent overfitting in neural networks.

5. Decision Trees:

   Techniques like pruning can be considered a form of regularization for decision trees, where branches of the tree that contribute little to improving the model's performance are pruned to avoid overfitting.

6. Ensemble Methods (Random Forests, Gradient Boosting Machines):

   Ensemble methods often integrate regularization techniques implicitly through the combination of multiple weak learners (trees). Regularization can also be applied through hyperparameters controlling tree depth, number of trees, learning rate, etc.

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

Answer : the term "error" refers to the discrepancy or difference between the predicted values (output) generated by the linear regression model and the actual observed values (target) in the dataset. This error is also known as the residual. The error in the context of linear regression represents the difference between the predicted and actual values of the dependent variable (or target variable), and minimizing this error is the primary objective when training a linear regression model.