MACHINE LEARNING

In Q1 to Q11, only one option is correct, choose the correct option:

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: D) Both A and B

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) Linear regression is sensitive to outliers

3. A line falls from left to right if a slope is _____?
A) Positive
B) Negative
C) Zero
D) Undefined
Answer: B) Negative

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: C) Both of them Regression and Correlation

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) Low bias and high variance

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) Predictive modal

7. Lasso and Ridge regression techniques belong to _____?
A) Cross validation
B) Removing outliers

C) SMOTE
D) Regularization
Answer: D) Regularization

8. To overcome with imbalance dataset which technique can be used?
A) Cross validation
B) Regularization
C) Kernel
D) SMOTE
Answer: D) SMOTE

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) TPR and FPR

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) False

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) Construction bag of words from a email

In Q12, more than one options are correct, choose all the correct options:

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) We don't have to choose the learning rate.
D) It does not make use of dependent variable.

ASSIGNMENT – 39

MACHINE LEARNING

Q13 and Q15 are subjective answer type questions, Answer them briefly.

13. Explain the term regularization?

Answer: Regularization is a technique used in machine learning to prevent overfitting and improve the generalization of a model. Overfitting occurs when a model learns the training data too well. Regularization techniques modify the learning process to penalize models with high complexity. This encourages the model to learn simpler, more generalizable patterns from the data.

14. Which particular algorithms are used for regularization?

Answer: Here are some commonly used regularization algorithms in machine learning:

1. **L1 Regularization (Lasso Regression):** Adds a penalty term equal to the sum of the absolute values of the model's coefficients (weights). Encourages sparsity, potentially driving some coefficients to zero, effectively performing feature selection. Useful for identifying the most important features and reducing model complexity.

2. **L2 Regularization (Ridge Regression):** Adds a penalty term equal to the sum of the squared values of the model's coefficients. Shrinks the coefficients towards zero but doesn't eliminate them entirely. Reduces overfitting and improves generalization, especially in cases with high multicollinearity.

3. **Dropout:** Randomly drops a certain percentage of neurons (and their connections) from the network during each training iteration. Prevents co-adaptation of neurons and forces them to learn more robust features independently. Commonly used in deep neural networks to reduce overfitting and improve generalization.

4. **Early Stopping:** Monitors the model's performance on a validation set during training. Stops training when validation performance starts to degrade, even if training performance continues to improve. Prevents overfitting by halting the learning process before the model becomes too complex and specific to the training data.

5. **Elastic Net:** Combines L1 and L2 regularization, offering a balance between feature selection and coefficient shrinkage.
Provides flexibility in controlling the degree of sparsity and regularization strength.

6. **Weight Decay:** Gradually reduces the magnitude of weights during training, similar to L2 regularization. Prevents weights from growing too large and helps control model complexity.

7. **Data Augmentation:** Artificially expands the training dataset by creating modified versions of existing data (e.g., rotations, flipping, cropping images). Exposes the model to more diverse examples, reducing overfitting and improving generalization.

8. **Bagging and Boosting:** Ensemble methods that combine multiple models trained on different subsets of data or with different hyperparameters. Can reduce overfitting and improve overall model performance by averaging or combining predictions from multiple learners.

9. **Bayesian Regularization:** Incorporates prior knowledge about the model's parameters into the learning process. Can lead to more robust and generalizable models by constraining the values that parameters can take.

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

Answer: The term "error" refers to the difference between the observed (actual) values and the values predicted by the linear regression model. This difference is often denoted as the residual, and the process of fitting a linear regression model involves minimizing these residuals to find the best-fitting line.

In other words, the "error" in the linear regression equation refers to the discrepancies between the actual observed values and the values predicted by the model, and the goal is to minimize these errors to obtain an accurate and reliable model.