

ASSIGNMENT – 39

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

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

A) Linear regression is sensitive to outliers

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

B) Negative

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

B) Correlation

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

C) Low bias and high variance

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

B) Predictive modal

7. Lasso and Ridge regression techniques belong to _____?

D) Regularization

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

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

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

B) False

11. Pick the feature extraction from below:

B) Apply PCA to project high dimensional data

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.

ASSIGNMENT – 39

MACHINE LEARNING

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

13. Explain the term regularization?

Regularization is a technique used to prevent overfitting by adding a penalty to the model's complexity. This is done by adding a regularization term to the loss function, which discourages the model from fitting the noise in the training data.

14. Which particular algorithms are used for regularization?

There are two algorithms used commonly in the regularization.

LASSO:

A regression model which uses the L1 Regularization technique is called LASSO (Least Absolute Shrinkage and Selection Operator) regression. Lasso regression also helps us achieve feature selection by penalizing the weights to approximately equal to zero if that feature does not serve any purpose in the model.

Regularization Term:

The regularization term is the sum of the absolute values of the coefficients multiplied by a tuning parameter,

$$LOSS FUNCTION = RSS = \gamma \sum_{n=1} p|B_j|$$

Where RSS is the Residual Sum of Squares, B_j are the coefficients, and λ controls the amount of shrinkage.

Feature Selection:

Lasso can shrink some coefficients to exactly zero, effectively performing feature selection. This means it can select a simpler model that only includes the most important label.

Bias-Variance Tradeoff:

By increasing λ , we introduce more bias into the model but reduce variance, leading to potentially better performance on unseen data.

Ridge:

It is similar to the LASSO but it doesn't eliminate the feature that does not give importance to the label, Ridge regression retains all features in the model, reducing the impact of less important features by shrinking their coefficients.

Regularization Term:

The regularization term is the sum of the absolute values of the coefficients multiplied by a tuning parameter λ :

$$LOSS FUNCTION = RSS = \gamma \sum_{n=1}^p B_j^2$$

Where RSS is the Residual Sum of Squares, B_j^2 are the coefficients, and λ controls the amount of shrinkage.

No Feature Selection:

Ridge does not perform feature selection. Instead, it shrinks the coefficients towards zero but does not set any of them to exactly zero. All features remain in the model but with reduced impact.

Bias-Variance Tradeoff:

Similar to Lasso, increasing λ increases bias but reduces variance. This can improve the model's performance on new data.

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

An error term is a residual variable produced by a statistical or mathematical model, which is created when the model does not fully represent the actual relationship between the independent variables and the dependent variables. Although the error term and residual are often used interchangeably, while an error term represents the way observed data differs from the actual population, a residual represents the way observed data differs from sample population data.