<u>MACHINE LEARNING ANSWERS</u>

1. A least square
2. B Linear regression is sensitive to outliers
3. B Negative
4. B Correlation
5. C Low bias and high variance
6. B Predictive Model
7. D Regularization
8. D SMOTE
9. A TPR and FPR
10. B False
11. A Construction bag of words from a email
12. A We don’t have to choose the learning rate.

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

13. Explain the term regularization?

<u>Answer:</u>

Regularization is a crucial technique in machine learning used to prevent overfitting, which occurs when a model learns the noise in the training data rather than the underlying patterns. Here's how it works:

Complexity Control: Regularization adds a penalty term to the loss function, discouraging the model from assigning too much importance to individual features or coefficients. This helps control model complexity and results in better generalization to new data.

Preventing Overfitting: By penalizing large coefficients, regularization prevents the model from becoming overly complex and memorizing the training data. Instead, it learns underlying patterns.

Balancing Bias and Variance: Regularization strikes a balance between model bias (underfitting) and model variance (overfitting), leading to improved performance.

Feature Selection: Some regularization methods promote sparse solutions, automatically selecting important features while excluding less important ones.

Handling Multicollinearity: When features are highly correlated, regularization stabilizes the model by reducing coefficient sensitivity to small data changes.

In summary, regularization helps models generalize better by avoiding overfitting and maintaining a balance between complexity and performance.

14. Which particular algorithms are used for regularization?

Answer:

Regularization techniques are essential for preventing overfitting in machine learning models. Here are some commonly used algorithms:

1. **L1 Regularization (Lasso):** L1 regularization adds a penalty term to the loss function by constraining the sum of absolute feature coefficients. It encourages sparse solutions, effectively selecting important features while excluding less relevant ones.

2. **L2 Regularization (Ridge):** L2 regularization penalizes the sum of squared feature coefficients. It helps control model complexity by preventing large coefficients and promoting better generalization to new data.

3. **Elastic Net:** Elastic Net combines L1 and L2 regularization. It balances the trade-off between feature selection (L1) and coefficient shrinkage (L2), providing a flexible approach to regularization.

Remember, regularization aims to strike a balance between bias and variance, leading to improved model performance. It’s a powerful tool for preventing overfitting and enhancing generalization.

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

Answer:

In the context of linear regression, the error term (also known as the residual or disturbance term) represents the difference between the actual observed values and the predicted values from the regression model. Here are some key points about the error term:

Definition:

- The error term accounts for the variability in the dependent variable (response) that cannot be explained by the independent variables (predictors) included in the model.

- It captures all other factors affecting the dependent variable beyond those explicitly considered in the regression equation.

Interpretation:

- When the regression model predicts an outcome (e.g., stock price), the error term represents the deviation between the predicted value and the actual observed value.

- Ideally, the error term should be close to zero, indicating that the model accurately captures the relationship between the variables.

Mathematically:

- In a simple linear regression model, the equation is:

$$y = \beta 0 + \beta 1\ x + \epsilon$$

Here, (y) is the dependent variable, (x) is the independent variable, (\beta_0) and (\beta_1) are the regression coefficients, and (\epsilon) is the error term.

- The error term (\epsilon) represents the combined effect of unobserved variables, measurement errors, and random fluctuations.

Purpose:

- Researchers use regression models to estimate the relationship between variables and make predictions.

- By analyzing the error term, we assess how well the model fits the data and understand the precision of predictions.

Remember that a smaller error term indicates a better fit of the model to the data, while a larger error term suggests room for improvement. If the error term is consistently large, it may indicate omitted variables or other model deficiencies.