MachineLearning_assignment

Answers

1. A) Least Square Error
2. A) Linear regression is sensitive to outliers
3. B) Negative
4. B) Correlation
5. C) Low bias and high variance
6. A) Descriptive model
7. A) Cross validation
8. D) SMOTE
9. A) TPR and FPR
10. B) False
11. B) Apply PCA to project high dimensional data
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?

Ans)

Regularization is a technique used in machine learning to prevent overfitting and improve the generalization of models by adding a penalty term to the loss function of the model. The goal of regularization is to encourage the model to learn simpler patterns that generalize better to new data, rather than memorizing the training data.

In essence, regularization adds an additional constraint on the model's parameters, forcing them to take on smaller values. This is achieved by adding a regularization term to the loss function of the model, which penalizes large parameter values. The two most common forms of regularization are L1 and L2 regularization.

L1 regularization, also known as Lasso regularization, adds a penalty term to the loss function that is proportional to the absolute value of the coefficients. This encourages the model to learn sparse representations, where many of the coefficients are set to zero, resulting in a simpler and more interpretable model.

L2 regularization, also known as Ridge regularization, adds a penalty term to the loss function that is proportional to the square of the coefficients. This encourages the model to learn smaller but non-zero coefficients, which results in a smoother decision boundary and reduces the effect of small variations in the input data.

14. Which particular algorithms are used for regularization?

Ans)

There are several algorithms that can be used for regularization in machine learning, including:

1. L1 Regularization (Lasso Regression): It adds an L1 penalty term to the cost function, which encourages the model to have sparse weight values by shrinking some of them to zero.
2. L2 Regularization (Ridge Regression): It adds an L2 penalty term to the cost function, which encourages the model to have small weight values, but not necessarily zero.
3. Elastic Net Regularization: It is a combination of L1 and L2 regularization, which provides a balance between both methods.
4. Dropout: It randomly drops out some of the neurons during training, which helps to prevent overfitting and encourages the network to learn more robust features.
5. Early Stopping: It stops the training process once the validation error starts increasing, which prevents the model from overfitting the training data.
6. Data Augmentation: It increases the size of the training data by applying random transformations, such as rotation, cropping, and flipping, which can help prevent overfitting.
7. Batch Normalization: It normalizes the output of each layer to have zero mean and unit variance, which can help prevent the model from overfitting and improve its generalization ability.

These techniques are widely used in machine learning and deep learning to prevent overfitting and improve the generalization performance of the models.

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

Ans)

In linear regression, the error (also known as the residual) is the difference between the predicted output of the model and the actual output of the data point.

The linear regression equation can be represented as:

$y = mx + b + \varepsilon$

where:

- y is the dependent variable or the output that we want to predict
- x is the independent variable or the input feature that we use to make the prediction
- m is the slope of the line, which represents the relationship between x and y
- b is the y-intercept of the line, which represents the value of y when x is 0
- ε is the error term, which represents the deviation between the predicted value of y and the actual value of y for a given data point.

The error term captures the difference between the predicted output of the model and the actual output, which is not explained by the input features. The goal of linear regression is to minimize the sum of the squared errors (SSE) between the predicted and actual values of y over the entire dataset.

Minimizing the SSE can be achieved through various optimization techniques, such as the ordinary least squares (OLS) method, gradient descent, or closed-form solution. By minimizing the error term, the linear regression model can find the best-fitting line that describes the relationship between the input and output variables in the data.