SD For AI Report

Lokesh Punati

11514358

Task-2

Q1) According to you, why do overfitting and underfitting occur, and how resolve them? What is the difference between them?

underfitting: When a statistical model or machine learning algorithm is unable to recognize the underlying pattern in the data, or when it only performs well on training data but badly on testing data, this is referred to as underfitting. Our machine learning model's accuracy is destroyed by underfitting. Its recurrence merely indicates that our model or method does not adequately suit the data. It typically occurs when we try to develop a linear model with fewer non-linear data or when we have insufficient non-linear data to build an accurate model.

Some of the techniques to reduce underfitting are

- increase model complexity

- Increase feature count by doing feature engineering

- Data noise should be removed.

- To improve outcomes, increase the period of training or the number of epochs.

Overfitting: When a statistical model fails to produce reliable predictions on test data, it is said to be overfitted. A model begins to learn from the noise and erroneous data entries in our data set when it is trained with such a large amount of data. And when using test data for testing yields high variance. Due to too many details and noise, the model fails to appropriately identify the data.

Some techniques to reduce overfitting are

- Increase the training data.

- simplify the model.

- During the training phase, quitting too soon

- Regularization of the Ridge and the Lasso

Q2) What kind of pattern did you analyze in the Train and Test score while running the code of overfitting?

From the Code we can see that initially degree 1 of a linear function polynomial is insufficient to fit the training samples. And The true function is almost exactly approximated by a polynomial of degree 4. Higher degrees, however, will cause the model to overfit the training data, which means it will pick up on the noise in the training data.

Q3) What is cross-validation, and what did you analyze in a different type of validation that you performed?

A statistical technique called cross-validation is used to evaluate the effectiveness or accuracy of machine learning models. It serves as a precaution against overfitting in predictive models, especially when the available data may be limited. In cross-validation, the data is divided into a predetermined number of folds or divisions, the analysis is conducted on each fold, and the overall error estimate is then averaged.

Cross-validation is a technique used to prevent overfitting and evaluate the model's performance using new data. You can choose the value of k for your dataset using a variety of strategies.

Some of cross validation techniques are;

## K fold: One way to improve the holdout method is K-fold cross validation. This approach ensures that our model's score is independent of how we chose the train and test sets. The holdout approach is done k times after dividing the data set into k sections.This technique's disadvantage is that it requires k times as much computing to perform an evaluation because the training algorithm must be run from scratch each time.

## Stratified K Fold: It can be difficult to apply K Fold to a classification problem. The probability that we will obtain very uneven folds from randomly rearranging the data before dividing it into folds increases the possibility that our training will be biased.

## Leave-one-out : The value of p is set to one in the common requirements of Leave-P-Out cross validation. Now that we have n number of choices for n data points and p = 1, the approach is significantly less exhaustive.

Q4) Explain the analysis from generated ROC and validation curve and what they represent?

Roc stands for Receiver Operating Characteristics

The relation or trade-off between clinical sensitivity and specificity for each potential cut-off for a test or set of tests is typically portrayed graphically using ROC curves. Additionally, the ROC curve's area under the curve provides insight into the advantages of using the test(s) in question.

The areas under ROC curves are used to compare the utility of tests since they represent a measure of a test's general usefulness, with a larger area indicating a more valuable test.









