Software Development for Artificial Intelligence

Report

Name : Divi Reshmi Chowdary

ID : 11520230

**Task\_2**

### 1. According to you, why do overfitting and underfitting occur, and how resolve them? What is the difference between them

A: When your model is underfit, it produces accurate but initially erroneous predictions. In this instance, both the train error and the val/test mistake are significant. When your model is overfit, it produces inaccurate predictions. In this instance, val/test error is significant but train error is relatively modest.

Model performance might suffer from both overfitting and underfitting. Overfitting, however, is by far the most typical issue with applied machine learning.

Because the performance of machine learning algorithms on training data, which is what matters most to us, is different from how well the algorithm performs on unlabeled data, overfitting is such a problem.

Simply put, a validation dataset is a portion of your training data that you withhold from machine learning algorithms until the very end of your project. You can evaluate the learnt models on the validation dataset once you've chosen and fine-tuned your machine learning algorithms on your training dataset to get a final, objective picture of how the models would perform on untested data.

A gold standard in applied machine learning for determining model accuracy on unknown data is cross validation. Underfitting and overfitting, the two main problems with machine learning, are represented by these two variables. When the training error is significant, underfitting occurs. Overfitting occurs when there is a significant difference between the two, or when the testing error is higher than the training error.

The issue of overfitting can be solved using many ways, such as:

Cut back on the features.

Regularization: Include the penalty parameter in the cost function to increase the penalty for an overfitted model and hence penalize the model.

Cross-validation entails dividing the training data into k groups or folds at random, each of roughly similar size. The model is fitted on the remaining k - 1 folds with the initial fold acting as a validation set. This process is repeated k times, with each iteration treating a different collection of observations as a validation set. Then, an estimation can be created by averaging the k outcomes from the folds.

### 2. What kind of pattern did you analyze in the Train and Test score while running the code of overfitting?

A: An approach for examining how and why a certain model is overfitting on a particular dataset is called an overfitting analysis. At each stage of training, the model's performance on the train and test sets can be calculated, and graphs can be produced. This graph, which displays one curve for model performance on the training set and another for the test set for each learning increment, is frequently referred to as a learning curve plot. The data which is in starting should increasing after it starts decreasing.

Examining validation metrics like accuracy and loss can reveal overfitting. When the model is impacted by overfitting, the validation measures often grow until a point where they plateau or begin to decline. The model searches for a good fit during an upward trend, and when it finds one, the trend begins to decline or stagnate.

### 3. What is cross-validation, and what did you analyze in a different type of validation that you performed?

A: Cross validation is a method for evaluating the generalizability of a statistical analysis to a different data set. It is a method for assessing machine learning models that entails training various models on different subsets of the input data and then comparing the results. There is a good likelihood that we can easily identify over-fitting using cross-validation.

There are several cross validation techniques such as :-

1. K-Fold Cross Validation

2. Leave P-out Cross Validation

3. Leave One-out Cross Validation

4. Repeated Random Sub-sampling Method

5. Holdout Method

1. K-Fold Cross Validation:

Separate the test data set and solely use it for the assessment's conclusion. Therefore, cross-validation will be carried out on the training set.

One strategy to enhance 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. The drawback of this approach is that the evaluation process requires k times as much computing because the training algorithm must be done from scratch each time.

2.Repeated K-Fold Cross-Validation:

This means that a different splitting of the dataset into k-folds can be used each time the procedure is performed, which will change the distribution of performance scores and the mean estimate of the model's performance.

3.Leave -one-out cross validation:

When N, the number of data points in the set, is used as K, K-fold cross validation is brought to its logical conclusion. This signifies that a forecast is made for one point out of the N points for which the function approximator has been trained. The average error is once again computed and used to assess the model. Leave-one-out cross validation error (LOO-XVE) provides a good evaluation, although it appears to be highly expensive to compute at first glance. Thankfully, locally weighted learners are able to produce LOO predictions just as quickly as they can make conventional predictions. As a result, computing the LOO-XVE is substantially more efficient than computing the residual error and takes no longer.

### 4. Explain the analysis from generated ROC and validation curve and what they represent?

A: The relationship or trade-off between clinical sensitivity and specificity for each potential cut-off for a test or set of tests is usually depicted graphically using ROC curves. The area under the ROC curve also provides insight into the advantages of using the test(s) in question.

In clinical biochemistry, ROC curves are used to determine the best cut-off value for a test. The highest true positive rate and lowest false positive rate are found in the best cut-off.

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.

The term ROC stands for Receiver Operating Characteristic.

You must understand the ideas of true positive, true negative, false positive, and false negative in order to create a ROC curve.

**Task\_1**

A screenshot of a computer

Description automatically generated with medium confidence

Graphical user interface, chart, application

Description automatically generated

Text

Description automatically generated

Text

Description automatically generated

**Text

Description automatically generated**

**A screenshot of a computer

Description automatically generated with medium confidence**

**Text

Description automatically generated**

**Text

Description automatically generated**

**Text

Description automatically generated**

**Text

Description automatically generated**