Practical Machine Learning

Auto_sklearn Fail Report

Sanjeeb Humagain

May 6, 2024

There could several reasons that causes the reduction in accuracy of a ML model. In the provided code fail.py, accuracy of RandomForestClassifier is 0.67 with default hyperaparameters whereas the accuracy of AutoSklearnClassifier is 0.65 which runs for 5 minutes to find a better hyperparameters. This indicates that AutoSklearnClassifier failed to give a better result after utilizing considerable amount of time as compared to RF. One thing that immediately comes into my mind is overfitting. But we have to evaluate accuracy by changing different parameters to understand what is causing this failure.

Without using cross validation following accuracy score was achieved for training data and test data.

Training Accuracy of automl without CV: 0.8940
Test Accuracy of automl without CV: 0.5875

After introduceding cross validation of 2 and the following results in accuracy score was observed.

AutoML Training Accuracy with CV: 0.6005
AutoML Test Accuracy with CV: 0.6725

According to the observation made, the accuracy is much higher for the training set without cross validation as compared to the training set with cross-validation, it could be because of overfitting. This is because, while training the model without CV, all the data are being used for training and the model can potentially memorize the data it was trained on thereby resulting in a high accuracy on the training set. However, the accuracy for unseen test data, the accuracy will be worse.

On the other hand, if we look at the test data, the accuracy for test data without cross validation (default as in provided code) is 0.5875 but after introducing cross validation, the accuracy score was improved significantly to 0.6725. This suggest that, with cross validation ML model can perform better on unseen data. Additionally, cross validation helps to reduce the overfitting problem and hence increases the model performance.

Moreover, while comparing with the accuracy of the Random Forest Classifier, we can not compare two ML models without analyzing different values of hyperparameters. We have been using cross validation in several exercises. This exercise helped to understand the importance of cross validation in more depth by comparing the results of two ML models.