Practical Machine Learning

Auto-sklearn fail

Bimal Pandey

Introduction: Auto-sklearn is an automated machine learning (AutoML) tool that aims to simplify the process of building machine learning models. Auto-sklearn automates the selection of the best machine learning model, data preprocessing, and hyperparameter tuning. This automation saves time and effort, particularly for non-machine learning professionals. Auto-sklearn uses Bayesian optimization to efficiently search the hyperparameter space incorporates meta-learning to learn from previous experiments and apply this knowledge to new datasets. Meta-learning helps auto-sklearn to initialize the search space efficiently and guide the optimization process.

Analysis of the code: In the given code the the performance of Autosklearn in terms of accuracy is 0.63 which is outperformed by the Random Forest with its default hyperparameters gaining accuracy of 0.67. I found that the code in the fail.py uses default resampling strategy and search space is not specified. However, Autosklearn uses large space by default which might not be the actual reason behind the poor performance. Then, I implemented the resampling strategy using cross-validation.

The noticeable thing is that the Autosklearn achieved the training accuracy of 0.78 and test accuracy of 0.61 without using cross validation. But the training and test accuracy were 0.6 and 0.65 respectively with cross validation which potentially indicates overfitting.

Furthermore, the number of folds in cross validation is varied to examine the performance. After some trails, I found that the performance of model was changed when varying the number of folds. I noticed the following changes when varying the number of folds.

3- fold= 0.66

5- fold= 0.698

6 -fold= 0.68

10- fold= 0. 6789

Therefore, cross validation can help to reduce the overfitting in the model since it provides better evaluation on the unseen data that might be crucial for reducing or preventing the overfitting. Also the hyperparameter search need to be adjusted according to the dataset and their nature so that better performance can be achieved. Sometimes, the large space for hyperparameters can also be less efficient or may reduce the performance of Autosklearn.

Additionally, I tried measuring the performance using standard Scaler and one hot encoder but it didn’t prove to be beneficial or had minimal impact on performance. From my experiment and knowledge, I can conclude that the dataset do not have scaling problem. Similarly, I noticed time also changes the performance of this model.