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. I got different outcomes when varying the number of folds. For example, accuracy of 0.66 for 3 fold, 0.698 for 5 fold, 0.68 for 6 fold and 0.6789 during 10 folds.

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 after using cross validation indicating overfitting in the model. When training a model without CV, all data is used, which can cause the model to memorize the training set and achieve high accuracy. However, the precision of previously unseen test data will become worse.

By training and validating across multiple subsets of data, cross-validation minimizes the risk that the model will only perform well on a specific segment of the data. Cross-validation can help to address this issue with numerous training and validation sets, allowing the model to function consistently across previously unknown data points. Consistent exposure to various data features aids in refining the model to recognize larger patterns rather than memorizing the minutiae of the training set. From this observation, cross validation helps ML model to perform well in unseen data. 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 suggesting initial feature set was already suited for the model. Similarly, I noticed time also changes the performance of this model.