Machine Learning Lab 12

Grid Search and Pipeline

In machine learning, hyper parameter optimization or tuning is the problem of choosing a set of optimal hyperparameters for a learning algorithm.
Hyper-parameters are parameters that are not directly learnt within estimators and thus the best way to learn hyperparameters and optimize the performance of a machine learning algorithm is through hyper parameter optimization. There are various methods of hyperparameter optimization namely Grid Search, Random Search, Bayesian Search, Cross validation score etc.

A machine learning pipeline is used to help automate machine learning work-flows. They operate by enabling a sequence of data to be transformed and correlated together in a model that can be tested and evaluated to achieve an outcome, whether positive or negative. Scikit-learn's Pipeline class is designed as a manageable way to apply a series of data transformations followed by the application of an estimator. Ultimately, this simple tool is useful for
* Convenience in creating a coherent and easy-to-understand workflow
* Enforcing workflow implementation and the desired order of step applications
* Reproducibility
* Value in persistence of entire pipeline objects (goes to reproducibility and convenience)

The dataset

The dataset used to perform this experiment is the wine quality dataset, it is a combination of data on two types of wine variants, namely red wine and white wine, of the portuguese “Vinho Verde” wine. The dataset contains information on the parameters for fixed acidity, volatile acidity, citric acid, residual sugar, chlorides, free sulfur dioxide, total sulfur dioxide, density, pH, sulphates, alcohol.

Experiment

In this experiment I used the sklearn’s grid search functionality and pipelines to predict the quality of a wine and improve the performace of these classification models.

Using the pandas library in I loaded the red wine into the memory and then with the help of pandas.Dataframe.info() function I projected the information about the various columns in the dataset.

For performing the experiment I started with plotting the bar graphs for some of the features in the dataset with other features, this helped to find distribution of the dataset between the different features of the dataset.

Next I binned the target values of the dataset between ‘bad’ and ‘good’ so as to remove the skewness in the dataset and the used the label encoder to encode the target values of the dataset.

Next I trained random forest classifier, SGD classifier and support vector classifier on the data set to obtain an accuracy of 87%, 83% and 86% respectively.

Next I created two pipeline the first one contains a standard scaler, a PCA and an support vector classifier in that order and the second one contains only a standard scaler and support vector classifier. I set the parameters for grid search and on training the pipelines over the data I achieved an accuracy of 89% and 90% respectively.

Finally I use a cross validation over random forest and SGD to achieve an accuracy of 91%.

The code and plots can be found in the accompanying jupyter notebook.