Machine Learning Lab 5

Decision Tree

Decision tree is one of the most popular machine learning algorithms used all along. A decision tree is a decision support tool that uses a tree-like model of decisions and their possible consequences, including chance event outcomes, resource costs, and utility. It is one way to display an algorithm that only contains conditional control statements

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 decision tree algorithms to predict the quality of a wine.

Using the pandas library in I loaded the red wine and white wine datasets into the memory from their respective csv files and then merged the two datasets into one single pandas dataframe.

Using the pandas.Dataframe.describe() function in pandas I calculated the various statistical measures of each of the columns of the dataset.

For performing the experiment I started with plotting the scatter plot for each of the features in the dataset with every other feature, this helped to find if there were any features which were linearly separable. In the case of my dataset they were not.

Next used a decision tree with the training data containing all the 11 features from the dataset, the tree uses gini index and has a max depth of 3. With this decision tree I was able to obtain a training accuracy of 0.817 and a test accuracy of 0.816

Looking at the visualization of decision tree it is evident that the decision tree is overfitting and thus we would need something better than a single decision tree and this would be a random forest.

Using the random forest classifier in sklearn I calculated the feature importance for the features of the dataset and used those featured in a random forest classifier to achieve a training accuracy score of 1.0 and a test accuracy score of 0.89 which are much greater than those obtained with a decision tree.

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