Machine Learning Lab 4

Logistic Regression

In machine learning, the logistic model is a widely used statistical model that, in its basic form, uses a logistic function to model a binary dependent variable; many more complex extensions exist. Logistic Regression is a classification algorithm. It is used to predict a binary outcome (1 / 0, Yes / No, True / False) given a set of independent variables. To represent binary / categorical outcome, we use dummy variables. It can also be thought of as a special case of linear regression when the outcome variable is categorical, where we are using log of odds as dependent variable. In simple words, it predicts the probability of occurrence of an event by fitting data to a logit function.

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 logistic regression 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 random forests to find the importance of features in the dataset and as to how much each feature contributes to the importance. The two most important features are the fixed acidity and the volatile acidity.

Finally, I applied logistic regression to the dataset using both l1 and l2 penalty and the using the saga and newton-cg solver. I was able to achieve and accuracy of 52 and 53 percent respectively. The reason for this low accuracy was the dataset being skewed with just 20 examples in class 3 and more than 2000 examples in class 6 and this skewness in the dataset was the reason for a bad performance of the logistic regression.

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