Machine Learning Lab 1

Data Preprocessing

Data preprocessing is a data mining technique that involves transforming raw data into an understandable format. Real world data is often incomplete, inconsistent, and/or lacking in certain behaviors or trends, and is likely to contain many errors. Data preprocessing is a proven method of resolving such issues. Data preprocessing prepares raw data for further processing. Data preprocessing steps include but are not limited to data cleaning, data integration, data transformation, data reduction and data discretization.

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 have performed operations to clean the data, done statistical analysis of the data and then used various visualization tools to visualize the data in different ways which in turn reveals different information about the data which are otherwise not easily discernible.

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.

* The dataset has a total of 4898 rows.
* The means for each of the columns are calculated as:
    - Fixed acidity -> 6.85
    - Volatile acidity -> 0.27
    - Citric acid -> 0.33
    - Residual sugar -> 6.39
    - Chlorides -> 0.045
    - Free sulphur dioxides -> 35.30
    - Total sulphur dioxides -> 138.36
    - Density -> 0.99
    - pH -> 3.18
    - Sulphates -> 0.48
    - Alcohol -> 10.51

Using the pandas.Dataframe.dtypes method gives the datatypes of all the columns in the dataframe.

The pairplot function in seaborn library in python I was able to plot each column vs every other column in the dataset in the form of a scatter plot. And for also look at the values of each column in the form of a histogram.

Using the violin plot function in seaborn I am able to plot a violin plot for a column is the dataset. Violin plots are similar to box plots except that they also show the probability density of the data at different values. Thy also include a marker for median of the data and a box indicating the inter quartile range.

Using the violin plots I was able to infer that the median for the density lies between 1.00 and 0.99, and that for citric acid lies between 0.25 and 0.5, and for sulphates it lies between 0.6 and 0.4.

Seaborn library also allows to plot box plot for a dataset with its boxplot function. A boxplot depicts groups through their quartiles. It has whiskers indicating variability outside the upper and lower quartiles. Outliers are plotted as individual points as can be seen in the diagram in the jupyter notebook.

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