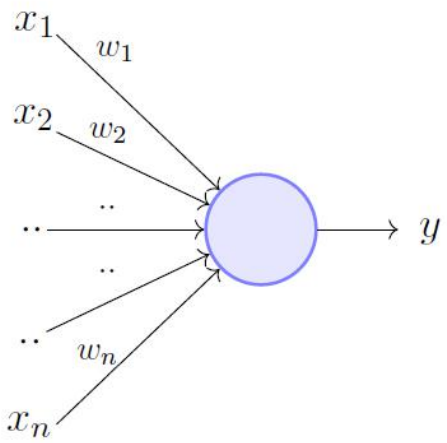
Machine Learning Lab 2

Perceptron Learning Algorithm

In machine learning, the perceptron is an algorithm for supervised learning of binary classifiers. It takes an input, aggregates it (weighted sum) and returns 1 only if the aggregated sum is more than some threshold else returns 0. A single perceptron can only be used to implement linearly separable functions. It takes both real and boolean inputs and associates a set of weights to them, along with a bias (the threshold thing I mentioned above). We learn the weights, we get the function.



$$y = 1 \quad if \sum_{i=1}^{n} w_i * x_i \geq \theta$$

$$= 0 \quad if \sum_{i=1}^{n} w_i * x_i < \theta$$

Rewriting the above,

$$y = 1 \quad if \sum_{i=1}^{n} w_i * x_i - \theta \geq 0$$

$$= 0 \quad if \sum_{i=1}^{n} w_i * x_i - \theta < 0$$

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 implemented a perceptron from scratch and and then used the matplotlib visualization library in python to plot the decision surface generated by the perceptron. This reveals how perceptron are very useful in cases where the data is linearly separable however they fail in the cases where is data is not linearly separable.

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

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 target values for each data point in the form of a histogram.

Using the numpy library I implemented the Perceptron class along with the helper function for fitting a perceptron to a dataset and predicting the output from a data.

Initally I went ahead with fitting all the 11 parameters of the dataset but since plotting 11 dimensional data point is not possible, hence, for sake of visualization I chose the two most important features from the dataset fixed acidity and the volatile acidity.

From the visualization of the decision boundary it is clear that since the dataset was not linearly seperable the perceptron was not able to map a very clear boundary between the red and blue i.e the good and the bad wine samples. Examples can be seen with overlapping decision boundaries and thus one can conclude that the perceptron learning algorithm performs poorly on this task.

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