ML Lab Assignment 1

Notebook:	Machine Learning Lab		
Created:	16/08/2018 22:31	**Updated:**	05/09/2018 12:46
Author:	Prateek Chanda		
Tags:	Ass1		

Machine Learning Lab Assignment 1

In this assignment we used Naive Bayes Classifier as a classification algorithm and analyzed its performance on proposed data sets.

We used two data sets

- House Votes Data
- Breast Cancer Data

to analyze our implementation of Naive Bayes Classifier algorithm.

Using Weka

We first collected the data sets and analyzed the features present under corresponding data set using the tool Weka.
Then we classified each of the data sets by applying Naive Bayes classification algorithm and using 5-fold cross-validation and analyzed the results obtained on Weka.

Breast Cancer Data

Naive Bayes Classification - 0.949 (Acuracy)

House Votes Data

Naive Bayes Classification - 0.93 (Acuracy)

Implementation in Python

Now we go on to implement our Naive Bayes Algorithm in python.
Since we need certain in-built functionalities for our implementation purpose, we use the scikit-learn package. We simply import the module by running the following on the terminal

```
pip install scikit-learn
```

Once we have imported the corresponding modules, we read the data from the data sets stored as csv files using python.
Now, we built our **Naive Bayes Classifier Algorithm** using the corresponding implementation from `scikit-learn` called `GuassianNB()`

```
naive_clf = GaussianNB()
```

Using the concept of 5-fold cross-validation, we then divide the whole dataset into training set of 80% and test set of 20% data

```
cross_validate(clf_nb,train_x,train_y,cv=5)
```

We then computed the accuracy and repeated the above experiment for 10 times.

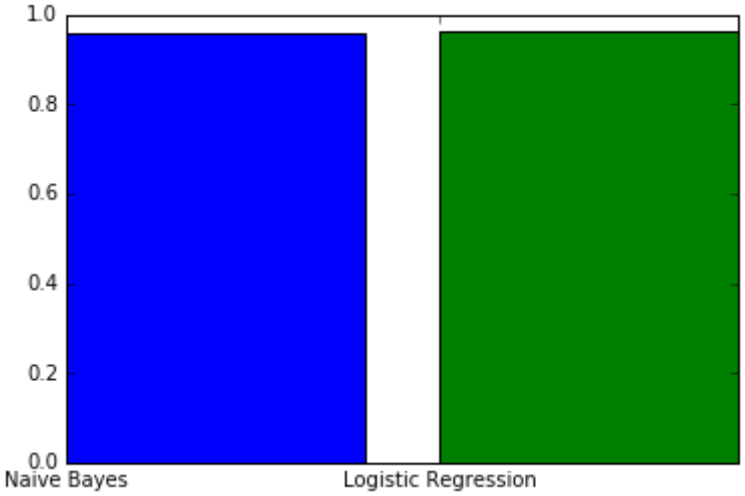
Finally we computed the average of all the accuracy values obtained and calculated the mean accuracy for that particular dataset.

We now also compute the accuracy in a similar manner using **Logistic Regression** as a classification model and compared our results obtained with respect to Naive Bayes Classification results.

Results Obtained

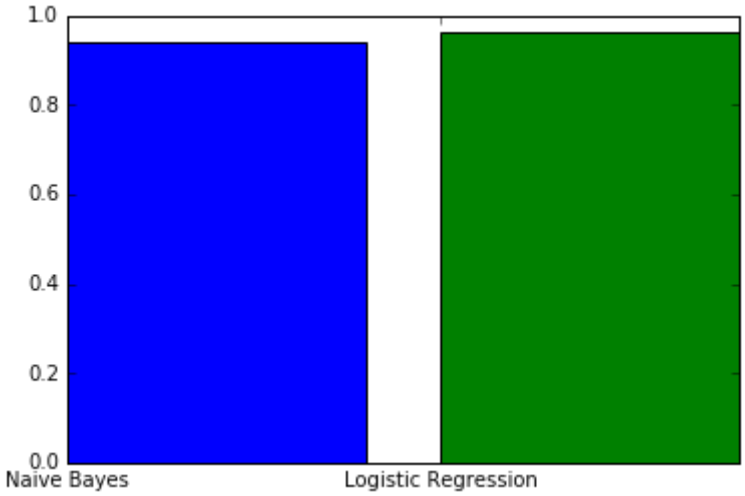
Breast Cancer Data

Naive Bayes Classification - 0.954 (Acuracy)
Logistic Regression - 0.963 (Accuracy)



House Votes Data

Naive Bayes Classification - 0.935 (Acuracy)
Logistic Regression - 0.960 (Accuracy)



Conclusions

As we can see Logistic Regression Performs much better than Naive Bayes for both the data sets **House Votes** and **Breast Cancer** according to our implementation and also according to weka tool.