NAÏVE BAYES CLASSIFICATION AND LOGISTIC REGRESSION

Here we use Gaussian Naïve Bayes and Logistic Regression to classify the Spam base data from UCI Machine Learning repository.

Part I – Classification with naïve Bayes

We split the data into training set and test set, each having about 2300 examples. Both Training and test set has about 40 percent of spam instances and 60 percent of non-spam instances, to reflect the statistics of the full data set. Here the target 0 is used to represent not-spam data and 1 is used to represent spam data. We then compute the prior probability for each class, 1 and 0 in the training data. For each 57 features we compute mean and standard deviation in the training set of the values given each class. Gaussian naïve Bayes algorithm is used to classify the instances in the test set. It is done using mean and standard deviation of each class:

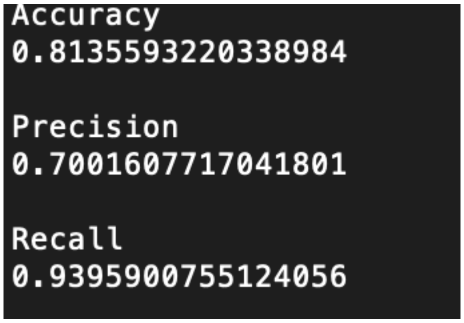
$$class_{NB}(\mathbf{x}) = \underset{class \in \{+1,-1\}}{\operatorname{argmax}} \; P(class)\prod_i P(x_i \mid class)$$

$$P(x_i \mid c) = N(x_i; \boldsymbol{\mu}_{i,c}, \boldsymbol{\sigma}_{i,c})$$

where

$$N(x;\boldsymbol{\mu},\boldsymbol{\sigma}) = \frac{1}{\sqrt{2\boldsymbol{\pi}}\boldsymbol{\sigma}} e^{-\frac{(x-\boldsymbol{\mu})^2}{2\boldsymbol{\sigma}^2}}$$

Accuracy, Precision and Recall on the test set:



```
Accuracy
0.8135593220338984

Precision
0.7001607717041801

Recall
0.9395900755124056
```

We see that the precision is low for the dataset using this gaussian algorithm.
Compared to Support vector machine, we see Gaussian naïve bayes algorithm shows less accuracy range. Accuracy of the dataset using svm with linear kernel is 0.933 whereas accuracy using gaussian naïve bayes is 0.813, which is low compared to svm.
The precision value is also low in gaussian naïve bayes, in svm we could see 0.914.
Only the recall value remains little higher in gaussian naïve bayes than the one in svm.

Confusion matrix for the test set:

```
[[1001  373]
 [  56  871]]
```

Almost 373 of the 0’s are predicted as 1’s, which is more compared to the actual 1’s predicted as 0’s. Highest predicted class is not-spam class.

All the attributes in the spam base dataset are independent of each other and can provide better separability over many kinds of spam
No, Naïve bayes does not do well in this problem compared to svm and logistic regression because the accuracy range is only around 80 percent.
Feature selection would help to get more better accuracy.
Another common technique for handling continuous values is to use binning to discretize the feature values.
It would be good to consider dimensionality reduction to make sure we have most, though not all, features to have similar distribution.

Part II – Classification with Logistic Regression

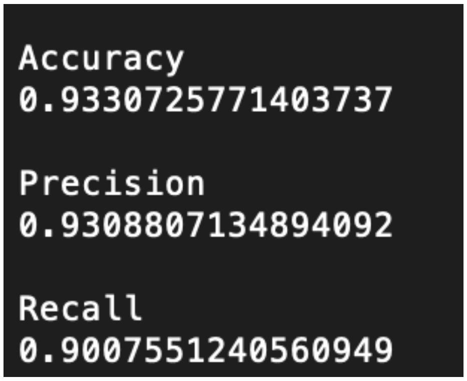
I used Scikit learn library to implement Logistic regression.

from sklearn.linear_model import LogisticRegression
lr=LogisticRegression()

Default parameters values used in running logistic regression are:

Parameters: {'warm_start': False, 'C': 1.0, 'n_jobs': None, 'verbose': 0, 'intercept_scaling': 1, 'fit_intercept': True, 'max_iter': 100, 'penalty': 'l2', 'multi_class': 'warn', 'random_state': None, 'dual': False, 'tol': 0.0001, 'solver': 'warn', 'class_weight': None}

Accuracy, Precision and recall values of the learned model on the test set:



```
Accuracy
0.9330725771403737

Precision
0.9308807134894092

Recall
0.9007551240560949
```

Accuracy and precision of this learned model is high compared to that of gaussian naïve bayes model.

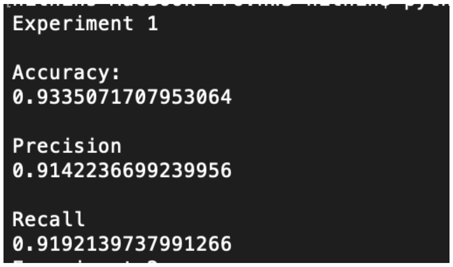
Confusion matrix for the test set:

```
[[1312   62]
 [  92  835]]
```

Almost 92 of the 1’s are predicted as 0’s, which is more compared to the actual 0’s predicted as 1’s. Highest predicted class is not-spam class for both logistic regression and gaussian naïve bayes.

Comparision:

Support vector machine results:



```
Experiment 1

Accuracy:
0.9335071707953064

Precision
0.9142236699239956

Recall
0.9192139737991266
```

From the results of Logistic regression, gaussian naïve bayes and support vector machine, we see Accuracy of support vector machine is the highest, Precision of Logistic regression model is the highest and Recall of the naïve bayes is the highest.
As a result, we could say Logistic regression is best compared to naïve bayes considering the performance and confusion matrix.