Machine Leaning Assignment-2

Lalith Satya Srinivas Kaladi
2017CS10340

March 2020

1 Part 1

1.1 Question A

Implemented Naive-Bayes with Laplace smoothening on the training data provided and able to achieve accuracy of 79.66 percent on test dataset.

1.2 Question B

The test accuracies obtained by various prediction methods are mentioned in the table below.

Prediction Type	**Accuracy**
Naive Bayes	79.66
Random Prediction	48.19
Majority Prediction	50.70

1.3 Question C

The confusion matrix with rows representing to the predicted value (first row representing 0, while the other representing 4). And columns representing the original values.

$$\text{Confusion Matrix} = \begin{bmatrix} 144 & 40 \\ 33 & 142 \end{bmatrix}$$

As it can be seen that there are more misses in case of negative tweets in the data as there are 40 tweets that are predicted positive when they are not, where as that count for positive tweets is 33. And there are more correct predictions when the tweet is positive than when negative.

1.4 Question D

The data is now trained after removing stop-words, usernames and stemming. The accuracy is now measured by applying this trained model on the test data. The accuracy that's achieved was 81.38 percent. This increase in percentage is expected as stop-words will not contribute to predicting the tweet, similarly stemming helps in grouping similar words which differ by the suffixes.

1.5 Question E

1.5.1 Bi-gram

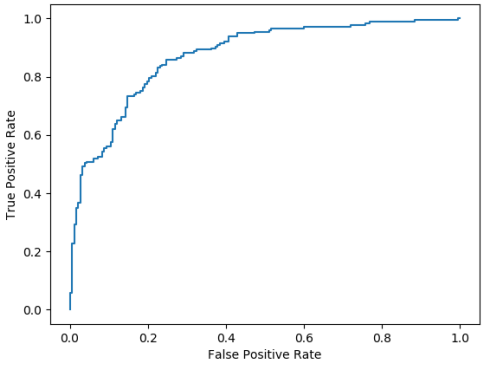
Bi-gram technique involves pairing up two consecutive words in the tweet and using them for prediction. The accuracy achieved by using this technique 83.79 percent. The accuracy has increased.

1.6 Question F

- The accuracy obtained by using sklearn vectorizer, by selecting random 100,000 vocabulary (because od memory issues) and running GaussianNB (partially fitting by dividing dataset into batches of size 10,000) on it is 51 percent.
- The accuracy obtained by using sklearn vectorizer, by selecting top 0.5 percentile words as vocabulary (because od memory issues) and running GaussianNB (partially fitting by dividing dataset into batches of size 10,000) on it is 65 percent.

1.7 Question G

Following is the ROC curve obtained.



2 Part 2

2.1 Binary Classification

My Entry-Number ends with 0, So my data contains images belonging to article classes 0, 1.

2.1.1 Question A

Using CVXOPT library to solve the QP optimisation problem with Linear kernel for the given data and measured the accuracies on test, validation. The accuracies were 98.2, 97.4 percent respectively.

2.1.2 Question B

Using CVXOPT library to solve the QP optimisation problem with Gaussian kernel for the given data with C = 1.0, γ= 0.05. The accuracies measured on test, validation datasets are 99.3, 97.8 percentage respectively.
For an $\xi = 10^{-4}$ 825 support vectors were obtained.
Comparing the accuracies observed through this model with the Linear kernel, the Gaussian model is able to achieve better accuracies with both test, validation datasets.

2.1.3 Question C

1. The accuracies Obtained with Scikit SVM packages are mentioned below

Prediction Type	**Dataset**	**Accuracy**
Linear	Test	98.9
	Validation	98.6
Gaussian	Test	99.3
	Validation	97.8

2. *b, nSV's* obtained from the trained models using Linear, Gaussian kernel are given in the table below.

Prediction Type	**Parameter**	**Value**
Linear	b	-1.63
	nSV	198
Gaussian	b	8.23×10^{-8}
	nSV	825

3. Time taken for Gaussian in sore than time taken for the linear kernel. It took 121 seconds in the case of gaussian kernel where as the linear kernel took just 70 seconds. This observation is expected as calculation of gaussian kernel takes more time (because of the dot product involved). But this time is worth spending as it resulted in a greater accuracy (we have seen above).

2.2 Multi-Class Classification

2.2.1 Question A

This question involves training the data model using CVXOPT, making $\binom{k}{2}$ classifiers using gaussian kernels with C = 1, $\gamma = 0.05$. It took around 3.5 hours to train the model and compute required parameters and an accuracy of 65.3, 66.4 was observed with training, validation datasets respectively.

2.2.2 Question B

This involves prediction using Scikit's SVM package. The accuracies observed with test, validation datasets are 88.08, 87.92 respectively. And the time taken to train using this package is far less than that of the first part mentioned above. It took around 20 minutes for this model.

2.2.3 Question C

All the confusion matrices has rows as real and columns as predicted.

- The confusion matrix for test data with CVXOPT is given below.

$$\begin{bmatrix} 21 & 0 & 8 & 12 & 0 & 0 & 454 & 0 & 5 & 0 \\ 1 & 444 & 6 & 2 & 0 & 0 & 47 & 0 & 0 & 0 \\ 2 & 0 & 351 & 2 & 5 & 0 & 134 & 0 & 6 & 0 \\ 1 & 12 & 2 & 279 & 1 & 0 & 204 & 0 & 1 & 0 \\ 0 & 1 & 93 & 11 & 76 & 0 & 317 & 0 & 2 & 0 \\ 0 & 0 & 0 & 0 & 0 & 477 & 1 & 7 & 2 & 13 \\ 4 & 1 & 50 & 4 & 0 & 0 & 434 & 0 & 7 & 0 \\ 0 & 0 & 0 & 0 & 4 & 0 & 0 & 483 & 10 & 3 \\ 8 & 0 & 8 & 1 & 0 & 15 & 35 & 0 & 439 & 2 \\ 0 & 0 & 0 & 0 & 0 & 5 & 0 & 6 & 1 & 488 \end{bmatrix}$$

By observing the table above it can be seen that the class 6 was predicted more often when it is not the real class. This might have happened because of the similarities between those classes. Class 6 belongs to images of shirts and classes 0, 3, 4 belongs to classes T-shirt, dress, coat respectively. As we can see that they have similarities with shirt class this can be explained.

- The confusion matrix for validation data with CVXOPT is given below.

$$\begin{bmatrix} 13 & 2 & 1 & 6 & 0 & 1 & 226 & 0 & 1 & 0 \\ 1 & 240 & 2 & 1 & 0 & 0 & 5 & 0 & 1 & 0 \\ 1 & 0 & 194 & 0 & 3 & 0 & 48 & 0 & 4 & 0 \\ 0 & 7 & 0 & 157 & 2 & 0 & 82 & 0 & 2 & 0 \\ 1 & 2 & 51 & 3 & 45 & 0 & 147 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 & 0 & 239 & 2 & 1 & 0 & 7 \\ 2 & 0 & 29 & 0 & 0 & 0 & 219 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 29 & 0 & 199 & 0 & 22 \\ 3 & 0 & 2 & 0 & 0 & 5 & 12 & 2 & 226 & 0 \\ 0 & 0 & 0 & 0 & 0 & 5 & 2 & 4 & 1 & 238 \end{bmatrix}$$

- The confusion matrix for test data with SVM is given below.

$$\begin{bmatrix} 433 & 0 & 1 & 8 & 0 & 0 & 26 & 0 & 3 & 0 \\ 0 & 482 & 3 & 7 & 0 & 0 & 2 & 0 & 1 & 0 \\ 5 & 0 & 411 & 3 & 18 & 0 & 13 & 0 & 5 & 0 \\ 6 & 0 & 0 & 457 & 8 & 0 & 9 & 0 & 1 & 0 \\ 1 & 1 & 24 & 8 & 399 & 0 & 15 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 & 0 & 473 & 0 & 2 & 1 & 5 \\ 34 & 0 & 28 & 3 & 19 & 0 & 315 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 8 & 2 & 471 & 1 & 11 \\ 0 & 0 & 1 & 1 & 1 & 0 & 1 & 2 & 489 & 0 \\ 0 & 0 & 0 & 0 & 0 & 6 & 0 & 8 & 1 & 474 \end{bmatrix}$$

  The explanation here for the confusion matrix is similar to that of the above.

- The confusion matrix for validation data with SVM is given below.

$$\begin{bmatrix} 212 & 0 & 1 & 8 & 0 & 0 & 26 & 0 & 3 & 0 \\ 0 & 237 & 3 & 7 & 0 & 0 & 2 & 0 & 1 & 0 \\ 5 & 0 & 206 & 3 & 18 & 0 & 13 & 0 & 5 & 0 \\ 6 & 0 & 0 & 228 & 8 & 0 & 9 & 0 & 1 & 0 \\ 1 & 1 & 24 & 8 & 200 & 0 & 15 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 & 0 & 241 & 0 & 2 & 1 & 5 \\ 34 & 0 & 28 & 3 & 19 & 0 & 165 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 8 & 2 & 230 & 1 & 11 \\ 0 & 0 & 1 & 1 & 1 & 0 & 1 & 2 & 244 & 0 \\ 0 & 0 & 0 & 0 & 0 & 6 & 0 & 8 & 1 & 235 \end{bmatrix}$$

2.2.4 Question D

C values and corresponding accuracies are mentioned in the table below.

C	Accuracy within	Test Accuracy
10^{-5}	11.53	57.36
10^{-3}	9.32	56.13
1	87.86	88.02
5	88.2	88.10
10	88.9	88.24

The accuracy in both the cases peaked at C = 10. And the plot with logarithmic X-axis scaling is shown below.

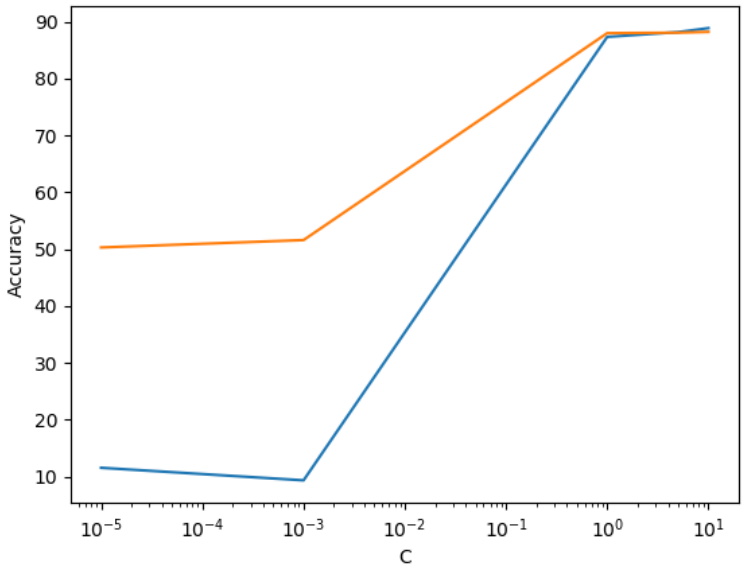


Figure 1: Accuracies with varying C