Assignment 2
Machine Learning
Sagar Goyal
2015cs10253

Q1: Text Classification using Naïve Bayes

(a). As the text file that was supplied contained a lot of noise and irrelevant data, only alphanumeric terms were separated from the text and separated by any other character. This was done using regex code.
Finally, a vocabulary was made that contained unique words in the text. These words were stored in a file and a dictionary was made from that, which stored a index value associated with the word.

Accuracy over training data: 68.836 %
Accuracy over test data: 38.672 %

(b). Test accuracy by randomly guessing: 9.9 %
Test accuracy by majority guessing: 20.0 %

The improvement obtained by our algorithm is 28.7 % over random baseline and 18.6 % over majority guessing baseline.

(c). **The confusion matrix** is as follows:

```
[[4363 1647 1423 1085    0    0  409  435  327  773]
 [  49   33   43   40    0    0    8    8    2   10]
 [ 131  152  193  186    0    0   65   52   23   34]
 [ 234  268  478  649    0    0  261  162   86   98]
 [   0    0    0    0    0    0    0    0    0    0]
 [   0    0    0    0    0    0    0    0    0    0]
 [  35   49  114  204    0    0  364  292  128  148]
 [  57   51  119  238    0    0  534  692  451  516]
 [  12    4   11   17    0    0   48   91   74  120]
 [ 141   98  160  216    0    0  618 1118 1253 3300]]
```

Category 1 has the highest value of the diagonal entry and the value is : 4363. This means that the most number of correct predictions were made of the category 1 and also means that the model learnt is better trained to detect category 1.

In a lot of cases the model predicted the class to be 10 even when the actual classes were 7, 8 and 9 and similarly the model also predicted the class to be 1 a lot of times even when the correct answers were 2, 3 and 4. Thus the weightage given to these 2 classes are clearly more in the features learnt than the other classes. Also there is no data of the classes 5 & 6.

(d). In this, both the test data and the train data was processed using the script given. A new vocabulary was made and the entire algorithm was run again.

The accuracy obtained was: 38.32 %

There is a very slight decrease in the accuracy obtained as compared to when we did not use stemming and stop word removal. But decrease cannot be accounted for and is mainly for the reason that the data has a lot of noise and extra words that cannot describe the quality of a review.

(e). The new features added was Bigram in one case, the test accuracy obtained for that was 38.58 % which was slight improvement from earlier data.

Then another feature using POS tags was modelled over the bigrams and accuracy obtained= 39.13 %

Q2. MNIST Handwritten digit classification

(a). and (b).
The Pegasos algorithm was implemented for the data with, Batch size=100 and the number of iterations to be set to 10000

Test accuracy obtained: 89.67 %
Training accuracy obtained: 88.71 %

(c).
Linear Kernel Test Accuracy :- 92.78
Gaussian Kernel Test Accuracy :- 97.23

There was a difference of almost 2% between our model and the model predicted by Linear SVM model of LIBSVM.

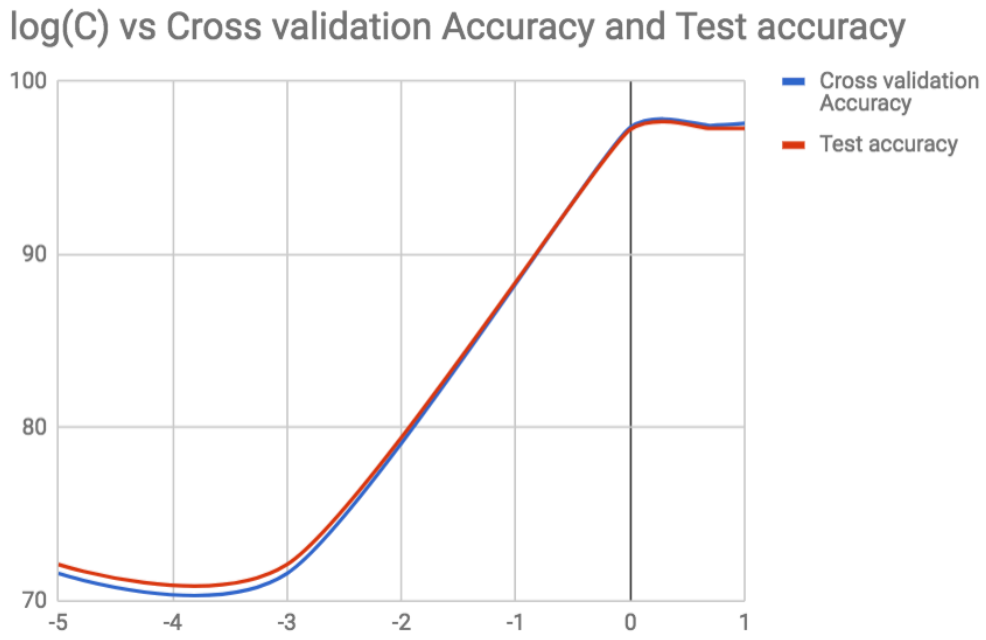
D.

Value of C	Cross-Validation Accuracy	Test Accuracy

0.00001	71.59	72.11
0.001	71.59	72.11
1	97.355	97.23
5	97.455	97.29
10	97.575	97.29

Best Cross-Validation Accuracy is obtained for C = 10.
C = 5 and C = 10 both give the highest test accuracy.

We can clearly see that on increasing the value of C the cross-validation Accuracy increases and so does the Test Accuracy.



(d). This is the confusion matrix obtained when predicting through SVM

Actual values on top, Predicted values on left of the confusion matrix

	0	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
0	969	0	4	0	1	2	5	1	4	4
1	0	1122	0	0	0	0	4	4	0	4
2	1	3	1000	8	4	3	0	20	3	3
3	0	2	4	985	0	6	0	2	10	8
4	0	1	2	0	962	1	3	3	1	9
5	3	2	0	4	0	866	4	0	5	4
6	4	2	1	0	5	7	940	0	3	0
7	1	0	6	7	0	1	0	986	3	9
8	2	2	15	5	2	5	2	2	942	11
9	0	1	0	1	8	1	0	10	3	957

The class 9 is the most difficult to predict as a lot of entries that were 9, were not predicted correctly which can be observed by the large numbers in the final column of 9.

Misclassified examples:

Actual: 9, Predicted: 8

Actual:7, predicted: 4

Actual:3, predicted: 5