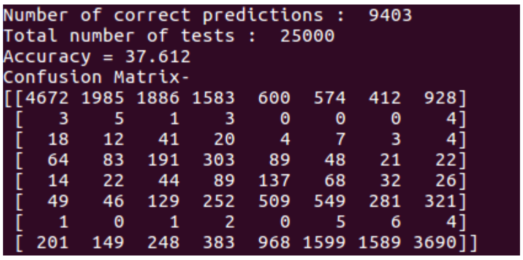
<u>COL-774: Assignment-2</u>
<u>Manikaran Kathuria</u>
<u>2017MCS2077</u>

Q1. Text Classification

a) The reviews are classified into one of the 8 categories. The accuracy of the Naive Bayes algorithm on the training data is found to be 68.876 % and on test data it is found to be 37.612 %.

b) The accuracy obtained using the random prediction = 12.528%
If most occured class in training set is used for prediction i.e., majority prediction, the accuracy achieved = 20.088%
Our Naive Bayes algorithm gives 37.612% accuracy on the test data which improves the accuracy by three fold as compared to random prediction.

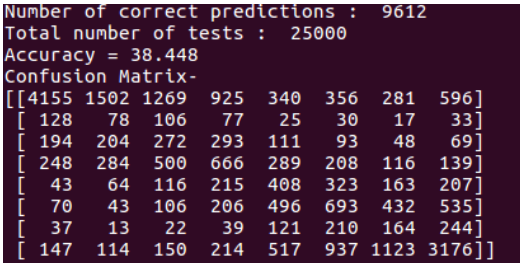
c)



```
Number of correct predictions :  9403
Total number of tests :  25000
Accuracy = 37.612
Confusion Matrix-
[[4672 1985 1886 1583  600  574  412  928]
 [   3    5    1    3    0    0    0    4]
 [  18   12   41   20    4    7    3    4]
 [  64   83  191  303   89   48   21   22]
 [  14   22   44   89  137   68   32   26]
 [  49   46  129  252  509  549  281  321]
 [   1    0    1    2    0    5    6    4]
 [ 201  149  248  383  968 1599 1589 3690]]
```

For class = 1, the diagonal entry in the confusion matrix is found to be highest. It means that the classifier has predicted movie reviews in the class 1 most correctly. The classifier had to classify reviews in one of the 8 classes. One of the reasons of getting low accuracy is that the words used in lower ratings are relatively same. Therefore, it can be observed from the first row of the confusion matrix that our classifier has predicted class = 1 for 1985 class 2 reviews, 1886 class 3 reviews and 1583 class 4 reviews which constitutes 5454 wrong predictions. It can also be observed from the last 4 enteries of the first row that the number of reviews of classes 7,8,9 or 10 predicted in class 1 are small in number.
Class label = 10 has got the second highest value in the diagonal. The behaviour similar to class 1 predictions can also be observed here. Number of reviews predicted to be in classes 1,2,3 or 4 are small in number.
The number of reviews predicted to be in class 2 or class 9 are very small in number. This result can be supported by the fact that the words used in reviews belonging to class 2 are very similar to those in class 1 and similarly, the words used in reviews belonging to class 9 are very similar to those in class 10.

d)



```
Number of correct predictions :  9612
Total number of tests :  25000
Accuracy = 38.448
Confusion Matrix-
[[4155 1502 1269  925  340  356  281  596]
 [ 128   78  106   77   25   30   17   33]
 [ 194  204  272  293  111   93   48   69]
 [ 248  284  500  666  289  208  116  139]
 [  43   64  116  215  408  323  163  207]
 [  70   43  106  206  496  693  432  535]
 [  37   13   22   39  121  210  164  244]
 [ 147  114  150  214  517  937 1123 3176]]
```

Though, the number of correct predictions corresponding to class 1 and class 10 has reduced, the accuracy of the classifier after stemming is increased to 38.448%. There is a very slight increase in the accuracy of the classifier. Stemming tries to cut off details like exact form of a word and produce word bases as features for classification. The same words in different forms of verb are treated different by the normal classifier. After stemming, the words in different verbs form are reduced to their base form because of which overall accuracy of the classification has increased.

e)
I have tried bigram, bigram + stemming, bigram + unigram and bigram+unigram stemming as feature. It was found that bigram + unigram stemming resulted in highest accuracy (40.596 %) on test set. Use of bigram along with the unigram increases the test accuracy because it fetches the context better. It is better than the test accuracy obtained in part (a) and part (d). I have eliminated unigrams and bigrams whose count is less than 7 and 6 respectively. This resulted in lesser number of features and also the accuracy of the classifiers has increased.

Q2. MNIST Handwritten digit Classification

b) In Pegasos algorithm, we randomly permute training data before the start of every epoch. So, while training the classifier, I have run the algorithm multiple times and used the parameters giving the maximum accuracy over the test data. The accuracy of the training data was found to be 95.78% and on the test data, the classifier achieved 92.73% accuracy.

c)
The test set accuracy of Linear Kernel = 92.76%
The test set accuracy of Gaussian Kernel =
The accuracy achieved using Pegasos algorithm is similar to the test set accuracy of the Linear Kernel. In Pegasos algorithm, we trained 45 one vs one classifers. An example x belongs to class 1 if $w^Tx + b > 0$ otherwise, x is predicted to be in class -1. So, the decision boundary here is linear. Similarly, in Linear Kernel the decision boundary is also linear. Therefore, the difference in accuracies achieved in both the cases is just 0.03%.

d)

Value of C	Average Validation Set Accuracy
0.00001	71.555 %
0.001	71.54 %
1	97.4 %
5	97.535 %
10	97.46 %

Table 1. The average validation set accuracy corresponding to different values of C

Value of C	Test Set Accuracy
0.00001	72.1 %
0.001	72.1 %
1	97.23 %
5	97.29 %
10	97.29 %

Table 2. The test set accuracy corresponding to different values of C

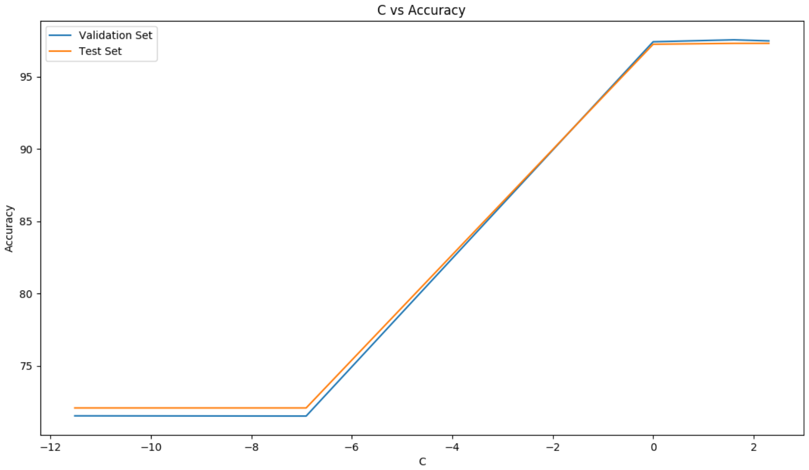


Fig. Plot of C vs Accuracies

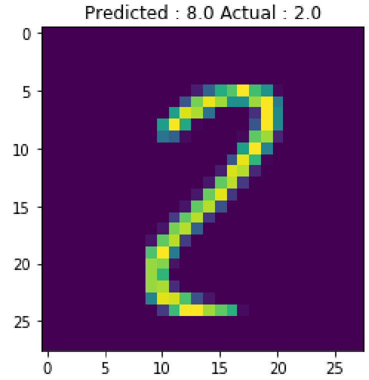
C = 5 gives the best value of average accuracy on validation set. It also gives the best accuracy on test set. C = 10 gives the same accuracy on test set but lesser average accuracy on validation set. It has been observed that the average validation set accuracy is near to the test set accuracy. Cross validation is a model evaluation method. In the assignment, we have used K-Cross validation for K = 10. The training set was divided into 10 folds and training is done on 9 folds whereas 1 fold is used for validating the model. Each fold acts as the validation set once in 10 iteration and then, the average accuracy is reported. This helps in finding the best value of the parameters. Since, the examples in 1 fold were unseen to classifiers, so the average validation accuracy is near to the test set accuracy for all values of C.

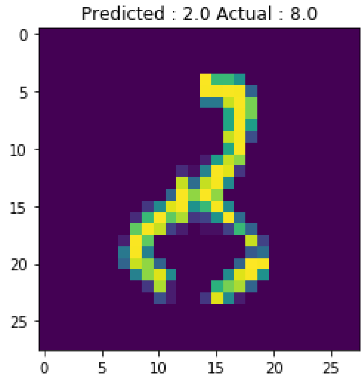
e)

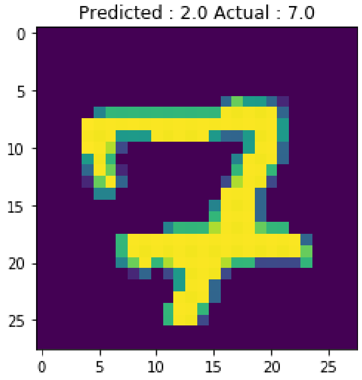
Confusion Matrix

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

From the confusion matrix, it is clear that most of the wrong predictions occurred in classification of 7 and 2. 15 hand-written 2s were classified as 8 and 20 hand-written 7s were classified as 2. The reason for this misclassification could be the shape of these digits.
Images of a few misclassified examples:


Predicted : 8.0 Actual : 2.0


Predicted : 2.0 Actual : 8.0


Predicted : 2.0 Actual : 7.0