COL-774
Assignment 2

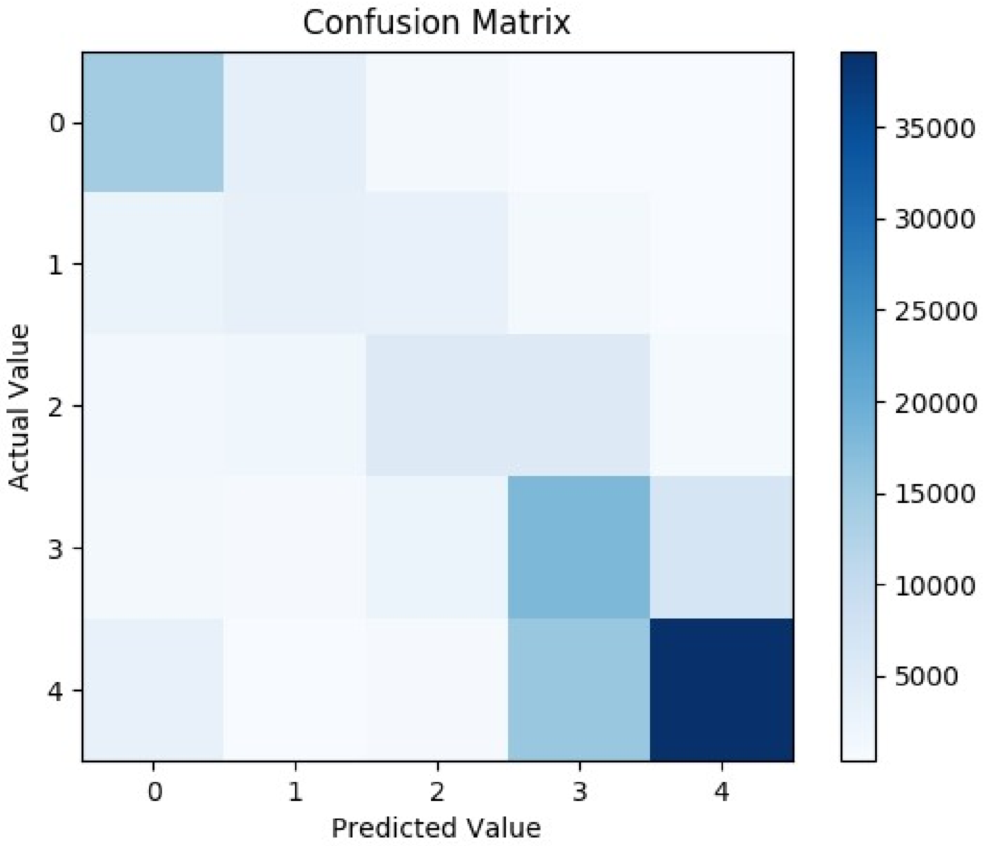
-Shantanu Verma
(2016CS10373)

Q1. a)

Accuracy over train set = 62.94664891787194 %
Accuracy over test set = 59.8902167247491 %

b) Accuracy for random prediction= 20.240356571291823 %
Accuracy for major prediction = 43.9895900327555 %

c)



Confusion Matrix

```
[[14266  3837  1122   473   471]
 [ 2905  3259  3055  1132   487]
 [ 1459  1709  5048  5182  1133]
 [ 1219   670  2624 17429  7416]
 [ 3305   338   660 15128 39391]]
```

From the matrix as well as figure we can very well see that 5 stars has the maximum diagonal entry.

This implies that 5 stars has been predicted correctly most of the times among rest stars.

Diagonal values are maximum in all the rows hence, correct ratings are predicted most of the times.

We also observe that 3 stars are predicted as 4 stars almost equal number of times as correctly predicted. Same is the case with 2 and 3 stars.

d) Accuracy with Stemming = 59.37345757489643 %

We observed that there is no significant difference in accuracies with stemming and without stemming but the time taken with stemming is much more as compared to raw data. Hence, stemming is not feasible.

e) Two features added are :

1) Lemmetization
2) Bigrams

Accuracy with added features = 61.485213454359 %

f) F1 score for best performing model =
[0.65858782 0.31562636 0.37337278 0.5073797 0.73135908]

Macro F1 score = 0.5172651473997585

Macro F1 score is better suited for multi-class prediction than accuracy because it calculate metrics for each label, and find their unweighted mean.

g) Accuracy with full training set = 62.55654436949401 %

F1 score =

[0.66466925 0.34053367 0.42280367 0.52484199 0.7390766]

Macro F1 score = 0.5383850378586144

Q2.

1)

a) The set of support vectors are provided in supportvectos.txt

Accuracy over training set = 100 %
Accuracy over test set = 99.49799196787149 %

b) Accuracy = 99.8995983935743 %

Since the accuracy was already quite high, with gaussian kernel accuracy is increased but the difference isn’t significant.

c)

Accuracy over training set = 99.498%
Accuracy over test set = 99.5482%

number of support vectors using LIBSVM for linear kernel = 134
number of support vectors using LIBSVM for gaussian kernel = 444

number of support vectors using LIBSVM for linear kernel = 144
number of support vectors using LIBSVM for gaussian kernel = 637

The computation time of LIBSVM is considerably fast (~8 seconds) while the computation time of our implementation is significantly high (~170 seconds).

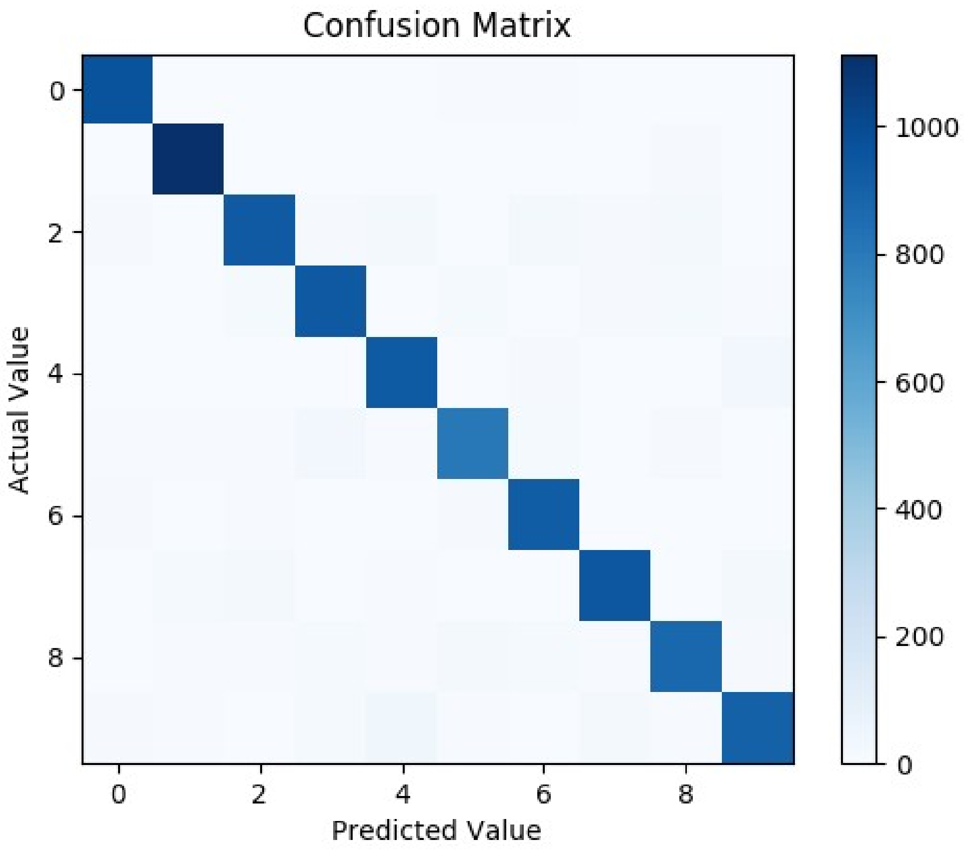
2)

a) Accuracy over training set = 98.34 %
Accuracy over test set = 97.65 %

b) Accuracy over training set using LIBSVM = 93.36 %
Accuracy over test set using LIBSVM= 93.25 %

The accuracy of my implementation of Multi-Class SVM is better but the time taken to train the model is significantly high. (4.8 hours for my implementation and 4-5 mins for LIBSVM) c)

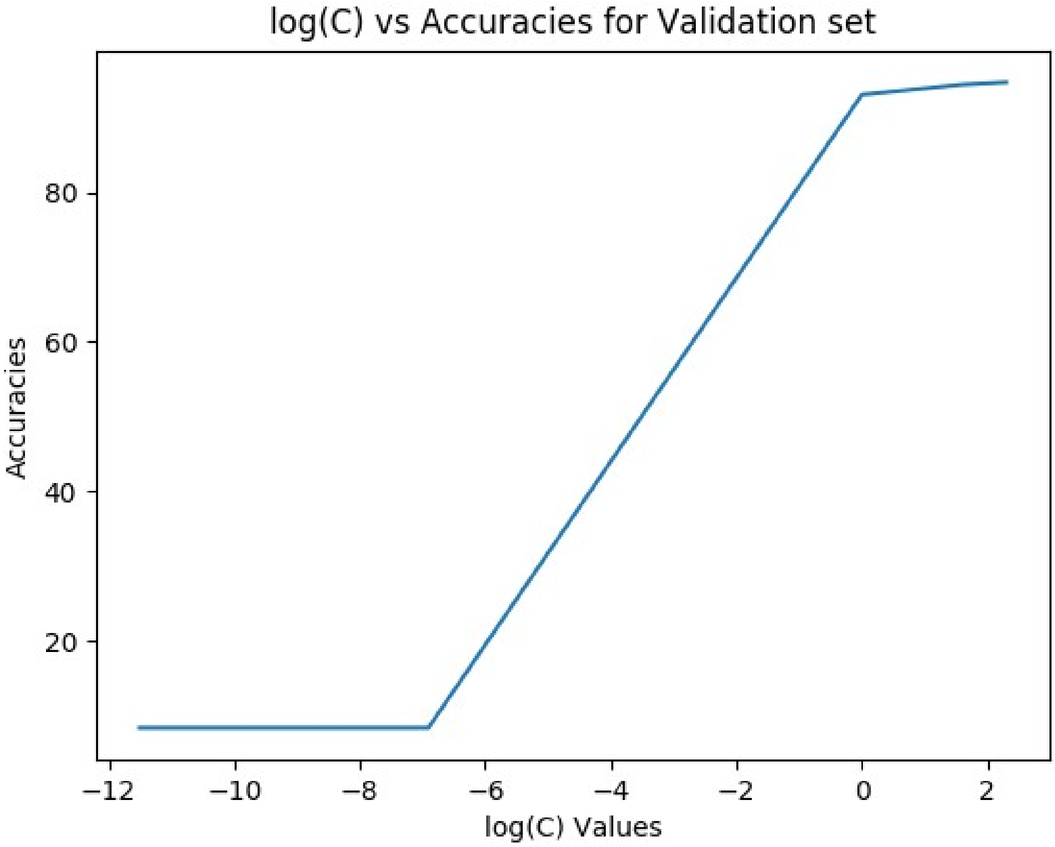
The confusion matrix is drawn here

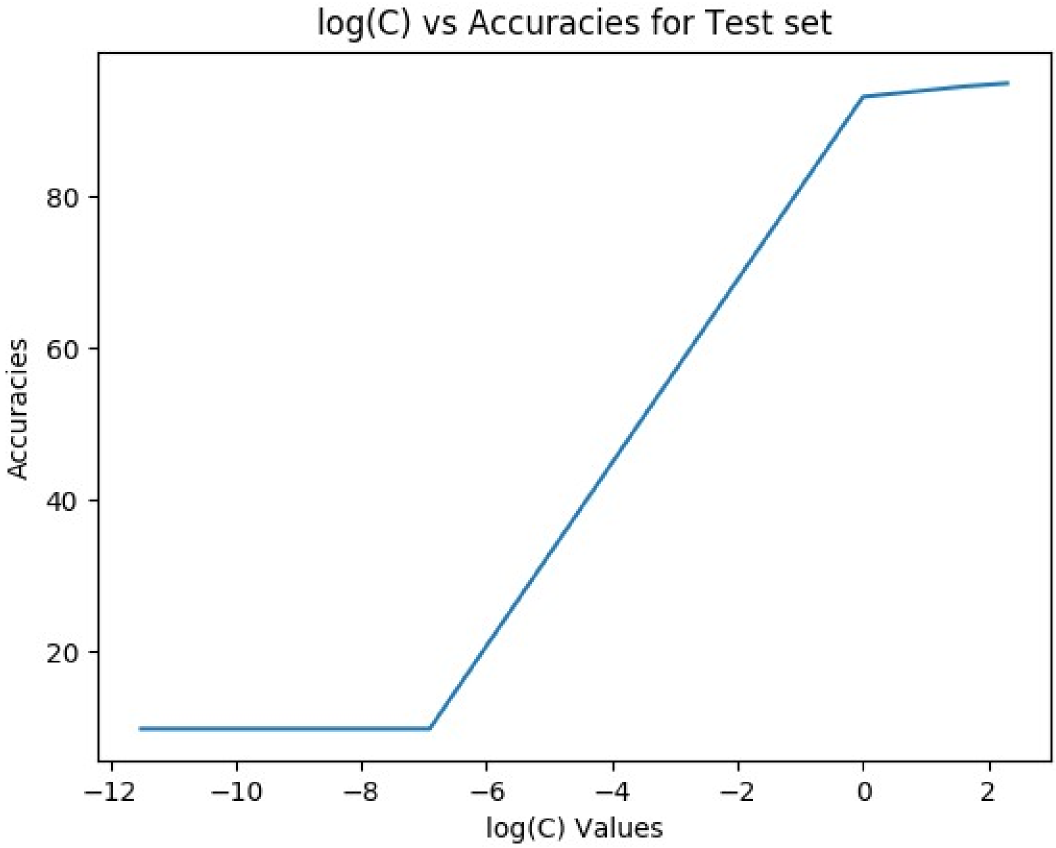


The most wrongly classified digit is 9 which is classified as 4 maximum number of times.

d)

Validation accuracies for validation set are
[8.35, 8.35, 93.05, 94.4, 94.7]
Validation accuracies for Test set
[9.8, 9.8, 93.18, 94.56, 94.96]





Hence the best suited value of C is 10 as the accuracy increase by increasing value of C.