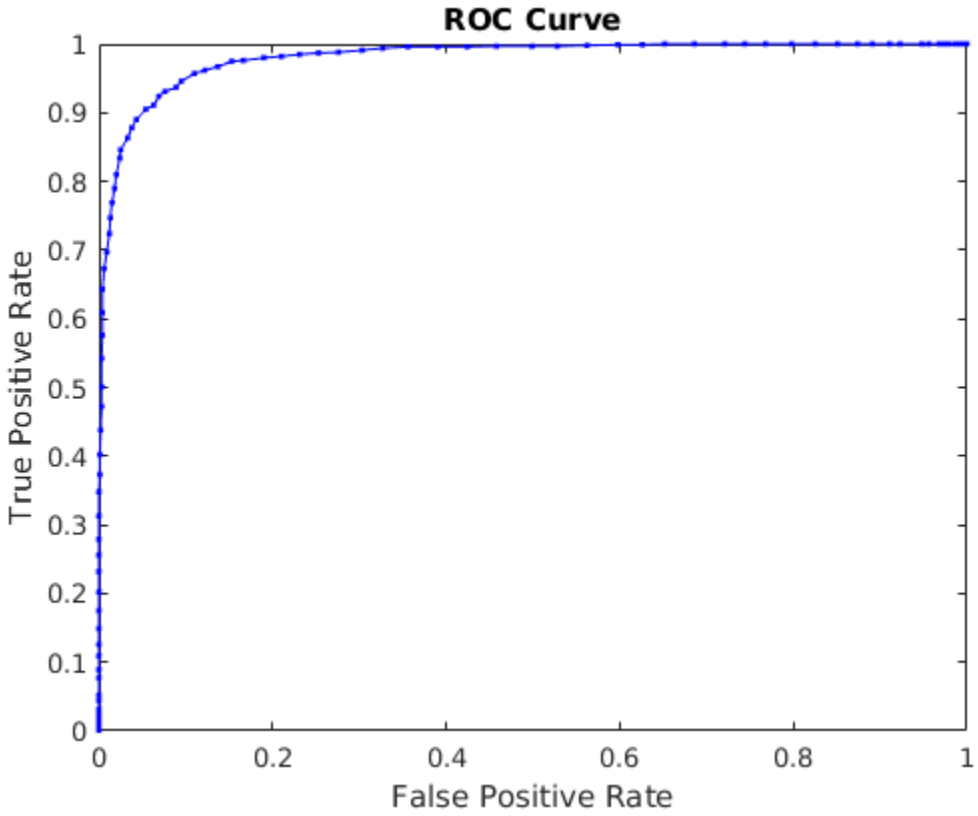
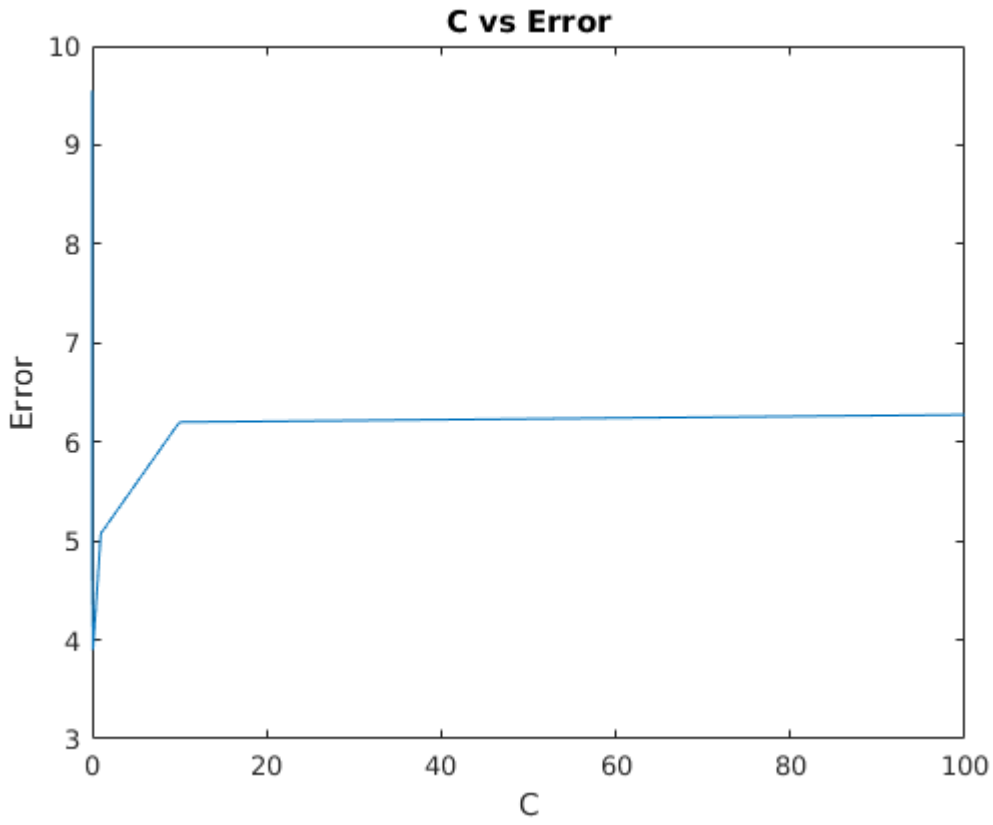
ML ASSIGNMENT 3 REPORT

Part 1: Linear SVM:
Best C = 0.1
Accuracy = 96.1 %



C vs Average Cross Validation:
Search space for C = [0.0001,0.001,0.1,1,10,100];

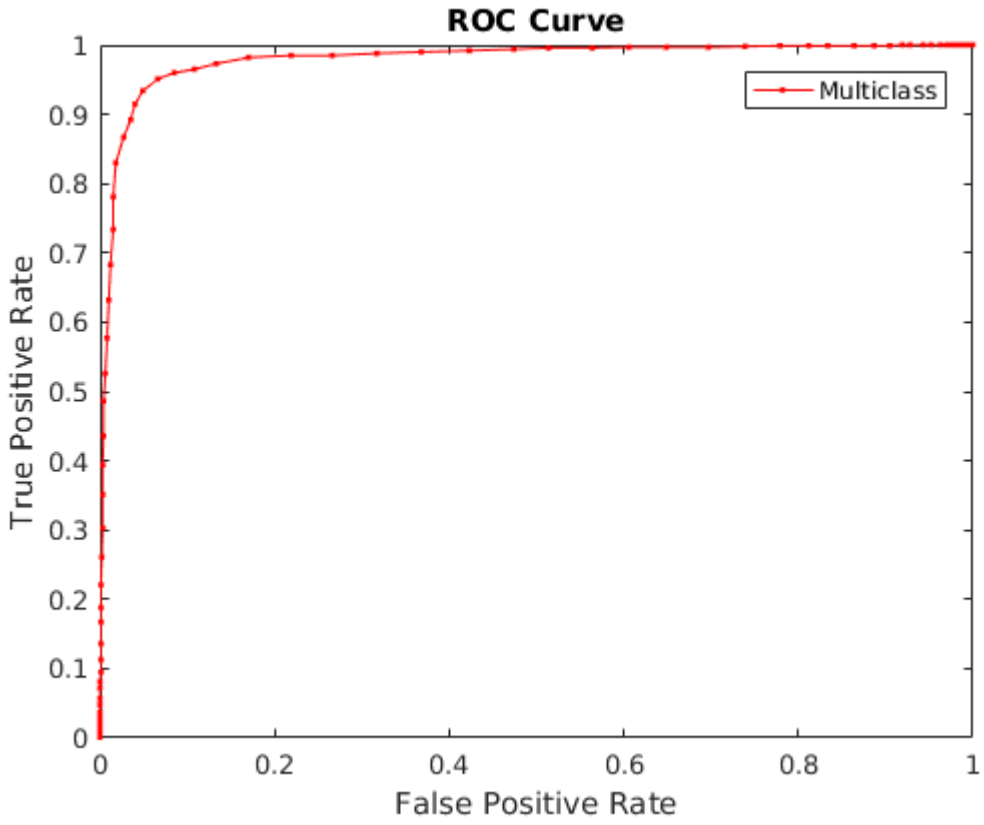
C	Accuracy(%)
0.0001	90.45
0.001	95.3
0.1	96.1
1	94.925
10	93.8
100	93.725



Steps performed:

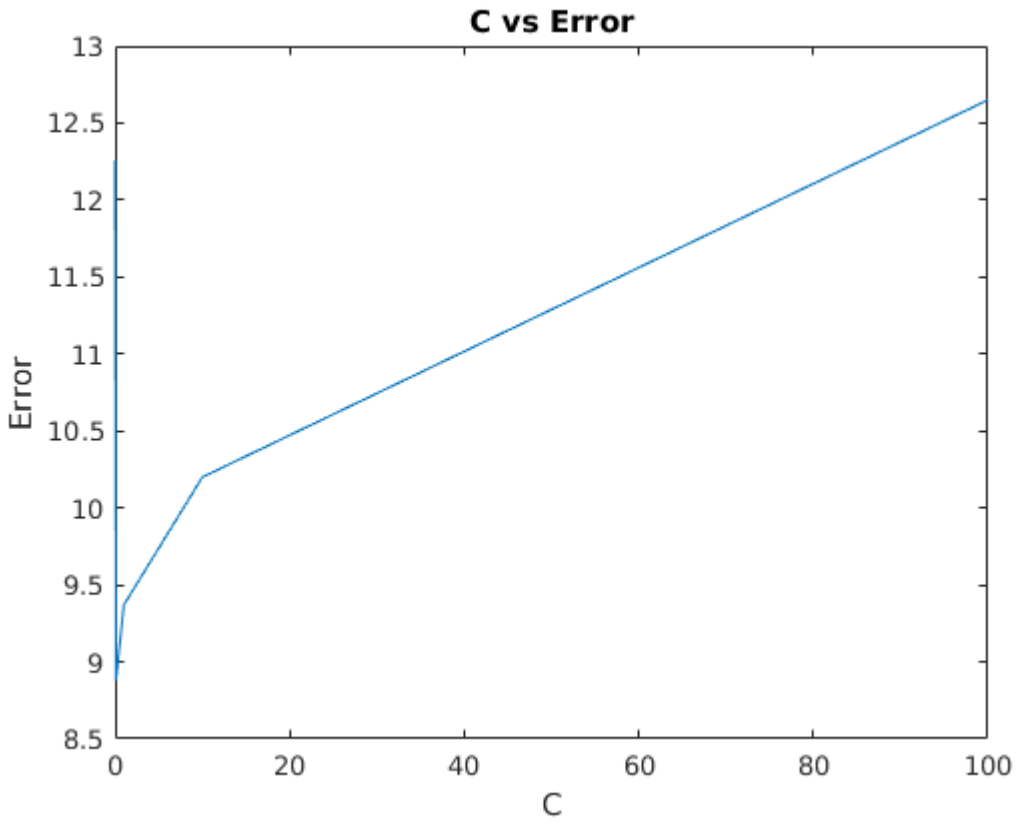
1. Filtering the training data to keep only labels 3 and 8.
2. Setting labels as 1 for 3 and -1 for 8.
3. Performing the grid search for best C.
4. Training using the entire training data.
5. Testing on the entire test data.
6. Plot ROC:
    a. Flipping labels
    b. Training again
    c. Testing again and obtain dec_values.
    d. Creating the scoreMatrix and trueLabels and plot ROC

Part 2: Multiclass SVM:
Best C = 0.1
Accuracy = 90.08%



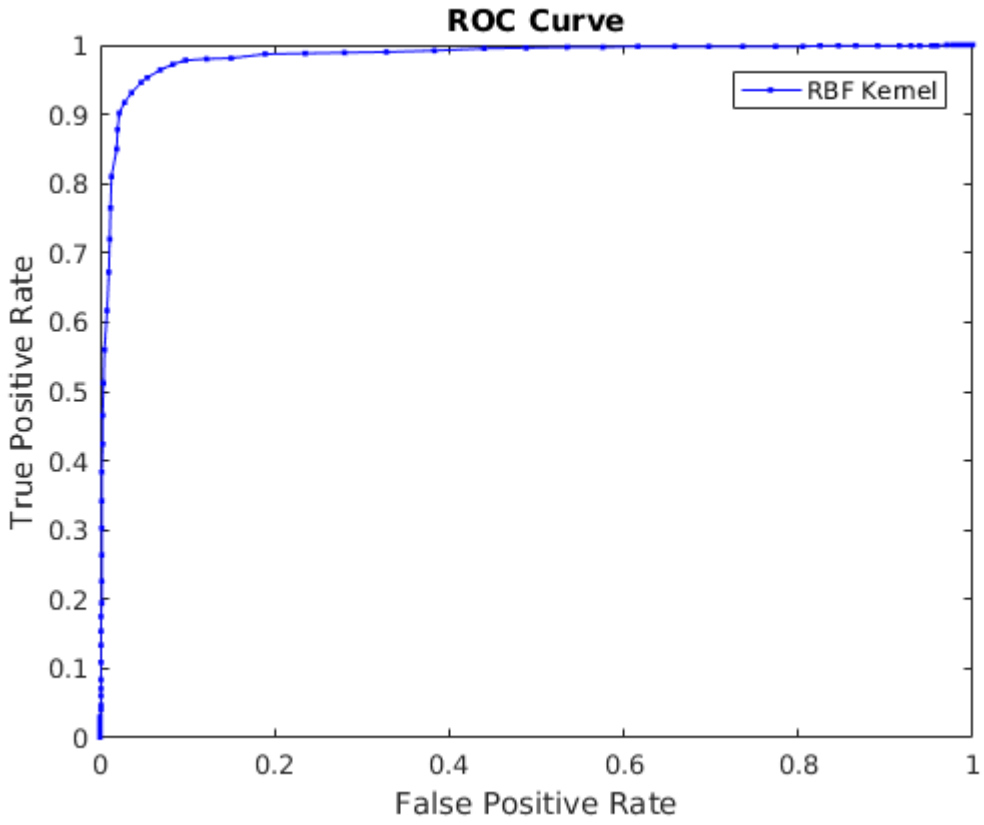
C vs Average Cross Validation:
Search space for C = [0.001,0.1,1,10,100];

C	Accuracy(%)
0.001	87.74
0.1	91.12
1	90.63
10	89.8
100	87.35



Steps performed:
1. Cross validation for each value of C.
2. Find out the best model by highest C.
3. Train using entire training data to obtain the model.
4. Use this model to test the entire training set and obtain dec_values.
5. Use dec_values and labels to create trueLabels and scoreMatrix and plot ROC curve.

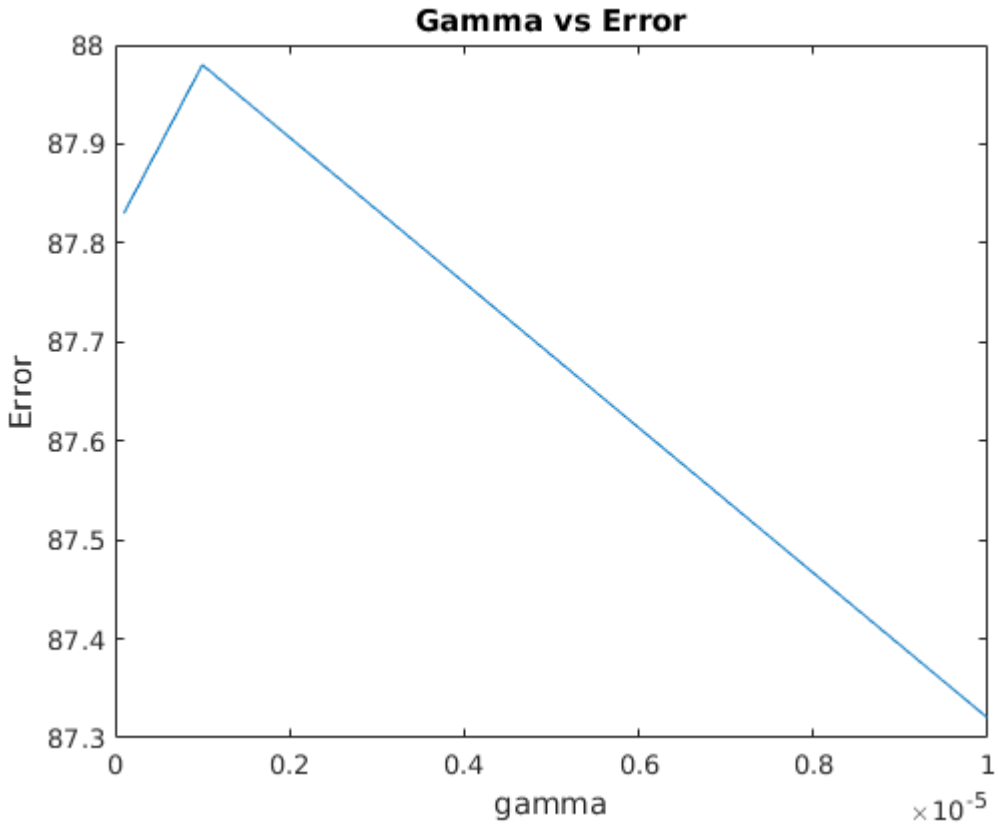
Part 3: RBF Kernel SVM:
Best C =10, gamma = 0.000001
Accuracy = 91.18%

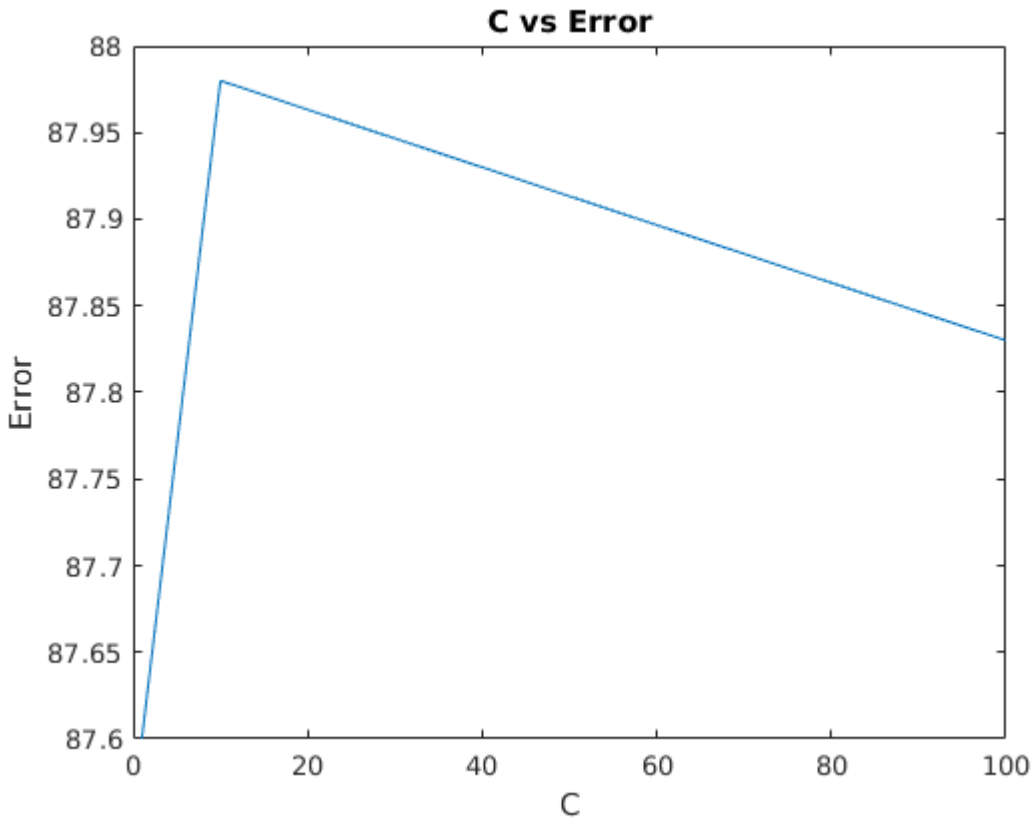


I could not complete cross validation on entire dataset since it was taking too long (I kept it running for 4 hours before quitting).

So to find out the value of gamma and C, I first ran the test on feasibly smaller dataset to get an idea of optimum C and gamma and then ran it on the larger dataset.

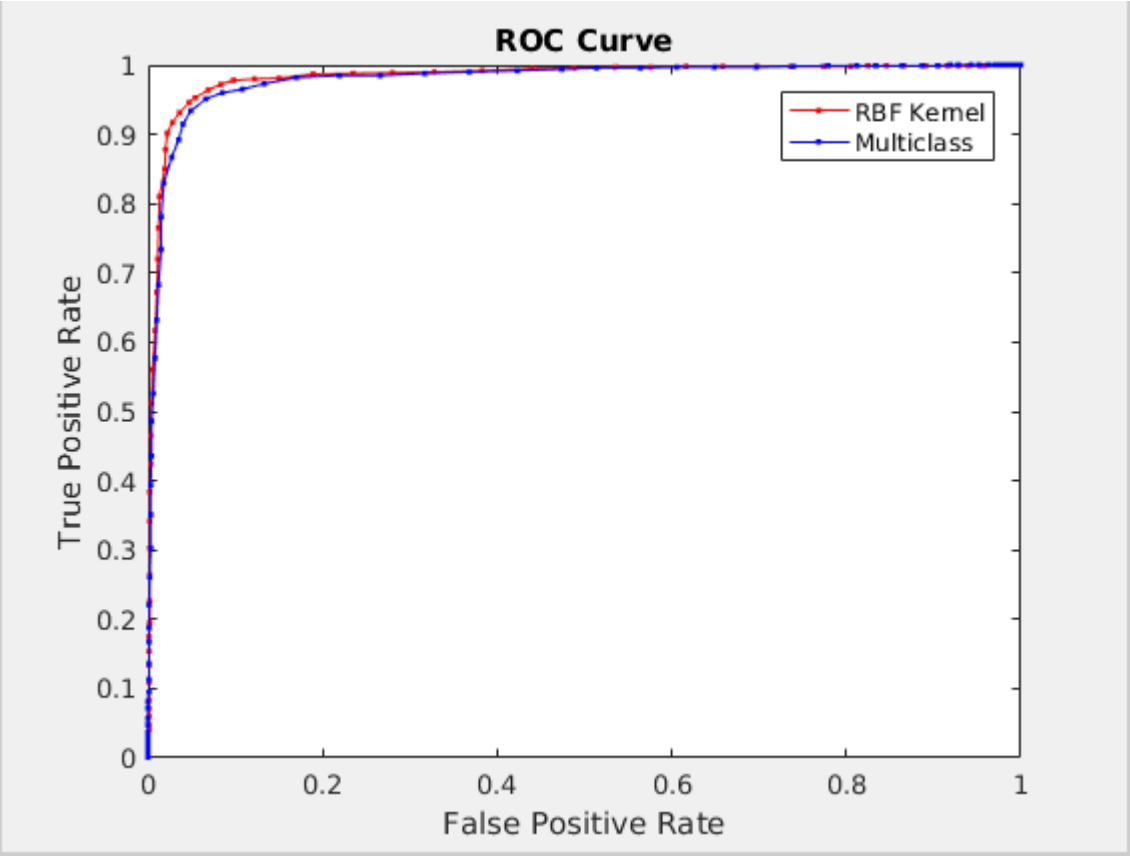
C	***gamma***	***Accuracy***
1	*0.0000001*	*85.45*
10	*0.0000001*	*86.14*
100	*0.0000001*	*87.83*
1	*0.000001*	*87.60*
10	*0.000001*	*87.98*
100	*0.000001*	*86.16*
1	*0.00001*	*84.22*
10	*0.00001*	*86.74*
100	*0.00001*	*87.32*


Gamma vs Error
88
87.9
87.8
87.7
87.6
87.5
87.4
87.3
Error
0
0.2
0.4
0.6
0.8
1
gamma
×10^-5



Steps performed:
1. Cross validation for each value of C and gamma.
2. Find out the best model by highest C and gamma.
3. Train using entire training data to obtain the model.
4. Use this model to test the entire training set and obtain dec_values.
5. Use dec_values and labels to create trueLabels and scoreMatrix and plot ROC curve.

ROC Curve of Multi class vs RBF Kernel:



So, we notice that accuracy(= area under curve) is better for RBF kernel.
This is because it transforms the data to a higher dimension where the data is separable. Also, it can instantiate the Gaussian at each of the training points for a better prediction,