

# Assignment-2

## ELL409

Submitted By:

Ritik Agrawal (2017EE10482)

### Part-A

Standard Library chosen-

sklearn (Uses one vs one method)

#### 1. Convex optimisation package (cvxopt) -

Labels chosen for binary classification- (0,1)

##### a) Linear svm-

i) Hyperparameter-  $C = 0.001$

ii) Accuracy using standard library- 99.58 %

iii) Accuracy using cvxopt- 99.58 %

iv) Threshold for Lagrange multipliers-  $1e-5$

v) Number of support vectors using cvxopt- 137

vi) Number of support vectors using standard library- 137

vii) Weight vector using cvxopt- [-0.29173095 0.05910209 -0.05906562

0.04191261 0.04250484 0.02141655 0.01635956 -0.0241487 0.00807477

-0.02572151 0.01983177 0.03253638 0.0517182 0.04262058 -0.00697337

0.03460874 -0.00750425 -0.04114124 -0.01700473 -0.01772738 0.0214405

-0.00430899 0.02441536 0.00455086 -0.00542265]

viii) Weight vector using standard library- [-0.29176996 0.05907868

-0.05905626 0.04190865 0.0425314 0.02141111 0.01636301 -0.0241671

0.00806411 -0.02570811 0.01984043 0.03250874 0.05169157 0.04261874

-0.00698916 0.03462023 -0.00749906 -0.04112123 -0.01697879 -0.01769426  
0.02145376 -0.00431263 0.02441905 0.00455932 -0.00541434]

ix) Intercept using cvxopt- -0.11075067231925902

x) Intercept using standard library- -0.05833678

## b) Non-linear svm (Rbf) -

i) Hyperparameters- C= 1, Gamma- 0.01

ii) Accuracy using standard library- 99.79 %

iii) Accuracy using cvxopt- 99.79 %

iv) Threshold for Lagrange multipliers-  $1e-5$

v) Number of support vectors using cvxopt- 37

vi) Number of support vectors using standard library- 36

ix) Intercept using cvxopt= -0.492712535915936

x) Intercept using standard library- -0.4611137

## Observations-

i) The accuracy turned out to be exactly the same for both linear and non-linear svm using the standard svm library and cvxopt.

ii) The number of support vectors calculated using the standard library as well as the cvxopt were nearly the same, they differ because of the threshold defined on finding the lagrange coefficients.

iii) The values of intercept and weight vectors were nearly the same using both methods.

## 2. Binary Classification-

### i) Using 10 features-

a) Classes- (0,1)

Kernel	C	Gamma	Degree	Accuracy
Linear	0.001	-	-	99.83
Polynomial	0.1	1	3	99.83
Rbf	1	0.01	-	99.83
Sigmoid	0.1	0.01	-	99.83

b) Classes- (2,3)

Kernel	C	Gamma	Degree	Accuracy
Linear	1	-	-	93.5
Polynomial	1	0.1	3	96.33
Rbf	2	0.1	-	96.83
Sigmoid	1000	0.001	-	93.66

c) Classes- (5,6)

Kernel	C	Gamma	Degree	Accuracy
Linear	1	-	-	95
Polynomial	1	0.1	3	98
Rbf	2	0.1	-	97.3
Sigmoid	10	0.01	-	95.33

ii) Using 25 features-

a) Classes- (0,1)

Kernel	C	Gamma	Degree	Accuracy
Linear	0.001	-	-	99.83
Polynomial	0.1	0.1	3	99.83
Rbf	1	0.01	-	99.83
Sigmoid	0.1	0.01	-	99.83

b) Classes- (2,3)

Kernel	C	Gamma	Degree	Accuracy
Linear	0.01	-	-	95.88
Polynomial	0.1	0.1	3	98.66
Rbf	2	0.1	-	98.33

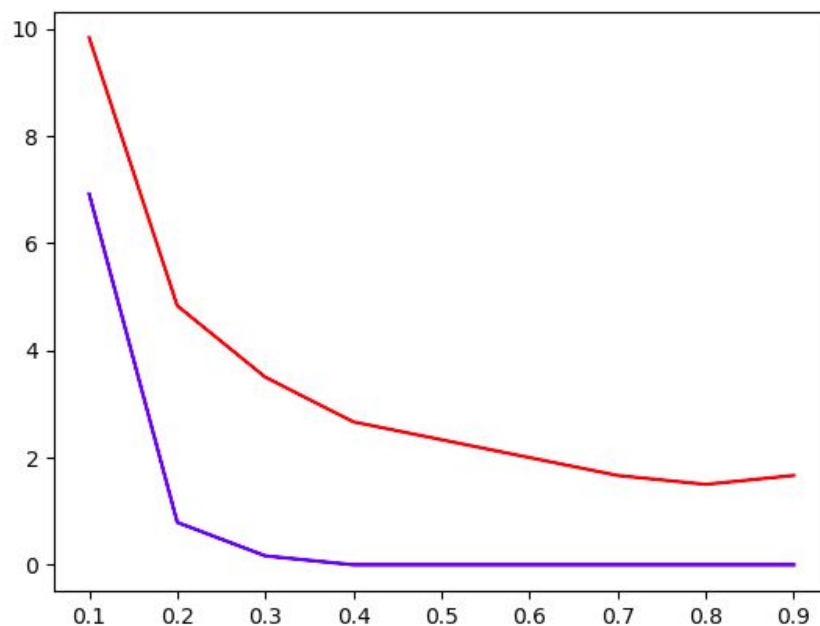
Sigmoid	1000	0.001	-	95.3
---------	------	-------	---	------

c) Classes- (5,6)

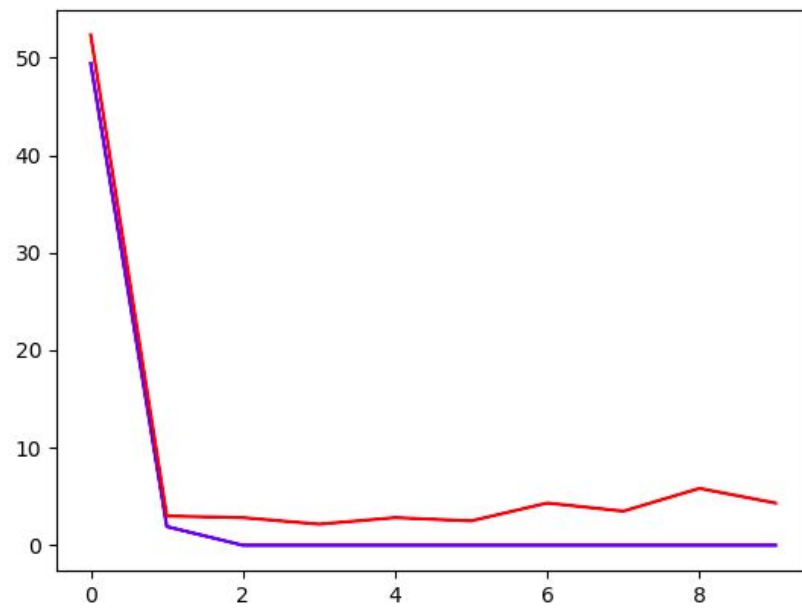
Kernel	C	Gamma	Degree	Accuracy
Linear	1	-	-	97
Polynomial	1	0.1	3	98
Rbf	2	0.1	-	98.16
Sigmoid	10	0.01	-	96.33

iii) Plots-

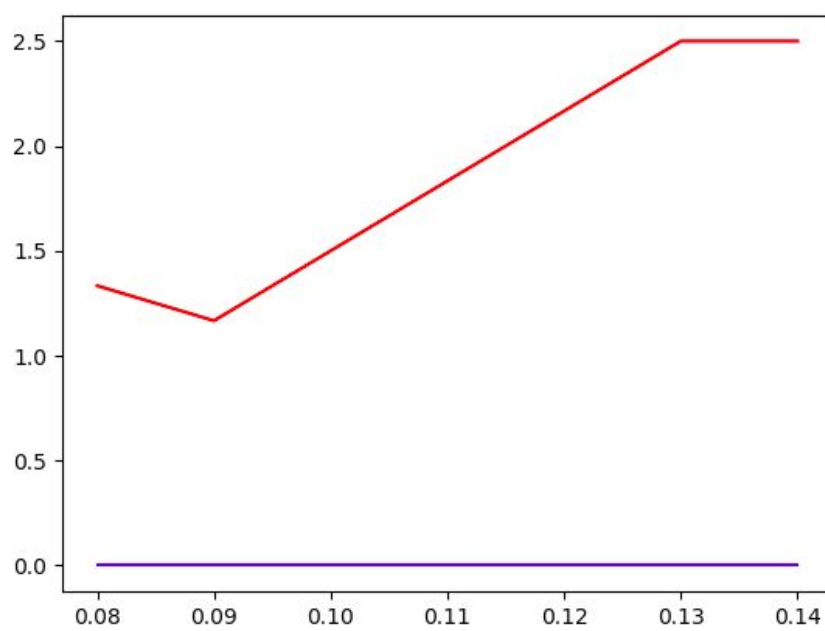
a) Error vs C-



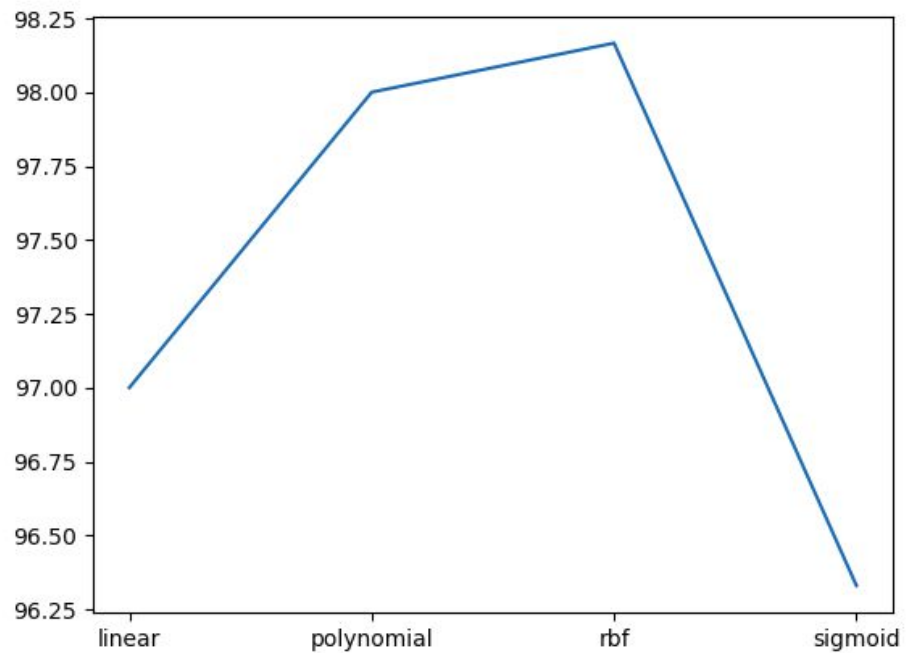
b) Error vs Degree-



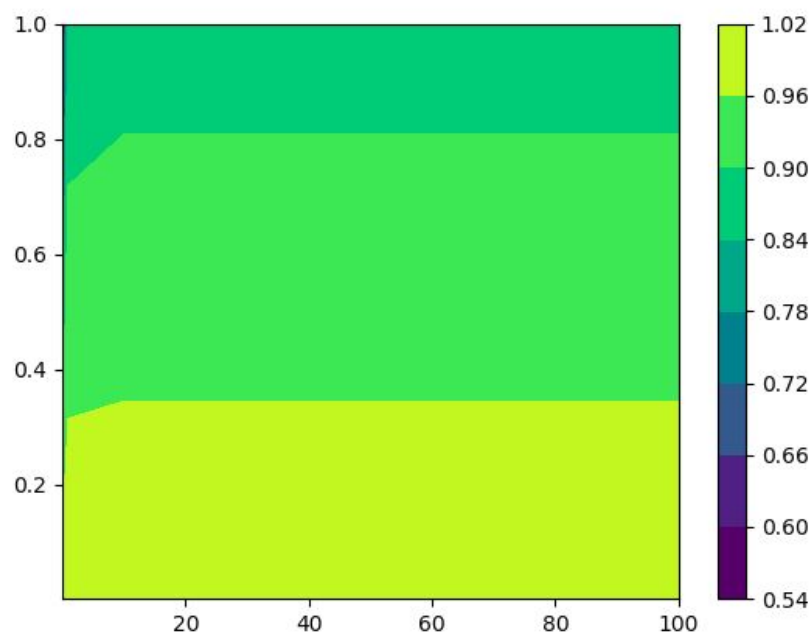
c) Error vs Gamma-



d) Accuracy vs Kernel-



e) Contour Plot (Accuracy vs C and Gamma)-



## Observations-

i) Effect of changing the hyperparameters- On varying the various hyperparameters, such as C, gamma and degree, we could see that as they increased the model complexity increased with due to which model showed overfitting once we went ahead of the optimal points. This is due to the fact that a higher C provides lower bias and high variance because it increases the penalty on the misclassified points, and a higher gamma affects the curvature of the decision boundary due to which overfitting is caused.

ii) Effect of changing the choice of kernel- We could observe that mostly the rbf kernel and the polynomial kernel provided the maximum accuracy, but when the data itself was linearly separable, all the four kernels provided the same accuracy, which was the case for classes with labels- (0,1). But generally, a higher dimensional kernel provides a better accuracy than a lower dimensional one.

iii) Effect of changing the number of features- We could see that as the number of features increased the accuracy came out to be slightly higher which means that additional number of features were required to explain the label completely. But in some cases it didn't affect which indicated that the extra features were redundant and were not needed for classification of the data. Also the values of hyperparameters didn't vary a lot for which the good fit was achieved.

## 3) Multiclass Classification-

### i) Using 10 features-

Kernel	C	Gamma	Degree	Accuracy
Linear	1	-	-	82.36
Polynomial	0.1	0.1	3	86.86
Rbf	3.1	0.07	-	89.1



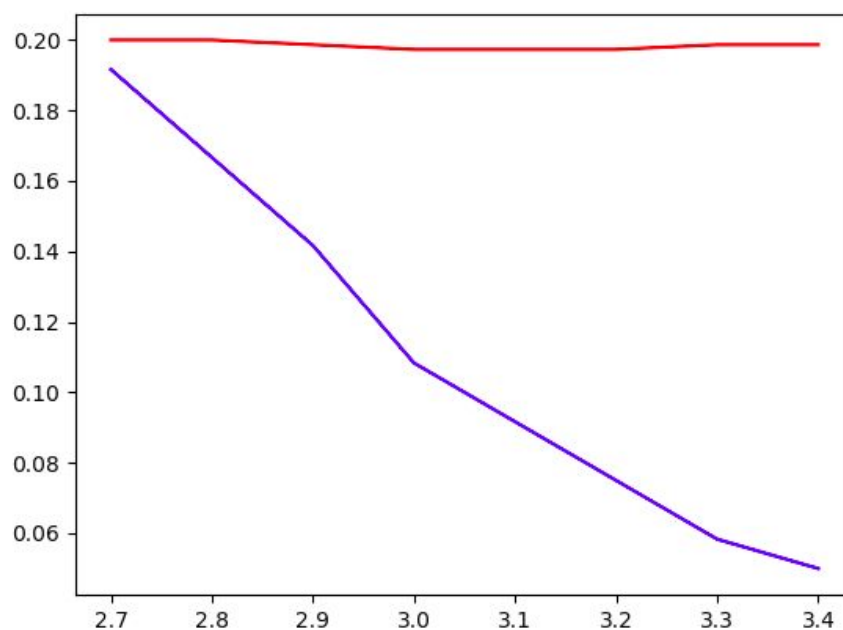
Sigmoid	1000	0.0001	-	82.3
---------	------	--------	---	------

ii) Using 25 features-

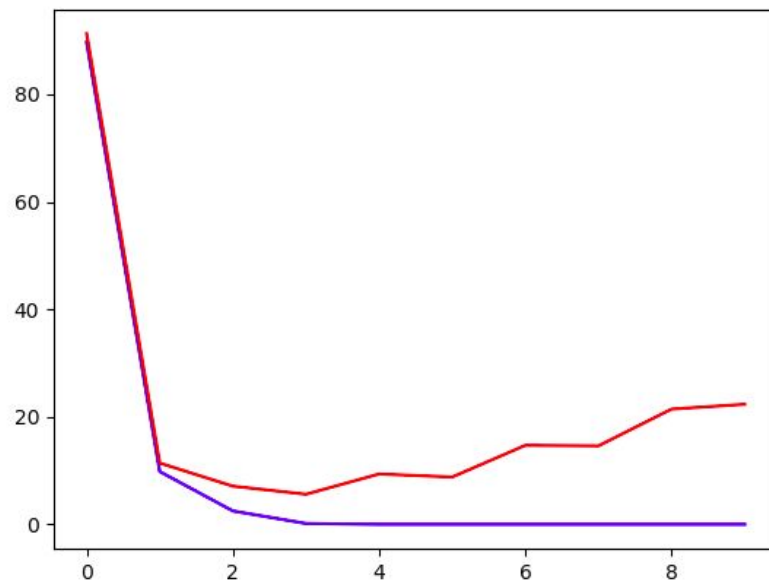
Kernel	C	Gamma	Degree	Accuracy
Linear	0.1	-	-	89.76
Polynomial	0.1	0.1	3	93.13
Rbf	3.1	0.04	-	94.93
Sigmoid	1000	0.0001	-	89.76

iii) Plots-

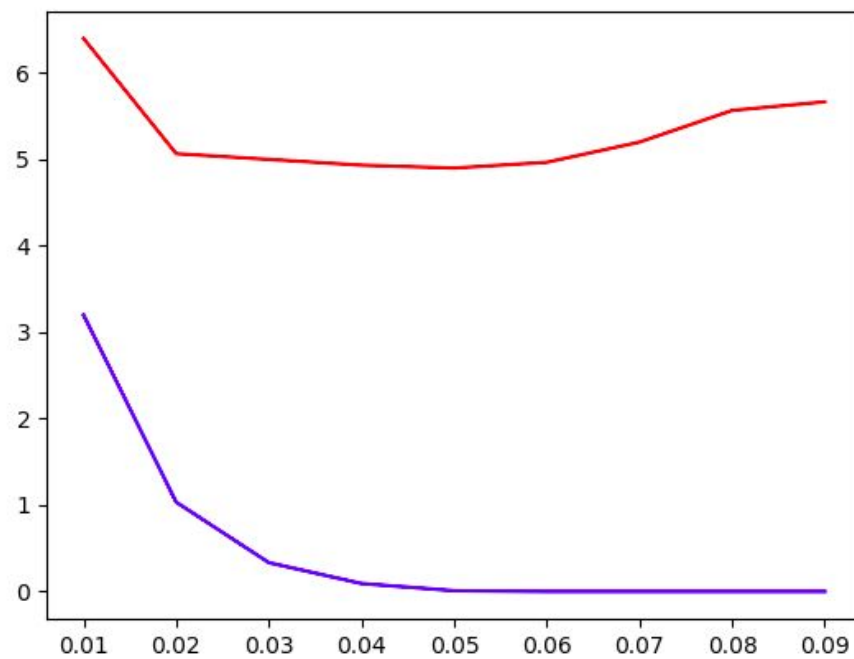
a) Error vs C-



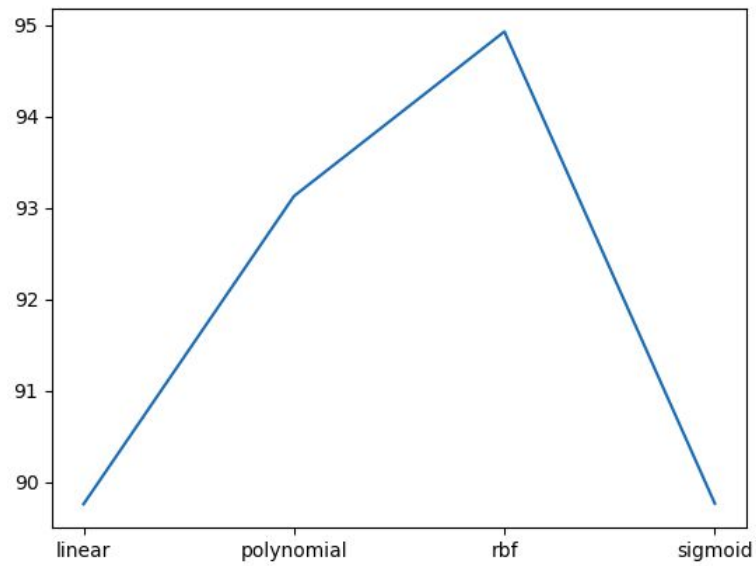
b) Error vs Degree-



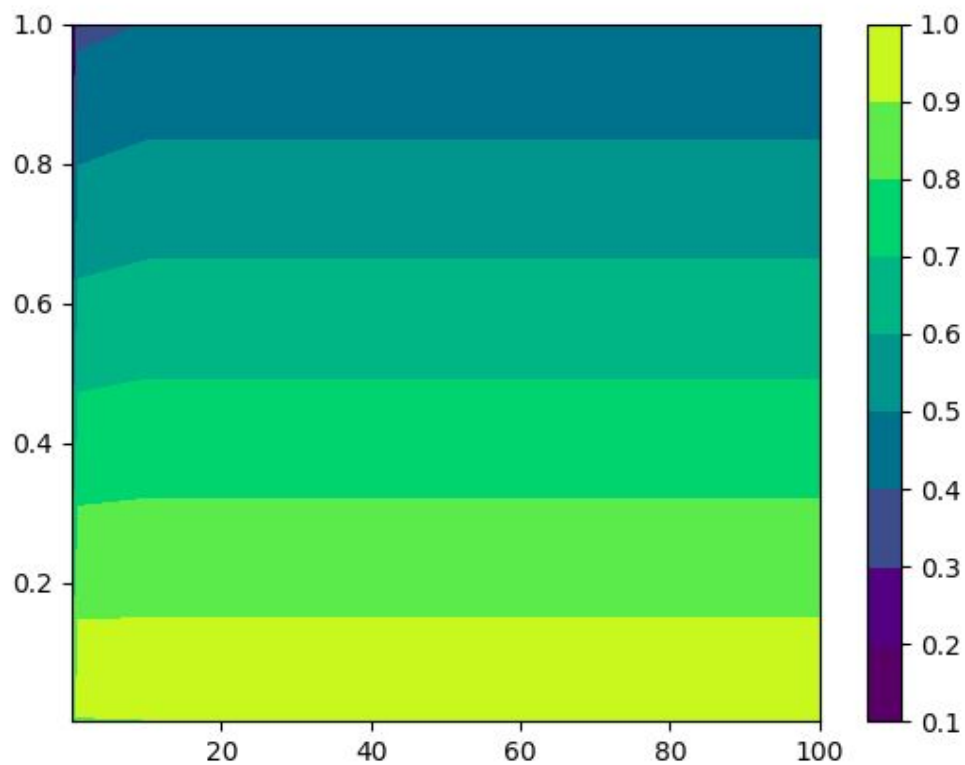
c) Error vs Gamma-



d) Accuracy vs Kernel-



e) Contour Plot (Accuracy vs C and gamma)-



## Observations-

i) Effect of changing the hyperparameters- On varying the various hyperparameters, such as C, gamma and degree, we could see that as they increased the model complexity increased with due to which model showed overfitting once we went ahead of the optimal points. This is due to the fact that a higher C provides lower bias and high variance because it increases the penalty on the misclassified points, and a higher gamma affects the curvature of the decision boundary due to which overfitting is caused.

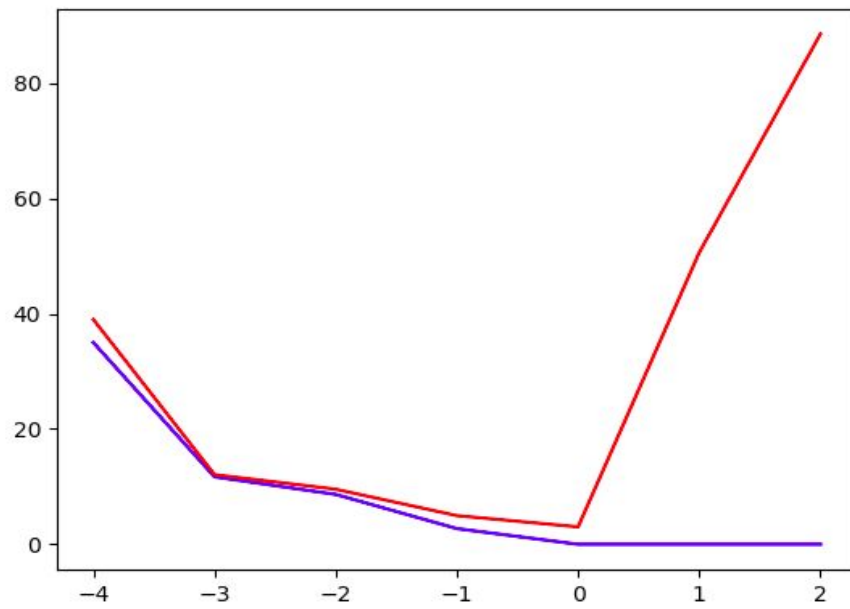
ii) Effect of changing the choice of kernel and classification type- We could observe that mostly the rbf kernel and the polynomial kernel provided the maximum accuracy. Also, when we switched to multiclass classification the values of C and gamma varied slightly due to the fact that we need a higher dimension to classify as number of labels have now increased and to cater to that, the values of C and gamma have increased accordingly.

iii) Effect of changing the number of features- We could see that as the number of features increased the accuracy came out to be much higher which means that additional number of features were required to explain the label completely.

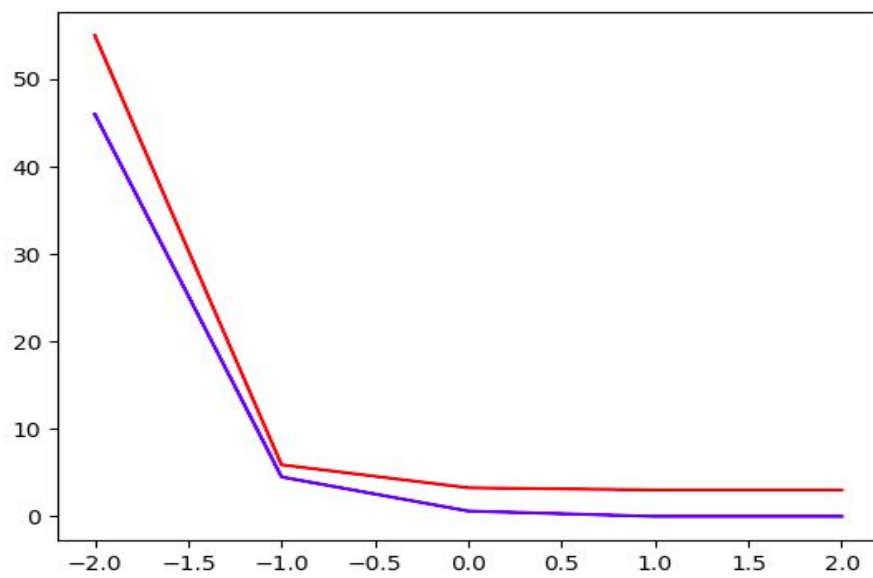
## Part-B

### 1) Hyperparameter tuning-

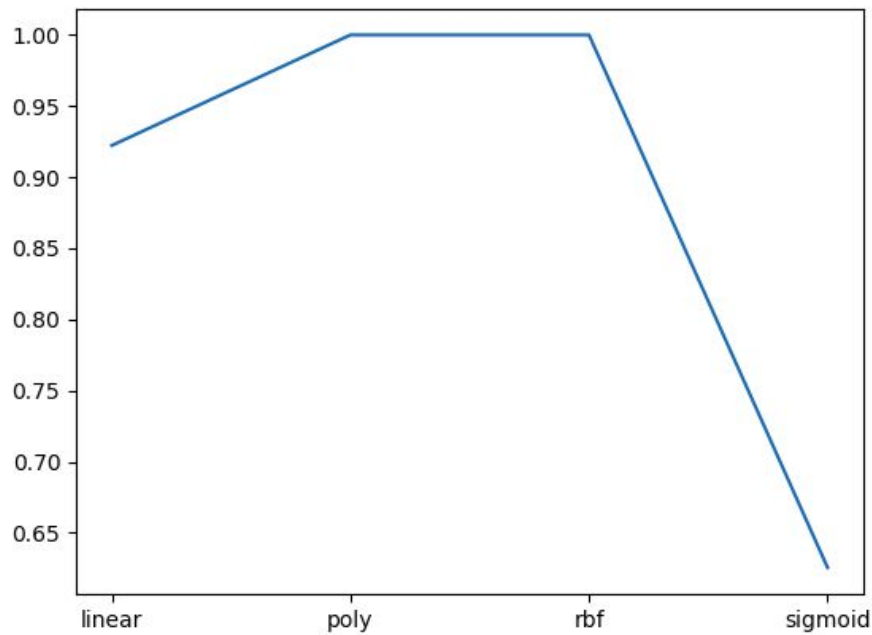
#### a) Error vs Gamma-



b) Error vs C-



## 2) Choice of Kernel- Rbf



3) Optimal Hyperparameters-  $C=10$ ,  $\text{Gamma}=1$

4) Feature Scaling-  $X = (X - X_{\min}) / (X_{\max} - X_{\min})$

5) Cross-Validation Accuracy- 96.984