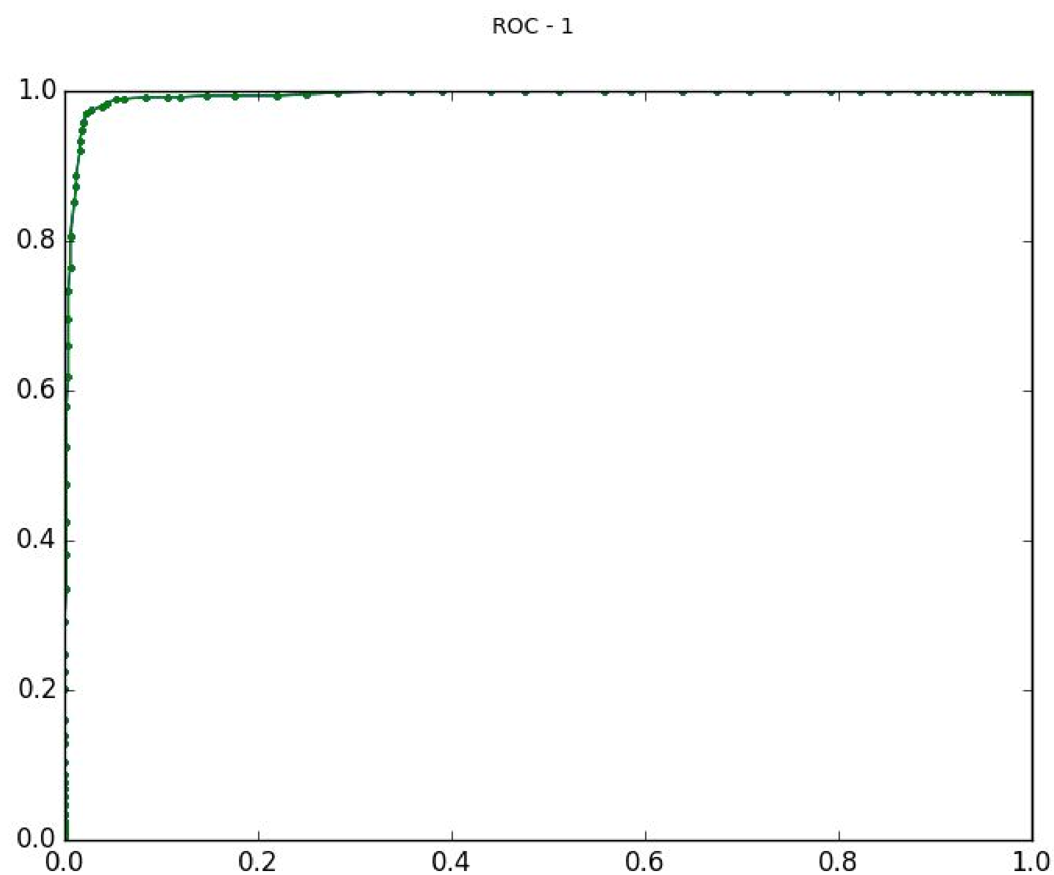
Assignment 3
Report

2)

Steps performed for linear SVM for binary classification

- Converted the data file into .csv file
- Picked 2000 samples of 3 and 8 each from the data set.
- Permute the rows to increase randomness. The above steps were performed for both testing and training files.
- For soft margin SVM formulation apply 5 fold cross validation with Grid search to get an approximate value of C.
- I performed 5 fold cross validation to find the optimal value of C. For my program, C=0.03125.
- Used the above value of C to trained the model and saved it
- Test the test data using the saved model “model_linear.model”.
- Plot roc curve for it.

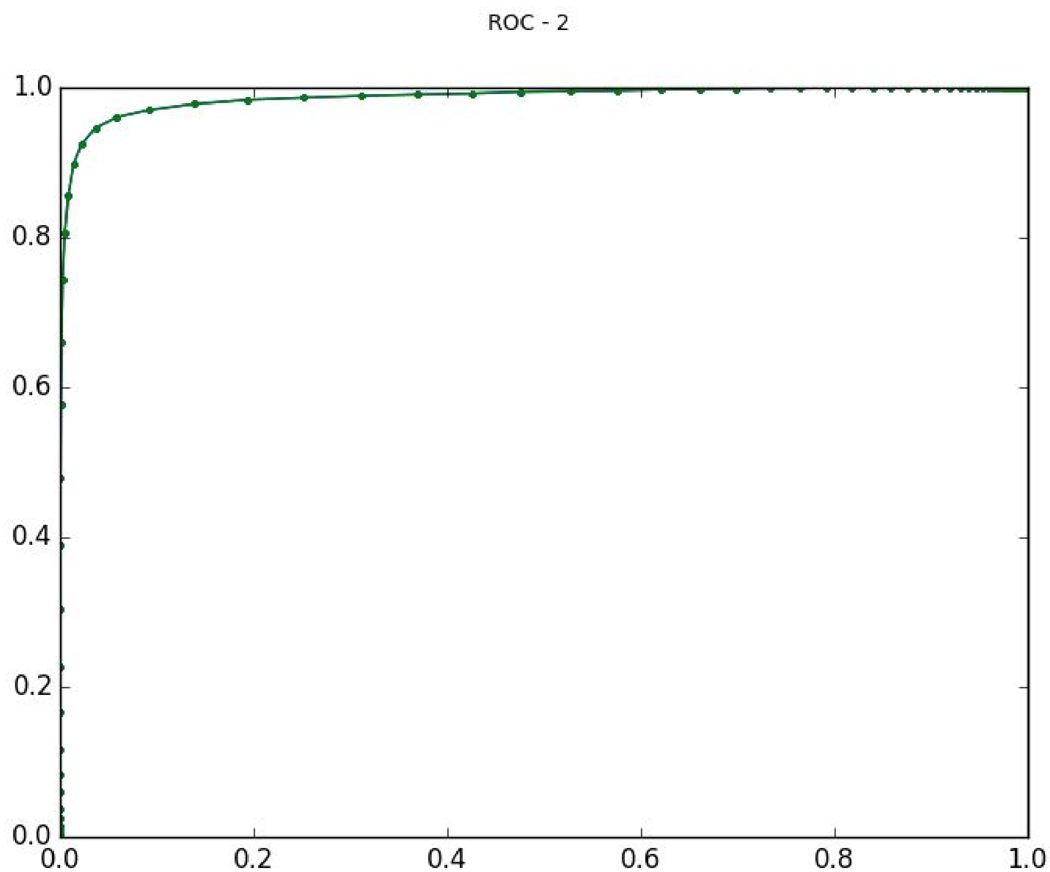
ROC curve obtained:



Linear SVM for multi-class classification

Steps Performed:

- Converted the data file into .csv file
- Picked 2000 samples of 3 and 8 each from the data set.
- Permute the rows to increase randomness. The above steps were performed for both testing and training files.
- Run loop for one vs all approach such that for each label we have different trained model.
- Test the various models on the subset of 5000 samples.
- Plot roc curve for it.

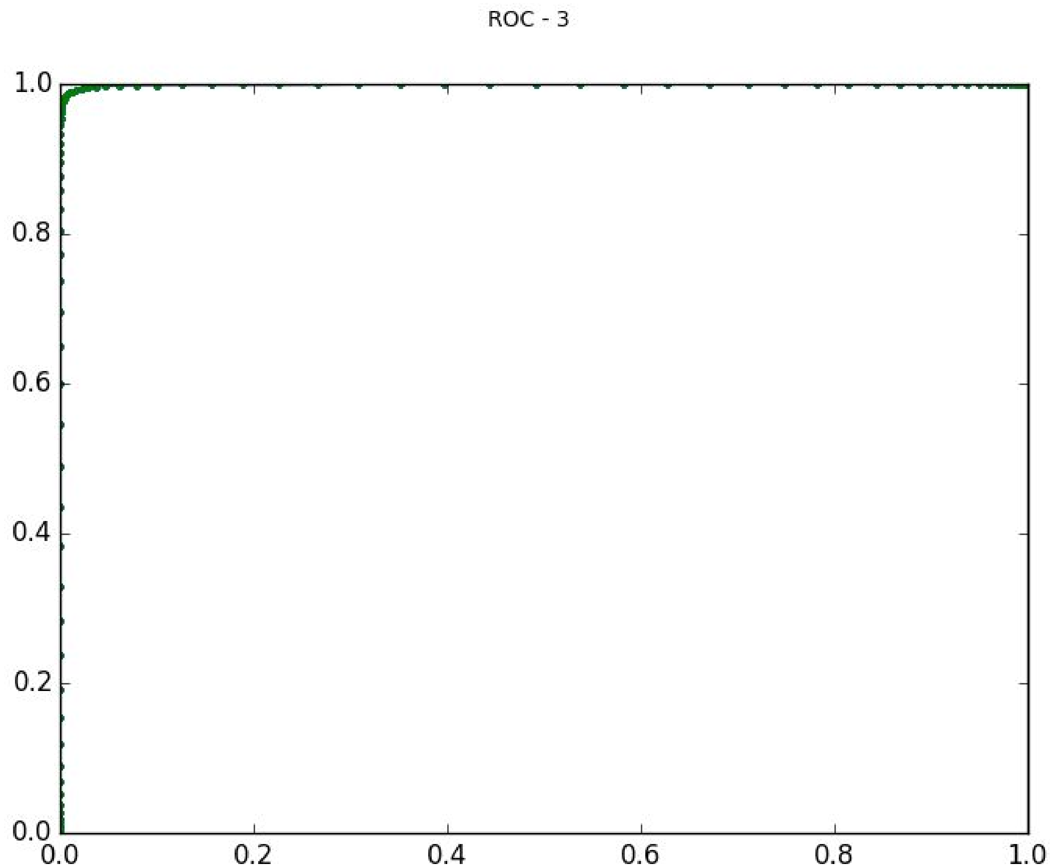


RBF Kernel for multi-class classification

Steps:

- Converted the data file into .csv file
- Picked 2000 samples of 3 and 8 each from the data set.
- Permute the rows to increase randomness. The above steps were performed for both testing and training files.

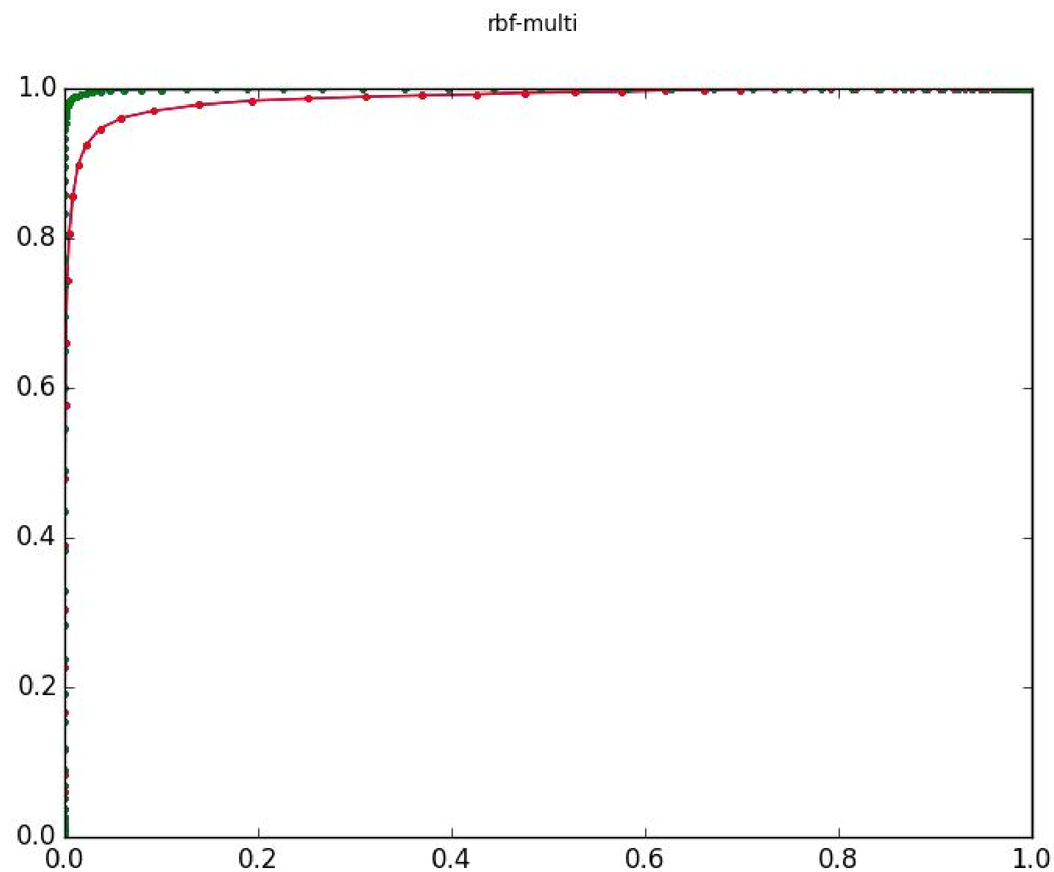
- For soft margin SVM formulation apply 5 fold cross validation with Grid search to get an approximate value of C.
- I performed 5 fold cross validation to find the optimal value of C. For my program, C=10 and gamma=0.01
- Run loop for one vs all approach such that for each label we have different trained model.
- Test the various models on the subset of 5000 samples.
- Test the test data using the saved model “multi2.model”.
- Plot roc curve for it.



KPCA and k-nearest neighbor classification

Steps performed:

- Select 150 random samples from the training set of MNIST data set.
- Create test set using 100 samples from test set.
- Randomised both the dataset
- Construct the kernel matrix and perform KPCA with k=3
- Use 5 fold cross validation with grid search to estimate gamma for RBF kernel.
- Perform KPCA + kNN classification on the 1000 test samples and check the accuracy.



Results:

- For linear, after grid search and cross validation, the value of C was found to be 0.03125
- This value can be used for linear models in binary classification and one-versus-all approach.
- For RBF, the value of gamma and C found by 5 fold cross validation with grid search are C = 10, gamma = 0.01.
- **Accuracy** :
  Binary classification : 92.9%
  One-versus-all : 89.3%
  Rbf kernel : 95%
  KPCA+kNN : 50%

Theory Questions

Name - Prerna Singh
Roll no - 2013149

Evergreen
Page No.
Date: / /201

Machine Learning Assignment

(1) $p(y_i = 1 | f(x_i)) = \frac{1}{1+e^{-f(x)}}$ where $f(x_i) = w^T x_i + b$

$p(y_i = -1 | f(x_i)) = 1 - p(y_i = 1 | f(x_i))$

$= 1 - \frac{1}{1+e^{-f(x_i)}}$

$= \frac{1+e^{-f(x_i)} - 1}{1+e^{-f(x_i)}} = \frac{e^{-f(x_i)}}{1+e^{-f(x_i)}}$

$= \frac{\frac{1}{e^{f(x_i)}}}{1+\frac{1}{e^{f(x_i)}}} = \frac{\frac{1}{e^{f(x_i)}}}{\frac{e^{f(x_i)}+1}{e^{f(x_i)}}}$

$= \frac{1}{e^{f(x_i)}+1}$

$\therefore$ General form of probability can be written as :-

$p(y_i | f(x_i)) = \frac{1}{1+e^{-yf(x_i)}}$ — (1)

→ (1) satisfies both :-

$p(y_i = 1 | f(x_i)) = \frac{1}{1+e^{-f(x_i)}}$

$p(y_i = -1 | f(x_i)) = \frac{1}{1+e^{f(x)}}$

Since we have datapoints (x_i, y_i) for $i = 1, \dots N$ and $x_i \in R^d$

$\therefore p(Y | f(x)) = \prod_{n=1}^{N} p(y_i | f(x_i))$

$= \prod_{n=1}^{N} \frac{1}{1+e^{-y_i f(x_i)}}$

Taking log of the above equation

$\log p(y | f(x)) = \log \prod_{n=1}^{N} \frac{1}{1+e^{-y_i f(x_i)}}$

$$= \sum_{i=1}^{N} \log\left(\frac{1}{1+e^{-y_i f(x_i)}}\right)$$

$$= \sum_{i=1}^{N} \log 1 - \sum_{i=1}^{N} \log\left(1+e^{-y_i f(x_i)}\right)$$

$$= 0 - \sum \log\left(1+e^{-y_i f(x_i)}\right)$$

$$= -\sum_{i=1}^{N} \log\left(1+e^{-y_i f(x_i)}\right)$$

$\therefore$ log likelihood with addition of regularization term is given by

$$-\sum \log\left(1+e^{-y_i f(x_i)}\right) + \lambda \|w\|^2$$

$\therefore$ -ve log likelihood would be :-

$$\sum_{i=1}^{N} \log\left(1+\exp(-y_i f(x_i))\right) + \lambda\|w\|^2 \quad \text{--- (2)}$$

$\therefore$ (2) can be rewritten as :-

$$\sum_{i=1}^{N} E_{LR}(y_i, f(x_i)) + \lambda\|w\|^2$$

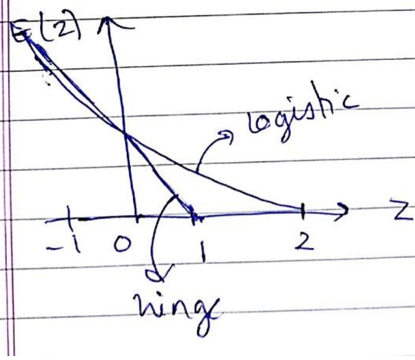
where $E_{LR}(y_i, f(x_i)) = \log\left(1+e^{-y_i f(x_i)}\right)$

$\therefore$ Hence proved

① b) Contrast between the logistic error function with the hinge error function typically used in SVMs:-

The key difference is that the flat region in hinge error function defined by $[1-y_n f(x)]_+$ for the positive part leads to sparse solution.

- By sparse solution, we mean that α_i are zero.
- Also, the smoothness is another contrasting point.
- Following are the graphs for logistic error, ~~[illegible]~~ rescaled by the hinge error function



④ $L(w, b, \varepsilon_i, \ldots \varepsilon_N, \hat{\varepsilon}_i \ldots \hat{\varepsilon}_N) = C \sum_{i=1}^{N} (\varepsilon_i + \hat{\varepsilon}_i)$

$+ \frac{1}{2}\|w\|^2 - \sum_{i=1}^{N} (\mu_i \varepsilon_i + \hat{\mu}_i \hat{\varepsilon}_i)$ ~~$+ \frac{1}{2}\|w\|^2$~~

$- \sum a_i(\epsilon + \varepsilon_i + f(x_i) - y_i) - \sum \hat{a}_i (\varepsilon + \hat{\varepsilon}_i - f(x_i) + y_i)$

Substituting $f(x_i) = w^T \phi(x) + b$ and then differentiating wrt to $w, b, \varepsilon_i, \hat{\varepsilon}_i$ we get:-

$$\frac{dL}{dw} = \|w\| - \sum_{i=1}^{N} a_i \phi(x) + \sum_{i=1}^{N} \hat{a}_i \phi(x)$$

$$\|w\| = \phi(x)\left[\sum_{i=1}^{N} a_i - \sum_{i=1}^{N} \hat{a}_i\right] \quad \text{—— (1)}$$

$$\frac{dL}{db} = -\sum_{i=1}^{N} a_i(1) + \sum_{i=1}^{N} \hat{a}_i(1) = 0$$

$$\sum_{i=1}^{N} a_i - \sum_{i=1}^{N} \hat{a}_i = 0 \quad \text{—— (2)}$$

$$\frac{dL}{d\varepsilon_i} = C\sum_{i=1}^{N} 1 - \sum_{i=1}^{N} \mu_i - \sum_{i=1}^{N} a_i = 0$$

$$CN = \sum_{i=1}^{N} \mu_i + \sum_{i=1}^{N} a_i \quad \text{—— (3)}$$

$$\frac{dL}{d\hat{\varepsilon}_i} = C\sum_{i=1}^{N} 1 - \sum_{i=1}^{N} \hat{\mu}_i - \sum_{i=1}^{N} \hat{a}_i$$

$$CN = \sum_{i=1}^{N} \mu_i + \sum_{i=1}^{N} \hat{a}_i \quad \text{—— (4)}$$

Rearranging the given equation of $L(w, b, \varepsilon_i \ldots \varepsilon_n, \hat{\varepsilon}_i, \ldots \hat{\varepsilon}_n)$,

$$\underbrace{C\sum_{i=1}^{N} \varepsilon_i - \sum_{i=1}^{N} \mu_i \varepsilon_i - \sum_{i=1}^{N} a_i \varepsilon_i}_{(1)} + \underbrace{C\sum_{i=1}^{N} \hat{\varepsilon}_i - \sum_{i=1}^{N} \hat{\mu}_i \hat{\varepsilon}_i - \sum_{i=1}^{N} \hat{a}_i \hat{\varepsilon}}_{(2)}$$

$$+ \underbrace{\frac{1}{2}\|w\|^2 - \sum_{i=1}^{N} a_i w^T \phi(x) + \sum_{i=1}^{N} \hat{a}_i w^T \phi(x)}_{(3)}$$

$$- \underbrace{b\left(\sum_{i=1}^{N} a_i - \sum_{i=1}^{N} \hat{a}_i\right)}_{(4)} - \epsilon \sum_{i=1}^{N} (a_i + \hat{a}_i)$$

$$+ \sum_{i=1}^{N} (a_i - \hat{a}_i) y_i$$

Finding the value of above 4 terms,

i) From differentiation and equation (3), we have

Date: / /201

$$CN = \sum_{i=1}^{N} \mu_i - \sum_{i=1}^{N} a_i$$

or $C = \mu_i - a_i$

Multiplying by $\xi_i \Rightarrow C\xi_i = \mu_i \xi_i - a_i \xi_i$

Do summation

$$C\sum_{i=1}^{N} \xi_i = \sum_{i=1}^{N} \mu_i \xi_i - \sum_{i=1}^{N} a_i \xi_i$$

$$C\sum_{i=1}^{N} \xi_i - \sum_{i=1}^{N} \mu_i \xi_i - \sum_{i=1}^{N} a_i \xi_i = 0$$

Hence ~~from~~ term ① $\longrightarrow 0$

Similarly, from eqn ④, we have

$$CN = \sum_{i=1}^{N} \hat{\mu}_i + \sum_{i=1}^{N} \hat{a}_i$$

or $C = \hat{\mu}_i + \hat{a}_i$

multiplying by $\hat{\xi}_i$,

$$C\hat{\xi}_i = \hat{\mu}_i \hat{\xi}_i + \hat{a}_i \hat{\xi}_i$$

Summation,

$$C\sum_{i=1}^{N} \hat{\xi}_i = \sum_{i=1}^{N} \hat{\mu}_i \hat{\xi}_i + \sum_{i=1}^{N} \hat{a}_i \hat{\xi}_i$$

$$C\sum_{i=1}^{N} \hat{\xi}_i - \sum_{i=1}^{N} \hat{\mu}_i \hat{\xi}_i - \sum_{i=1}^{N} \hat{a}_i \hat{\xi}_i = 0$$

Clearly term ② $\longrightarrow 0$

Term ④, $\rightarrow 0$ since $\sum_{i=1}^{N} a_i - \sum_{i=1}^{N} \hat{a}_i = 0$ from eqn ②

Finally ③, $\frac{1}{2}\|w\|^2 - \sum_{i=1}^{N} a_i w^T \phi(x) + \sum_{i=1}^{N} \hat{a}_i w^T \phi(x)$

$$= \frac{1}{2}\|w\|^2 - w^T \phi(x) \left[\sum_{i=1}^{N} a_i - \sum_{i=1}^{N} \hat{a}_i\right]$$

$$= \frac{1}{2}\|w\|^2 - \|w\|^2 = -\frac{1}{2}\|w\|^2$$

$$= -\frac{1}{2} w^T w$$

$\hat{a}_i)$

$) =$

o

fu

i

if

c

Page No.

Date : / /201

$$= -\frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{N} (a_i - \hat{a}_i)(a_j - \hat{a}_j)\, \phi(x_i)^T \phi(x_j)$$

Hence finally, the equation becomes

$$-\frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{N} (a_i - \hat{a}_i)(a_j - \hat{a}_j)\, K(x_i, x_j)$$

$$- t \sum_{i=1}^{N} (a_i + \hat{a}_i) + \sum_{i=1}^{N} (a_i - \hat{a}_i)\, y_i$$

where $K(x_i, x_j) = \phi(x_i)^T \phi(x_j)$