Homework 3
Harshvardhan Kalra
2014043

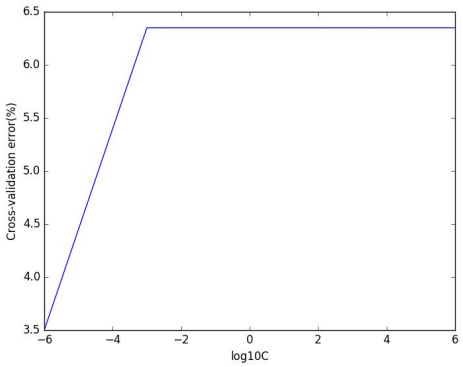
The general steps done in the assignment were:

a) Parsing the ubyte files and generating the training and test subsets. The references used for parsing the files are mentioned in the code base.
b) Choosing values for C and gamma. These were done mainly by hit and trial, but since the computation time was too long, I took some arbitrary values for C and gamma, based on certain facts such as lower C giving higher training accuracy with a penalty of overfitting. A similar analysis was used for predicting the gamma values.

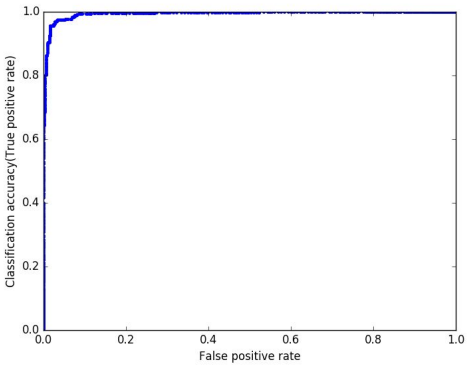
   Values used for grid search:
   i) C : [1e-6, 1e-3, 1, 1e3, 1e6] for Linear binary classifier
   ii) C : [1e-7, 1e-3, 1, 10, 1e5] for Linear multiclass classifier
   iii) C : [1e-7, 1e2] and gamma : [1e-6, 1] for RBF multiclass classifier

   The computational overhead prevented me from considering more than 2 values apiece for gamma and C for the RBF kernel.
c) Using sklearn's inbuilt functions, I computed the cross-validation error across 5 folds, and used the best estimator for testing. Each run was for 3000 iterations.

a) 3 vs 8 binary classifier: Since a lower C would be giving higher scores in general, C = 1e-6 gave the best score, and hence the least error. The test accuracy varied between 96% to 97%, while the training (cross-validation) mean accuracy was around 93%.

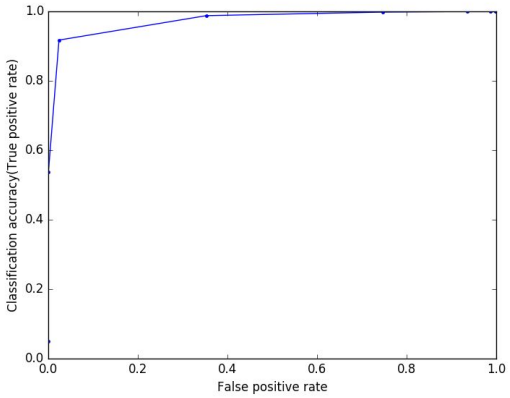


i) C vs errors

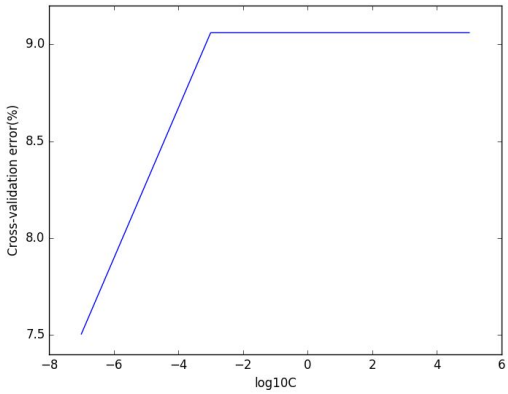


ii) ROC curve

b) Linear multiclass classifier : Similar to the previous case, C = 1e-7 gave the best accuracy of 82%. Fine-tuning the hyperparameter might lead to better results.



i) ROC curve



ii) C vs errors

c) RBF multiclassifier : The values C = 100, and gamma = 1e-6 gave the most accuracy of 91%.

Unable to procure graphs for RBF and RBF vs Linear.

1) $f(x_i) = w^T x_i + b \quad \forall i \in 1, 2 \ldots N$
$y_i \in \{-1, 1\}$ and $x_i \in R^d$

If $p(y_i = 1 | f(x_i))$ is defined as:

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

- we have,

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

where σ denotes the sigmoid function.

$$\Rightarrow p(y | f(x)) = \sigma(yf(x))$$
$$= \frac{1}{1 + e^{-yf(x)}}$$

Taking ~~ln~~ Applying $-\ln$ on both sides, we g

$$-\ln(p(y | f(x))) = -\ln\left(\frac{1}{1 + e^{-yf(x)}}\right)$$
$$= -[\ln 1 - \ln(1 + e^{-yf(x)})]$$
$$= \ln(1 + e^{-yf(x)})$$

$\Rightarrow$ ~~$-\ln p(y)$~~ The negative of the log-likelih becomes, with a regularization coefficie

$$\sum_{i=1}^{N} \ln(1 + e^{-y_i f(x_i)}) + \lambda w^T w$$

Hence proved.

Now, the hinge error function is given by:

$$\sum_{i=1}^{N} [1 - y_i f(x_i)]_+ + \lambda w^T w$$

Comparing with the previous equation, we see that the hinge loss and logistic regression error are continuous approximations of the misclassification error, which we seek to reduce.

a) $L(w, b, \xi_1 \dots \xi_N, \hat{\xi}_1 \dots \hat{\xi}_N)$

$$= C \sum_{i=1}^{N} (\xi_i + \hat{\xi}_i) + \frac{1}{2} w^T w - \sum_{i=1}^{N} (\mu_i \xi_i + \hat{\mu}_i \hat{\xi}_i) - \sum_{i=1}^{N} a_i (\epsilon + \xi_i + f(x_i) - y_i)$$

$$- \sum_{i=1}^{N} \hat{a}_i (\epsilon + \hat{\xi}_i - f(x_i) + y_i), \quad f(x_i) = w^T \phi(x_i) + b$$

Here, $\xi, \hat{\xi} \geq 0$ are slack variables.

$$\frac{\partial L}{\partial w}$$

$$\Rightarrow \frac{\partial L}{\partial w} = C \cdot 0 + w - 0 - \sum_{i=1}^{N} a_i \phi(x_i) - \sum_{j=1}^{N} \hat{a}_j \phi(x_j) = 0$$

$$\Rightarrow w = \sum_{i=1}^{N} (a_i - \hat{a}_i) \phi(x_i) \quad -(I)$$

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

$$\Rightarrow \sum_{i=1}^{N} a_i = \sum_{i=1}^{N} \hat{a}_i \quad -(II)$$

$$\frac{\partial L}{\partial \xi_i} = 0 \rightarrow C \sum_{i=1}^{N} (1) + 0 - \sum_{i=1}^{N} \mu_i - \sum_{i=1}^{N} a_i(1) = 0$$

$$\Rightarrow C = a_i + \mu_i \quad \forall i \in \{1, 2 \ldots N\} \quad \text{(III)}$$

$$\frac{\partial L}{\partial \hat{\xi}_i} = 0 \Rightarrow C = \hat{a}_i + \hat{\mu}_i \quad \forall i \in \{1, 2 \ldots N\} \quad \text{(IV)}$$

(similar to previous eqn)

$$\Rightarrow L = \frac{1}{2} w^T w + C \sum_{i=1}^{N} (\xi_i + \hat{\xi}_i) - \sum_{i=1}^{N} (\mu_i \xi_i + \hat{\mu}_i \hat{\xi}_i)$$

$$- \sum_{i=1}^{N} a_i (\epsilon + \xi_i + f(x_i) - y_i) - \sum_{i=1}^{N} \hat{a}_i (\epsilon + \hat{\xi}_i - f(x_i) + y_i)$$

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

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

$$- \sum_{i=1}^{N} (\mu_i \xi_i + \hat{\mu}_i \hat{\xi}_i) - \sum_{i=1}^{N} (a_i \epsilon + a_i \xi_i + a_i f(x_i)$$

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

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

$$+ C \sum_{i=1}^{N} (\xi_i + \hat{\xi}_i) - \sum_{i=1}^{N} [(\mu_i + a_i) \xi_i + (\hat{\mu}_i + \hat{a}_i) \hat{\xi}_i]$$

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

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

Using I, III and IV, we get

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

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