Programming Assignment 3
Classification and Regression
CSE 574

Group 23 - Members :

- Rakesh Balasubramanian
- Maira Saboia Da Silva
- Ramya Rao
- Alok Asok

Abstract :

In this assignment, we have implemented Logistic Regression to predict handwritten digits. We have also learned the Support Vector Machine model and computed the accuracy of prediction with respect to training data, validation data and test data against several parameters.
The details of the observation are mentioned in this report. The data sample used is MNIST.

Logistic Regression :

This is a discriminative classifier wherein we directly model the posterior probability. Here the conditional distribution is Bernoulli distribution. This method gives better generalization even with limited training data. The results of our experiment are :
Training set Accuracy: 92.328 %
Validation set Accuracy: 91.46 %
Testing set Accuracy: 91.92 %

Support Vector Machine:

The results of our experiment are :
1) Using linear kernel (all other parameters are kept default).
Accuracy:
Training set Accuracy: 97.286%
Validation set Accuracy: 93.64%
Testing set Accuracy: 93.78%

2) Using radial basis function with value of gamma setting to 1 (all other parameters are kept default).
Accuracy:
Training set Accuracy: 100%
Validation set Accuracy: 15.48%
Testing set Accuracy: 17.14%

3) Using radial basis function with value of gamma setting to default (all other parameters are kept default).

Accuracy:

Training set Accuracy: 94.294%
Validation set Accuracy: 94.02%
Testing set Accuracy: 94.42%

4) Using radial basis function with value of gamma setting to default and varying value of C (1, 10, 20, 30, · · · , 100) and plot the graph of accuracy with respect to values of C in the report.

Accuracy:

Training set Accuracy:

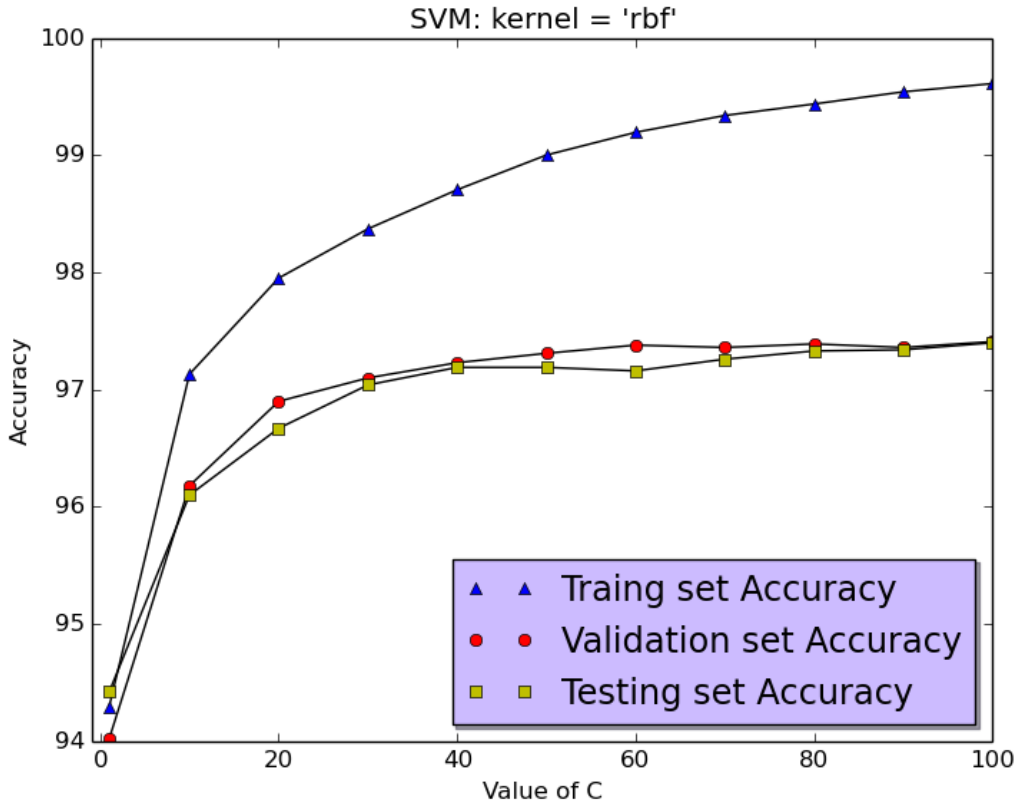
[94.42 97.13 97.95 98.37 98.70 99.00 99.19 99.34 99.43 99.54 99.61]%

Validation set Accuracy:

[94.02 96.18 96.9 97.1 97.23 97.31 97.38 97.36 97.39 97.36 97.41]%

Testing set Accuracy:

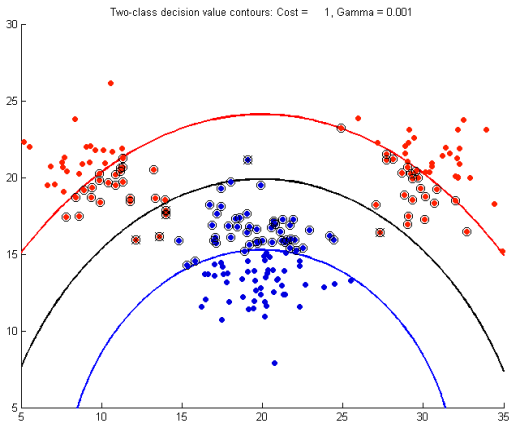
[94.42 96.1 96.67 97.04 97.19 97.19 97.16 97.26 97.33 97.34 97.4]%



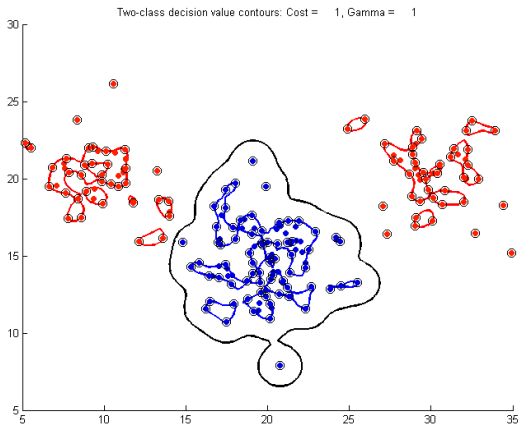
Gamma effect:

Gamma is related to the variance for the Gaussian in the RBF Kernel. Higher values of gamma means that each Gaussian has a small variance and so small are the influence of the training samples selected as support vectors. In the case 1) in which we use the value of gamma setting to 1, the model has adjusted too much to the training data and presented an **overfitting** situation, 100% accuracy for training data and 15% and 17% of accuracy for validation and testing data set, respectively.

If gamma is 0.0 then 1/n_features is used instead. In our case,it means that the value of gamma is around 0.001 (1 / 715). We can see in the figures below, how the model adjusts to the data with respect to the value of gamma = 0.001 and gamma = 1.0.
(Reference for below images : http://wiki.eigenvector.com/index.php?title=Svmda)



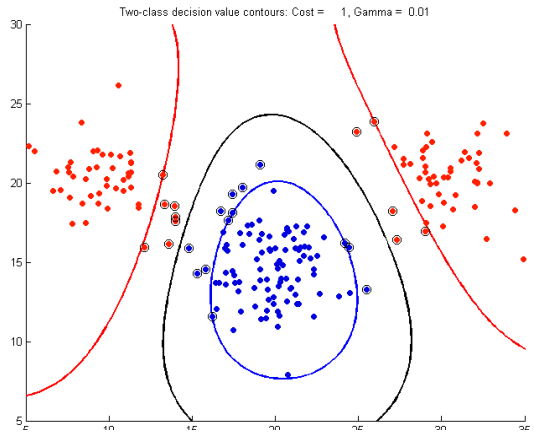
Gamma=0.001, Cost=1



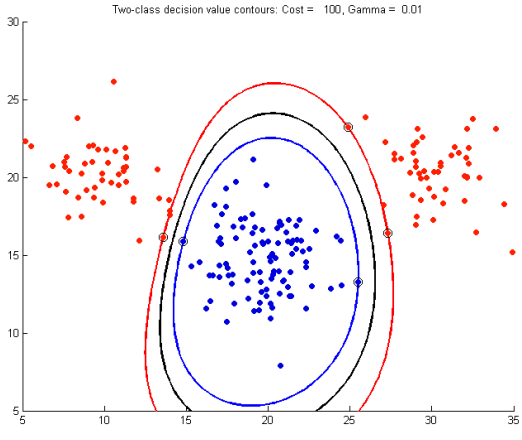
Gamma =1, Cost = 1

Cost effect:

The C parameter is the penalty associated with points that lie within the margins of the hyperplanes and indicate how much we want to avoid misclassification in the training samples. If C is large, the margins are small since the penalty is high. For very tiny values of C, we get wider margins and a simpler decision surface. In the figures given below, it is possible to visualize the effect of changing values of C.



Gamma=0.01 , Cost = 1



Gamma = 0.01, Cost = 100

(Reference for the images: http://wiki.eigenvector.com/index.php?title=Svmda)

Logistic Regression vs SVM:

The basis for the Support Vector Machine method is the idea that higher the dimension of the space , the easier it is to find a hyperplane that linearly separates the data. Here we map each vector in a much higher dimensional space, where the data can then be linearly separated.

Logistic Regression finds any solution that separates the two classes. Hard SVM finds "the" solution among all possible ones that has the maximum margin. Logistic Regression finds a hyperplane that corresponds to the minimization of some error. Soft SVM tries to minimize the error (another error) and at the same time trades off that error with the margin via a regularization parameter.

References:

1. http://wiki.eigenvector.com/index.php?title=Svmda
2. http://en.wikipedia.org/wiki/Support_vector_machine

Inference (Comparison of Experimental Results) :

Performing classification using logistic regression takes lesser time as compared to SVM method.
Accuracy with Logistic Regression method is slightly lesser than SVM method(except when gamma =1 which results in overfitting).