CSE 574

PROGRAMMING ASSIGNMENT - 3

CLASSIFICATION AND REGRESSION

COMPILED BY-

Team 51

Naina Nigam (50208030)

Surabhi Singh (50208675)

Vanshika Nigam (50208031)

LOGISTIC REGRESSION

Training Set Accuracy : 84.894%

Validation Set Accuracy : 83.75%

Test Set Accuracy : 84.2%

Direct Multi-Class Logistic Regression

Training Set Accuracy : 93.448%

Validation Set Accuracy : 92.48%

Test Set Accuracy : 92.55%

Logistic Regression vs Multi-class Logistic Regression using one-vs-all strategy

Logistic Regression is used for binary classification while Multi-class Logistic Regression is used for multi-class classification using one classifier that can classify n classes at the same time. The performance of Multi-class logistic regression is better than logistic regression when using one-vs-all strategy because in the former all classes are learned directly and the parameters of each class are estimated independently. The model so obtained is also more robust against outliers.

SUPPORT VECTOR MACHINES

1. **Linear Kernel**

   Training Set Accuracy : 97.286%

   Validation Set Accuracy : 93.64%

   Test Set Accuracy : 93.78%

2. **Radial Basis Function with gamma=1(all other parameters are kept default)**

   Training Set Accuracy : 100%

   Validation Set Accuracy : 15.48%

   Test Set Accuracy : 17.14%

3. <u>Radial Basis Function with gamma set to default (all other parameters are kept default)</u>

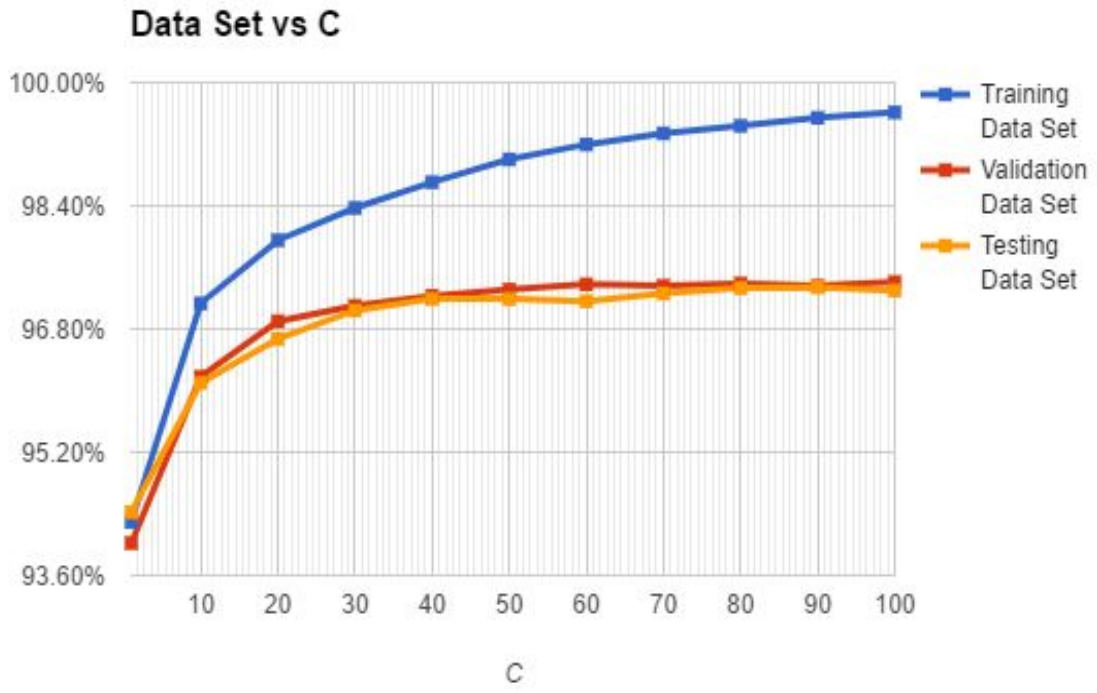
Training Set Accuracy : 94.294%

Validation Set Accuracy : 94.02%

Test Set Accuracy : 94.42%

4. <u>Radial Basis Function with gamma set to default and varying values of C (1,10,20,....,100)</u>

C	Training Data Set	Validation Data Set	Testing Data Set
1	94.29%	94.02%	94.42%
10	97.13%	96.18%	96.10%
20	97.95%	96.90%	96.67%
30	98.37%	97.10%	97.04%
40	98.71%	97.23%	97.19%
50	99.00%	97.31%	97.19%
60	99.20%	97.38%	97.16%
70	99.34%	97.36%	97.26%
80	99.44%	97.39%	97.33%
90	99.54%	97.36%	97.34%
100	99.61%	97.41%	97.30%



Inference : It is evident from the above plot that the accuracies for all three Training, Validation and Testing Data Sets increase fast for initial values of C but afterwards the increase is very slow(nominal). Also it can be observed that Validation Data set and Testing Data Set accuracies converge almost at the same point.

Linear Kernel vs Radial Basis Function Kernel

1. Solving optimization problem for a linear kernel is much faster.
2. Typically, the best possible predictive performance is better for a nonlinear kernel (or at least as good as the linear one).
3. Use linear kernel when number of features is larger than number of observations.
4. Use gaussian kernel when number of observations is larger than number of features.
5. If number of observations is larger than 50,000 speed could be an issue when using gaussian kernel; hence, one might want to use linear kernel.