CSE 574 – Introduction of Machine Learning
PROGRAMMING ASSIGNMENT - 3

GROUP 12

Prasanna Pai – 50132731
Gaurav Kshirsagar – 50134944
Vivin Rane - 50134206

Results –

1. **Logistic Regression**

   Training set Accuracy: 92.286%

   Validation set Accuracy: 91.47%

   Testing set Accuracy: 91.87%

   Runtime: 535.548 s

2. **SVM**
   a. **Linear Kernel**
      Training set Accuracy: 97.286%

      Validation set Accuracy: 93.64%

      Test set Accuracy: 93.78%

      Runtime: 1705.485 s

   b. **RBF with gamma = 1**
      Training set Accuracy: 100.0%

      Validation set Accuracy: 15.48%

      Test set Accuracy: 17.14%

      Runtime: 34893.482 s

   c. **RBF with all parameters at default**
      Training set Accuracy: 94.294%

      Validation set Accuracy: 94.02%

      Test set Accuracy: 94.42%

      Runtime: 3658.048 s

   d. **RBF with varying C**
      C=1
      Training set Accuracy: 94.294%

      Validation set Accuracy: 94.02%

      Test set Accuracy: 94.42%

C =10

Training set Accuracy: 97.132%

Validation set Accuracy: 96.18%

Test set Accuracy: 96.1%

C = 20

Training set Accuracy: 97.952%

Validation set Accuracy: 96.9%

Test set Accuracy: 96.67%

C = 30

Training set Accuracy: 98.372%

Validation set Accuracy: 97.1%

Test set Accuracy: 97.04%

C = 40

Training set Accuracy: 98.706%

Validation set Accuracy: 97.23%

Test set Accuracy: 97.19%

C = 50

Training set Accuracy: 99.002%

Validation set Accuracy: 97.31%

Test set Accuracy: 97.19%

C = 60

Training set Accuracy: 99.196%

Validation set Accuracy: 97.38%

Test set Accuracy: 97.16%

C = 70

Training set Accuracy: 99.34%

Validation set Accuracy: 97.36%

Test set Accuracy: 97.26%

C = 80

Training set Accuracy: 99.438%

Validation set Accuracy: 97.39%

Test set Accuracy: 97.33%

C = 90

Training set Accuracy: 99.542%

Validation set Accuracy: 97.36%

Test set Accuracy: 97.34%

C = 100

Training set Accuracy: 99.612%

Validation set Accuracy: 97.41%

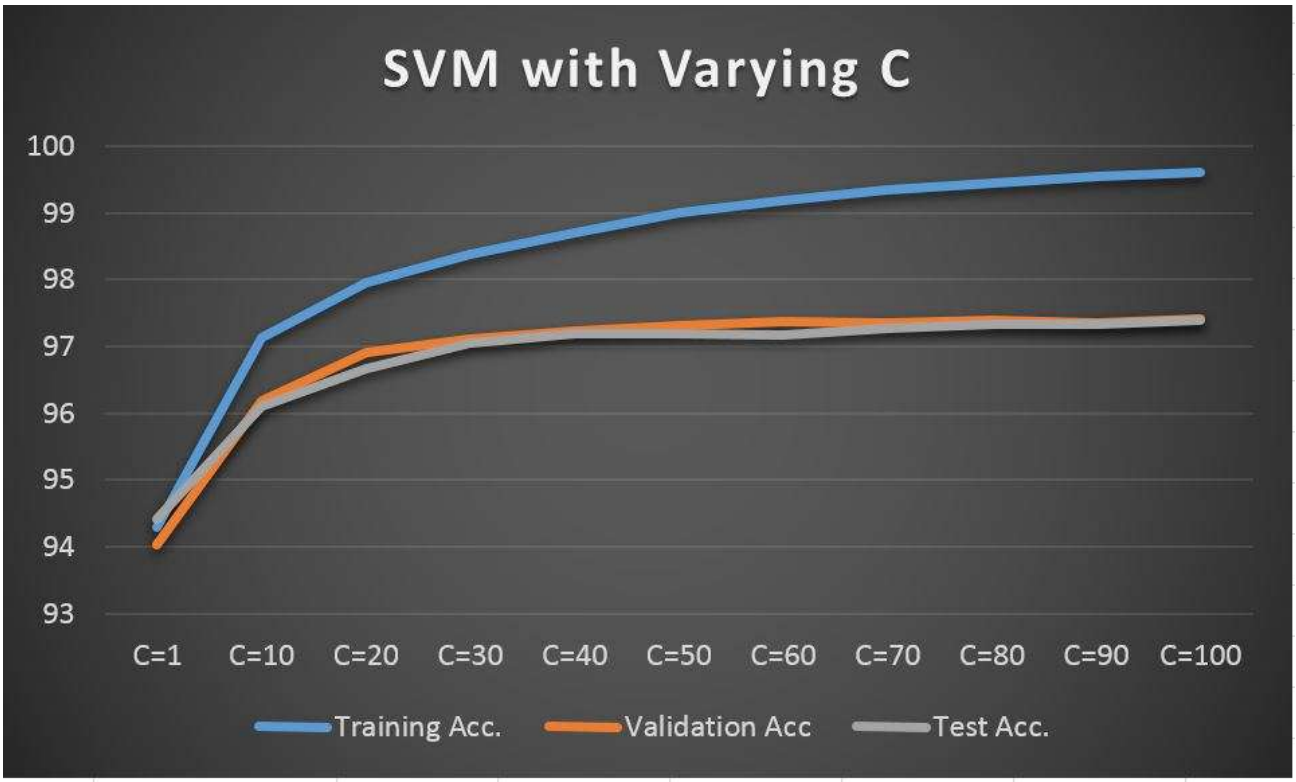
Test set Accuracy: 97.4%

Runtime: 16767.777 s

Runtime for entire SVM part: 57024.794 s

Runtime for entire code: 57560.342 s

Graph



Explanation:

The classification methods we test in this assignment are Logistic Regression and SVM. Logistic Regression works with two classes: one which has the maximum probability (the predicted class) and the other class (all other classes of input data with lower probabilities). This is not very good with high-dimensional data like we have. So, SVM does much better in this scenario since it uses kernels. Also by varying C which is the error penalty, we get a much better accuracy with higher C values which means that the objective function is well fit to classify the data of this type.

Conclusion:

In SVM, **For RBF with gamma = 1.0** the accuracies for validation and test data are much lower than that of training data. Training data accuracy is 100.0%, while validation set accuracy is 15.48% and test set accuracy is 17.14%. This is because there is over-fitting of data set when gamma value is 1.

For SVM with varying C values, graph for training, test and validation data for varying values of C = 1,10,20,30 … 100 is plotted. We find that as the value of C increases, the accuracy gradually increases and then maintains almost the same value.