CSE 474

Programming Assignment 3

Classification and Regression

Group 27

Group Member:

Xingtong Li (xli228)

Shuo Qiang (shuoqian)

Xuhui Lin (xlin44)

Introduction:

In this assignment, we are aim to implement Logistic Regression and use the Support Vector Machine toll to classify handwritten digit images and compare these methods.

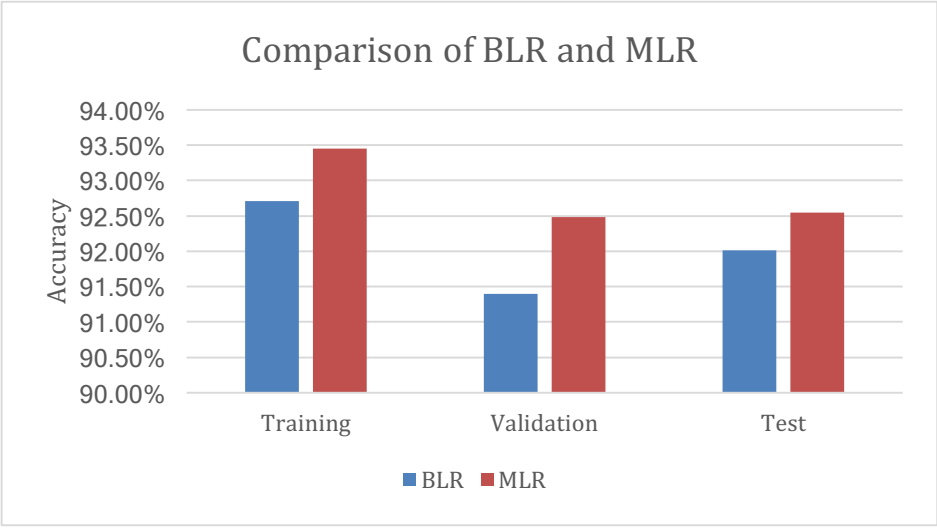
Part 1: Binary Logistic Regression (One-Vs-One)

Accuracy		
Training Set	Validation Set	Testing Set
92.71%	91.4%	92.01%

Part 2: Multi-class Logistic Regression (One-Vs-All)

Accuracy		
Training Set	Validation Set	Testing Set
93.448 %	92.48%	92.55%

In Binary Logistic Regression, we construct one classifier for each class. So we will have the same number of classes and classifiers. However, Multi-class Logistic Regression can classify all classifiers simultaneously. Since the given data is classified over ten classes, MLR obtains a better accuracy than BLR. The result we produced shows that each set's accuracy that MLR achieved is higher than BLR achieved. Therefore, MLR more efficient than BLR in general. Below is the comparison of BLR and MLR graph.



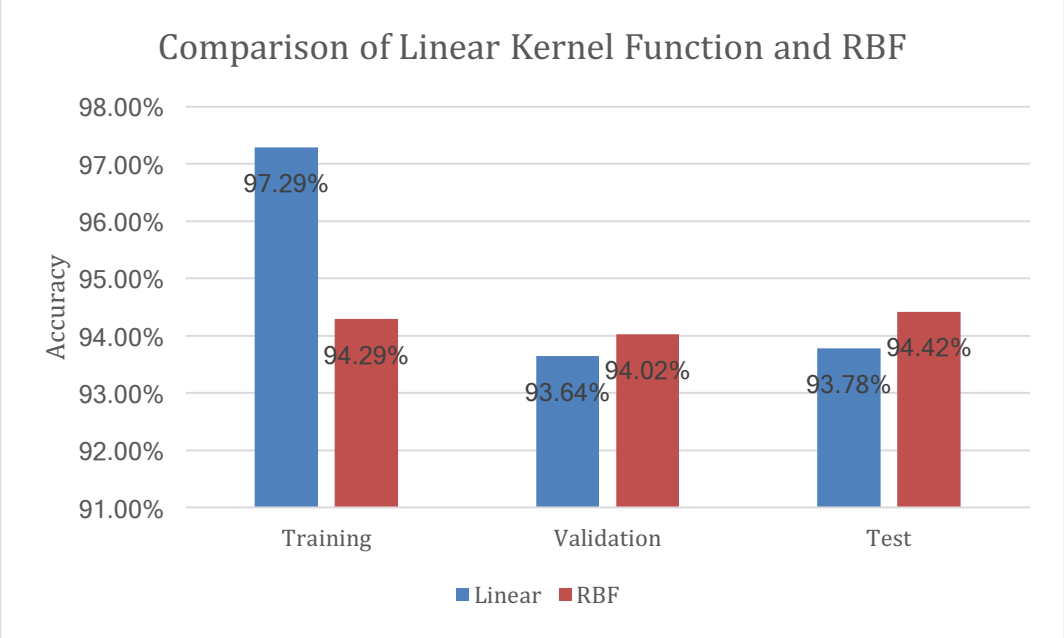
Part 3: Support Vector Machines

SVM with Linear Kernel		
Training Set	Validation Set	Testing Set
97.286%	93.64%	93.78%

Linear Kernel tends to performs better when the decision boundary is linear. It separates the points in a linear space.

SVM Accuracy (radial basis function, gamma = default)		
Training Set	Validation Set	Testing Set
94.294%	94.02%	94.42%

RBF Kernel performs better with non-linear boundary. Instead of a linear pattern, it utilized curves around the data sets. RBF gives better accuracy than linear kernel in general. Below is the graph of comparison of Linear Kernel Function and Radial Basis Function.



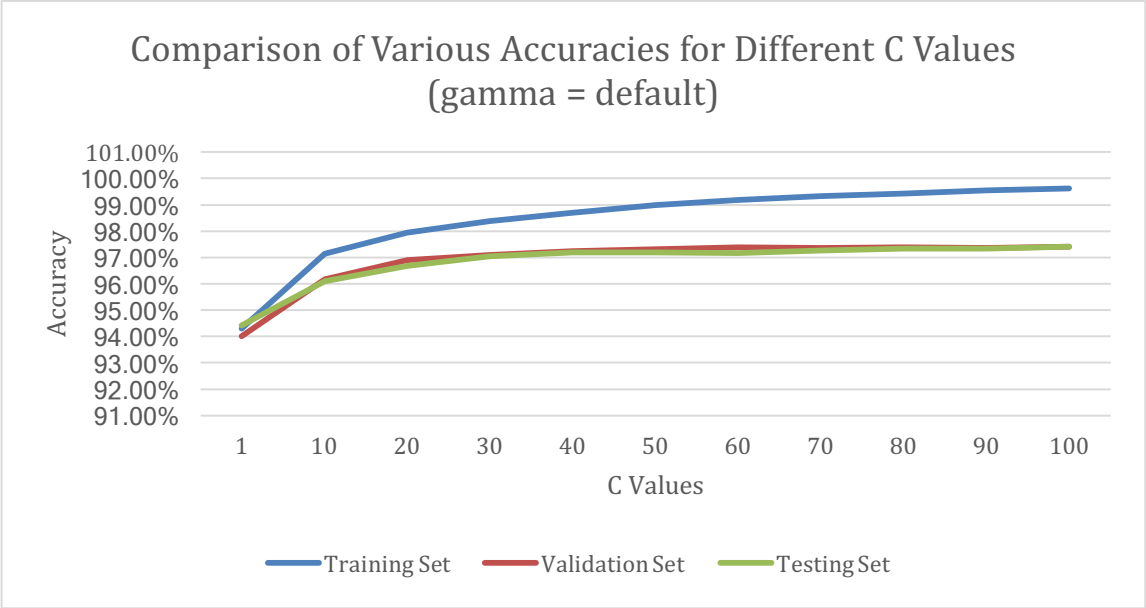
SVM Accuracy (radial basis function, gamma = 0.1)

Training Set	Validation Set	Testing Set
99.992 %	94.76 %	94.96 %

When gamma is set to 0.1, compared to when gamma is set to default, we observed that the training accuracy is 99.992 % which is much higher than gamma was default, even though other accuracies on test are lower (compared to the training set in this function), they still have better result than gamma setting to default. Hence, in this case, this model performs better with gamma setting to 0.1.

SVM Accuracy (radial basis function, gamma = default and different values of C)			
C	Training Set	Validation Set	Testing Set
1	94.294%	94.02%	94.42%
10	97.132%	96.18%	96.1%
20	97.952%	96.90%	96.67%
30	98.372%	97.1%	97.04%
40	98.706%	97.23%	97.19%
50	99.002%	97.31%	97.19%
60	99.196%	97.38%	97.16%
70	99.340%	97.36%	97.26%
80	99.438%	97.39%	97.33%
90	99.542%	97.36%	97.34%
100	99.612%	97.41%	97.40%

Below is the graph of comparison of various accuracies of different C values when gramma is set to default.



As the value of C increases, the accuracy also increases. A lower C makes tiny amount of error is acceptable during classification which results in more examples can be misclassified. As C increases, the training data will be classified better when the model select more samples from support vectors and lesser data will be misclassified by doing the training phase.

Besides, as the C value increasing, the accuracy of training set keep increasing but the accuracies of validation and testing set stay flat approximately after C equal to 30. This is a overfitting sign.

Hence, when C increase, the accuracy of the test set increases. Especially in the roughly range of 0 to 40.