CSE 574 Introduction to Machine Learning

Programming Assignment 3

Classification and Regression

Group 39

Rohith Doraiswamy Konoor (50169785)
Falguni Bharadwaj (50163471)
Malavika Reddy Tappeta (50169248)

In this assignment, we extend the first programming assignment in solving the problem of handwritten digit classification. We implement Logistic Regression and use the Support Vector Machine tool to classify hand-written digit images and compare the performance of these methods. We also implemented multi-class logistic regression and found that we got better accuracy for it as compared to normal logistic regression.

Logistic Regression:

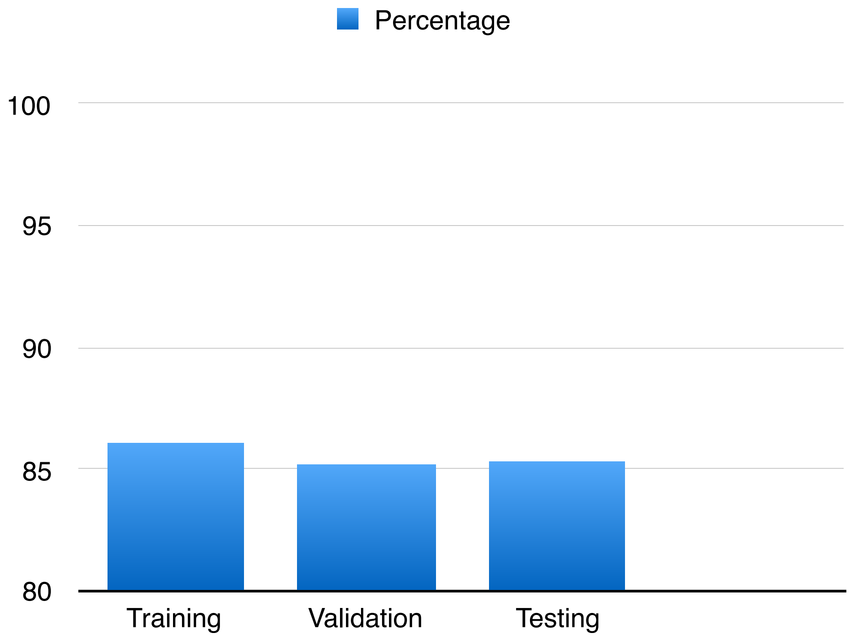
We implement the logistic regression model for handwritten digit recognition with the help of binary classifiers for each digit. Logistic regression simply uses binary classification as compared to linear regression. We code the blrObjFunction() and blrPredict() function to check the results generated using logistic regression. We run the module on training data, validation data and test data to generate the accuracy and performance of the module.

The accuracies that we obtained from the implementation are :

Training Set Accuracy : 86.086%

Validation Set Accuracy : 85.23%

Testing Set Accuracy : 85.31%



Support Vector Machine:

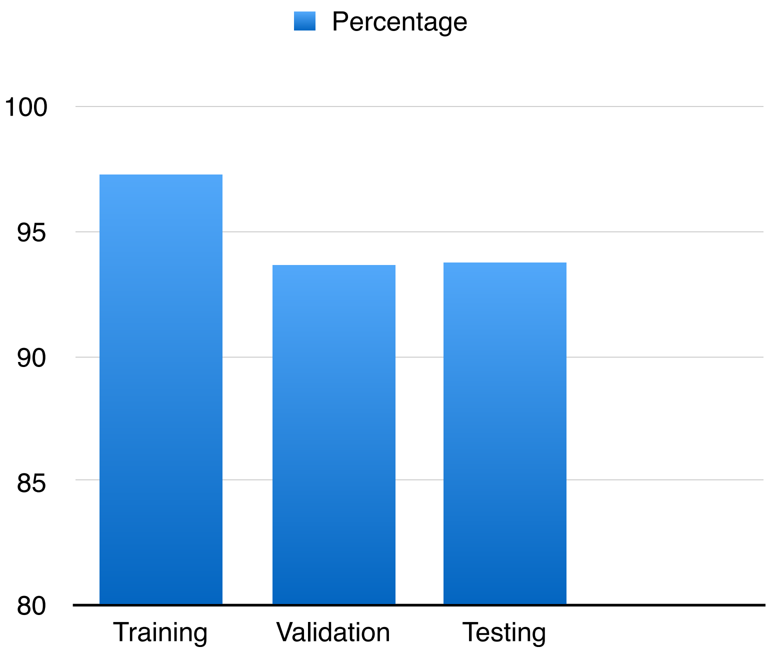
A support vector machine constructs a hyperplane or set of hyperplanes in a high- or infinite-dimensional space, which can be used for classification, regression, or other tasks. Intuitively, a good separation is achieved by the hyperplane that has the largest distance to the nearest training-data point of any class, since in general the larger the margin the lower the generalization error of the classifier.

Linear Kernel

Training Set Accuracy : 97.286%

Validation Set Accuracy : 93.64%

Testing Set Accuracy : 93.78



Radial Basis Function (Gamma = default)

Training Set Accuracy : 94.294%

Validation Set Accuracy : 94.02%

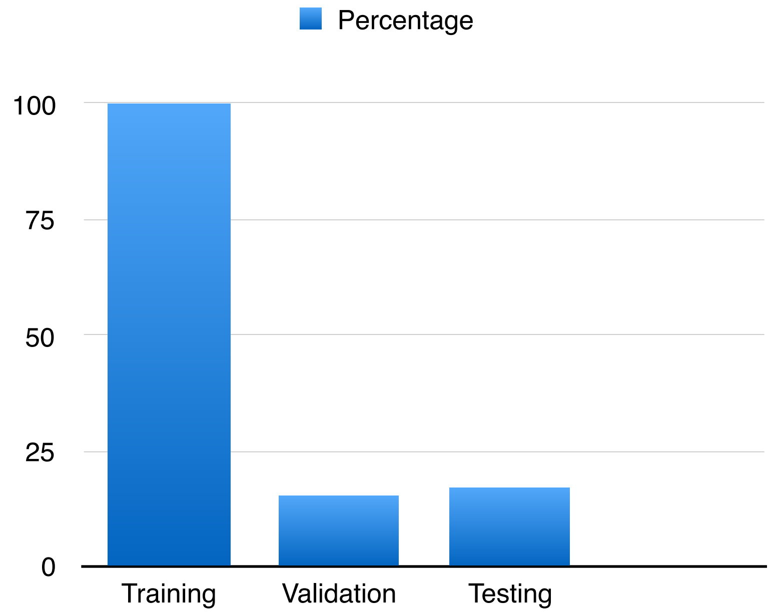
Testing Set Accuracy : 94.42%

Radial Basis Function (Gamma = 1)

Training Set Accuracy : 100.0%

Validation Set Accuracy : 15.48%

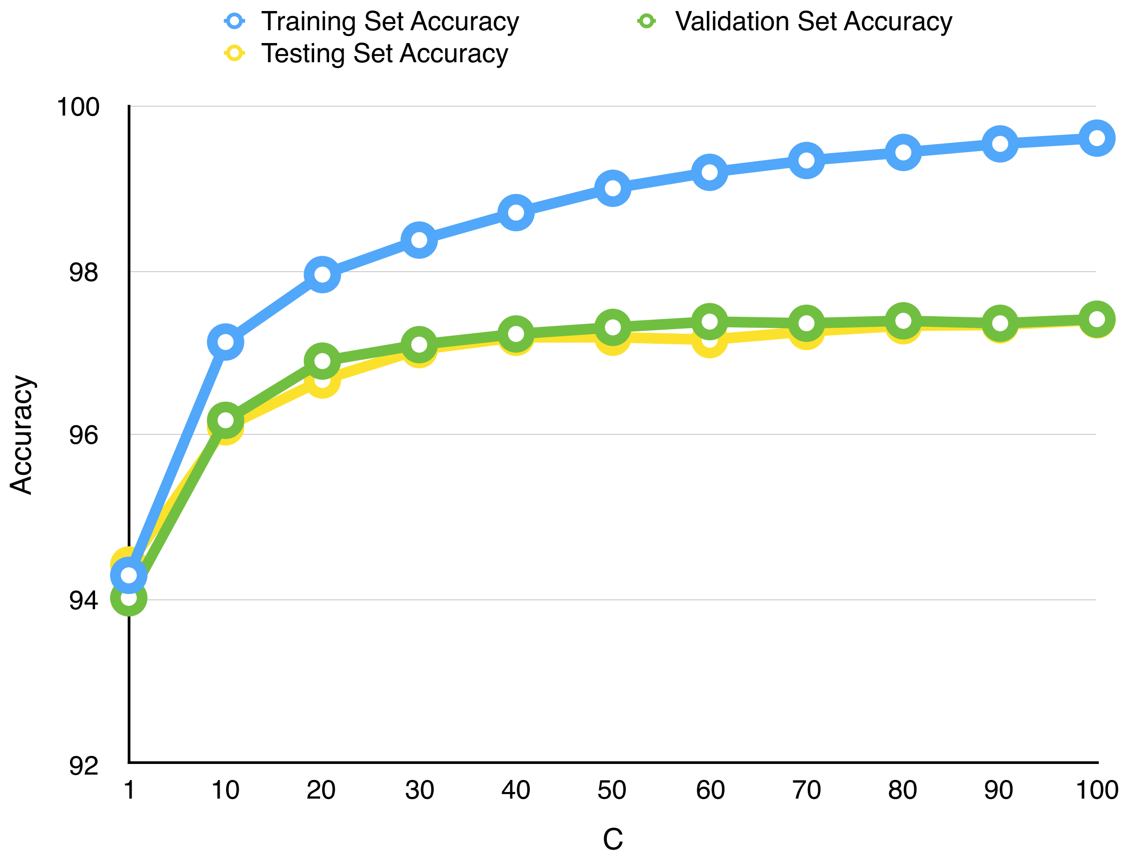
Testing Set Accuracy : 17.14%



Radial Basis Function (**Gamma = default, C = 1,10,20…100**)

	C=1	C=10	C=20	C=30	C=40	C=50	C=60	C=70	C=80	C=90	C=100
Training Set Accuracy %	94.294	97.132	97.952	98.372	98.706	99.002	99.196	99.34	99.438	99.542	99.612
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	97.36	97.33	97.34	97.4

Graph of accuracy with respect to values of C:



Multi Class Logistic Regression:

The multi class logistic regression model is implemented for the handwritten digit recognition in the functions mlrObjFunction() and mlrPredict() function. This model implements different discrete values instead of just binary values like logistic regression. We use the given softmax function instead of the sigmoid function in this phase. The results produced are the effect of the probabilities of the discrete values that are used to classify the testing data. We implement this method for the training data, validation data and test data to generate accuracy and performance for multi class logistic regression.

The accuracies that we obtained from the implementation are :

Training set Accuracy: 93.39%

Validation set Accuracy: 92.43%

Testing set Accuracy: 92.67%

Conclusion:

From the values obtained for accuracy we observe that Logistic Regression has low accuracy values when compared to SVM. When gamma is 1 and all others parameter values are set to default the accuracy for training dataset is 100% but for testing and validation dataset it is very low. However, the accuracy of SVM increases with increasing value of C and records very high accuracy, while gamma value is default. We also noticed that multi class regression has higher accuracy values as compared to logistic regression.