State University of New York at Buffalo

CSE 574 – Introduction to Machine Learning

Spring 2017

Programming Assignment -3

Classification and Regression Project Report

Group 32

Ajay Gandhi – 50207403 – agandhi3@buffalo.edu

Apoorva Hejib – 50206516 – apoorvah@buffalo.edu

Siddharth Pateriya – 50206348 – spateriy@buffalo.edu

Logistic Regression:

Logistic Regression is used to estimate/predict values taking into account all the data corresponding to dependant variables which can take 2 values (binary). **For binary logistic regression** on the MNIST dataset we obtained the following results:

Training set Accuracy: 86.176%

Validation set Accuracy: 85.25%

Testing set Accuracy: 85.41%

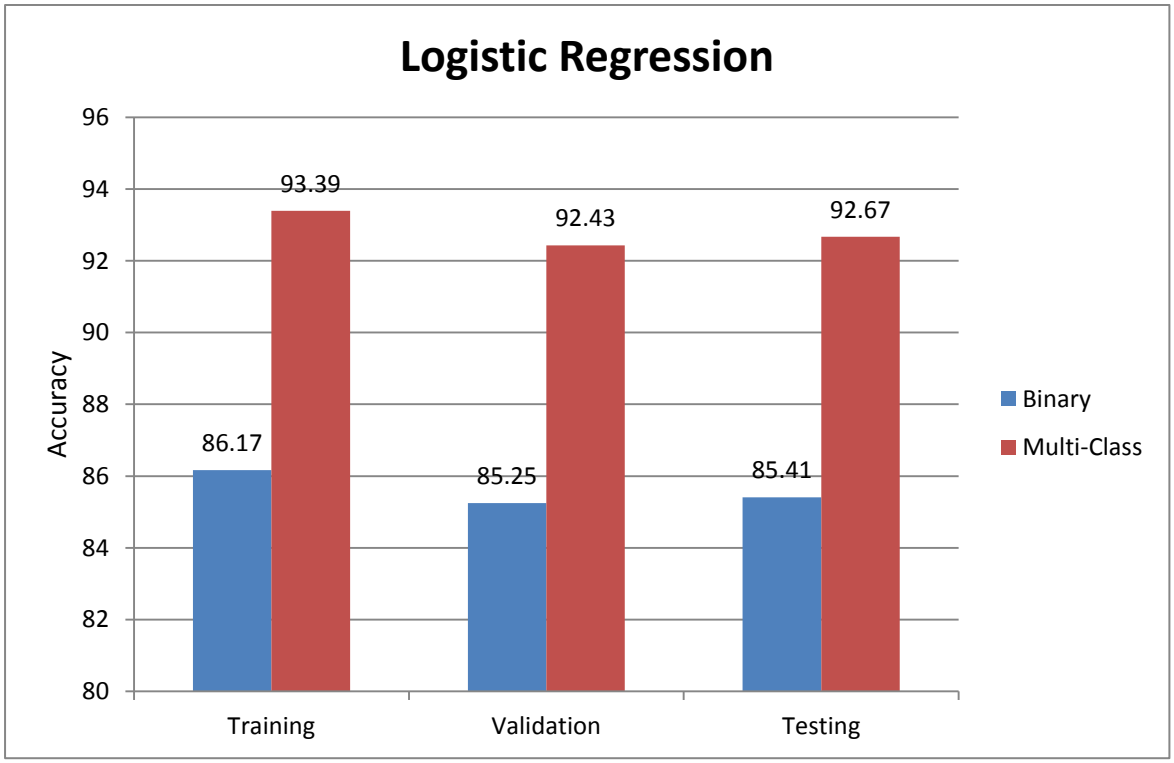
Direct Multi-class Logistic Regression (EXTRA CREDIT)

Training set Accuracy: 93.39%

Validation set Accuracy: 92.43%

Testing set Accuracy: 92.67%

We can observe from the results above that **Multi-Class Regression** approach gives a much better result than Binary Logistic Regression for the MNIST Data Set.



Support Vector Machines

The results of using SVMs on the MNIST data set with various different settings are as follows:

Using Linear Kernel

Training set Accuracy: 97.286%

Validation set Accuracy: 93.64%

Testing set Accuracy: 93.78%

We observe that the Linear Kernel SVM delivers a very good accuracy of 93.78%.

Using Radial Basis Function - Gamma = 1 (All other parameters kept default)

Training set Accuracy: 100.0%

Validation set Accuracy: 15.48%

Testing set Accuracy: 17.14%

We observe that with Gamma = 1, the SVM using the RBF kernel, the Training accuracy achieved is 100% but validation and testing accuracy is very low. This can be attributed to overfitting, since the model has been trained on the complete data set and does not respond to fluctuations in data very well. Gamma = 1 makes the model be influenced by all the training data examples.

Using Radial Basis Function - Keeping Gamma default

Training set Accuracy: 94.294%

Validation set Accuracy: 94.02%

Testing set Accuracy: 94.42%

Using Radial Basis Function – Default Gamma

Varying C values from 1 to 100

On varying values of C using the RBF Kernel and keeping Gamma as default(0), we observe an increase an accuracy with increase in the value of C on the Testing Data. The increase in accuracy becomes lesser in magnitude as C approaches 100, and finally we have an accuracy of 97.4% on Testing Data when C = 100.

SVM Accuracies with Default Gamma			
	Accuracies		
C	Training	Validation	Testing
1	94.294	94.02	94.42
10	97.132	96.18	96.1
20	97.952	96.9	96.67
30	98.372	97.1	97.04
40	98.706	97.23	97.19
50	99	97.31	97.19
60	99.196	97.38	97.16
70	99.34	97.36	97.26
80	99.438	97.39	97.33
90	99.54	97.36	97.34
100	99.612	97.41	97.4