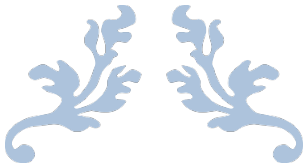


CSE 574: INTRODUCTION TO MACHINE LEARNING – ASSIGNMENT 3

HANDWRITTEN DIGITS CLASSIFICATION (CONTINUE)

TEAM MEMBERS:

Ankit Sarraf (5009 7190)

Karthick Krishna Venkatakrishan (5009 8126)

Rahul Singh Chauhan (5009 7213)

OVERVIEW OF THE PROJECT:

This project is the continuation of the Project 1 that was also based on the Handwritten Digit Recognition. In this project the classification of handwritten digits is done with the help of two Major techniques: The logistic regression and the Support vector Machines methods.

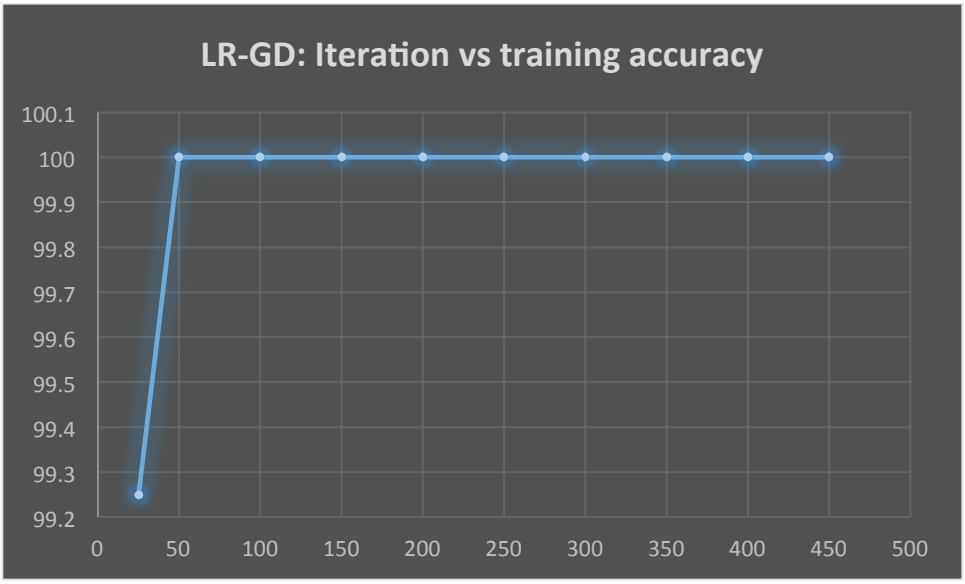
The first set of calculation is done by considering newdataset_MLR.mat

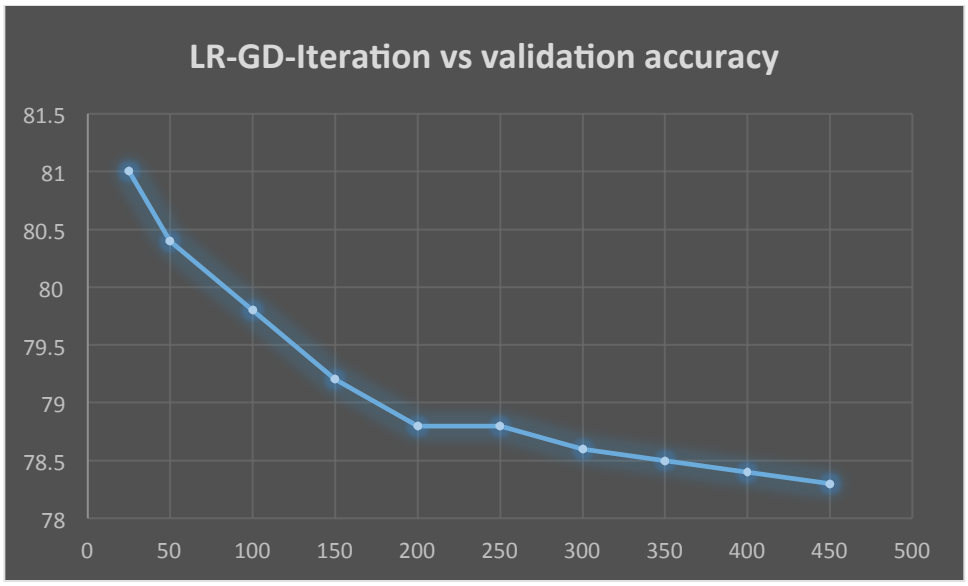
LOGISTIC REGRESSION METHODS:

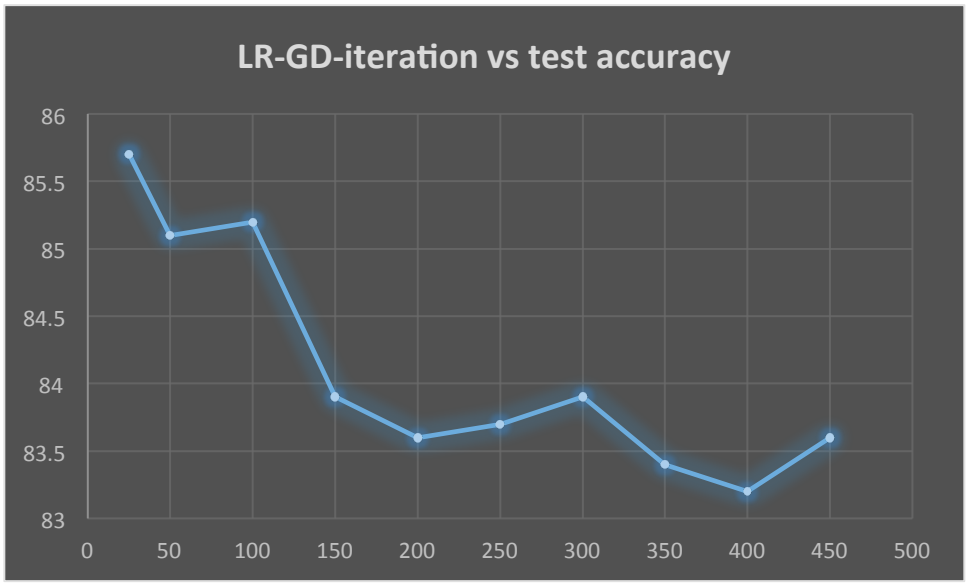
Logistic regression method classifies data into two classes 1 and 0. 1 signifies the correct classification and 0 signifies the incorrect classification. There are of two methods:

A) Logistic regression with gradient descent method:

MAXITER	TIME (sec)	TRAINING ACCURACY	VALIDATION ACCURACY	TESTING ACCURACY
25	7.216253	99.25	81	85.7
50	12.84138	100	80.4	85.1
100	25.9679	100	79.8	85.2
150	38.51027	100	79.2	83.9
200	51.21458	100	78.8	83.6
250	67.24676	100	78.8	83.7
300	73.93825	100	78.6	83.9
350	87.98598	100	78.5	83.4
400	94.0387	100	78.4	83.2
450	98.09731	100	78.3	83.6



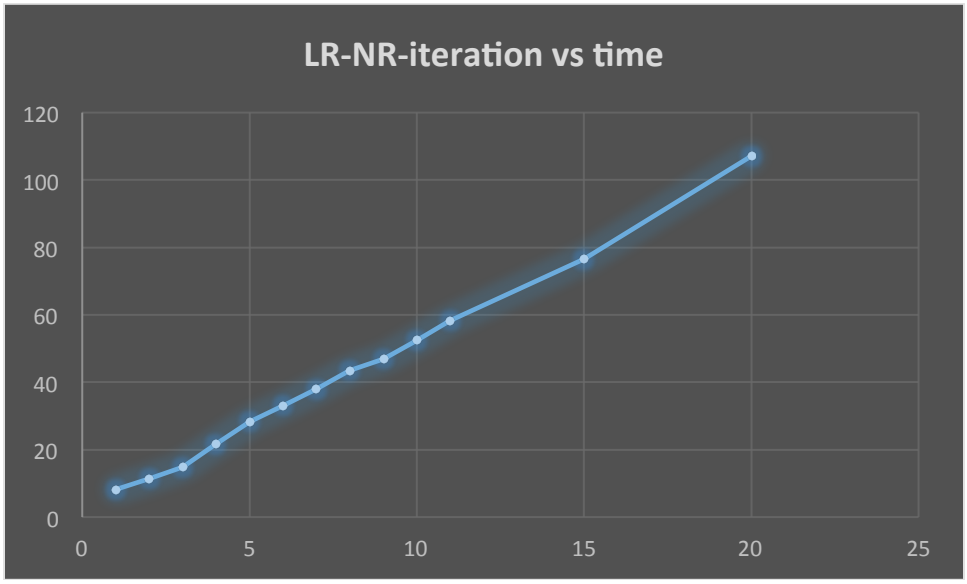


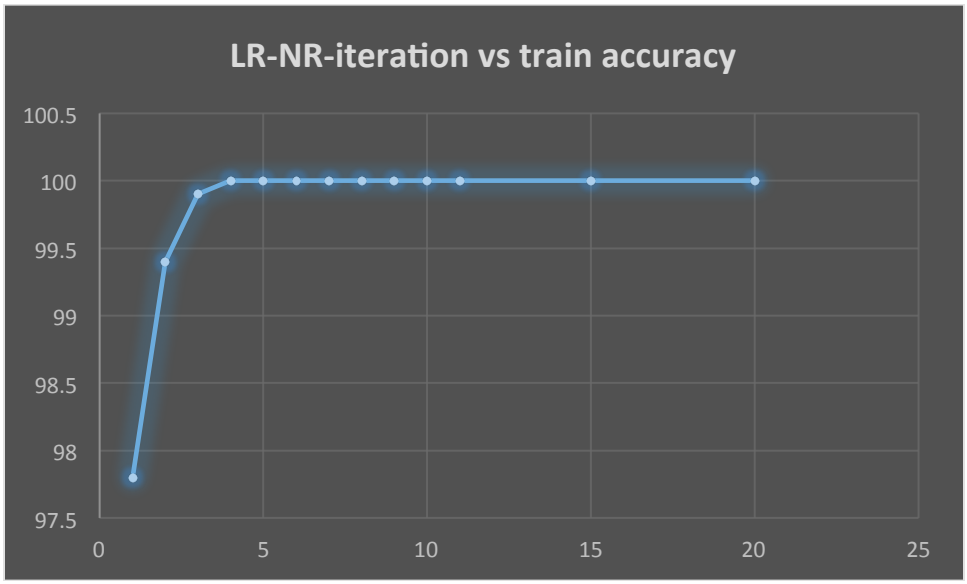


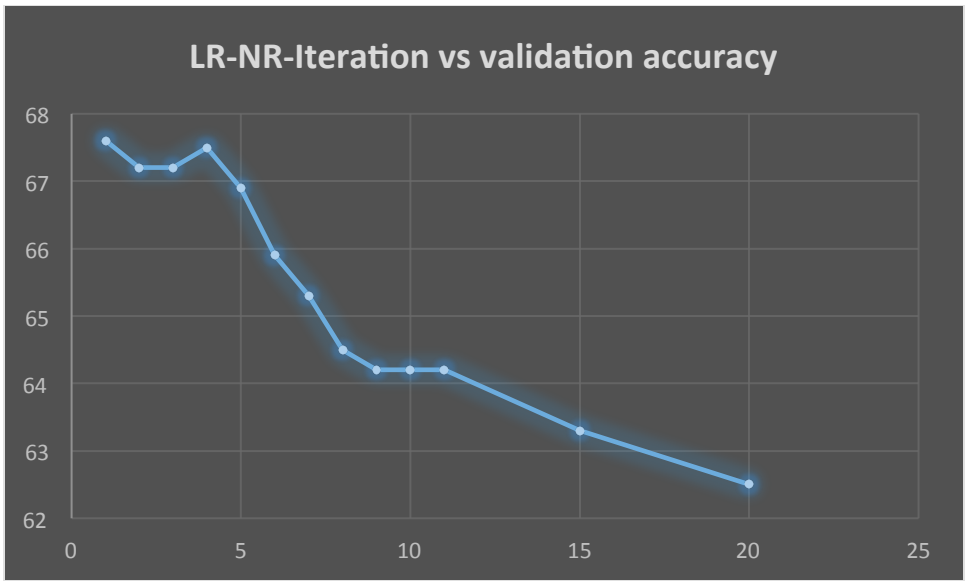
B) Logistic Regression with Newton Raphson Method:

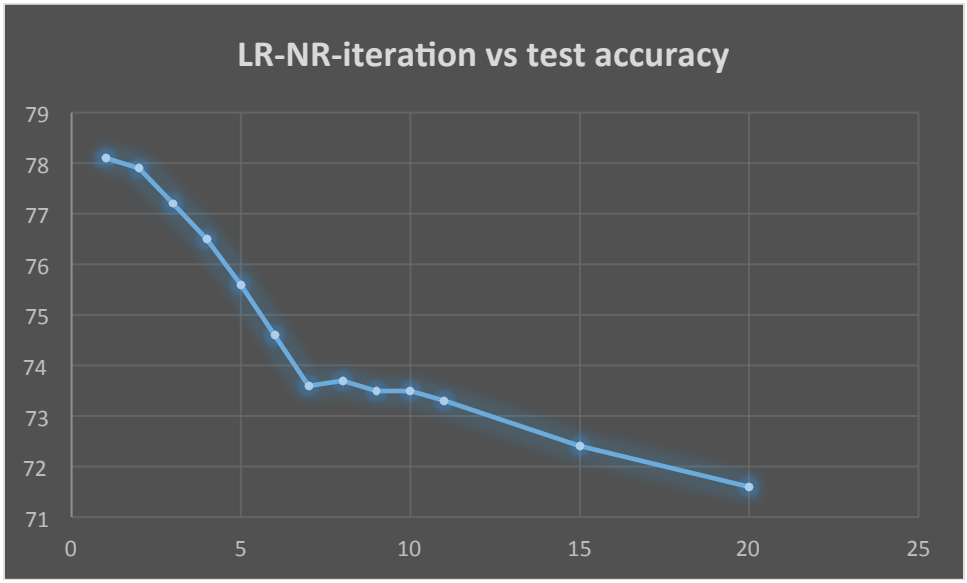
n_iter	time(seconds)	training accuracy	validation accuracy	Testing accuracy
1	8.216603	97.8	67.6	78.1
2	11.42066	99.4	67.2	77.9
3	14.73769	99.9	67.2	77.2
4	21.63068	100	67.5	76.5
5	28.21516	100	66.9	75.6
6	32.95533	100	65.9	74.6
7	37.91855	100	65.3	73.6
8	43.44961	100	64.5	73.7

9	46.99368	100	64.2	73.5
10	52.4059	100	64.2	73.5
11	58.31123	100	64.2	73.3
15	76.54192	100	63.3	72.4
20	107.049	100	62.5	71.6







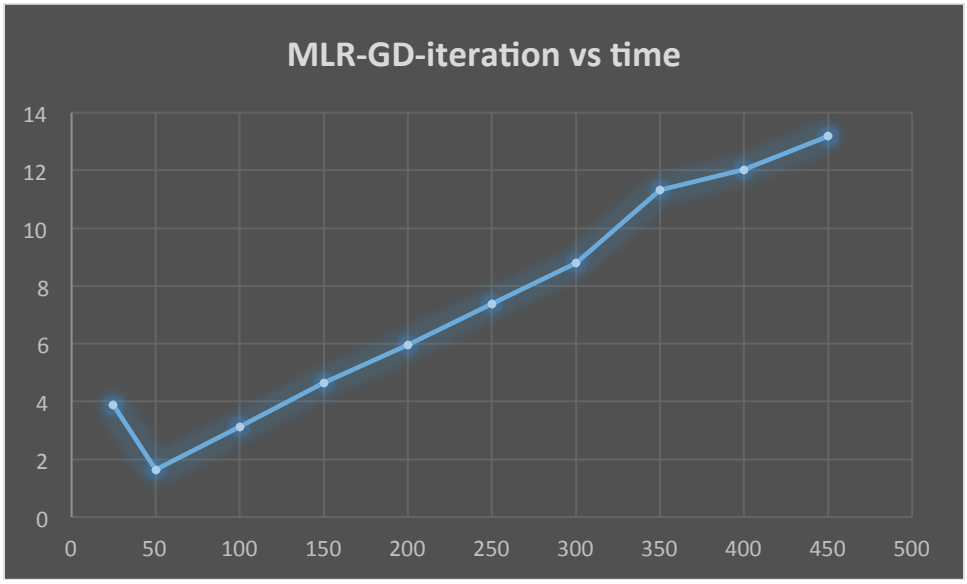


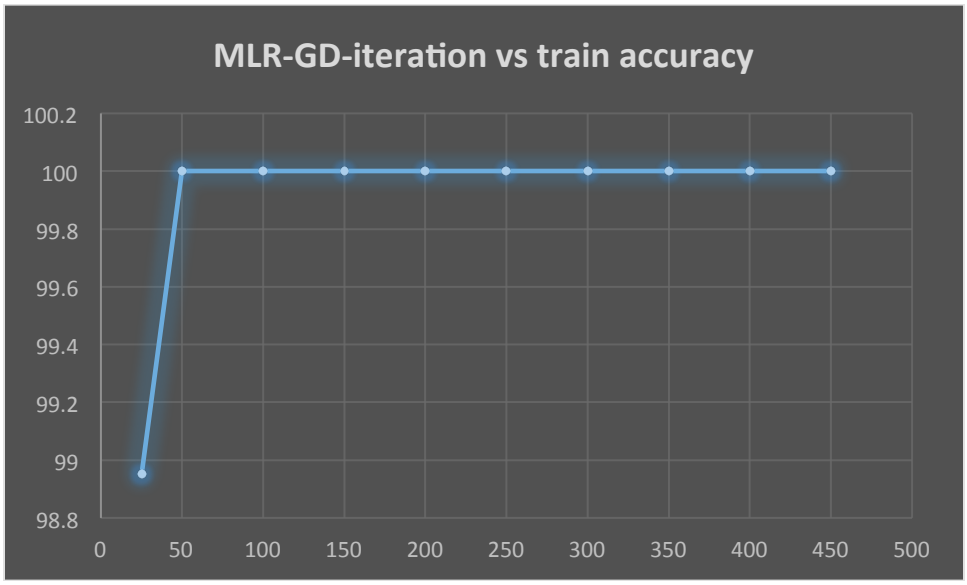
MULTICLASS LOGISTIC REGRESSION METHODS

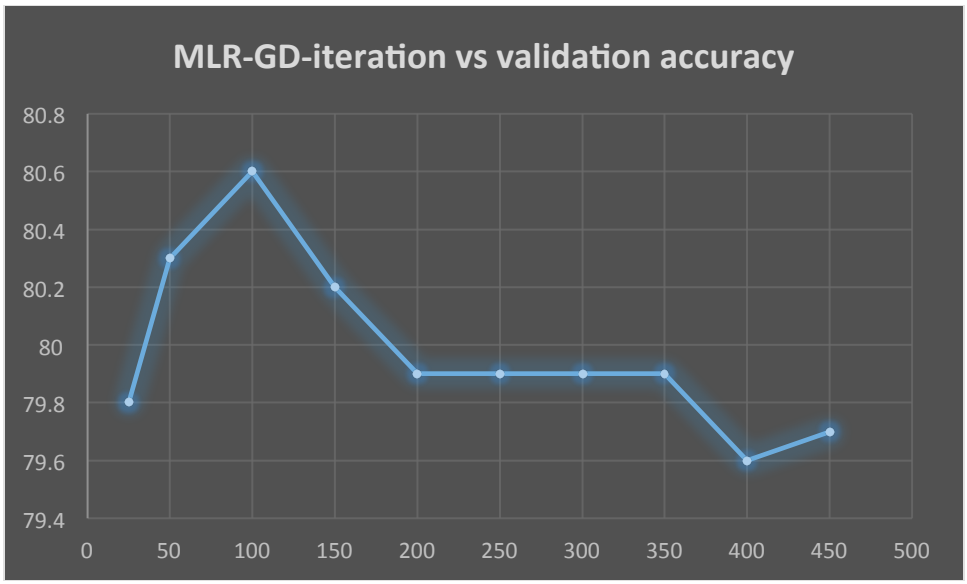
A) Multiclass Logistic Regression with Gradient Descent:

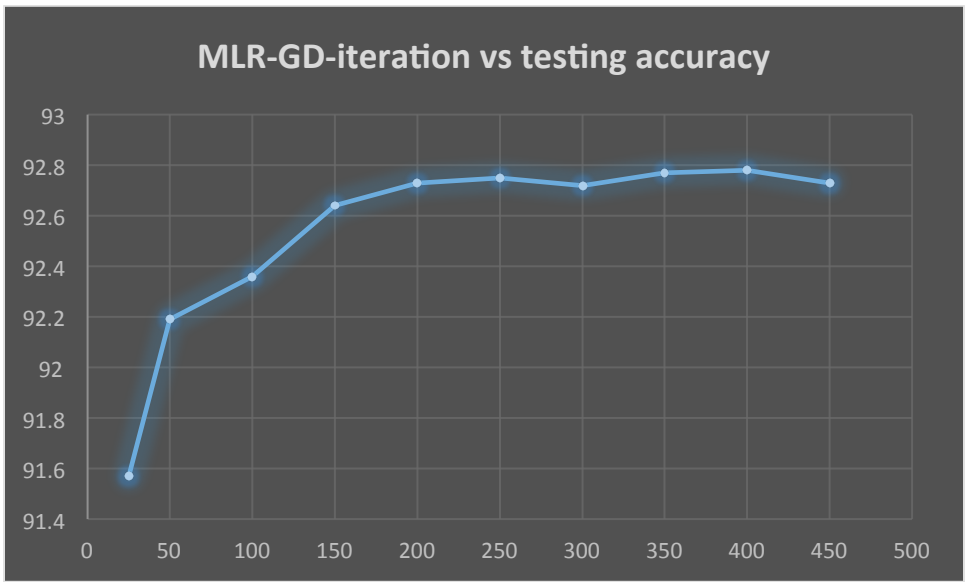
max_iter	time	train accuracy	validation accuracy	test accuracy
25	3.870692	98.95	79.8	85.7
50	1.641978	100	80.3	86
100	3.116163	100	80.6	86.6
150	4.648958	100	80.2	85.7
200	5.95051	100	79.9	85.6

250	7.384974	100	79.9	85.5
300	8.783945	100	79.9	85.7
350	11.32617	100	79.9	85.6
400	12.02484	100	79.6	85.6
450	13.17039	100	79.7	85.5









B) Multiclass Logistic Regression with Newton Raphson Method:

Following are the statistics for n_iter = 5:

Training Set Accuracy: 97.800000

Validation Set Accuracy: 67.600000

Test Set Accuracy: 78.100000

Total Processing Time: 1314.217903 seconds

Following is the second set of calculation is done by considering dataset.mat.

When the original data sets for the assignment 1, the following results were obtained:

A) Logistic regression with gradient descent:

MAXITER	TIME (sec)	TRAINING ACCURACY	VALIDATION ACCURACY	TESTING ACCURACY
25	125.6141	91.082	90.77	91.47
50	239.0117	92.144	91.71	91.65
100	456.6245	92.59	91.33	91.85
150	698.5539	92.818	91.52	91.85
200	934.169	92.942	91.43	91.85
250	1234.581	93.002	91.43	91.85
300	1415.345	93.112	91.45	91.89
350	1642.541	93.114	91.43	91.89
400	1791.608	93.136	91.4	91.83
450	2135.08	93.146	91.39	91.83

B) Logistic regression with Newton Raphson Method:

n_iter	time(sec)	training accuracy	validation accuracy	Testing accuracy
1	35.9226	85.79	85.3	86.05
2	68.37192	88.624	87.72	88.46
3	101.597	90.43	89.36	89.97
4	133.5143	91.69	90.21	90.72
5	179.566	92.48	90.55	91.13
6	201.2639	92.91	90.99	91.34
7	228.0712	93.156	91.13	91.38
8	255.4136	93.28	91.2	91.5
9	306.7942	93.34	91.04	91.42
10	344.9076	92.908	90.59	91.08
11	366.8767	58.136	58.09	57.99
15	502.5332	15.94	16.08	15.42
20	645.6863	11.408	11.63	11.55

C) Multiclass Logistic Regression with Gradient Descent:

MAXITER	TIME (sec)	TRAINING ACCURACY	VALIDATION ACCURACY	TESTING ACCURACY
25	18.59025	91.204	90.85	91.57

50	38.49595	92.188	91.94	92.19
100	72.17514	93.186	92.56	92.36
150	108.0829	93.368	92.44	92.64
200	133.3177	93.594	92.43	92.73
250	160.6679	93.816	92.37	92.75
300	190.2228	93.836	92.38	92.72
350	218.3292	93.868	92.38	92.77
400	248.9949	93.928	92.35	92.78
450	283.8206	93.998	92.31	92.73

D) Multiclass Logistic Regression with Newton Raphson Method:

The old data set was very large and hence it was not tested using Multiclass Newton Raphson method. It has been tested using only the new dataset as mentioned above in this report.

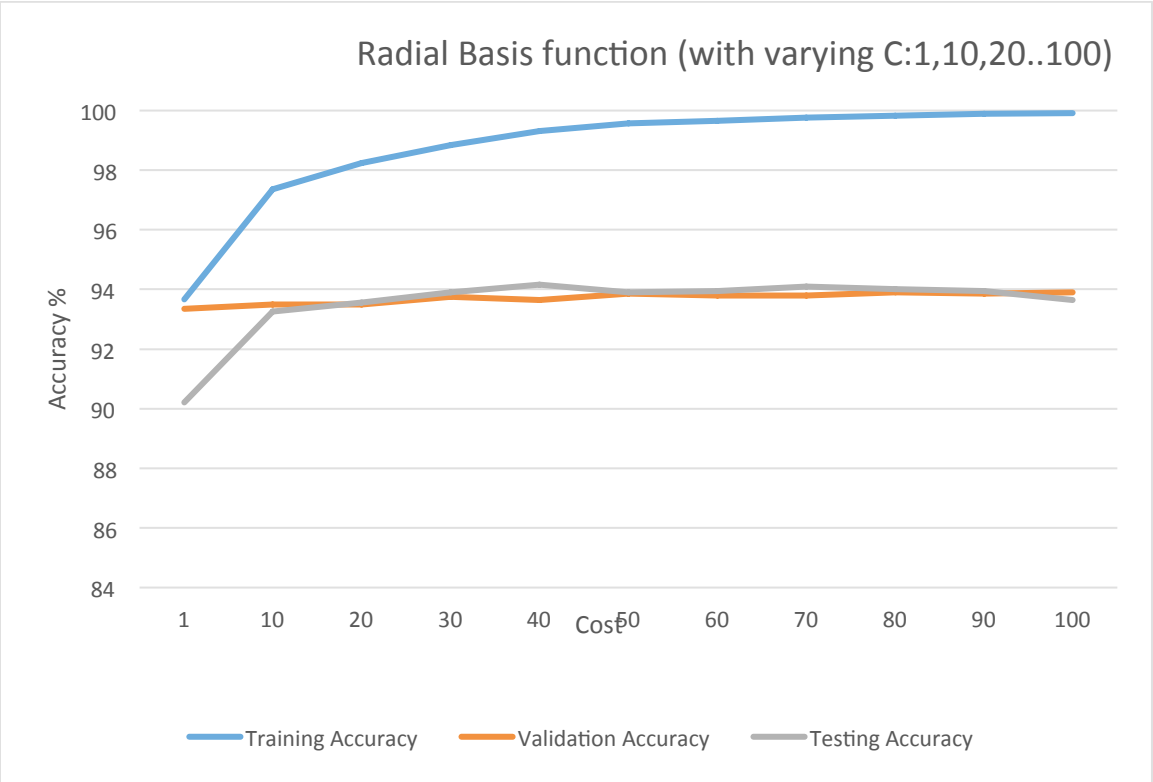
COMPARISION OF NEWTON RAPHSON AND GRADIENT DESCENT METHODS:

The statistics suggest that The Gradient descent method is better than the Newton-Raphson method since it gives a better efficiency and eliminates the need to find the second derivative.

Although the Newton Raphson method is of higher order (complexity is high) because the Hessian (Second order derivative of error function) has to be found, it requires lesser number of iterations as compared to the Gradient Decent methods.

Support Vector Machines

Support Vector Machine is based on the concept of decision places that define decision boundaries. A decision plane is one that separates between a set of objects having difference class memberships.



Max value of Test Accuracy = 94.13% is obtained at C = 40.
Max value of Validation Set accuracy = 93.9% is obtained at C = 80.

Accuracy Result for model_rbf_c:

Cost	Training Accuracy	Validation Accuracy	Testing Accuracy
1	99.84%	91.15%	90.85%
10	93.66%	90.35%	90.20%
20	98.25%	93.50%	93.55%
30	98.83%	93.75%	93.90%
40	99.32%	93.65%	94.15%
50	99.58%	93.85%	93.90%
60	99.65%	93.80%	93.95%
70	99.76%	93.80%	94.10%
80	99.83%	93.90%	94%
90	99.89%	93.85%	93.95%
100	99.91%	93.90%	93.65%

REFRENCES:

https://www.statsoft.com/textbook/support-vector-machines
Pattern Recognition and Machine Learning by Christopher M. Bishop

EXTRA POINTS TO BE NOTED:

'RESULTS DATA AND PARAMS' directory contains the files to represent the data that was gathered on the final execution on Mac 10.9 X OS (with 8 GB RAM and 1.7 Intel Processor).

Final execution of the SVM Part was done on Windows system (i7 Processor with 8 GB RAM).

These results might differ from the ones, which have been included in the Report. The testing was done on the Window (4 GB RAM with i5 Inter Intel Processor). Result from those executions has been included in the report.

Also, for the SVM, the path hasn't been added in the script. Although the necessary lib files has been included in the Base Code folder.

This directory also includes the params.mat that contains all the values of the variables specific to SVM.

params1.mat is the file which contains the 'W_blr' variable ONLY.