CSE 574: INTRODUCTION TO MACHINE LEARNING

PROGRAMMING ASSIGNMENT – 3:

Classification and Regression

Group – 40:

50413342 - Bodhith Edara
50409091 - Teja Reddy Alla
50417093 - Venkata Krishna Sai Sakhamuri

1. LOGISTIC REGRESSION

- **Task:**

  Implement Logistic Regression to classify hand-written digit images into correct corresponding labels (Binary Logistic Regression).

- **Classification results and accuracy.**

Data Set	Accuracy	Error
Training	92.69	7.31
Validation	91.44	8.56
Testing	91.99	8.01

- **Observations:**

  Logistic regression takes into account all data points in order to create a hyperplane that separates the data into classes. As a result, logistic regression performs well on data with a limited number of input features.

  The provided data is not linearly separable and contains more input features (dimensions). Across all three data sets, our experiment resulted an accuracy of approximately 92 %.

  And the training error (7.31) is less than the testing error (8.01) using binary logistic regression as it performs better on known data and any linear model behaves the same with regard to error.

2. MULTI-CLASS LOGISTIC REGRESSION

- **Task:**

Implement multi-class Logistic Regression. Logistic Regression is commonly used for binary classification. Logistic Regression, on the other hand, can be extended to solve multi-class classification problems. We don't need to build ten classifiers with this method. Instead, we only need to create one classifier that can classify ten different classes at the same time.

- **Classification results and accuracy.**

Data Set	Accuracy	Error
Training	93.138	6.862
Validation	92.54	7.46
Testing	92.53	7.47

- **Observations:**

Across all three data sets, our experiment yielded an accuracy of about 93 percent.

Here, the training error(6.86) is less than the testing error(7.47) using multi class logistic regression as it performs better on known data and any linear model behaves the same with regard to error.

BINARY LOGISTIC REGRESSION (One vs All)

VS

MULTI-CLASS LOGISTIC REGRESSION:

Data Set	MLR Accuracy (%)	BLR Accuracy (%)
Training	93.138	92.69
Validation	92.54	91.44
Testing	92.53	91.99

- **Inference**:

We can see that multi-class logistic regression outperforms binary logistic regression in terms of accuracy. Each input in our data set belongs to exactly one class and we estimate the parameters independently in multi-class logistic regression.

As a result, multi-class logistic regression outperforms binary logistic regression.

3. SUPPORT VECTOR MACHINES

- **Task:**

Use the Support Vector Machine tool in sklearn and SVM to perform classification on the data set.

- **Using Linear Kernel to perform classification**

Data Set	Accuracy (%)
Training	92.75
Validation	91.28
Testing	91.72

- **Observations:**

The results yielded by the linear kernel SVM are similar to the previously used linear model as its working is same as that of the linear model.

Using Radial Basis Function:

a) Radial basis function with Gamma = 1:

Data Set	Accuracy (%)
Training	100
Validation	10.03
Testing	11.42

Observations:

We can see that the model performs poorly on test data due to over-fitting caused by high gamma value and it is evident from 100% training accuracy that over-fitting has occurred.

b) Radial basis function with Gamma = default:

Data Set	Accuracy
Training	91.958
Validation	92.02
Testing	92.47

Observations:

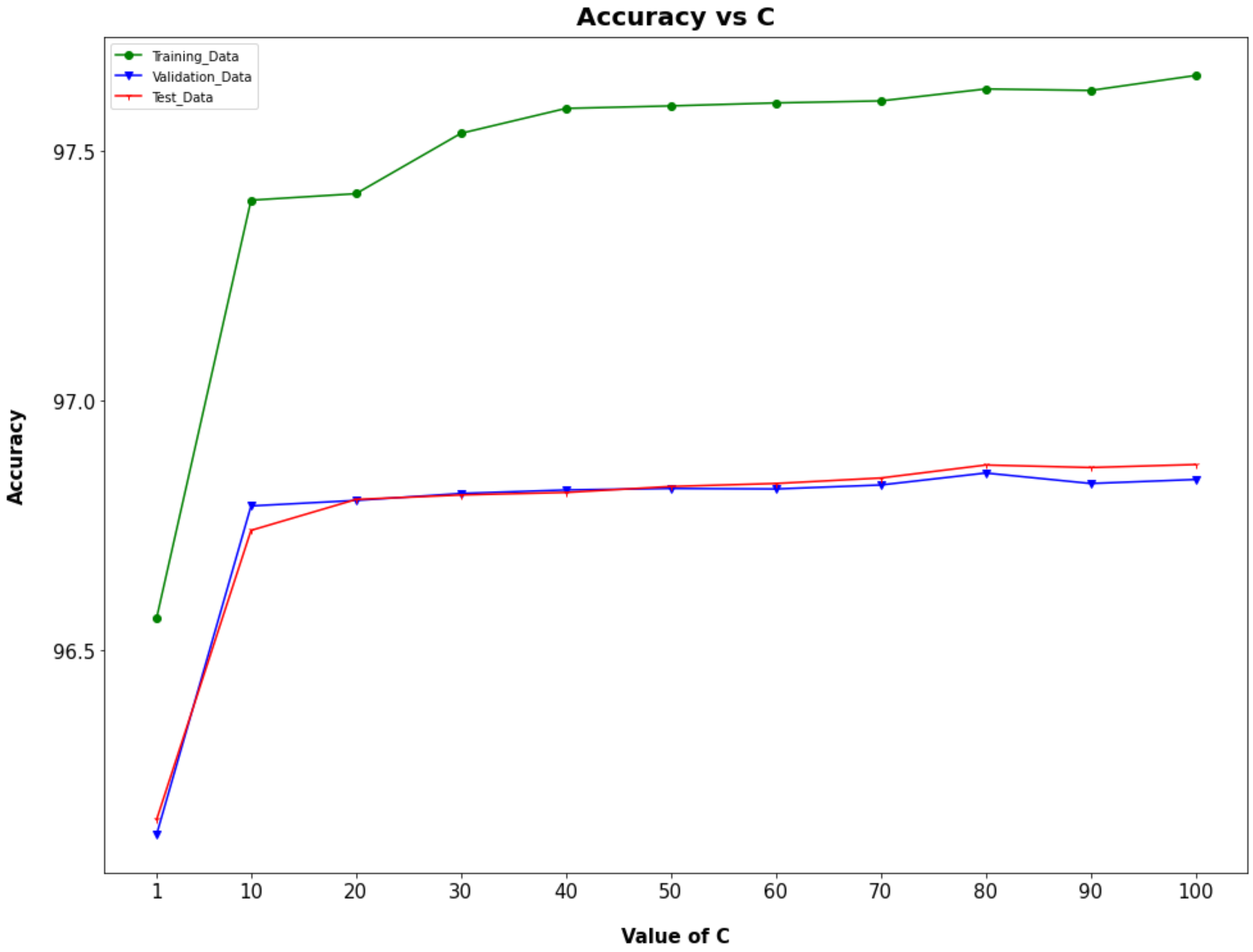
Here the model performs far better than the one with gamma=1 as it does not overfit.

c) Radial Basis Function with Gamma=default and with varying value of C (1, 10, 20, 30, …,100)

Here, we loop through the C values to find the ideal configuration, which we then use to test the entire dataset. This C variable determines how important the slack variable is to us. As a result, we can see a trade-off between margin width and the value of C.

C value	Training Accuracy	Validation Accuracy	Testing Accuracy
1	96.564	96.13	96.16
10	97.402	96.789	96.74
20	97.415	96.8	96.802
30	97.536	96.814	96.811
40	97.586	96.821	96.816
50	97.591	96.824	96.828
60	97.597	96.823	96.834
70	97.601	96.831	96.845
80	97.625	96.855	96.871
90	97.622	96.834	96.866
100	97.652	96.842	96.872

Plot for Accuracies (Training, Validation, Testing) vs C :



Observations:

Thus, we can see from the above table that the accuracies are optimal for C=80.

The accuracies for the whole dataset using optimal parameters (Gamma=default, C=80) are:

Training Accuracy:	99.33999999999999%
Validation Accuracy:	97.36%
Testing Accuracy:	97.26%

Inference: As we obtained optimal results using non-linear model, we can conclude that our dataset is non-linear in nature.