CSE 574 Introduction to Machine Learning

Programming Assignment 3

Classification and Regression

Group 4

Saurabh Bajoria	50208005
Sumedh Ambokar	50207865
Vidhi Shah	50207090

Task 1: Binary Logistic Regression

- **Logistic regression** is a statistical method for analyzing a dataset in which there are one or more independent variables that determine an outcome.
- **Binary Logistic regression** deals with situations in which the observed outcome for a dependent variable can have only two possible types, "0" and "1"

Below are the results and accuracies for test, train and validation data:

1. **Testing accuracy:** 84.534%
2. **Validation accuracy:** 83.7%
3. **Training accuracy:** 84.08%

Task 2: Multi-class Logistic Regression (Extra credit)

- **Multinomial Logistic** regression is a classification method that generalizes logistic regression to multiclass problems, i.e. with more than two possible discrete outcomes.
- In this method, we don't build 10 classifiers like in binary logistic regression. Instead we only built 1 classifier that classifies 10 classes simultaneously.

Below are the results and accuracies for test, train and validation data:

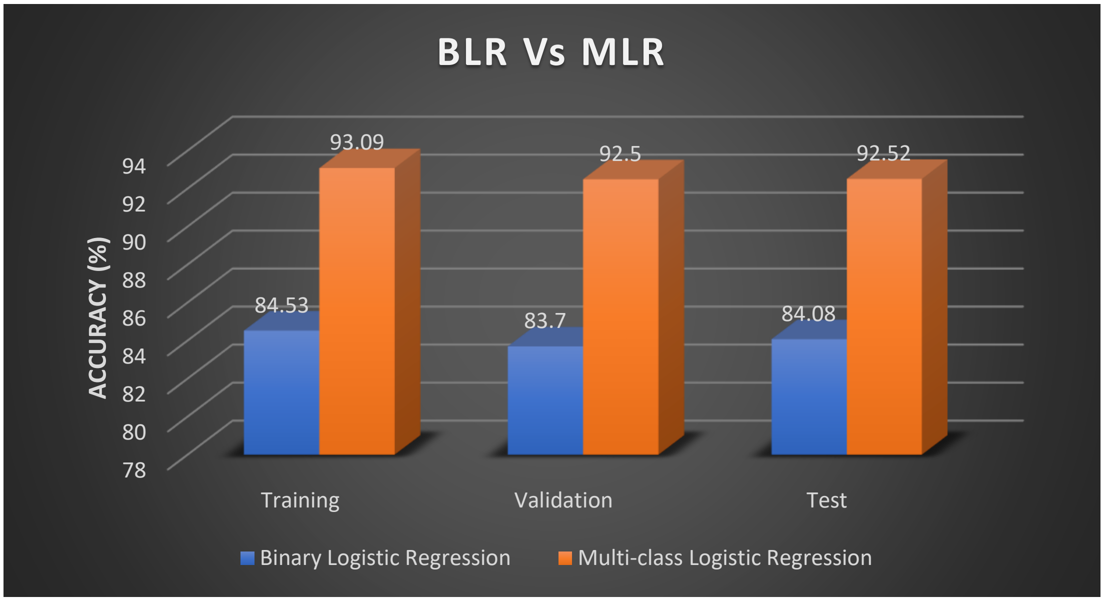
1. **Testing accuracy:** 93.09%
2. **Validation accuracy:** 92.5%
3. **Training accuracy:** 92.52%

Comparison of Binary Logistic regression and Multi-Class logistic regression

- Binary Logistic Regression is also known as One Vs One Classifier as it constructs one classifier per class. Since it requires more classifiers, this method is usually slower than one-vs-the-all because of its computational complexity. Also, Binary Logistic Regression is performed on a subset of data whereas Multiclass is performed on the entire dataset. Therefore, this method may be advantageous for algorithms which don't scale well with number of samples.

- Multi-Class Logistic Regression also known as one-vs-all, is implemented by fitting one classifier per class. Thus, it is computationally more efficient then binary regression. Since each class is represented by one and only one classifier, it is possible to gain knowledge about the class by inspecting its corresponding classifier. Also, it is faster as compared to Binary Logistic Regression.

- Below is the difference between the accuracies of Binary and Multi-class logistic regression.

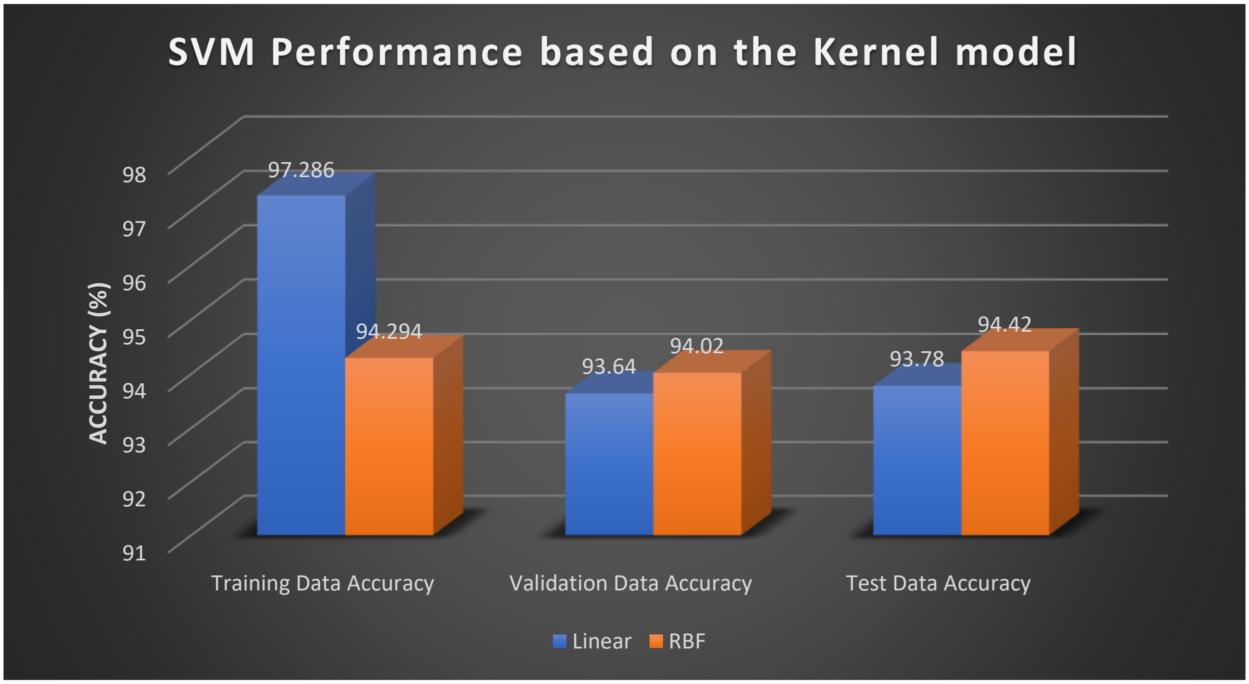


Task 3: Support Vector Model

Activity 1: Comparison between SVM Performance based on Kernel

- We have computed the accuracy for training, validation and testing data using SVM and used two different Kernels Linear and Radial Basis Function.

- A Linear Kernel is faster and performs at par with the RBF Kernel when the number of dimensions of the data is large. RBF Kernel in SVM, maps the data to multiple dimensions and provides better results in most of the cases when compared to the linear model. As the processing for RBF is high it takes longer time duration to compute. Thus, a kernel model should be chosen depending on the nature of the data.

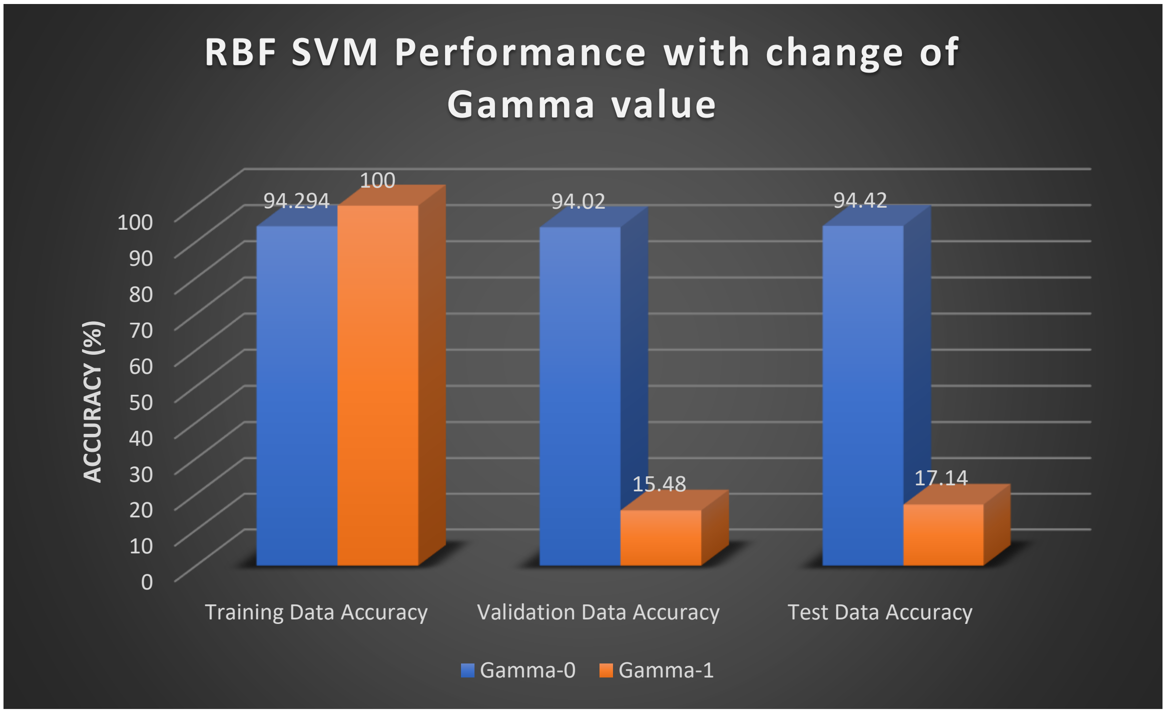


- For the given data, the number of dimensions is large and a non-linear model like RBF does not outperform Linear model as there is no need to map data to higher number of dimensions.

Activity 2: Comparison between Performance of SVM with RBF Kernel based on Gamma values.

- We have computed the accuracy for training, validation and testing data using SVM and a Radial Basis Function kernel, but modulated values of Gamma to compare the effect of the Gamma value on the outputs.

- The Gamma value is used to declare the extent of influence of every single training example. A higher value of Gamma represents that every single value of training example is important and should be considered while classifying the data. Thus, SVM with value of Gamma as 1 is a typical example of overfitting of data.



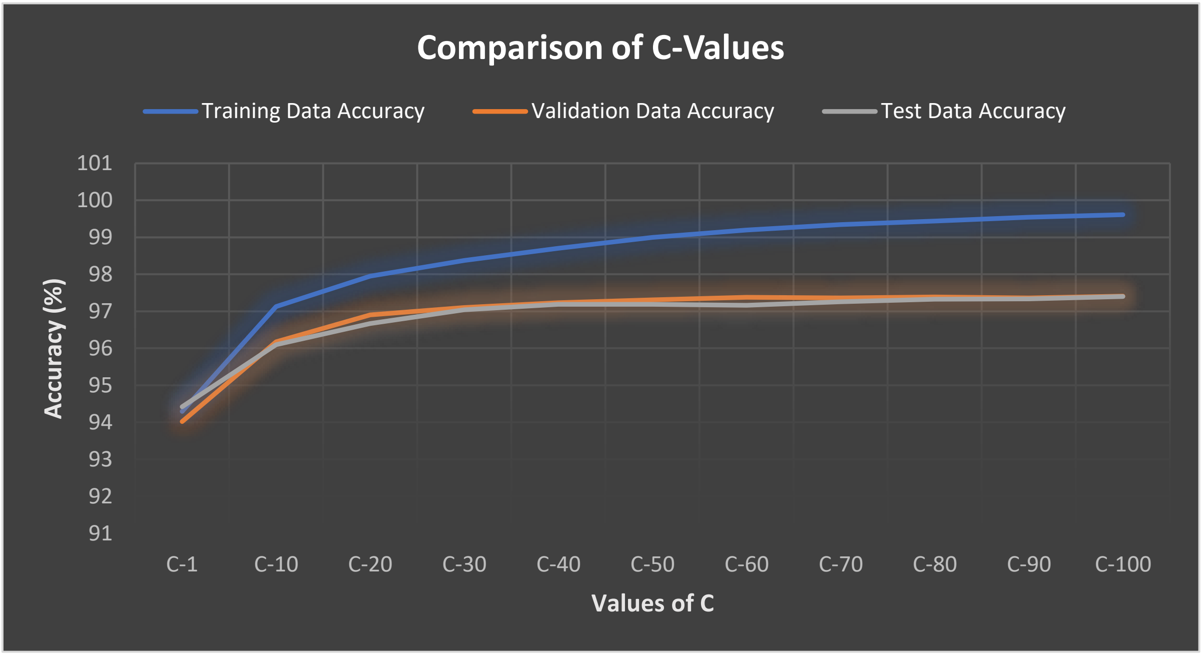
- The plotted graph points out the effect of higher value of Gamma resulting into overfitting. The Model works best for the data on which it was trained but fails drastically while classifying new data points.

Activity 3: Comparison between Performance of SVM with RBF Kernel based on Tolerance level [C] values.

- We have computed the accuracy for training, validation and testing data using SVM and a Radial Basis Function kernel, but modulated values of the tolerance factor ‘C’.

- C is used to indicate the tolerance level of misclassifying the training data. SVM tries to identify the appropriate margin for a given set of training data. With lower values of C, the SVM is forced to look for a larger margin even at the expense of misclassifying some of the training data points. For a higher value of C, the SVM tries to classify most of the data correctly and it ensures this by allowing a smaller margin separating the hyperplanes. A higher value of C will result into near perfect classification of the training data.

C	Training Data Accuracy	Validation Data Accuracy	Test Data Accuracy
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	99.002	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	99.542	97.36	97.34
100	99.612	97.41	97.4



- From the plotted graph, we can observe the effect of C on the computed SVMs. Higher the value of C, higher is the accuracy observed. But we can also notice that there is minimal change in accuracy after a threshold value of C is reached. Thus, value of C is dependent on the data which is used for computing and should be set experimentally.

References

1. https://en.wikipedia.org/wiki/Logistic_regression
2. https://en.wikipedia.org/wiki/Multinomial_logistic_regression
3. http://scikit-learn.org/stable/modules/multiclass.html