## CSE 474 / 574 Introduction to Machine Learning

## Programming Assignment 3

# Classification and Regression

**Project Report**

Group Number 50

May 3rd, 2017

*Team*

Muhammed Zaki Muhammed Husain Bakshi

Nikhil Prakash

Sunil Kunjappan Vasu

## Logistic Regression

We have implemented a Logistic Regression and the below accuracy was observed.

| OBSERVATION | |
|---|---|
| Training set Accuracy | 84.886% |
| Validation set Accuracy | 83.62% |
| Testing set Accuracy | 84.28% |

In Logistic Regression we employed the one-versus-all strategy. We have built 10 binary classifiers to distinguish between a given class from all other classes. From the result we could see that there was no overfitting and underfitting as the accuracy of all three observations i.e Training set, Validation set and Testing set were similar.

Logistic regression in general is used to predict the odds of being a case based on the values of the independent variables (predictors).

## Direct Multi-class Logistic Regression

A Direct Multi-class logistic regression was implemented on the handwritten data set and the below accuracy were obtained.

| OBSERVATION | |
|---|---|
| Training set Accuracy | 93.068% |
| Validation set Accuracy | 92.41% |
| Testing set Accuracy | 92.53% |

The multi-class Logistic Regression has a better accuracy than the Logistic Regression. Traditionally, Logistic Regression is used for binary classification. However, Logistic Regression can also be extended to solve the multi-class classification. With this method, we don’t need to build 10 classifiers like before. Instead, we now only need to build 1 classifier that can classify 10 classes at the same time. Due to this the accuracy is higher than binary Logistic Regression done in the above section.

The main comparison between Logistic Regression and Multiclass Logistic Regression is as follows

| | Logistic Regression | Multiclass Logistic Regression |
|---|---|---|
| 1 | Dependent Variable (DV) is categorical. | Dependent variable has more than two outcome categories |
| 2 | Model is used to estimate the probability of a binary response based on one or more predictor (or independent) variables (features). | Model that is used to predict the probabilities of the different possible outcomes of a categorically distributed dependent variable, given a set of independent variables (which may be real-valued, binary -valued, categorical-valued, etc. |

## Support Vector Machines

We have implemented code to learn the SVM model and compute accuracy of prediction with respect to training data, validation data and testing for the following cases and the accuracy are observed as below.

**CASE 1: Using linear kernel (all other parameters are kept default).**

| OBSERVATION | |
|---|---|
| Training set Accuracy | 97.286% |
| Validation set Accuracy | 93.64% |
| Testing set Accuracy | 93.78% |
| Run time | 0:16:08 |

The result obtained by linear kernel has a good accuracy in the validation and testing set. This also means that there are no significant overfitting.

**CASE 2: Using radial basis function with value of gamma setting to 1 (all other parameters are kept default) i.e RBF Kernel, Gamma-1.0.**

| OBSERVATION | |
|---|---|
| Training set Accuracy | 100.0% |
| Validation set Accuracy | 15.48% |
| Testing set Accuracy | 17.14% |
| Run time | 3:37:40 |

When the gamma is set as 1 it results in an ideal case of overfitting. As can be seen from the result the training set accuracy is 100% but the validation and test set accuracy is 15% and 17%. The overfitting of the training set is the cause for such an condition.

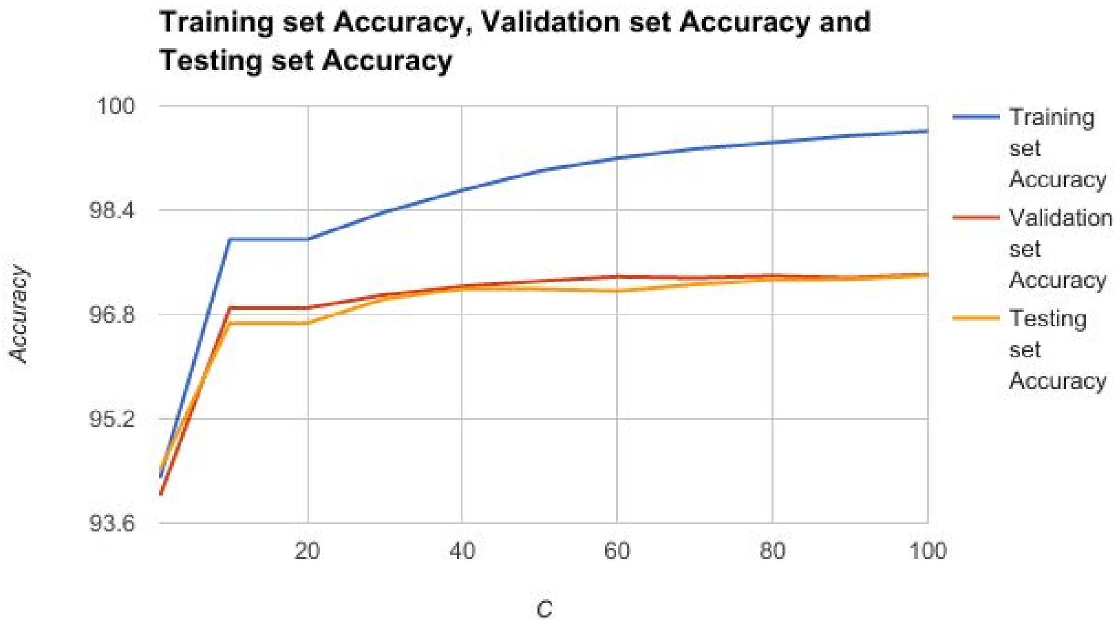
**CASE 3: Using radial basis function with value of gamma setting to default (all other parameters are kept default) i.e RBF kernel.**

This provides a better accuracy as the gamma is default. The overfitting that has occurred in the above case is avoided here. This provides us with a better accuracy for the training, validation and test set.

Gamma default means that 1/number of features will be used.

| OBSERVATION | |
|---|---|
| Training set Accuracy | 94.294% |
| Validation set Accuracy | 94.02% |
| Testing set Accuracy | 94.42% |
| Run time | 0:24:19 |

**CASE 4 : Using radial basis function with value of gamma setting to default and varying value of C (1, 10, 20, 30, · · · , 100) and plot the graph of accuracy with respect to values of C in the report.**



The Impact of the training example is determined by the parameter C thereby controlling the complexity of the learned hyperplane. As it can be observed from the above graph the higher values of C corresponds to higher accuracies. We could see a good validation and test set accuracy for different values of C. And as C is increased these accuracies also increase because C controls the penalty for error on each training examples.

**Comparing the selections of linear kernel and radial basis function kernel.**

Radial Basis Function maps the input features into infinite dimensional space thus giving better decision boundary albeit taking more time. The higher accuracy of the test data using RBF vs the accuracy using linear kernel proves this point.