CSE-574 Machine Learning

Programming Assignment 3

Classification and Regression

Spring 2017

University at Buffalo

Group Number: **22**

Members:

Hrishikesh Saraf	5020 5927
Nikita Konda	5020 7717
Shreya Nimje	5020 6520

Part 1: Binary Logistic Regression

Binary Logistic Regression was built on 10 binary-classifiers to distinguish a classifier from other classes. Implementation of Logistic Regression was performed using blrObjFunction() which computes the error and error gradient of Binary Logistic Regression; and blrPredict() which predicts the label of the given data and weights.

Training set accuracy: 84.868%
Validation set accuracy: 83.68%
Testing set accuracy: 84.17%

Part 2: Support Vector Machines

Using the Support Vector Machine tool in the sklearn.svm.SVC, the classifier was trained by using the fit() method of SVC. The accuracy of prediction on training set, validation set and testing set was determined by predicting the labels on that classifier by using the predict() method of SVC and then comparing it with the true label. By changing the parameters of SVC, different types of SVM model were learned and their accuracy results are as follows:

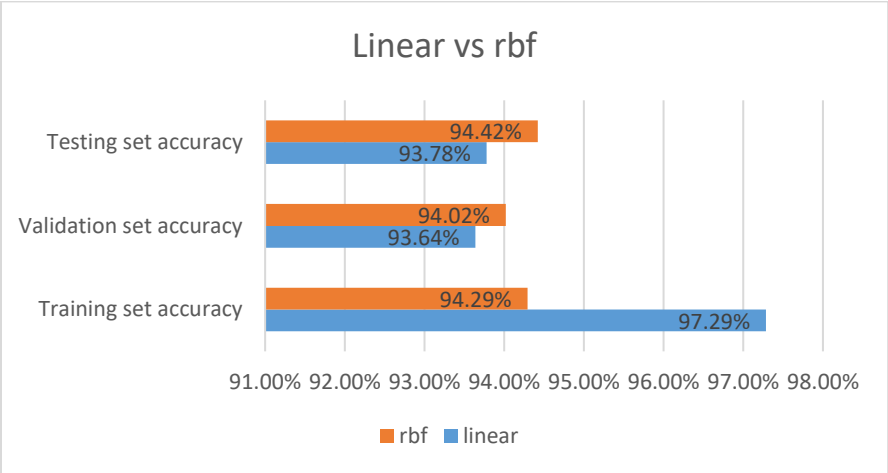
1. **Using linear kernel (all other parameters kept default).**
   Training set accuracy: 97.286%
   Validation set accuracy: 93.64%
   Testing set accuracy: 93.78%

2. **Using radial basis function with value of gamma setting to 1 (all other parameters are kept default).**
   Training set accuracy: 100.0%
   Validation set accuracy: 15.48%
   Testing set accuracy: 17.14%

   The gamma parameter defines how far the influence of a single training example reaches, with low values meaning ‘far’ and high values meaning ‘close’ [2]. As, we can see from the result, setting the parameter of gamma = 1 leads to overfitting of the data since all the training examples are considered. Hence, the accuracy of training data becomes 100% whereas the accuracy of validation and test data fall really low.

3. **Using radial basis function with value of gamma setting to default (all other parameters are kept default).**
   Training set accuracy: 94.294%
   Validation set accuracy: 94.02%
   Testing set accuracy: 94.42%

Comparison of linear kernel and radial basis function:



We can see that the accuracy of prediction on validation and test data increased slightly by using radial basis function with all the other parameters kept default. Hence, the prediction using radial basis function is better than the prediction using linear kernel.

4. **Using radial basis function with value of gamma setting to default and varying value of C (1, 10, 20, 30,…, 100).**

C is the penalty parameter of error term to avoid misclassifying each training example. For larger values of C we can achieve better prediction accuracies. The following graph plots the accuracy with respect to the values of C while keeping all the other parameters default. It is observed that with an increase in C the validation and test accuracies also increase but after C=30 it becomes stable.

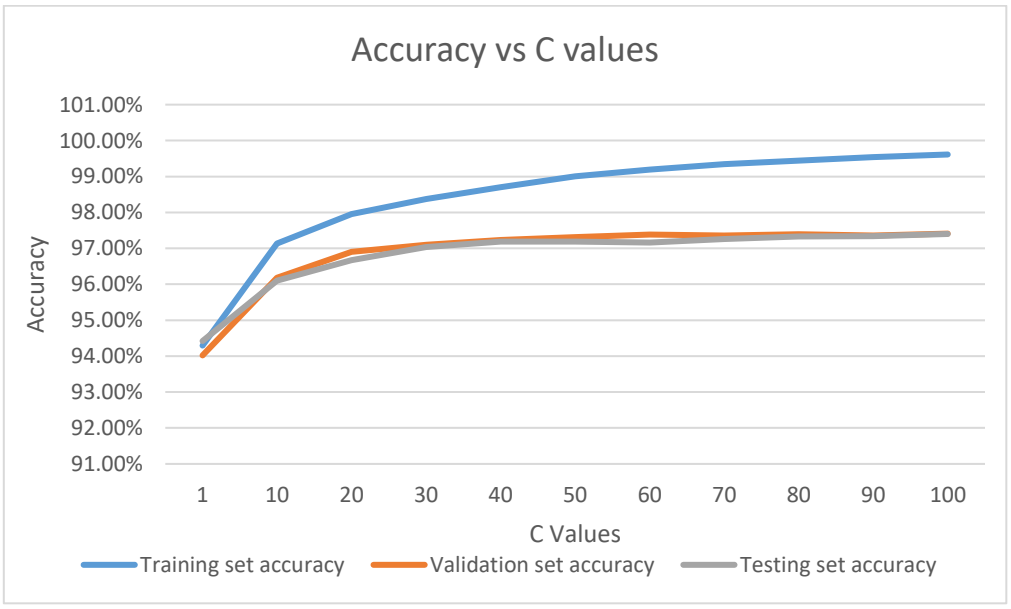


Fig 1: Accuracy vs C values for radial basis function kernel

So, overall in SVM we got the best results while using radial basis function and C =100 and all the other parameters set to default. The result is as follows:
Training set Accuracy: 99.612%
Validation set Accuracy: 97.41%
Testing set Accuracy: 97.4%

Part 3: Direct Multi-class Logistic Regression

Direct Multi-class Logistic Regression was built on 1 classifier (instead of 10) that classified 10 classes at the same time. The posterior probabilities were calculated using the softmax transformation of linear functions of feature variables. The likelihood function was estimated using 1-of-K encoding scheme.

The overall accuracy achieved in Direct Multi-class Logistic Regression was better than Binary Logistic Regression.
Training set accuracy: 93.204%
Validation set accuracy: 92.41%
Testing set accuracy: 92.49%

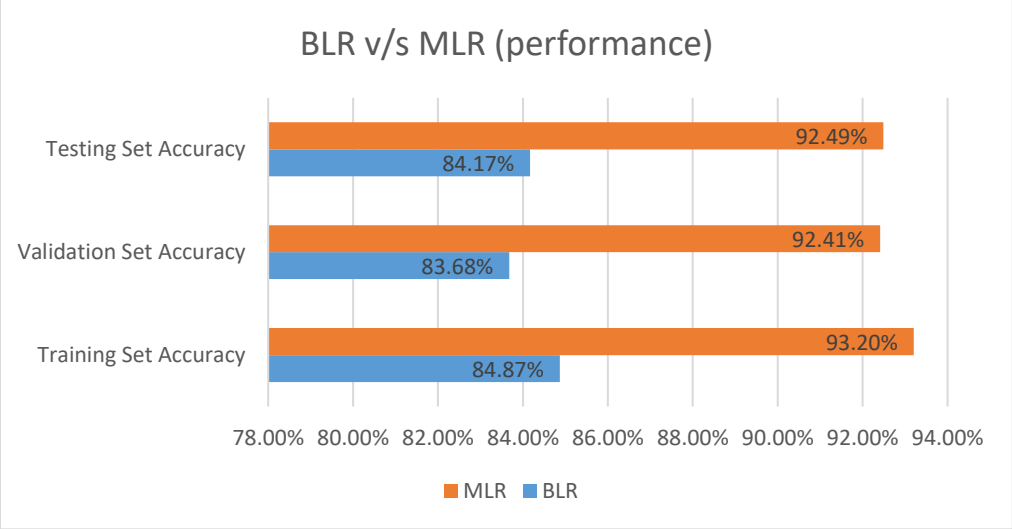


Fig 2: BLR Accuracy vs MLR Accuracy

From the results of both Binary Logistic Regression and Multi-Class Logistic Regression we conclude that Multi-Class Logistic Regression performs better than Binary Logistic Regression.

References:

1. http://scikit-learn.org/stable/modules/generated/sklearn.svm.SVC.html

2. http://scikit-learn.org/stable/auto_examples/svm/plot_rbf_parameters.html