A PROJECT REPORT ON

Classification and Regression

Abstract:

The main objective of the project is to evaluate the performance of various classification techniques like logistic regression and Support vector machines.

Problem 1: Logistic Regression

Implementation:

Using the Logistic Regression classification was performed on the training, validation and testing datasets. We have used blrObjFunction() and blrPredict() to implement the Logistic Regression using gradient descent where blrObjFunction() return error and error_grad.

In this project we have to classify handwritten digits so we divided the training data into 10 binary classifier and calculated posterior class probability of each group where 1 denotes that digit belong to that class and 0 denotes that it doesn't belong to that class.

We can calculate posterior class probability using this formulae.

$$P(C_1jx) = \sigma(w^T x)$$

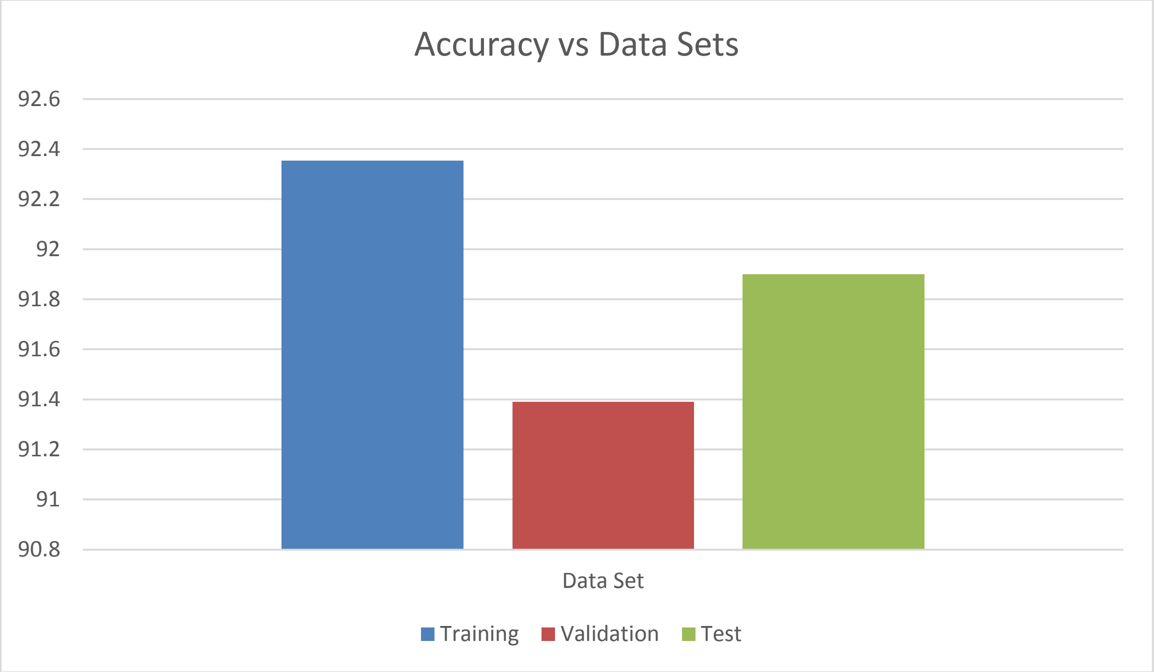
Error can be calculated as.

$$E(w) = \sum\{t_n \ln y_n + (1 - t_n) \ln(1 - y_n)\}$$

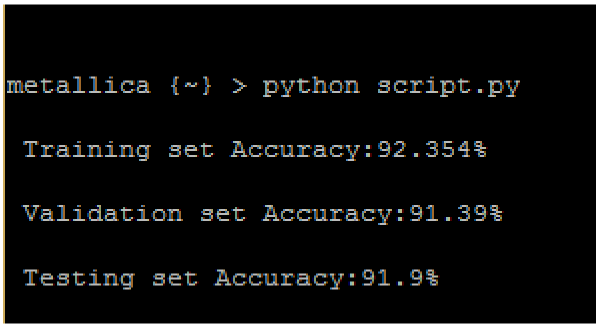
Now our blrObjFunction() will keep on modifying weights until it minimizes this error to a maximum of 50 times. After learning the final weight matrix we will pass this matrix to blrPredict() function so that using these weights we can predict the labels of test data.

Observations:

Data Set	Training	Validation	Testing
Accuracy	92.354 %	91.39 %	91.9 %



Result Screenshot:



```
metallica {~} > python script.py

 Training set Accuracy:92.354%

 Validation set Accuracy:91.39%

 Testing set Accuracy:91.9%
```

Problem 2: Support vector machines

Implementation:

Using the Support Vector Machine tool in sklearn.svm.SVM classification was performed on the training, validation and testing datasets.

Observations:

Following accuracies were observed for the training,validation and test datasets.

1. Using linear kernel (all other parameters are kept default)
   Training data accuracy:97.286%
   Validation data accuracy:93.64%
   Test data accuracy:93.78%

2. Using radial basis function with value of gamma setting to 1 (all other parameters are kept default).
   Training data accuracy:100%
   Validation data accuracy:15.48%
   Test data accuracy:17.14%
   The 100% accuracy of training data shows overfitting of data. The gamma parameter defines how far the influence of a single training example reaches, with low values meaning ‘far’ and high values meaning ‘close’. The gamma parameters can be seen as the inverse of the radius of influence of samples selected by the model as support vectors.

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

4. Using radial basis function with value of gamma setting to default and varying value of C from 1 to 100
   Following graph displays the accuracies obtained while varying value of C from 1 to 100.Thus as we increase the value of C accuracy goes on increasing. The C parameter trades off misclassification of training examples against simplicity of the decision surface. A low C makes the decision surface smooth, while a high C aims at classifying all training examples correctly by give the model freedom to select more samples as support vectors.

	Training set accuracy	Validation set 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

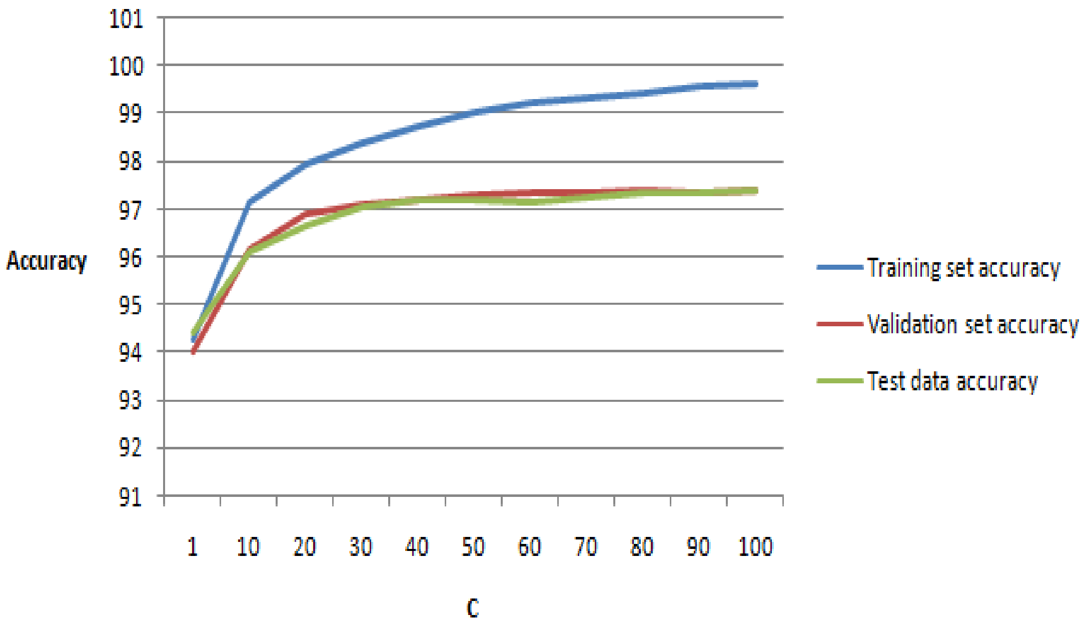


Fig:Accuracy vs C

Comparison of Logistic Regression and SVM:

After repeated experiments we can conclude that the performance of SVM and LR are approximtely close for the linear kernel where SVM exceeds performance by 3%-5%.
Accuracy from the Radial Basis function (with gamma parameter set to 1) are very low. It will be just too slow or the memory complexity will just be too high.

Accuracy of Logistic Regression can be increased by increasing the maxiter to a higher number than 50 but accuracy Vs time is a major factor. We have also performed the experiment when maxiter was 3, at that time we were getting accuracy about 79%. So to increase accuracy by next 12% we had to increase maxiter by 47. By this we can conclude that accuracy increases very slowly and is a very time taking process.