CSE 574 Introduction to Machine Learning

Classification and Regression

Team Number - 69

Aneesh Bhatnagar- 50208162
Moonis Javed - 50208261
Surabhi Singh Ludu - 50207139

Logistic Regression

Training set Accuracy:84.902%

Validation set Accuracy:83.73%

Testing set Accuracy:84.15%

Observation:

Logistic Regression accuracy is less as compared to other tests we ran because it regards all the points and then finds a hyperplane which separates data, not the best hyperplane. ALso Logistic Regression works better with low number of features.

Direct Multi-class Logistic Regression

Training set Accuracy:93.098%

Validation set Accuracy:92.41%

Testing set Accuracy:92.5%

Observation:

Multi-Class Logistic Regression performs better than one-vs-all Logistic Regression because, a single data field is simultaneously calculated for all the classes as compared to one class vs all classes in simple Logistic Regression. This leads to reduced run time for multi class Logistic Regression.

Support Vector Machines

Linear Kernel :

Training set Accuracy:97.286%

Validation set Accuracy:93.64%

Testing set Accuracy:93.78%

Radial basis function with value of gamma setting to 1 :

Training set Accuracy:100.0%

Validation set Accuracy:15.48%

Testing set Accuracy:17.14%

Radial basis function with value of gamma setting to default :

Training set Accuracy:94.294%

Validation set Accuracy:94.02%

Testing set Accuracy:94.42%

Radial basis function with value of gamma setting to default and varying value of C:

C Value	Train Accuracy	Validation Accuracy	Test 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%

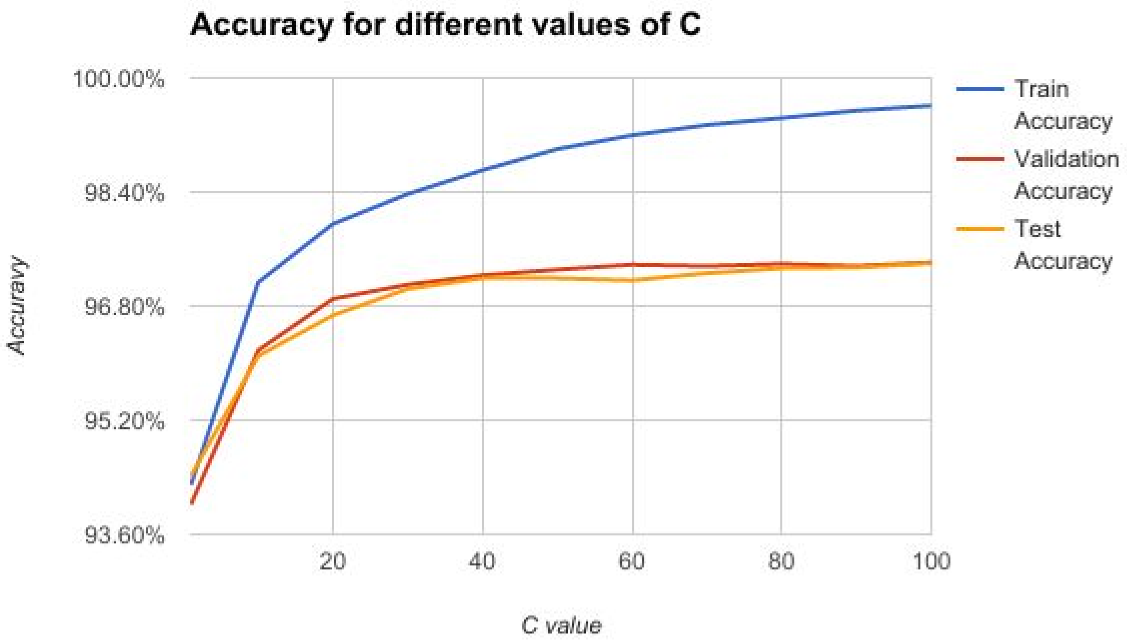

Fig 1. Graph for comparison of various accuracies for different values of C where gamma is default

Observation:

- In general, SVM performs better than Logistic regression.
- Linear-Kernel is more accurate for informative data which was not the case with the dataset provided which had digital images, as pixels are not very informative, so non-linear kernel performs better.
- When gamma is 1, we can see that train accuracy is 100% while other accuracies are extremely low, hence there is over-fitting in this case. The model does perform better with gamma default.
- We can see from the plot for gamma default and varying values of C that accuracy is higher for higher values of C. As C determines the penalty for error term for every training example. Thus, the weight of each error term is low when C is low, so even larger error values are accepted in training phase.

- But when C has a higher value , the weight of each error term is increased, so lower error values are accepted. So, a smaller margin hyperplane is built, but number of points misclassified are less, so the accuracy increases.
- We can see that there is a risk of overfitting for higher values of C as the train accuracy keeps increasing with bigger values of C while the test and validation accuracy stop increasing significantly post C=40.