CSE 574 INTRODUCTION TO MACHINE LEARNING

PROGRAMMING ASSIGNMENT 3

CLASSIFICATION AND REGRESSION

GROUP 31

ANUDEEP BULLA (50168954)

VALLABHANENI SUSMITHA CHOWDARY (50169236)

NARMADHA VISWANATHAN (50169758)

LOGISTIC REGRESSION

Logistic regression is a probabilistic statistical classification model. It is classified into two types namely *binomial* and *multinomial.* It includes all the points within a dataset to find the separating hyperplane. It works well when there are lesser number of input features. The output in case of binary logistic regression can take only two types, either True or False whereas Multinomial logistic regression the output can have more than two types.

RESULT

BINOMIAL LOGISTIC REGRESSION

Training data accuracy: 86.378%

Validation data accuracy: 85.57%

Testing data accuracy: 85.58%

MULTINOMIAL LOGISTIC REGRESSION

Training data accuracy: 92.486%

Validation data accuracy: 92.17%

Testing data accuracy: 92.17%

SUPPORT VECTOR MACHINE

Support vector machine constructs a set of hyper planes in an infinite dimensional space. SVM works well when the number of input features are high. The accuracy is obtained for training, testing and validation data by using various parameters listed as follows

1. Using linear kernel.
2. Using radial basis function with value of gamma setting to 1.
3. Using radial basis function with value of gamma setting to default.
4. Using radial basis function with gamma setting to default and varying the values of C.

LINEAR KERNEL FUNCTION

Linear kernel function is efficient in case of multi-dimensional data. The original data has to be very informative.

In case of Radial basis, the accuracy is low in case of gamma value =1 and this indicates an *overfitting* problem.

RESULT

	Training data accuracy	Validation data accuracy	Testing data accuracy
Linear kernel	97.286%	93.64%	93.78%
Radial basis Gamma=1	100%	15.48%	17.14%
Radial basis Gamma='auto'	94.288%	94.02%	94.41%

RADIAL BASIS FUNCTION WITH VARYING 'C' VALUES

C	Training data accuracy	Validation data accuracy	Testing data accuracy
1	94.288%	94.02%	94.41%
10	97.124%	96.18%	96.1%
20	97.95%	96.89%	96.66%
30	98.37%	97.1%	97.04%
40	98.704%	97.24%	97.19%
50	99%	97.31%	97.19%
60	99.198%	97.38%	97.17%
70	99.34%	97.36%	97.26%
80	99.438%	97.39%	97.33%
90	99.532%	97.36%	97.33%
100	99.608%	97.41%	97.4%

The accuracy increases with increase in the value of C. C indicates the tolerance in misclassifying data. A smaller-margin hyperplane is obtained for larger values of C. A larger-margin hyperplane is obtained for smaller values of C.

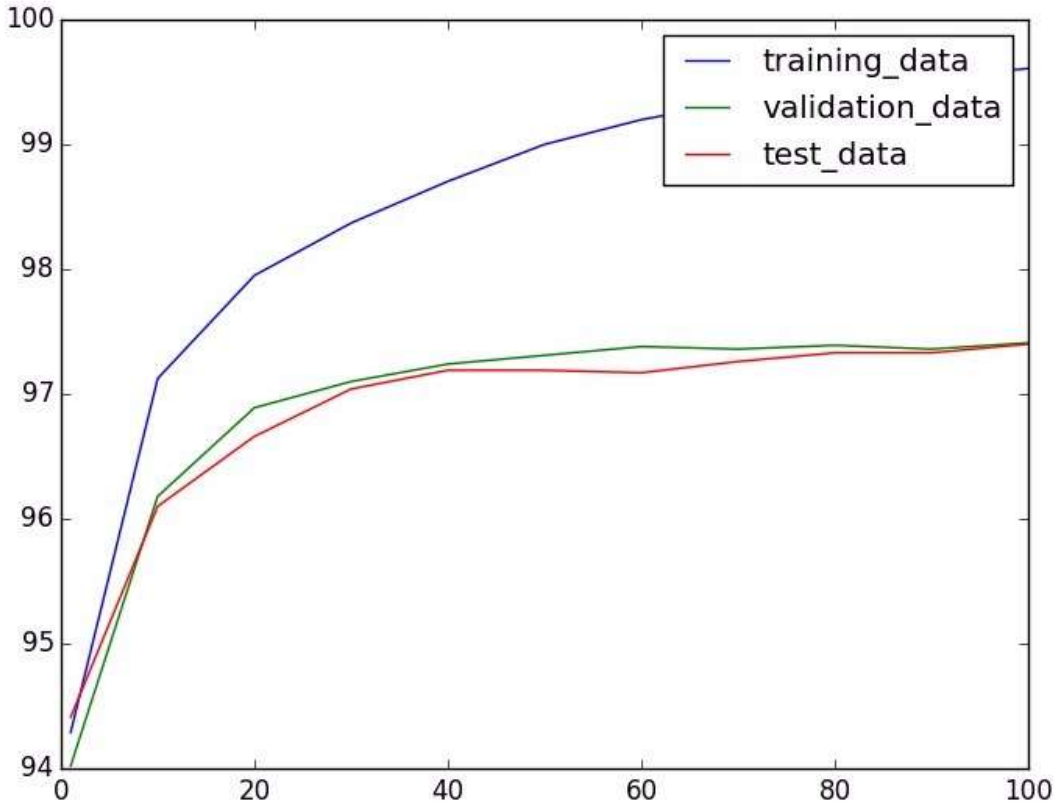

Fig 1: Radial basis with varying C values

CONCLUSION

Logistic regression works well when the training data has fewer features and SVM works well when the number of features are high. SVM works well in case of our MNIST dataset because the number of dimensions are higher, it has low noise and is not linearly separable.