Classification and Regression

Logistic Regression with Binomial Classifier:

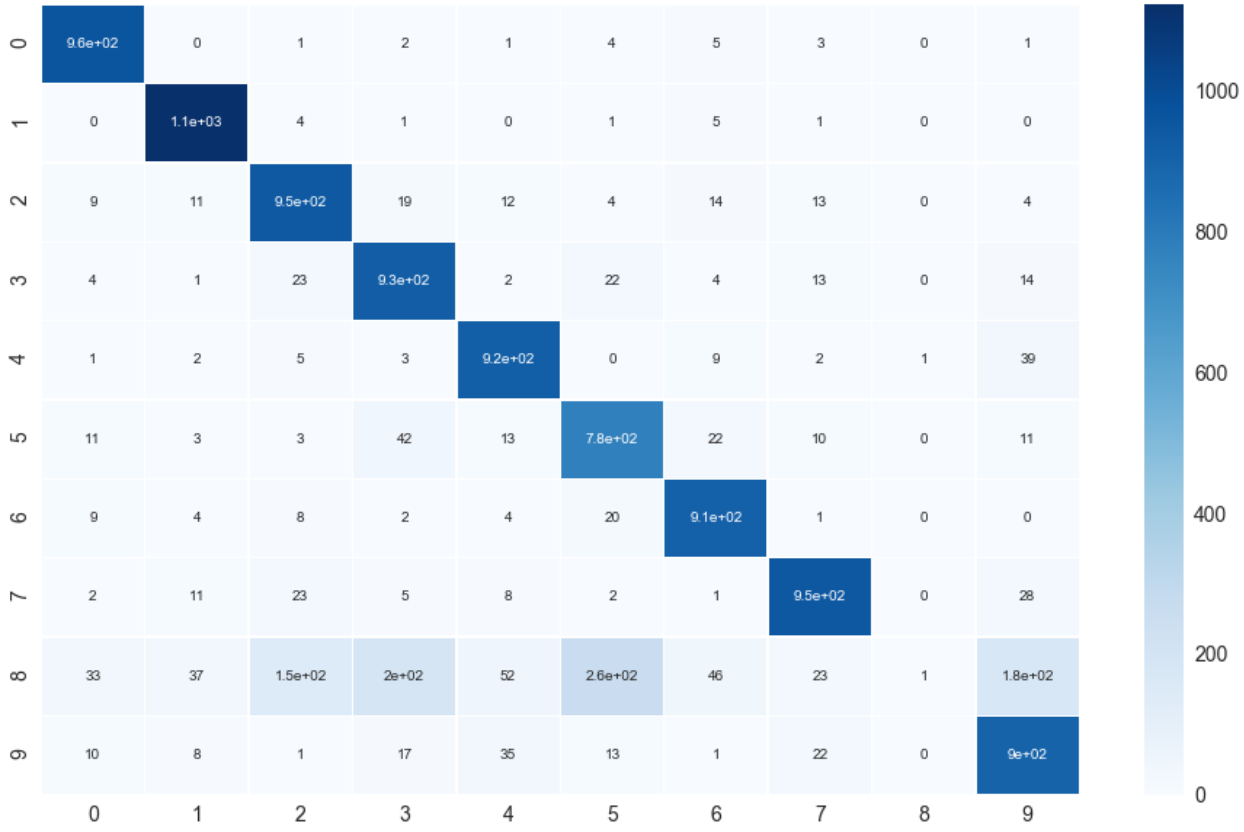
Logistic Regression is a predictive analysis, it works well when there are lesser number of input features and the VC dimension is smaller. The concept of Logistic Regression is about applying nonlinear transformation to the input data w.r.t weights.

The accuracy values obtained through Logistic regression is listed as follows:

Training data Accuracy: 84.926%

Validation data Accuracy: 83.69%

Testing data Accuracy: 84.12%



Multi - Class Logistic Regression:

The accuracy values obtained through Logistic regression is listed as follows:

Training data Accuracy: 93.11%

Validation data Accuracy: 92.54%

Testing data Accuracy: 92.55%

Support Vector Machines:

Support Vector Machines are supervised learning models with associated learning algorithms that analyze data used for classification and regression analysis.

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 value of gamma setting to default and varying the values of C.

	Training data accuracy	Validation data accuracy	Testing data accuracy
Linear Kernel	97.286%	93.64%	93.78%
Radial basis, Gamma = 1	100.0%	15.48%	17.14%
Radial basis, Gamma = 'default'	94.294%	94.02%	94.42%

Linear Kernel function is efficient in case of multi-dimensional data and the feature data must be very informative.

We performed SVM on the MNIST data which is a multi-dimensional data, but the feature data is less informative, so we got less accuracy when performed SVM with linear kernel compared with accuracies of Radial basis kernel which is a nonlinear kernel.

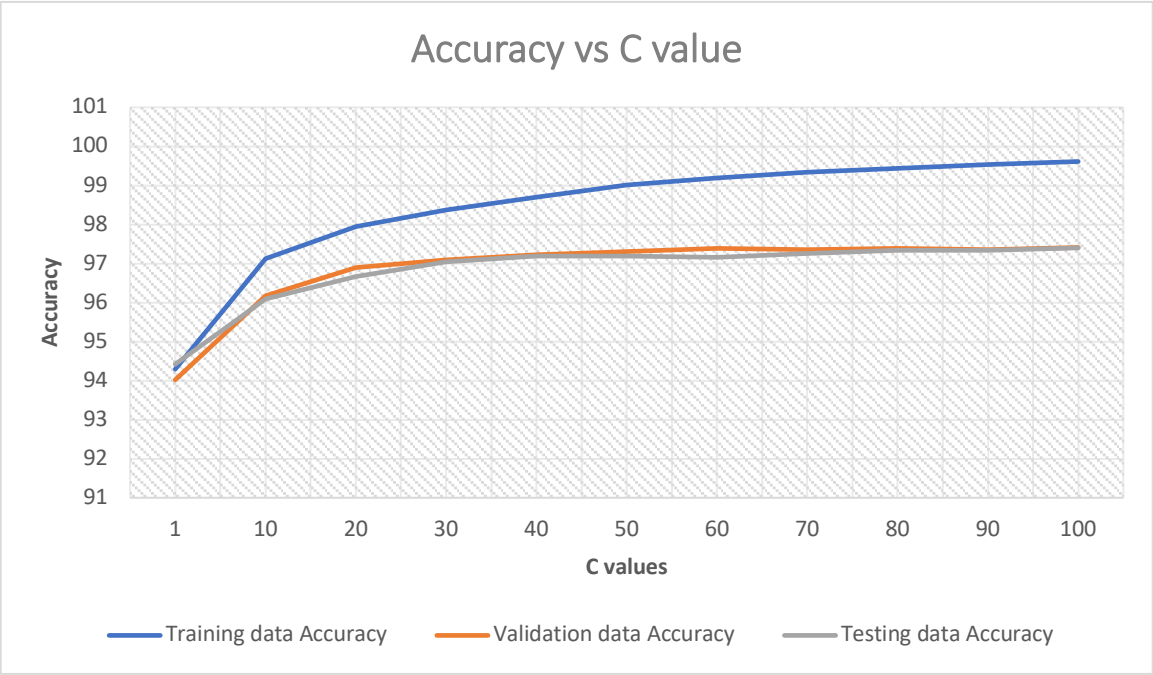
We can see that the Validation and Testing accuracy of Radial basis when gamma =1 is low, this indicates the over fitting problem since large values of gamma leads to high bias and low variance.

Radial Basis Function with Varying 'C' values

C	Training data accuracy	Validation data accuracy	Testing 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

C is a parameter for the soft margin cost function which controls the influence of each individual support vector.

C indicates the tolerance in misclassifying data, smaller-margin hyperplane is obtained for larger values of C and vice versa.



This figure plots the accuracy on training, test and validation data for different values of C.

As we can see from the plot, we get higher accuracy for higher values of C. This is because the penalty for the error term on each training sample is controlled by C.

The weight of each error term is low when the C is low and that is why a larger value during the training phase is accepted. With the expense of more samples being misclassified a larger margin hyperplane is built.

Conversely, the weight of each term increases as the C increases, that is why lower values will be accepted. This results in increase of the accuracy of data classification as less number of points will be misclassified because a smaller margin hyperplane will be built.

There is a high risk of overfitting for larger values of C. The reason for this is we can see how accuracy on the training set increases considerably when C grows, while the accuracy on the validation and test set saturates pretty fast and because of this we can say that increasing the complexity of the hyperplane may give overfitting.