Programming Assignment - 3

Classification and Regression

Team Members:
Pinaz Shaikh (50413528)
Janhavi Desale (50412898)
Krati Sharma (50416821)

Problem 1: Implementation of Logistic Regression

Binary Logistic Regression (BLR)

Set	Accuracy	Error
Training	92.6679	7.33
Validation	91.46	8.54
Testing	91.94	8.06

We can observe that the error is lower on training data but a little higher on tests. This is because it is a linear model and hence it performs better on seen data in comparison with unseen data.

Multi-class Logistic Regression (MLR)

Set	Accuracy	Error
Training	93.08	6.92
Validation	92.5	7.5
Testing	92.52	7.48

The results tabulated above are observed after running Multi-class Logistic Regression on Data. We can see that the Training error is slightly less than the Testing error. This is because it is a linear model and hence it performs better on seen data in comparison with unseen data.

Performance difference between multi-class strategy (MLR) with one-vs-all strategy (BLR) :

Set	MLR	BLR
Training	93.08	92.6679
Validation	92.5	91.46
Testing	92.52	91.94

We observed the accuracy of the MLR was better than the BLR classification. That's because parameters are estimated independently in multiclass which helps to prevent wrong classification.

Support Vector Machine (SVM) :

1. Linear Kernel

Set	Accuracy
Training	92.764
Validation	91.51
Testing	91.649

So, we can infer from the above results that the Linear Kernel works like a linear model, as the results are almost the same as the previous linear model we trained.

2. Radial Basis Function

    a. Radial Basis Function (Gamma = 1)

Set	Accuracy
Training	100.0
Validation	18.41
Testing	20.1

This model , though gives very high accuracy for training data, performs poorly for testing data.

b. Radial Basis Function (Gamma = default)

Set	Accuracy
Training	96.54
Validation	96.24
Testing	96.35

c. Using Radial Basis Function with value of gamma setting to default and varying value of C (1, 10, 20, 30, …,100)

Here we have introduced a new variable C. C variable controls the importance we are giving to the Slack variable.

C	Training	Validation	Testing
1	96.54	96.24	96.35
10	97.292	96.77	96.96
20	97.288	96.76	96.95
30	97.288	96.76	96.95
40	97.288	96.76	96.95
50	97.288	96.76	96.95
60	97.288	96.76	96.95
70	97.288	96.76	96.95

80	97.288	96.76	96.95
90	97.288	96.76	96.95
100	97.952	96.89	96.67

So, we can conclude that we are getting the best result by setting gamma to default and taking C = 20.

Kernel	C	Training	Validation	Testing
RBF	20	97.952	96.89	96.67

Plot of different parameters for various value of C :