CSE 574

Introduction to Machine Learning

Programming Assignment 3

Classification and Regression

Project report by: -

1) Siddhi Thakur (50365530)
2) Apurva Chavan (50365703)
3) Mehvish Shamshad (50374333)

Binary Logistic Regression (BLR)

```
Training set Accuracy:92.716%

Validation set Accuracy:91.44%

Testing set Accuracy:92.03%
```

Set	Accuracy	Error
Training	92.716%	7.284%
Validation	91.44%	8.56%
Test	92.03%	7.97%

- The result of the Binary Logistic Regression also knowns as ***'One vs Other'*** approach is given above.
- We can see that the training error is less than the testing error (7.284% < 7.97%).
- This tells us that the Binary logistic Regression model performs better on the seen data than on unseen data.

Multi-Class Logistic Regression (MLR)

```
Training set Accuracy:93.176%

Validation set Accuracy:92.46%

Testing set Accuracy:92.51%
```

Set	Accuracy	Error
Training	93.176%	6.824%
Validation	92.46%	7.54%
Test	92.51%	7.49%

- The result of the Multi-Class Logistic Regression also known as ***'One vs All'*** approach is given above.
- We can see that the Training Error is less than the Test error (6.824% < 7.49%).
- Again, we can infer that the Multi-Class Logistic Regression model performs better on the seen data than on unseen data.

Comparison between BLR and MLR.

Sets	BLR	MLR
Training	92.716%	93.176%
Validation	91.44%	92.46%
Test	92.03%	92.51%

- In Multi-Class Logistic Regression (MLR) we classify all 10 classes of the MNIST dataset at once, whereas in Binary Logistic Regression (BLR) we classify 1 class w.r.t all other classes at a time.
- Due to this MLR has less chances of overlapping than in BLR as it classifies all the classes together.

Support Vector Machine (SVM)

Randomly selecting 10000 samples from the dataset to train SVM with different Kernels

1) ***Using Linear Kernel:***

```
Training set Accuracy:92.456%

Validation set Accuracy:90.98%

Testing set Accuracy:91.72%
```

Accuracy on 10000 sampled train data

```
Training set Accuracy:99.72%
```

2) ***Radial Basis Function (RBF):***

a) Using RBF (gamma = 1)

```
RBF Kernel


Training set Accuracy:27.442%

Validation set Accuracy:10.0%

Testing set Accuracy:11.35%
```

Accuracy on 10000 sampled train data

```
Training set Accuracy:100.0%
```

b) Using RBF with value of gamma setting to default (all other parameters are kept default)

```
RBF Kernel


Training set Accuracy:96.414%

Validation set Accuracy:96.05%

Testing set Accuracy:96.36%
```

Accuracy on 10000 sampled train data

```
Training set Accuracy:98.82%
```

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

C	Training Accuracy	Validation Accuracy	Test Accuracy
1	96.486%	95.97%	96.25%
10	97.26%	96.52%	96.8%
20	97.26%	96.53%	96.81%
30	97.26%	96.53%	96.81%
40	97.26%	96.53%	96.81%
50	97.26%	96.53%	96.81%
60	97.26%	96.53%	96.81%
70	97.26%	96.53%	96.81%
80	97.26%	96.53%	96.81%
90	97.26%	96.53%	96.81%
100	97.26%	96.53%	96.81%

- So, from the above results we can conclude that we are getting the best possible result by setting gamma to default value and taking C = [20, 30, 40, 50, 60, 70, 80, 90, 100]
- Accuracy on 10000 sampled train data for c values [1,10,20…..100]

98.82	**99.99**	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**

- **Taking C=20, and gamma= default and training it on entire dataset.**

```
SVM with best parameters

Training set Accuracy:96.50999999999999%

Validation set Accuracy:96.04%

Testing set Accuracy:96.3%
```

- We observe that the accuracy of RBF kernel on test data is better than linear kernel. This is because in SVM, Kernel transforms the input data into any first as per the user requirements.
- Being a multi class prediction and data is non-linear, the RBF kernel works well. Linear kernel is faster, yet works well with linearly separable data.
- We observe that with RBF kernel and gamma set to 1, SVM does not perform well. Gamma

parameters determine their influence over a single training example, thus, overfitting the model.

Plot of Accuracy obtained on each of Training, Testing and Validation dataset w.r.t various values of C