CSE 574: Programming Assignment 3
Classification and Regression

Ajay Kumar Davuluri, UBIT ID: adavulur, Person No: 50168851
Lalith Vikram Natarajan, UBIT ID: lalithvi, Person No: 50169243
Srinath Goud Vanga, UBIT ID:srinathg , Person No:50169176

April 29, 2016

1 Objective

Implement Logistic Regression and use the Support Vector Machine tool in sklearn.svm.SVM to classify hand-written digit images and compare the performance of these methods.

2 Dataset

The MNIST dataset with a training set of 60000 examples and test set of 10000 examples. All digits have been size-normalized and centered in a fixed image of 28 x 28 size. In original dataset, each pixel in the image is represented by an integer between 0 and 255, where 0 is black, 255 is white and anything between represents different shade of gray.

3 Implementation

- preprocess() includes selecting features to reduce the computational overhead and shuffling data to ensure randomness in the experiment.
- blrObjFunction() computes 2-class Logistic Regression error function and its gradient.
- blrPredict() predicts the label of data given the data and parameter W of Logistic Regression
- mlrObjFunction() computes multi-class Logistic Regression error function and its gradient
- mlrPredict() predicts the label of data given the data and parameter W of Logistic Regression

4 Observations

4.1 Logistic Regression

- Training dataset accuracy - 86.244%
- Validation dataset accuracy - 85.41%
- Testing dataset accuracy - 85.45%

4.2 SVM Using Toolbox

4.2.1 Using linear kernel

- Training dataset accuracy - 97.286%
- Validation dataset accuracy - 93.64%
- Testing dataset accuracy - 93.78%

4.2.2 Using radial basis function with value of gamma setting to 1

- Training dataset accuracy - 100.0%
- Validation dataset accuracy - 15.48%
- Testing dataset accuracy - 17.14%

4.2.3 Using radial basis function with value of gamma setting to default

- Training dataset accuracy - 94.294%
- Validation dataset accuracy - 94.02%
- Testing dataset accuracy - 94.42%

4.2.4 Using radial basis function with value of gamma setting to default and varying value of C (1, 10, 20, 30, ..., 100)

C = 1

- Training dataset accuracy - 94.294%
- Validation dataset accuracy - 94.02%
- Testing dataset accuracy - 94.42%

C = 10

- Training dataset accuracy - 97.132%
- Validation dataset accuracy - 96.18%
- Testing dataset accuracy - 96.1%

C = 20

- Training dataset accuracy - 97.952%
- Validation dataset accuracy - 96.9%
- Testing dataset accuracy - 96.67%

C = 30

- Training dataset accuracy - 98.372%
- Validation dataset accuracy - 97.23%
- Testing dataset accuracy - 97.19%

C = 40

- Training dataset accuracy - 98.706%
- Validation dataset accuracy - 97.23%
- Testing dataset accuracy - 97.19%

C = 50

- Training dataset accuracy - 99.002%
- Validation dataset accuracy - 97.31%
- Testing dataset accuracy - 97.19%

C = 60

- Training dataset accuracy - 99.196%
- Validation dataset accuracy - 97.38%
- Testing dataset accuracy - 97.16%

C = 70

- Training dataset accuracy - 99.34%

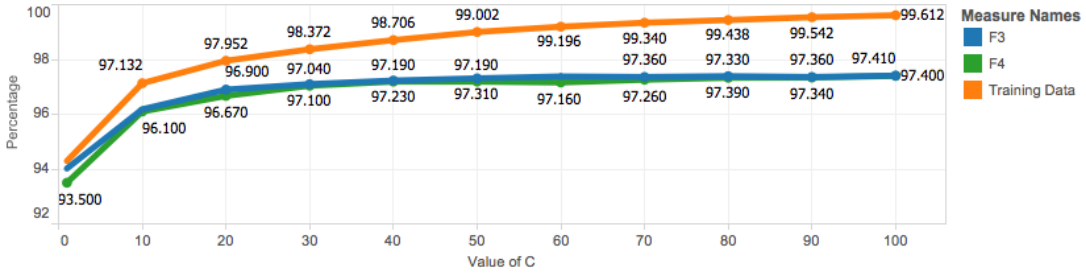

Figure 1: C vs Data Set accuracy

- Validation dataset accuracy - 97.36%
- Testing dataset accuracy - 97.26%

C = 80

- Training dataset accuracy - 99.438%
- Validation dataset accuracy - 97.39%
- Testing dataset accuracy - 97.33%

C = 90

- Training dataset accuracy - 99.542%
- Validation dataset accuracy - 97.36%
- Testing dataset accuracy - 97.34%

C = 100

- Training dataset accuracy - 99.612%
- Validation dataset accuracy - 97.41%
- Testing dataset accuracy - 97.4%

These values are then plotted with value of C on the X axis and accuracy percentage on the Y axis. The plots for different datasets are shown in the figure 1. You can see that the accuracies show sharp increase initially but then become stable or slow. The accuracies seem to converge for greater values of C.

4.3 Direct Multi-class Logistic Regression

- Training dataset accuracy - 93.39%
- Validation dataset accuracy - 92.43%
- Testing dataset accuracy - 92.67%