MACHINE LEARNING – CSE574

ASSIGNMENT – 3

Group – 5

Naga Vaishanavi Pakyala

Siddiq Syed

Bhanu Prasanth Yeeli

Implementation of Logistic Regression:

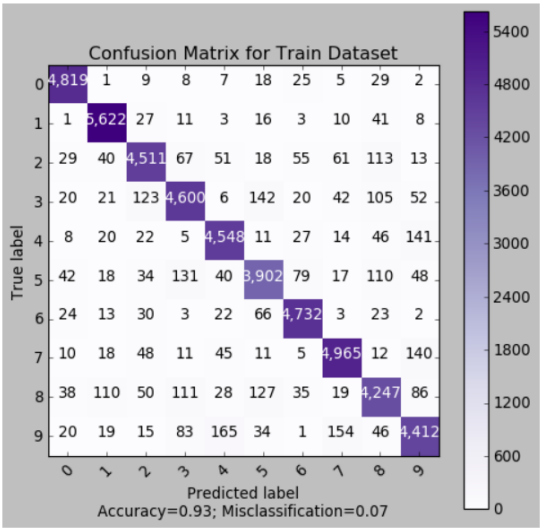
After coding the two functions to process the logistic regression, given MNIST dataset is processed and accuracies for training, validation and testing are as follows.

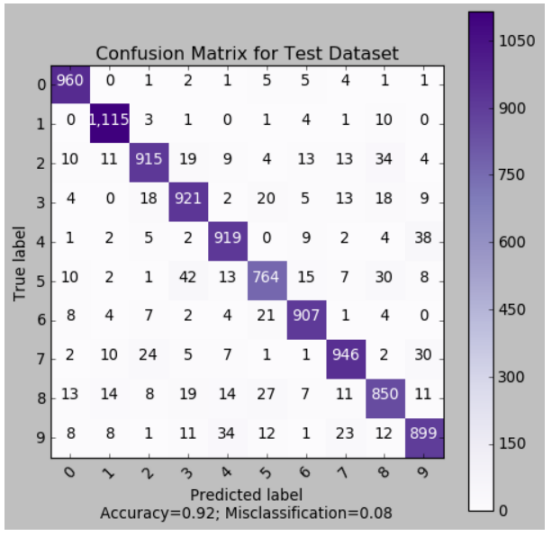
```
Training set Accuracy:92.716%

Validation set Accuracy:91.47999999999999%

Testing set Accuracy:91.96%
```

Looks like Training and Testing accuracies are pretty close. To compare the errors category wise let's look at the confusion matrix below for training and testing data.





For the above results of confusion matrix let us look at the report values.

Report for Training data

	precision	recall	f1-score	support
0	0.98	0.96	0.97	5011
1	0.98	0.96	0.97	5882
2	0.91	0.93	0.92	4869
3	0.90	0.91	0.91	5030
4	0.94	0.93	0.93	4915
5	0.88	0.90	0.89	4345
6	0.96	0.95	0.96	4982
7	0.94	0.94	0.94	5290
8	0.88	0.89	0.88	4772
9	0.89	0.90	0.90	4904

Report for Testing data

	precision	recall	f1-score	support
0	0.98	0.94	0.96	1016
1	0.98	0.96	0.97	1166
2	0.89	0.93	0.91	983
3	0.91	0.90	0.91	1024
4	0.94	0.92	0.93	1003
5	0.86	0.89	0.87	855
6	0.95	0.94	0.94	967
7	0.92	0.93	0.92	1021
8	0.87	0.88	0.88	965
9	0.89	0.90	0.89	1000

In both results we can see that 2,3,5,8,9 have little low precision/accuracy values when compared to other numbers and overall accuracy. This should be because these numbers more or less look a bit similar.

Also comparing training and test results we can see that testing accuraries are lesser than training accuracies. This can assumed because Logstic Regression algorithm builts the boundary/hyper plan considering all the values of the training dataset. Which will obviously perform better to the training dataset when performed on anyother data.

Multi-Class Logistic Regression:

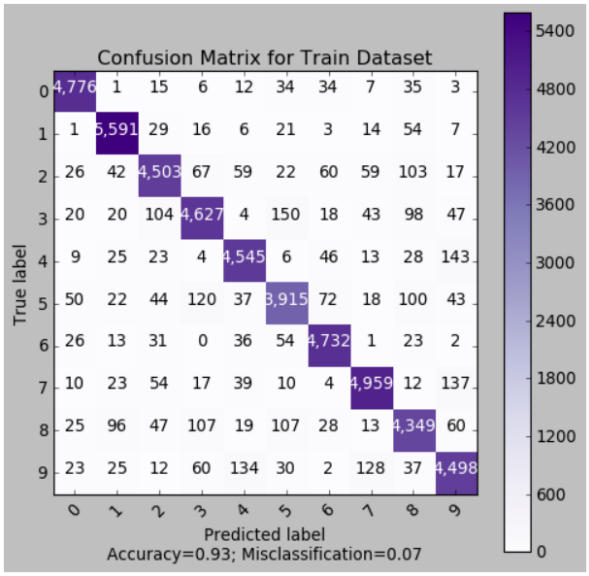
For MNIST dataset with Multi-Class Logistic Regression algorithm that was developed below are the accuracies of the training, validation and testing data.

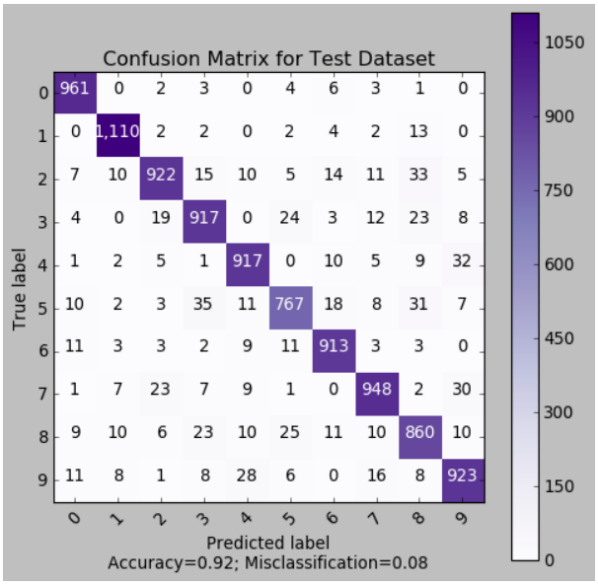
```
Training set Accuracy:92.99%

Validation set Accuracy:92.45%

Testing set Accuracy:92.38%
```

Looks like accuracies are slightly greater than Logistic Regression algorithm and this is because probabilities are calculated considering all the classes in the algorithm. Let us also compare the errors for training and testing data for this algorithm.





For the above results let us see the accuracies report class wise.

Report for Training data

	precision	recall	f1-score	support
0	0.97	0.96	0.97	4966
1	0.97	0.95	0.96	5858
2	0.91	0.93	0.92	4862
3	0.90	0.92	0.91	5024
4	0.94	0.93	0.93	4891
5	0.89	0.90	0.89	4349
6	0.96	0.95	0.95	4999
7	0.94	0.94	0.94	5255
8	0.90	0.90	0.90	4839
9	0.91	0.91	0.91	4957

Report for Testing data

	precision	recall	f1-score	support
0	0.98	0.95	0.96	1015
1	0.98	0.96	0.97	1152
2	0.89	0.94	0.91	986
3	0.91	0.91	0.91	1013
4	0.93	0.92	0.93	994
5	0.86	0.91	0.88	845
6	0.95	0.93	0.94	979
7	0.92	0.93	0.93	1018
8	0.88	0.87	0.88	983
9	0.91	0.91	0.91	1015

We can notice here that individual are more balanced compared to general version of the algorithm. We can also see that like in the previous algorithm even here 3,5,8 has lesser accuracies compared to other digits, but are more balanced as mentioned above.

Also testing accuracy is little less when compared to training accuracy and the reason will as mentioned in the previous case i.e. the model is built upon training dataset and logstic regression considers all the points in a dataset to build a model. So, the model can not [erform better than a training dataset accuracy.

SUPPORT VECTOR MACHINES:

When MNIST dataset classified with SVM algorithm and the same is executed for 10k samples of training dataset and also for complete dataset. The observed results are same in both the cases.

Using linear kernel:

Below are the accuracies for the same.

```
SVM Using linear kernel

 Training Accuracy:97.286%

 Validation Accuracy:93.64%

 Testing Accuracy:93.78%
```

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

Below are the accuracies on MNIST dataset.

```
SVM Using radial basis function with value of gamma setting to 1

Training Accuracy:100.0%

Validation Accuracy:15.479999999999999%

Testing Accuracy:17.14%
```

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

The default setting for the same will be gamma = 0 and below are the accuracies when run for MNIST dataset.

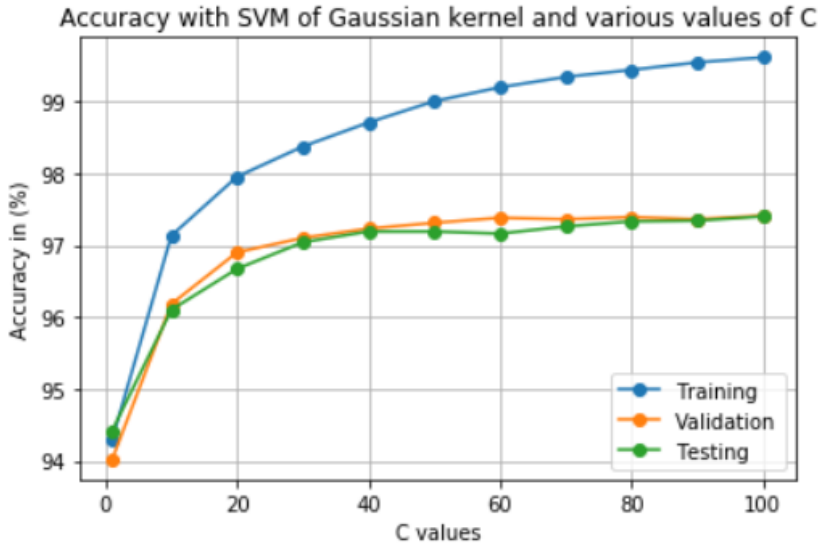
```
Training Accuracy:94.294%

Validation Accuracy:94.02000000000001%

Testing Accuracy:94.42%
```

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

Below is plot for C values Vs Accuracies on the MNIST dataset.



Corresponding values are included in the pickle file attached.

The SVM algorithm with C value 100 is performing the best among all the settings with accuracies of training set close to 100 and testing accuracy over 97%. Also the difference among different settings is clear with the above displayed accuracies.