Learning From Data - HW3 Report

Bengisu Güresti - 150150105

Yunus Güngör - 150150701

Dataset for this homework is taken from :

https://archive.ics.uci.edu/ml/datasets/Optical+Recognition+of+Handwri tten+Digits

In this dataset train set and test set are seperate. Train set consists of 3823 data and test set consists of 1797 data. Each data has 64 features and there are 10 different classes (classes: 0, 1, 2, 3, 4, 5, 6, 7, 8, 9) that each data can belong to.

Question 1: The dataset is classified using neural networks(MLP). The training of the data is done on the Train set and optimum weights are obtained by evaluating classification results using 10-fold cross validation. Then the performance of the classification is tested on the Test set.

The accuracy percentages of each fold cross validation of the Train set is given below:

accuracyP =

97.1429

99.2147

98.1675

96.8586

98.9529

97.6440

98.1675

95.8115

97.1204

96.5969

The accuracy percentage of this classification is their average, which is:

validationAccuracy = 97.5677

Then the performance of the classification is tested on Test set, and the accuracy percentage is found as :

accuracyPTest = 95.4925

The first figure below is the classification of the first 20 data in the Test set.

The second figure below is the classification of the whole Test set.

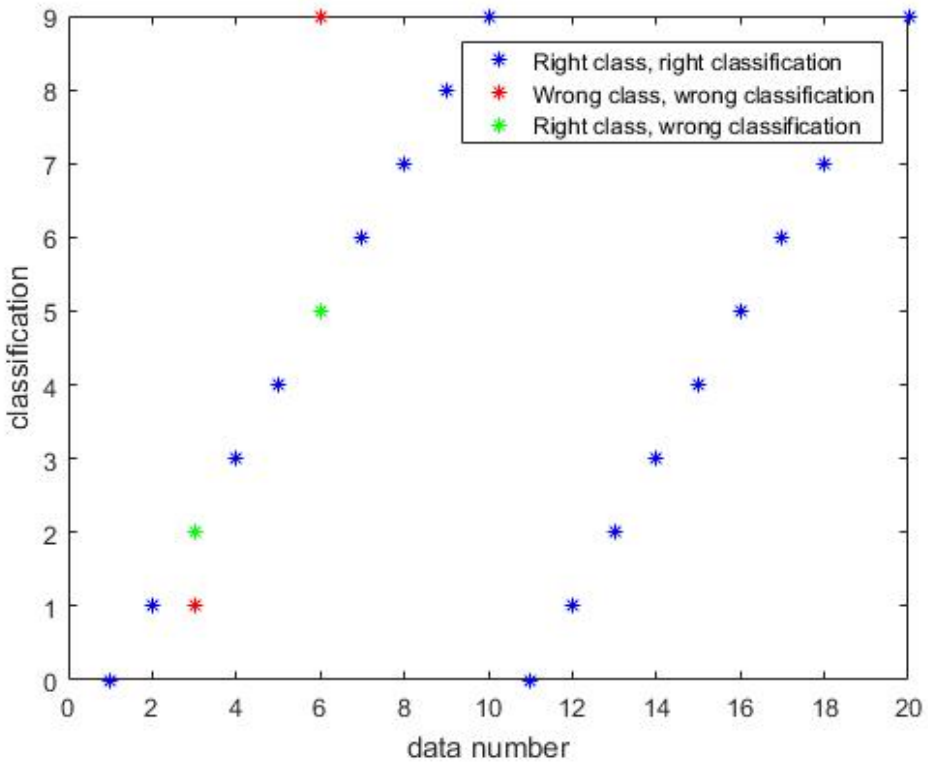
Blue : right result, right classification

Red : wrong result, wrong classification

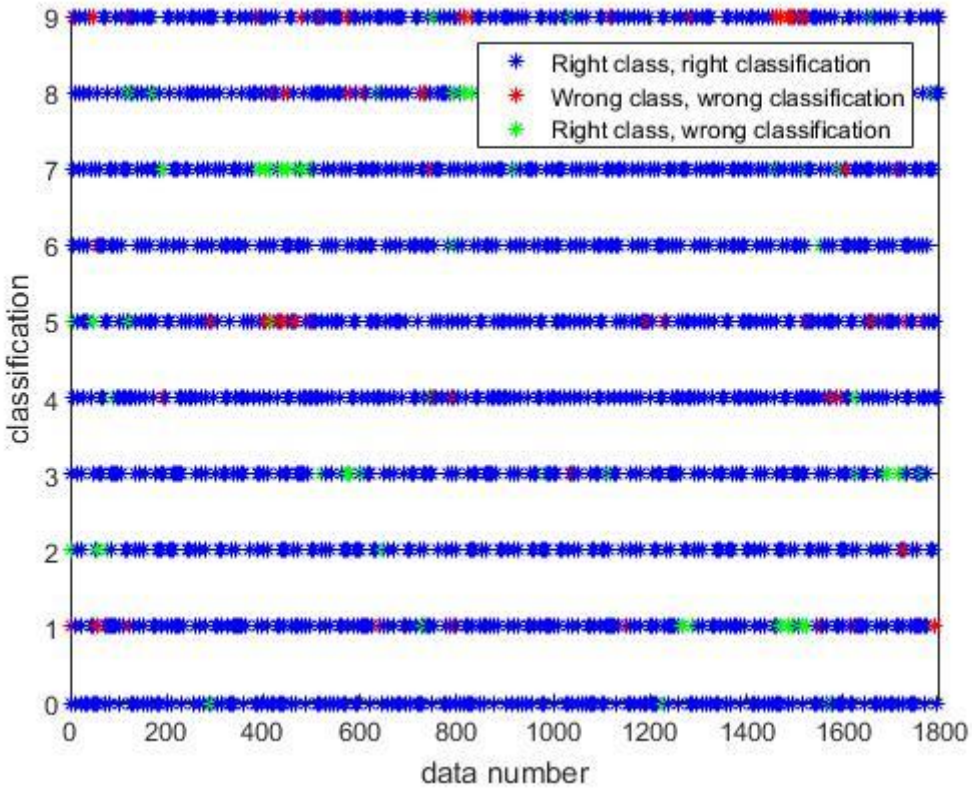
Green : right result, wrong classification

So, blue indicates right classification, red and green indicate wrong classification.

1)



2)



Question 2: The dimension of the data is reduced to 2 using PCA, as a result when the data set is classified using neural networks(MLP) only 2 features of the data are used. Dimensionality reduction is first applied to training dataset; that means, eigenvalues are computed on training set and then they were applied to test set. After the dimensionality reduction, the dataset is classified using neural networks(MLP). The training of the data is done on the dimensionality reduced Train set and optimum weights are obtained by evaluating classification results using 10-fold cross validation. Then the performance of the classification is tested on the dimensionality reduced Test set.

The accuracy percentages of each fold cross validation of the Train set is given below:

accuracyP =

62.5974

58.6387

58.6387

58.3770

61.7801

59.9476

59.1623

53.1414

58.9005

56.2827

The accuracy percentage of this classification is their average, which is:

validationAccuracy = 58.7467

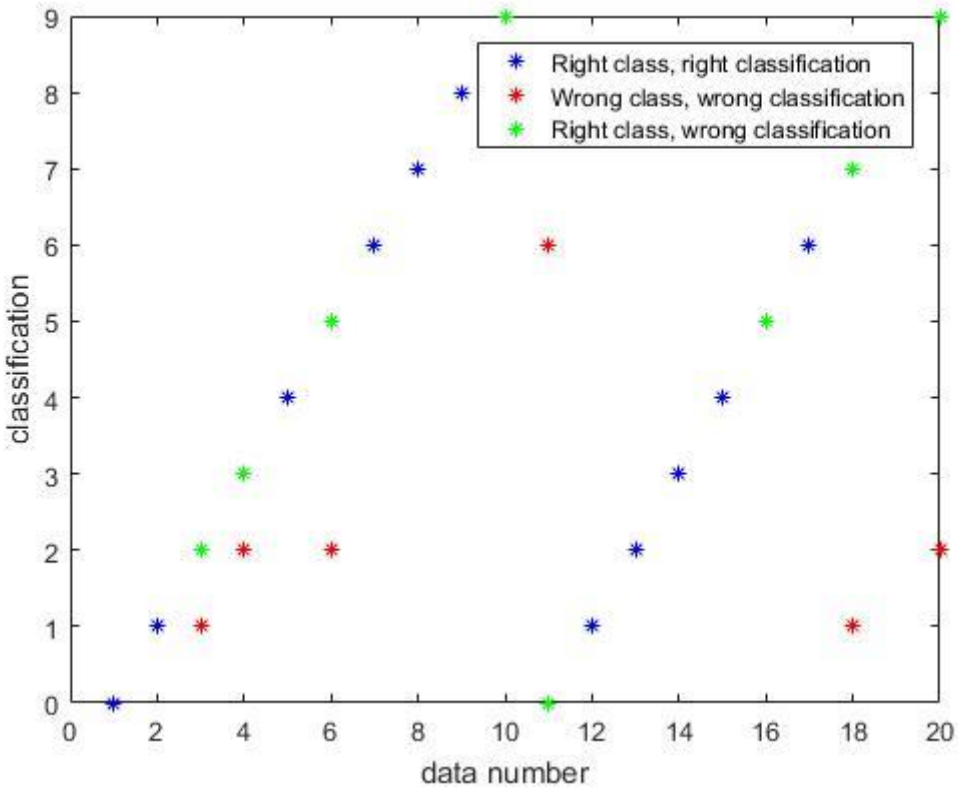
Then the performance of the classification is tested on the dimensionality reduced Test set, and the accuracy percentage is found as :

accuracyPTest = 57.0395

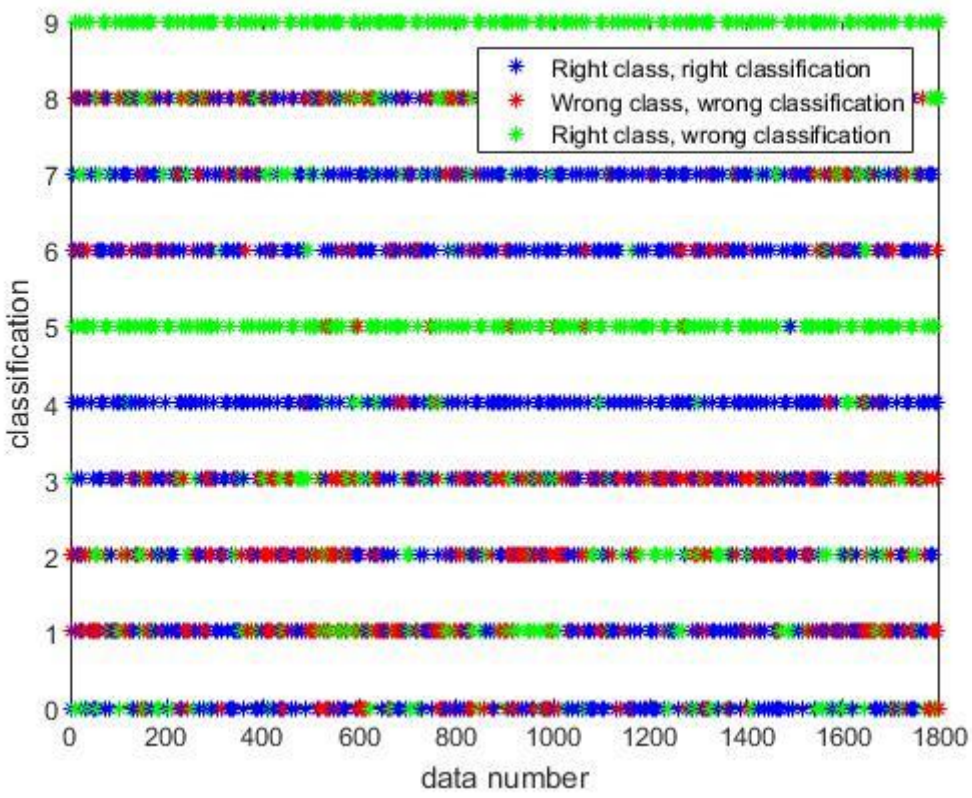
The first figure below is the classification of the first 20 data in the dimensionality reduced Test set.

The second figure below is the classification of the whole dimensionality reduced Test set.

1)



2)



Comparison between Q1 and Q2:

The plots of the result of the first and second question are given. The results of the 1. question is different than the results of the 2. question. It is easy to see that there are many more misclassification in the result of the 2. question since there are many more red and green dots there. Because there are many more misclassifications in the classification of dimensionality reduced data, it can be inferred that 2 features are not enough to classify this data correctly. 1. classification is by all means better.

Comparison between Q2 and Q3:

Accuracy of Q2 is higher because the red and green dots in the plot of Q3 is more. 2 features are not enough for Q2 or Q3 but dimensionality reduction in Q2 works better for this case.

Question 3: The dimension of the data is reduced to 2 using autoencoder, as a result when the data set is classified using neural networks(MLP) only 2 features of the data are used. Dimensionality reduction is first applied to training dataset; that means, eigenvalues are computed on training set and then they were applied to test set. After the dimensionality reduction, the dataset is classified using neural networks(MLP). The training of the data is done on the dimensionality reduced Train set and optimum weights are obtained by evaluating classification results using 10-fold cross validation. Then the performance of the classification is tested on the dimensionality reduced Test set.

The accuracy percentages of each fold cross validation of the Train set is given below:

accuracyP =

22.8571

20.4188

24.0838

21.2042

19.6335

24.6073

18.5864

22.7749

24.0838

20.4188

The accuracy percentage of this classification is their average, which is:

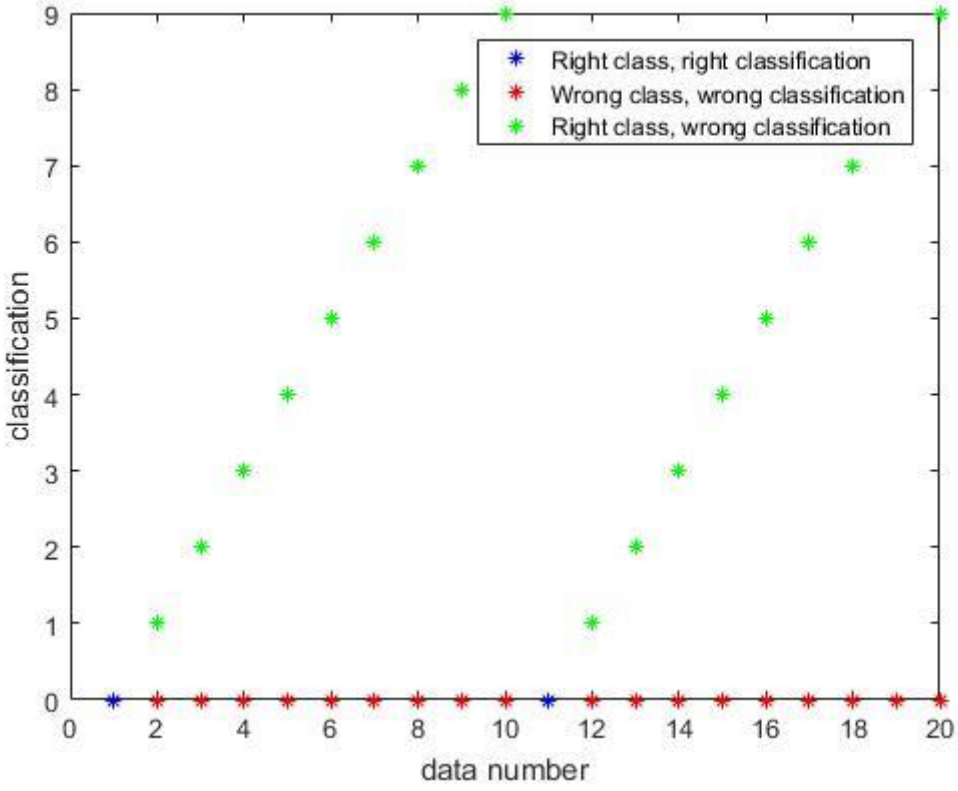
validationAccuracy = 21.8669

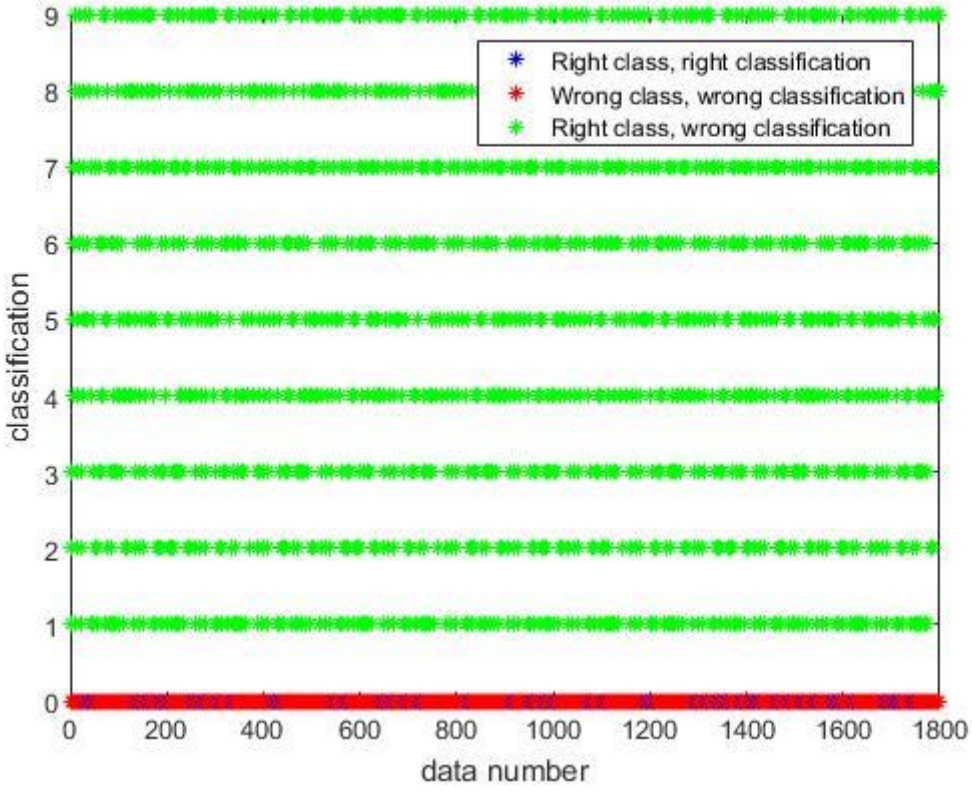
Then the performance of the classification is tested on the dimensionality reduced Test set, and the accuracy percentage is found as :

accuracyPTest = 9.9054

The first figure below is the classification of the first 20 data in the dimensionality reduced Test set.

The second figure below is the classification of the whole dimensionality reduced Test set.


Right class, right classification
Wrong class, wrong classification
Right class, wrong classification
classification
data number
0
1
2
3
4
5
6
7
8
9
0
2
4
6
8
10
12
14
16
18
20


Right class, right classification
Wrong class, wrong classification
Right class, wrong classification
classification
data number
0
1
2
3
4
5
6
7
8
9
0
200
400
600
800
1000
1200
1400
1600
1800

The results in Q3 are different from the results in Q2 and Q1. Dimensionality reduction in Q3 increases the error. The results of Q1 and Q2 are more accurate than Q3.