Statistical Machine Learning Task

Nearest Neighbors Method for Digit Recognition

Agbatan Fiacre Luc KOUDERIN

African Institute for Mathematical Sciences (AIMS - Rwanda)

December 22, 2024

AIMS | African Institute for Mathematical Sciences RWANDA

Outline

- Multi-Class Classification (MCC)
- Binary Classification (BC)

MCC : Data description

- We will use the MNIST data available in R in the library. *dslab*
- We will take 5% of the whole data and then our data can be describe as :

```
Training set size: 3002
Test set size: 502
Training class proportions:
ytrain
         0          1          2          3          4          5          6          7
0.10193205 0.12225183 0.08627582 0.09427049 0.10526316 0.08727515 0.09893404 0.10726183
         8          9
0.10426382 0.09227182
Test class proportions:
ytest
         0          1          2          3          4          5          6          7
0.07370518 0.14143426 0.09960159 0.10358566 0.09561753 0.08764940 0.08964143 0.10956175
         8          9
0.10358566 0.09561753
```

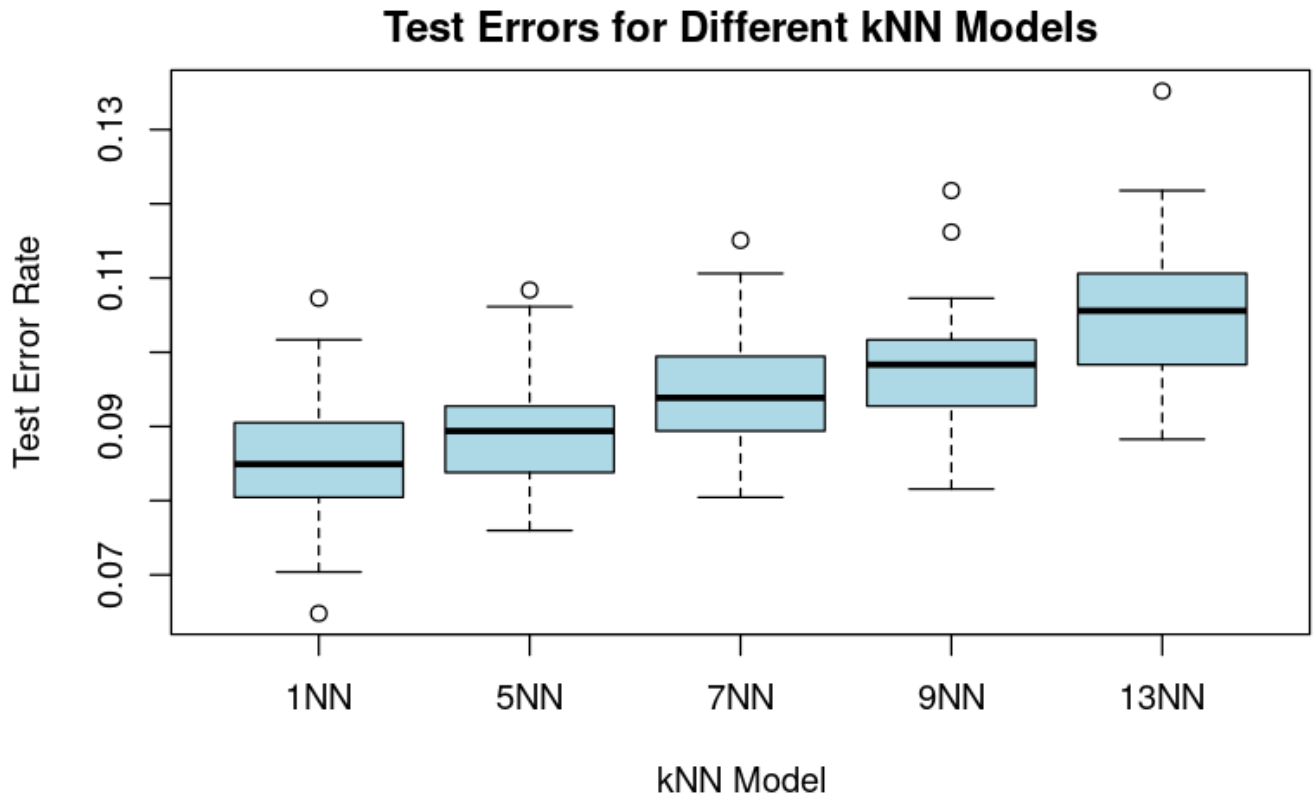
MCC : Models fitting and comparison

- We are using kNN machine with 5 differents values of k.
- Using 50 as the number of random splits of the data into 70% training and 30% test, we have this summary of the errors :

```
[1] "Summary of test errors:"
      1NN               5NN               7NN               9NN              13NN
 Min.   :0.06480   Min.   :0.07598   Min.   :0.08045   Min.   :0.08156   Min.   :0.08827
 1st Qu.:0.08073   1st Qu.:0.08380   1st Qu.:0.08939   1st Qu.:0.09274   1st Qu.:0.09860
 Median :0.08492   Median :0.08939   Median :0.09385   Median :0.09832   Median :0.10559
 Mean   :0.08592   Mean   :0.08914   Mean   :0.09444   Mean   :0.09763   Mean   :0.10532
 3rd Qu.:0.09050   3rd Qu.:0.09274   3rd Qu.:0.09916   3rd Qu.:0.10168   3rd Qu.:0.11034
 Max.   :0.10726   Max.   :0.10838   Max.   :0.11508   Max.   :0.12179   Max.   :0.13520
```

MCC : Models fitting and comparison

- The boxplot presenting this errors is given by :



MCC : Models fitting and comparison

- The confusion matrix of the best machine is given by :

```
[1] "Confusion Matrix for the Last Split (1NN):"
          Reference
Prediction  0   1   2   3   4   5   6   7   8   9
         0 88   0   1   0   0   0   1   0   0   1
         1  0 110   0   0   0   0   0   0   0   0
         2  1   2  67   2   2   0   1   0   0   2
         3  0   0   0  79   0   1   0   2   1   1
         4  0   0   0   0  89   0   0   0   0   5
         5  1   0   0   4   0  66   1   0   5   1
         6  1   1   0   0   0   0  87   0   0   0
         7  0   3   1   0   0   0   0  86   0   6
         8  1   5   0   1   1   2   2   0  79   2
         9  0   0   1   0   3   1   0   3   0  75
```

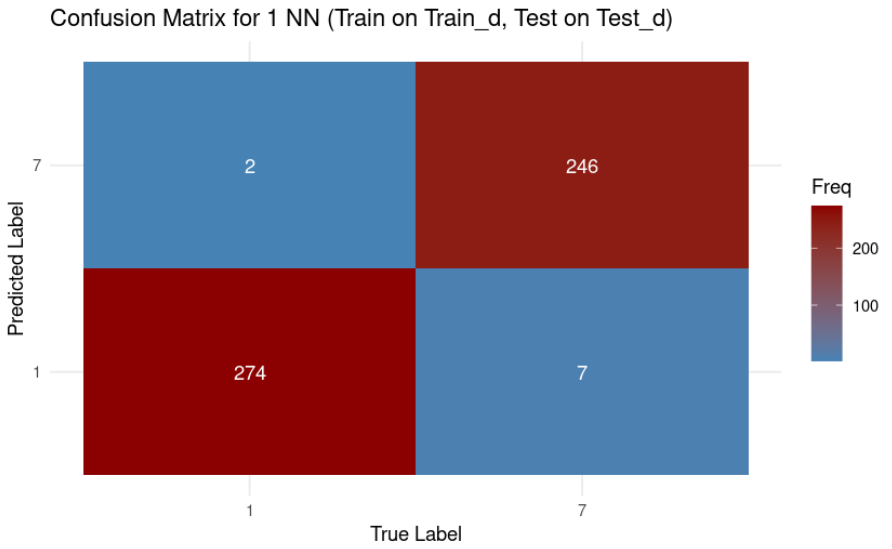
- Most digits misclassified : 7 and 9 , 4 and 9 , 5 and 8.

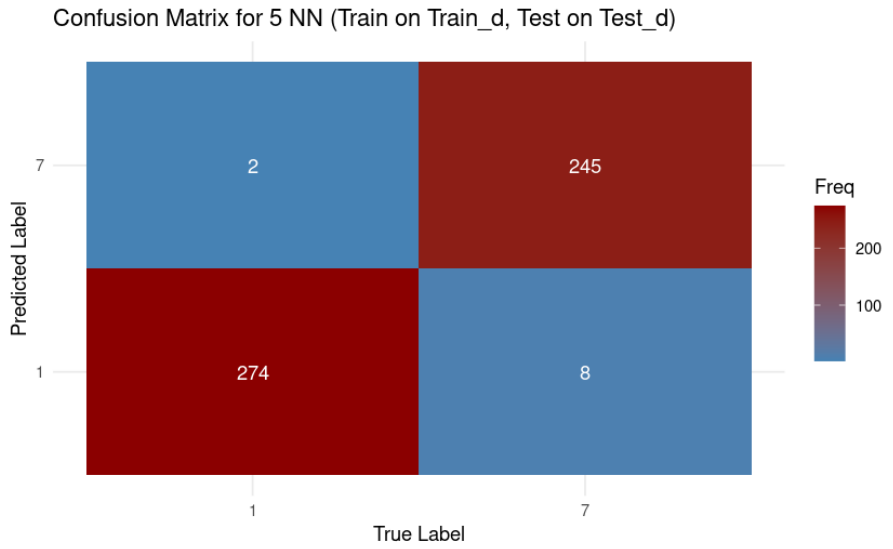
BC : Data description

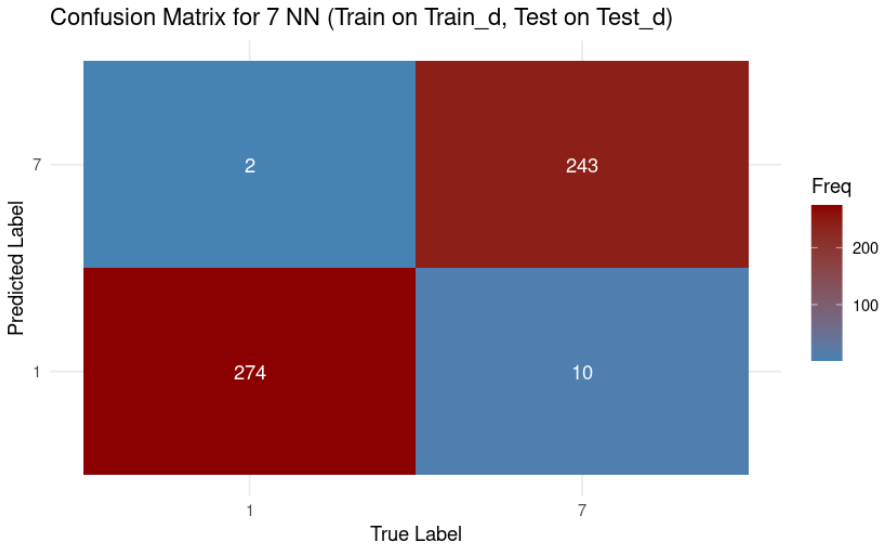
- Here, we are going to classify digit '1' against digit '7'
- Our new data can be described as :

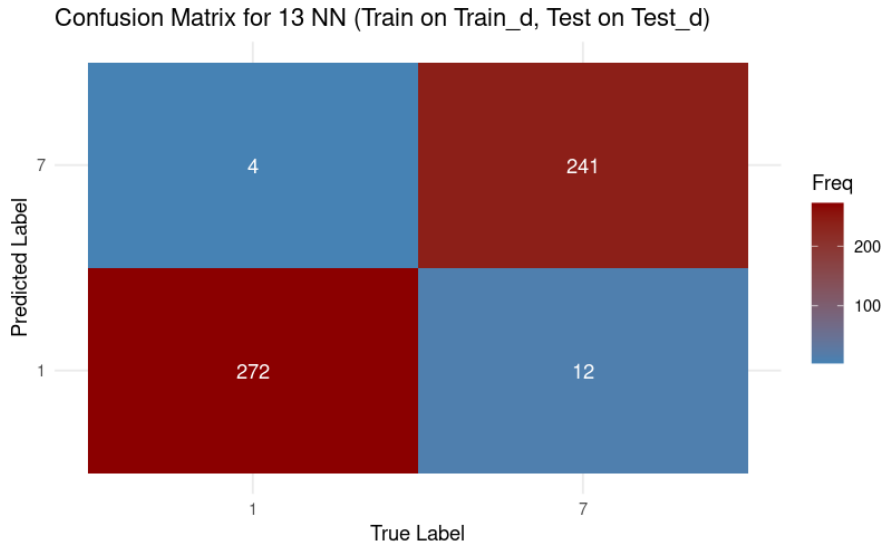
```
Training set size: 2120
Test set size: 529
Training class proportions:
ytrain_d
        1         7
0.5221698 0.4778302
Test class proportions:
ytest_d
        1         7
0.5217391 0.4782609
```

BC : Confusion matrix

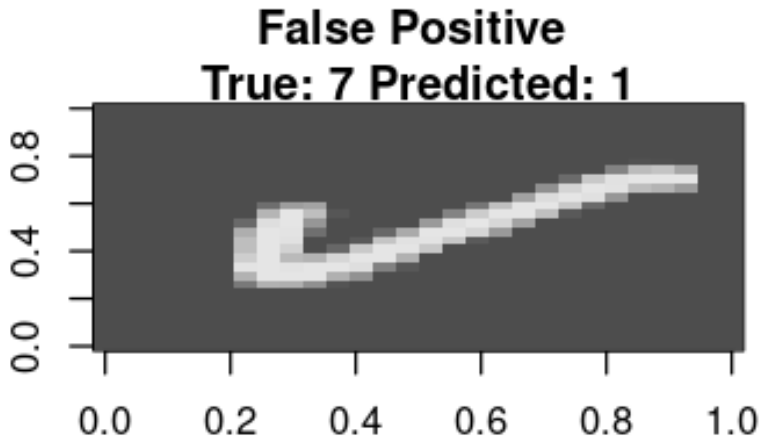


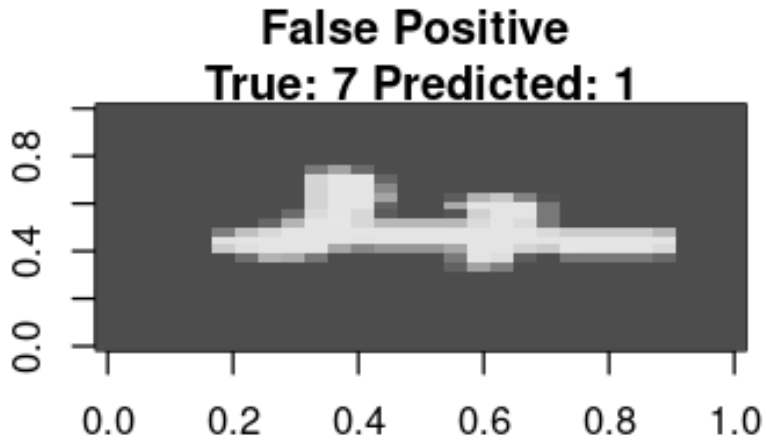


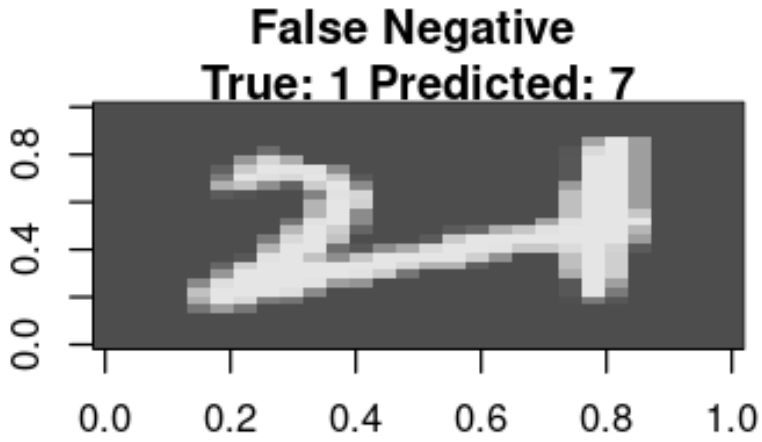


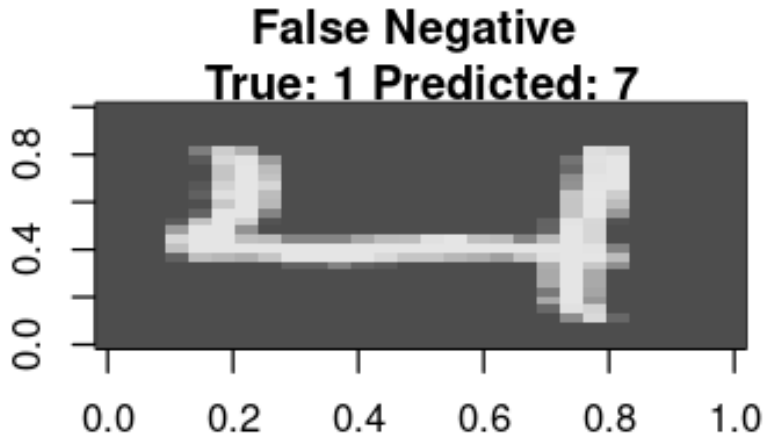


BC : Examples of misclassification









THANK YOU FOR YOUR ATTENTION