Computer Lab 1 block 1

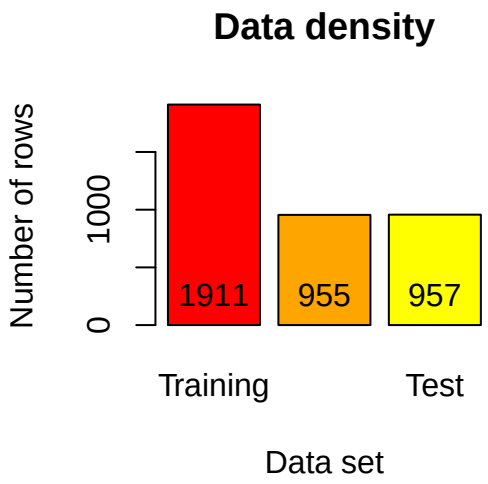
Kerstin(kerni714), Tugce() and Alan(alaca734)

2022-11-16

Assignment 1 Handwritten digit recognition with Knearest neighbors.

1.1 Data set Partition

The original data set **"optdigits.csv"** has a dimension of **3823 rows and 65 columns.** The last column[65] is the target, and represent which number the each row is. After dividing the data set with the proportions indicated the number of rows for each sub set are: **Training: 1911, Validation: 955 and Test: 957**.



1.2 Training the model (Confusion matrices and Misclasification errors)

The Misclassification rate is calculated as follows:

$MR = \frac{FP+FN}{TotalCases}$

- **FP**= False Positives
- **FN** = False Negatives
- **FP + FN**= Incorrect Predictions

The Miscclassification Rate of the Training data set is 4.50 % The Miscclassification Rate of the Test data set is 5.33 %

From the confusion matrices of Testing and Training data sets, it can be calculated the Missclasification rate for each class. For the Training the numbers 1, 4 and 9. For the Testing the numbers 4, 5 and 8, are the ones with highest error.

The Misclasifiation error from training and test are similar which it indicates that the model is too simple, but it will not overfit.

```
##    pred_test
##       0   1   2   3   4   5   6   7   8   9
##   0  77   0   0   0   1   0   0   0   0   0
##   1   0  81   2   0   0   0   0   0   0   3
##   2   0   0  98   0   0   0   0   0   3   0
##   3   0   0   0 107   0   2   0   0   1   1
##   4   0   0   0   0  94   0   2   6   2   5
##   5   0   1   1   0   0  93   2   1   0   5
##   6   0   0   0   0   0   0  90   0   0   0
##   7   0   0   0   1   0   0   0 111   0   0
##   8   0   7   0   1   0   0   0   0  70   0
##   9   0   1   1   1   0   0   0   1   0  85
```

```
##    pred_train
##       0   1   2   3   4   5   6   7   8   9
##   0 202   0   0   0   0   0   0   0   0   0
##   1   0 179  11   0   0   0   0   1   1   3
##   2   0   1 190   0   0   0   0   1   0   0
##   3   0   0   0 185   0   1   0   1   0   1
##   4   1   3   0   0 159   0   0   7   1   4
##   5   0   0   0   1   0 171   0   1   0   8
##   6   0   2   0   0   0   0 190   0   0   0
##   7   0   3   0   0   0   0   0 178   1   0
##   8   0  10   0   2   0   0   2   0 188   2
##   9   1   3   0   5   2   0   0   3   3 183
```

```
## Misclasification rate per class - Train
```

```
##    0    1    2    3    4    5    6    7    8    9
## 0.00 8.21 1.04 1.60 9.14 5.52 1.04 2.20 7.84 8.50
```

```
## Misclasification rate per class- Test
```

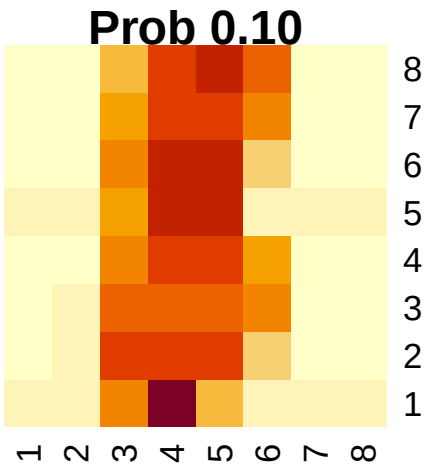
```
##     0     1     2     3     4     5     6     7     8     9
##  1.28  5.81  2.97  3.60 13.76  9.71  0.00  0.89 10.26  4.49
```

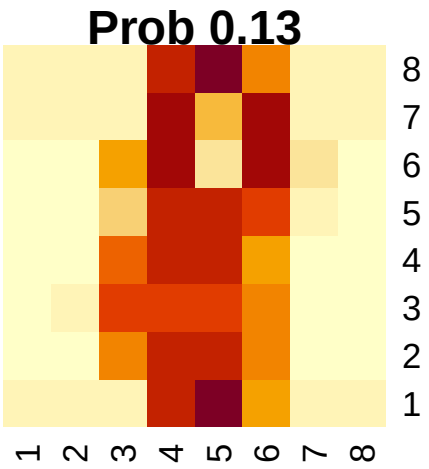
Add the percentage error per class and explain

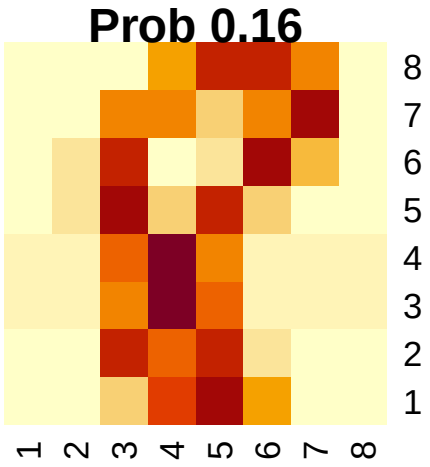
1.3 Easiest and Hardest cases for digit "8"

The heatmap plot shows that the three cases with lowest probability or the hardest to predict, have shapes which the two circles of the 8 figure is not define in comparison with the easiest to predict that both circles are well define. Another detail is that the cases with highest probability have darker colors on the perimeter of the shape, this indicates that there is more pixels in that position and is easier for the model to identify.

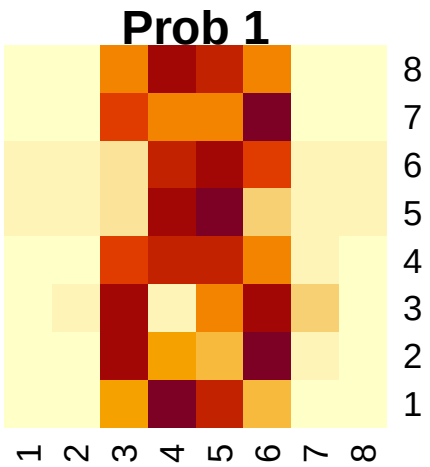
Hardest

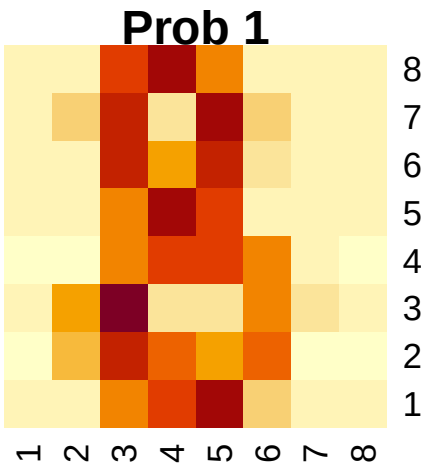






Easiest



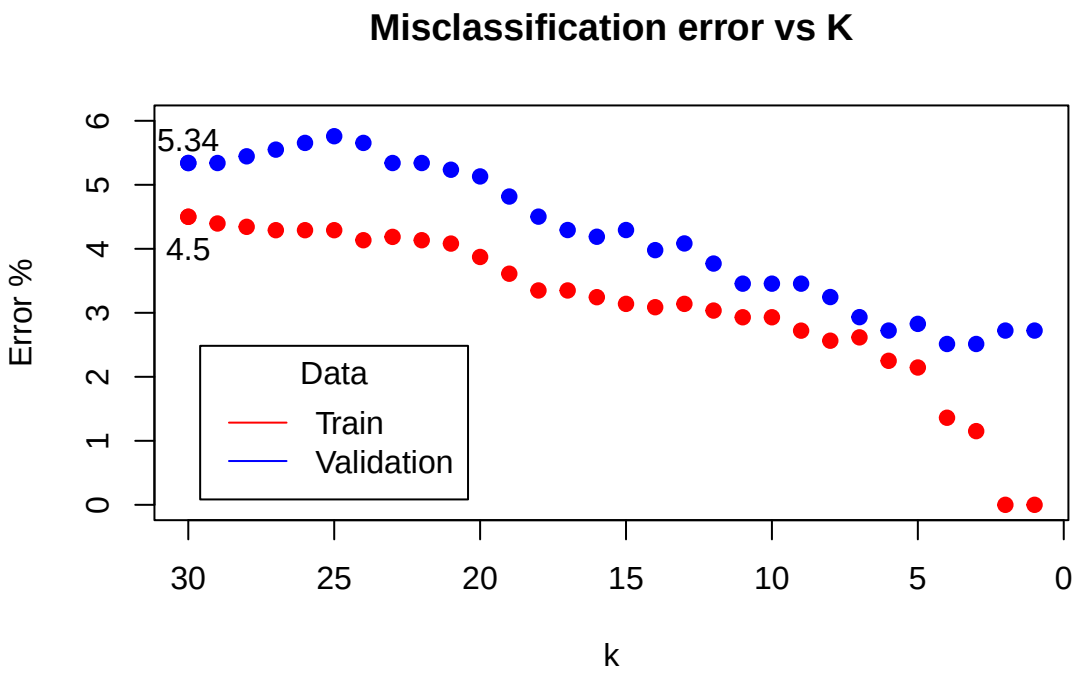


1.4 Model Complexity

The complexity of the models increases as the number of K or numbers nearest neighbors decreases. The optimal K is number 4 because it has the lowest Misclassification Rate. Compare with K equals 3, it has the same error but we chose 4 because it has smaller complexity.

From the plot it can be seen that K equals to 1 and 2, for the training data set, resulted in a Misclassification rate equals to 0. For K equals to 1, the model will just identified each observation as only point to consider while training, so identifies each observation as single case which could lead to over fitting.

After training different models changing the value of k= [1:30], with the training and validation data. From the plot there is a substantial difference between the models, the model train and test with the training data show a smaller Misclassification error than the Validation model.



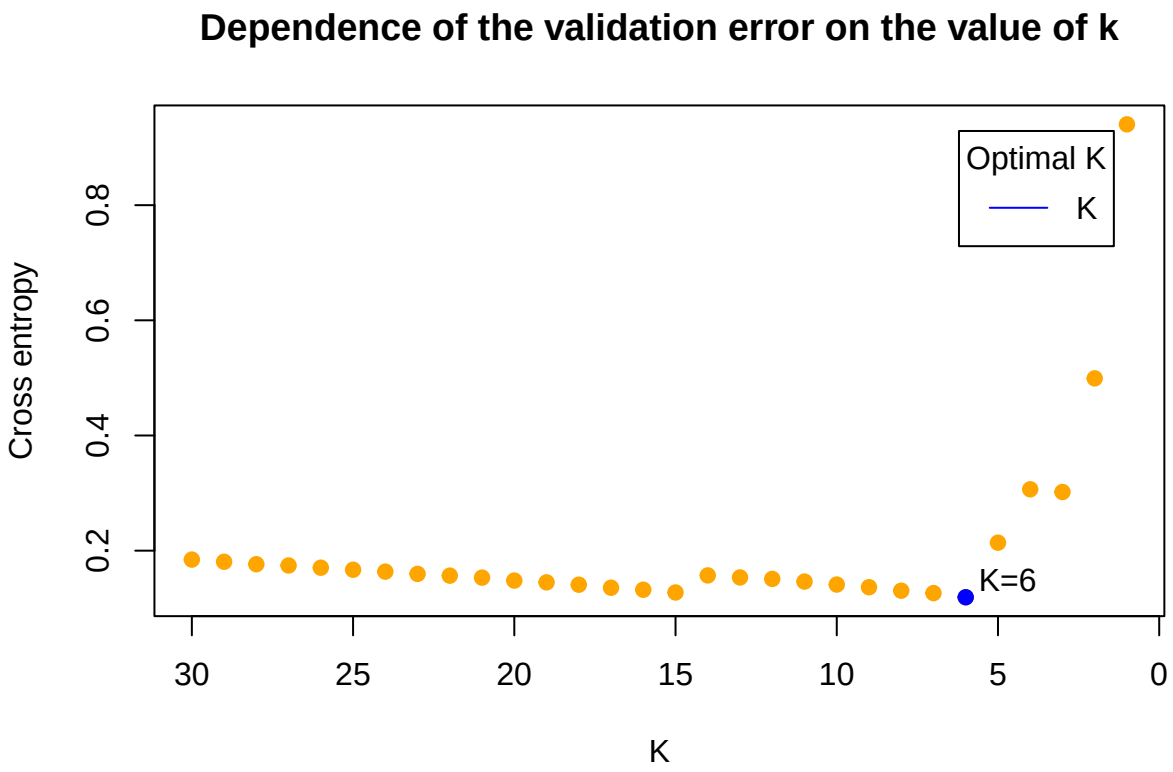
1.4.1 Misclassification error with Test data set, with optimal K The overall Misclassification error of the test data has a similar value to the model with the validation data. Looking at the error rates per class, we can see that the highst error is 5.5%, which could be thought as a good model.

```
##    0    1    2    3    4    5    6    7    8    9
## 1.28 2.33 0.99 1.80 5.50 1.94 0.00 1.79 5.13 4.49
```

```
##  Missclasification rate trainning % Missclasification rate validation %
##                                1.360                                2.513
##       Missclassification_rate_Test
##                              2.510
```

5. Cross entropy for training data

The optimal value of K is 6, because is the one with lowest cross entropy. Cross entropy is a better error performance metric than Misinterpretation rate for multiclass classification; Cross entropy penalize the lower probabilities on each observation, so it gives a better sense of how is performing your model.



Statement of Contribution

Assignment 1 was contributed by Alan Cacique. Assignment 2 was contributed by Tugce Izci. Assignment 3 was contributed by Kerstin Nilsson. All three assignments procedures and results were review and discussed before the creation of the final report to ensure the understanding of the subjects of each person of the team. A long side each team member went individually on each assignment.