Computer Lab 1 block 1

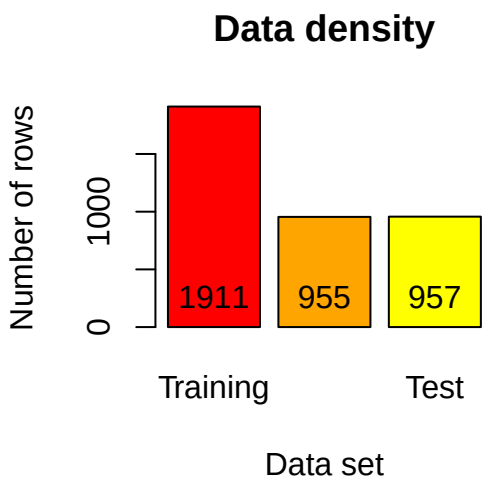
Kerstin(kerni714), Tugce() and Alan(alaca734)

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

$$R(Y, \hat{Y}) = \frac{1}{N}\sum_{i=1}^{N} I(Y_i \neq \hat{Y}_i)$$

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

From the confusion matrices of Testing and Training data sets, we calculated the prediction error for each class as the number that was erroneously classified divided by the total number of observations for that class. For the Training the numbers 1, 4 and 9. For the Testing the numbers 4, 5 and 8, are the ones with the highest error. The error rates for the test set ranges from 0-14%, thus the highest prediction error for the individual digits is almost three times as high as the overall error rate.

The Misclassification error from training and test are similar which indicates that the model might be too simple.

```
##    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
```

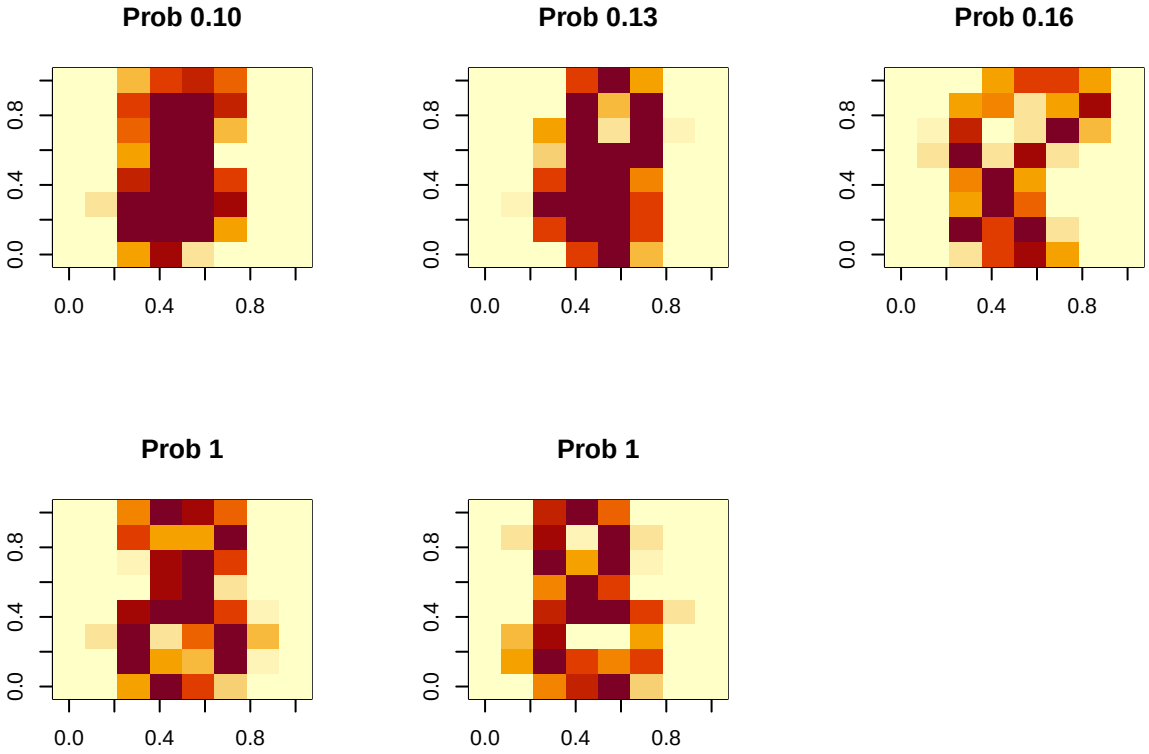
```
## Misclasification rate per class- Test
```

```
##     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
```

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

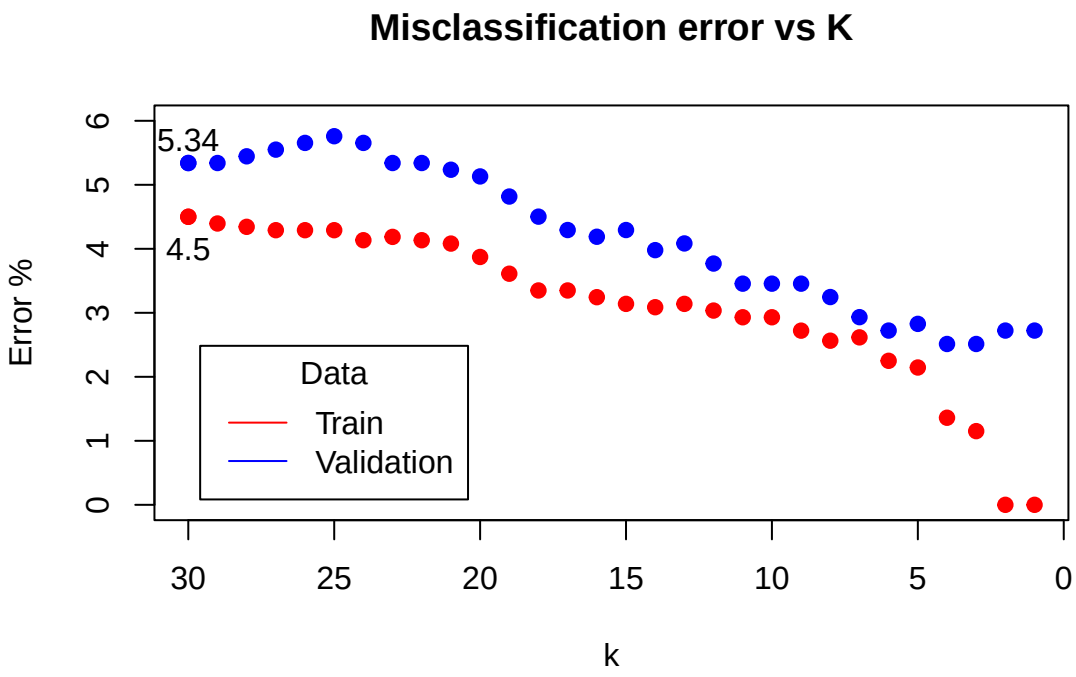
are well defined. Another detail is that the cases with highest probability have darker colors on the perimeter of the shape, this indicates that there is more pixels in that position and is easier for the model to identify.



1.4 Model Complexity

The complexity of the models increases as the number of K or numbers of nearest neighbors decreases. For the training set the error decreases with increasing complexity, down to 0 for K equals 1 and 2. The validation error also decreases generally until K equals 4. The Validation error then starts to increase from K equals 2. The optimal K is number 4 because it has the lowest Misclassification Rate and compared with K equals 3, that has the same error we choose K=4 because it has lower complexity.

From the plot it can be seen that for K equal to 1 and 2, for the training data set resulted in a Misclassification rate equal to 0. For K equal to 1 for the training set, the model will just identify for each observation itself as the closest data point, meaning the prediction error will be 0 for the training set, and the model will highly likely lead to overfitting.



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, around 2.5%, which is now higher than for training dataset (error 1.4%). Looking at the error rates per class, we can see that the highest error is 5.5%, which might be thought of as a good model (but it also depends on the application).

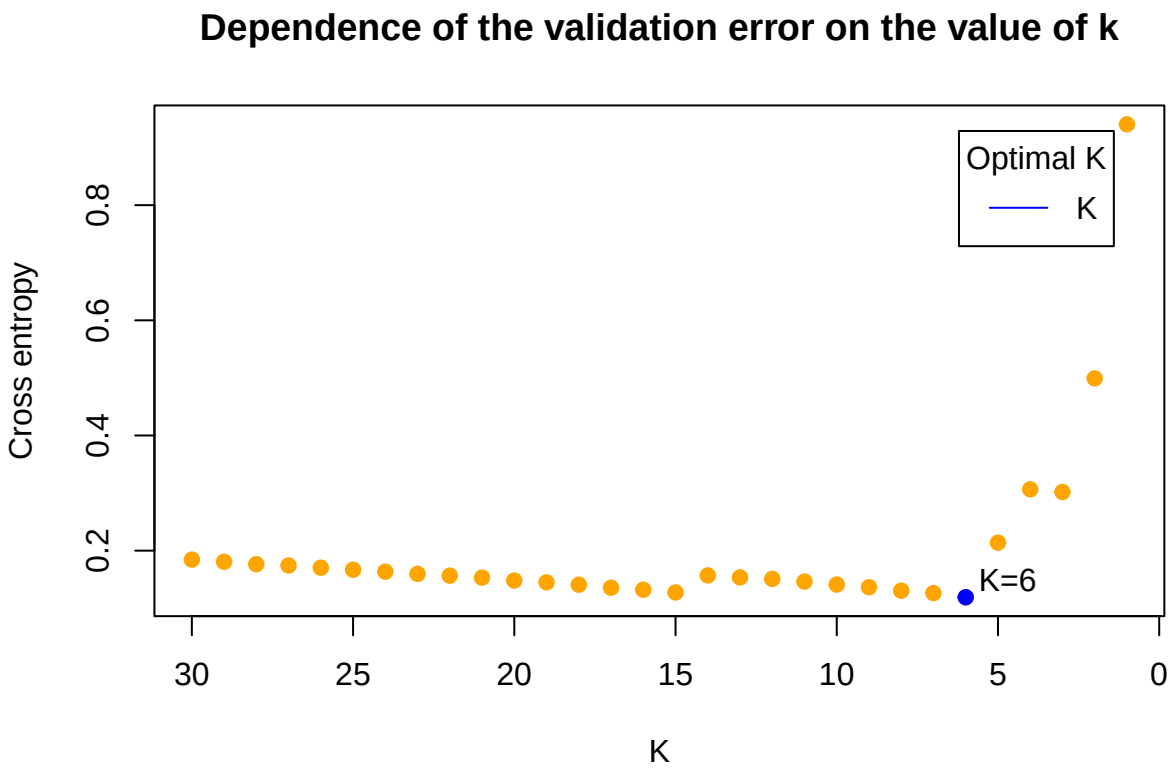
```
##  Misclasification rate trainning % Misclasification rate validation %
##                                1.360                                2.513
##       Missclassification_rate_Test
##                              2.510

## By class error rates:

##    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
```

5. Cross entropy for training data

The optimal value of K is 6, because it 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 for the targets more (penalizes higher loss to predictions), 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.