Exercise 3 – Clustering

Reading material for this exercise.

- Chapter 9 (Finding Groups of Data - Clustering with K-Means) of "Machine Learning with R" (Second Edition) by Brett Lantz, Packt Publishing Ltd., 2015
- Chapter 10.3 (Clustering Methods) of "An Introduction to Statistical Learning with Applications in R" from G. James, D. Witten, D., T. Hastie, R. Tibshirani. Springer 2013.

Exercise 3.1 K-means clustering:

3.1.1 Try to improve the performance of two persons training data (disjunct). Perform K- means clustering of each cipher individually for the training set, in order to represent the training data as a number of cluster centroids. Now perform the training of the k-NN using the centroids of these clusters. You can try with different cluster sizes and see the resulting performance.

```
set.seed(2345)

cipher_cluster <- c()
label_cluster <- c()

for( i in 0:9) {
  clusterData <- kmeans(id_train[ id_train_labels == i, ], 200)
  cipher_cluster[[i + 1]] <- clusterData$centers
  label_cluster[[i + 1]] <- c(1:200)*0 + i
  }

train_lab <- factor(unlist(label_cluster))
train_dat <- do.call(rbind, cipher_cluster)
```

3.1.2 Compare your KNN performance based on the raw training data and based on the cluster centroids of the training data. During the comparison you should also consider the run times of the algorithm. As the generation of clusters is based on random starting points cross-validation should be performed.

3.1.3 Perform K-means clustering on each cipher individually for the training data from all the available datasets (disjunct). Represent the training data as a number of cluster centroids and compare performance, try multiple cluster sizes.

Exercise 3.2: Hierarchical clustering

3.2.1 Show a low level dendrogram containing 5 instances of each digit (one person).

3.2.2 Use K-Means clustering to compress each digit into 5 clusters, as done in 3.1.1, and perform hierarchical clustering to show a low level dendrogram of this (one person).

3.2.3 Discuss the results and relate them to the cross validation tables from k-NN classification.

Exercise 3.3: Evaluation methods of k-NN:

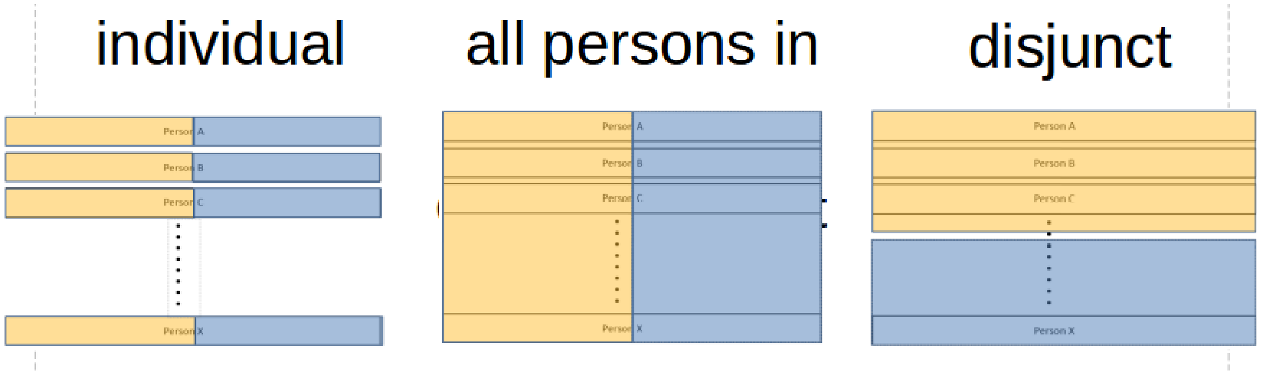
As seen in the hierarchical clustering plot we often get different labels when finding the nearest neighbors of different ciphers. This indicates that we are not completely sure about our estimation. Until now, in k-NN we have simply used the one with most votes. But we can also exclude predictions which does not have enough of the same labels.

In k-NN we can set the “l” to the minimum number of “k” nearest neighbors of the strongest label to accept a match.

3.3.1 Plot the precision-recall curves for 1 to 13 “k” with “l” values up to the “k” value. Here, the results should be one plot containing “k” lines, and each one have “k” datapoints.

3.3.2 Plot the maximum F1 values for each of the k in a plot together. With F1 score on the y- axis and “k”-value on the x-axis.

3.3.3 Discuss the results from 3.3.1 and 3.3.2. What do you think would be the most important part of a digit recognition system. Precision or recall? Please discuss in what situations would the different factors be more important?


individual
all persons in
disjunct
Person A
Person B
Person C
Person X
Person A
Person B
Person C
Person X
Person A
Person B
Person C
Person X