Exercise Sheet 3: Clustering

Exercise 3.1 K-means clustering:

By using K-means clustering the dataset can be sparsified, and timing and accuracy performance can be improved.

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.

Code example of clustering is in the SVN (k-clustering.R).

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 the entire class (disjunct), or a large part of it, e.g. 30 persons. 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:

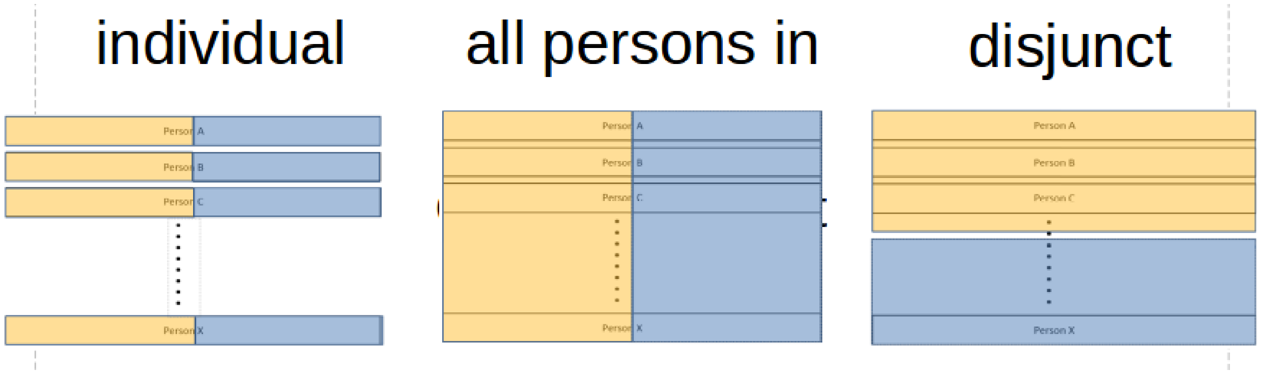
As seen in the hierarchical clustering plot we often get different labels when finding the nearest neighbors of different ciphers. This could indicate that we are not completely sure about our estimate. Until now, in k-NN we have simply used the one with most votes. But we can also exclude predictions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

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

3.3.1 Plot the 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, who each 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, in what situations would the different things be 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