Exercise 1 – Data preparation and KNN

This exercise consist of the following tasks: Downloading the data, setup R and perform K-nearest neighbor classification. There will be included both a guide to do it in Windows and Ubuntu.

It is important to **read** the provided material:

- Chapter 1 (Introducing Machine Learning)
- Chapter 3 (Lazy Learning - Classification using Nearest Neighbours) of "Machine Learning with R" (First Edition) by Brett Lantz, Packt Publishing Ltd., 2013

1.1 Downloading Data

1.1.1 Download data files located on itslearning: Resources/Exercises/Data (e.g. id100.Rda)

1.2 Installing R and libraries

This tasks concerns the installation of R and Rstudio, respectively the programming language and the graphical front-end. Additional software packages used in the course are also installed. It is important to remember to install missing packages.

1.2.1 Install R and Rstudio.

Windows & Mac:

http://cran.r-project.org/bin/windows/base/,

get Installer (not tarball) from http://www.rstudio.com/products/rstudio/download/

Ubuntu:

In terminal write:

```
“sudo apt-get install r-base r-base-dev”
```

get Installer (not tarball) from http://www.rstudio.com/products/rstudio/download/ and use software manager

1.2.2 Now open R-Studio.

You also need to get some packages, missing packages can be installed by the command install.packages(), e.g. to install gmodels, type:

```
install.packages("gmodels", dependencies=TRUE)

install.packages("class", dependencies=TRUE)

install.packages("caret", dependencies=TRUE)

install.packages("swirl", dependencies=TRUE)
```

1.2.3 To learn programming in R we will use “swirl”, type:

```
library(swirl)

swirl()
```

Follow the guide to complete:

- 1: R Programming
  - 1: Basic Building Blocks
  - 12: Looking at Data

1.2.4 Load the data as:

```
        load("id100.Rda")
```

The data is now loaded as “id” in your workspace.

1.2.5 Now analyse the data, analyse “id” using the tools you have learned in “12: Looking at Data”

1.2.6 Documentation on R can be found on http://www.r-project.org/

1.3 Performing K-nearest neighbor (Report 1)

The following exercise will concern the topic of today and it is required that the exercises are made into a report. The method will be tested on your created cipher data.

The dataset should be split into 50/50 for training and testing. Before splitting the dataset should be shuffled, remember to set a seed for reproducible results:

```
set.seed(423)
```

```
dataset_shuffle <- dataset[sample(nrow(dataset)),]
```

1.3.1 K-Nearest Neighbour: Using the methods learned in the Chapter 3 in “Machine Learning With R”, KNN can now be performed on our own generated dataset. First we will test on a single person. Remember to split between training and test set (split in two equally sized parts). Document the results. Can you explain the performance (computation time and test accuracy) on the training and test-set?

1.3.2 Performance of varying K: Analyse performance with varying K.

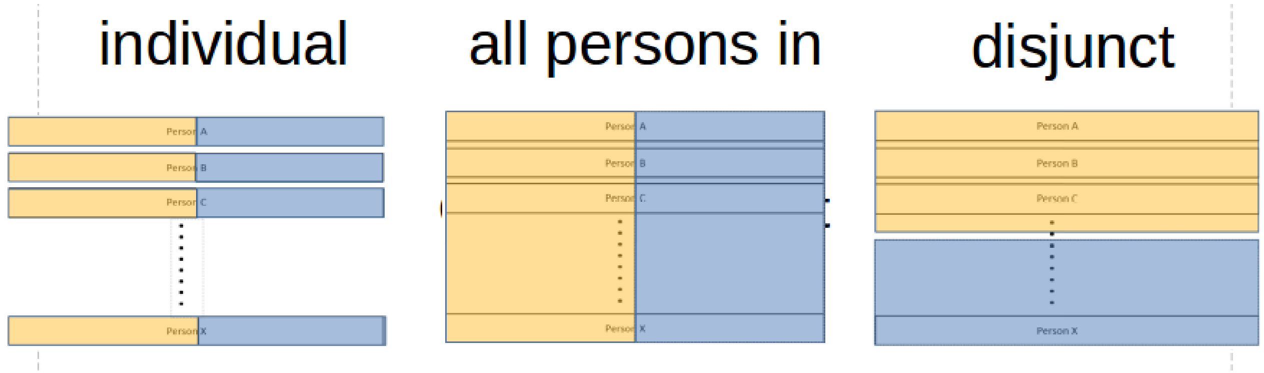
1.3.3 Cross validation: Perform a cross validation with a 90% / 10% split with 10 runs. Report mean and standard deviation of the performance.

```
folds <- createFolds(id$X1, k = 10)
```

1.3.4 Person independent KNN: Now try to apply k-nearest neighbor classification to the complete data set from all students attending the course. Distinguish two cases: Having data from all individuals in the training set and splitting the data according to individuals. Generate and explain the results.

```
load("idList-nc-100.Rdata")
id <- do.call(rbind, idList[1:10])
id <- as.data.frame(id)
id$X1 <- factor(id$X1)
```

1.3.5 Performance of sample size: Lastly report computational time of the prediction step for varying ‘k’ and using a small and large datasets. You don’t have to test every ‘k’ simply give an overview. Discuss how the accuracy changes with different sizes of the dataset, is ‘k’ dependent on the dataset size?


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