Introduction to Machine Learning
NPFL 054

http://ufal.mff.cuni.cz/course/npfl054

Barbora Hladká **Martin Holub**

{Hladka | Holub}@ufal.mff.cuni.cz

Charles University,
Faculty of Mathematics and Physics,
Institute of Formal and Applied Linguistics

Programming questions

- **Feature frequency**
  - Implement a function that receives a vector of 1s and 0s and returns the number of 1s.
- **MOV data set**
  - Run `https://ufal.mff.cuni.cz/~hladka/2021/docs/load-mov-data.R`
  - For each genre-related feature compute its feature frequency. Plot all the feature frequency values.
  - Produce side-by-side boxplots of the ratings of the movies rated 67 times. For each movie, plot a symbol for its average rating in the corresponding boxplot.

Programming questions

- **USArrests data set**
  - `d <- USArrests`
  - Print a vector of the state names from the highest `Assault` rate to the lowest `Assault` rate.
  - Produce a scatter plot of `Rape` and `Murder`.
  - Compute Pearson correlation coefficient for `Murder` and `Rape`, `Rape` and `UrbanPop`, `Murder` and `UrbanPop`.
- **Titanic data set**
  - `d <- read.csv("https://ufal.mff.cuni.cz/~hladka/2021/docs/train.csv")`
  - Get the number of missing values of the AGE feature
  - Create a contingency table for PCLASS and SEX and visualize it using a mosaic plot.
  - Draw a mosaic plot for PCLASS, SEX and SURVIVED.