Introduction to Machine Learning (NPFL054)

HW #1 – Data analysis and clustering in R

The exercises relate to the `mov.development.csv` data set, which you can download from the course webpage `https://ufal.mff.cuni.cz/course/NPFL054/materials`. Upload this data set into R using the `load-mov-data.R` code posted at the website as well.

Hint:

```
source("load-mov-data.R"); ls()
[1] "examples" "movies" "u" "users" "votes"
```

1. Work with `examples`. Compute conditional entropy H(OCCUPATION|RATING).

   Points: 1

2. Produce side-by-side boxplots of the ratings of the movies rated 67 times. For each movie, draw a point for its average rating in the corresponding boxplot. Provide an interpretation of the boxplots.

   Points: 2

   Hint: Get familiar with the functions `boxplot()` and `points()`.

   Solution: See Figure 1 below.

3. Cluster the users in `users`.

   (a) extend `users` with 5 new features, namely
   - ONE for the relative frequency of ratings 1 assigned by the user
   - TWO for the relative frequency of ratings 2 assigned by the user
   - THREE for the relative frequency of ratings 3 assigned by the user
   - FOUR for the relative frequency of ratings 4 assigned by the user
   - FIVE for the relative frequency of ratings 5 assigned by the user

   !!! Round the relative frequencies to 2 decimal places **!!!**

   (b) use the features AGE, ONE, TWO, THREE, FOUR, FIVE and perform hierarchical agglomerative clustering using average linkage

   (c) cut the dendrogram at a height that results in twenty clusters

   (d) explore the clusters
   - compute the number of users in each cluster
   - compute the average age of users in each cluster
   - for each cluster check whether there are some duplicates, i.e. users of the same age and with identical distribution of ratings

   Points: 7

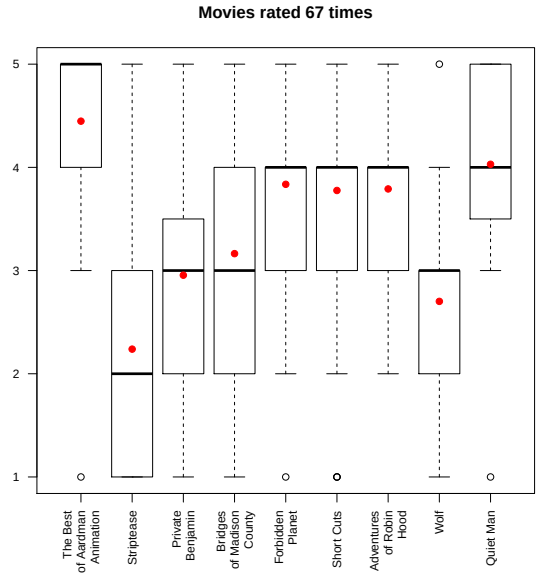

Figure 1: Solution of the exercise 2

How to submit your assignment

- Write your R code to get answers for the exercises and name it `YourLastName_YourFirstName_hw1.R`. Do not use diacritical marks!
- Write your answers into the template file `hw1.odt` posted at the course webpage. Do not change the structure of this file. Save the file as `YourLastName_YourFirstName_hw1.odt` and then export it as `YourLastName_YourFirstName_hw1.pdf`. Do not use diacritical marks!
- E-mail both files `YourLastName_YourFirstName_hw1.[R|pdf]` to the contact person specified in the homework assignment.