INTRODUCTION TO MACHINE LEARNING (NPFL054)
A template for Homework #1

Name: Milan Wikarski

School year: 2019/2020

- **Provide answers for the exercises (1) - (3).**
- **For each exercise, your answer cannot exceed one sheet of paper.**

1. Conditional entropy [1pt]

2. Boxplots of ratings of the movies rated 67 times [2pt]

There are exactly 9 movies rated 67 times. I have generated a boxplot for each of these movies and plotted a point for average rating of each movie.

Boxplot shows the median value (bold line), the Q1 value (bottom part of the box), Q3. value (top part of the box), MIN and MAX, taking into consideration IQR and outliners (if there are some) – values which are outside of the IQR.

We are going to use movie “Short Cuts” as an example. Using boxplot, we can find out, that 67 reviews have:

Median	4
Q1	3
Q3	4
Min (IQR)	2
Max (IQR)	5
Outliners	At least one outliner with value 1

3. Clustering the users [7pt]

Most of the process is explaind in the code using comments.

Althought it was not required I used several normalization methods and compared them. Different methods resulted in different cluster sizes with different average age.

When data was not normalized, the most important factor in clustering was age. This was because values of properties one, two, three, four and five were between 0 and 1, whereas the values of age were between 7 and 73 – therefore the distance was larger.

Using different normalization methods, the importance of age decreased, resulting in more diversity when it comes to the values of age in clusters. We can observe that, when normalizing data, it is more likely for two clusters to have very similar average age. Another very important thing we can observe, when normalizing data, is that the clustering method results in 1 very large cluster (more than half of the users) and 19 smaller clusters. When using no normalization, cluster sizes are more regular, but there seem to be very small clusters made out of users with a value of the age parameter than can be considered an ouliner (very young or very old).

I have created a dendrogram for raw data (not normalized) and for each normalization method, as well as one barplot showing the cluster sizes and one barplot showing the average age in cluster for raw data and for each normalization method. I have used boxes to visually cut the dendrogram into 20 clusters.