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

Name: Pavel Mikuláš

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]**

Using the chain rule property of entropy:

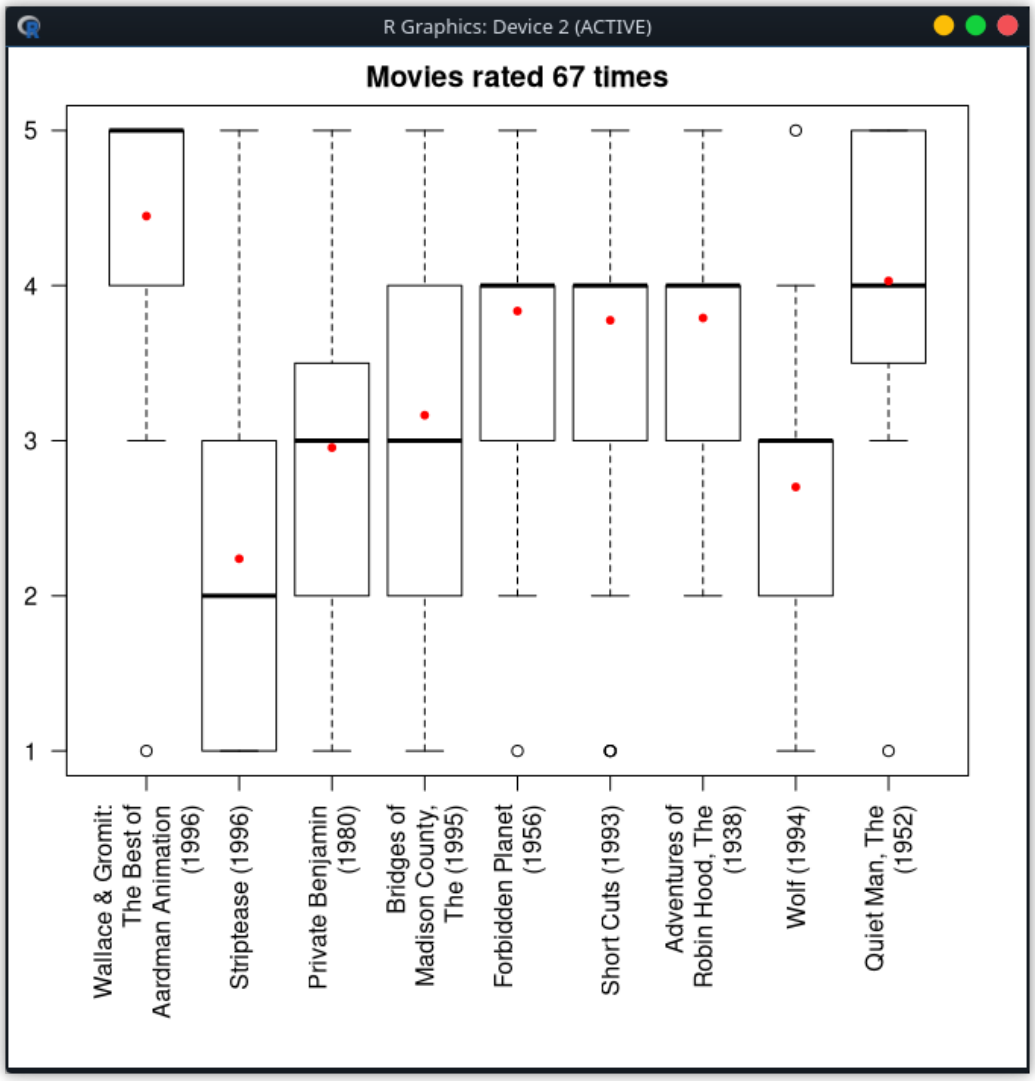
$H(X|Y) = H(Y, X) - H(Y)$

We get:

$H(\text{OCCUPATION}|\text{RATING}) = H(\text{RATING, OCCUPATION}) - H(\text{RATING})$

$= 5.877486 - 2.117503 = \underline{3.759984}$

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



From the ammount of data we had we can safely say that we are only really interrested in the data within the boxes(between 1. and 3. quartile). Because all the other results are really just deviations that tell us very little about the data.

We can conclude movies that Wallace & Gromit: The Best of Aardman Animation and The Quiet Man were really successful among users since they have high median and narrow midspread. While the movie Striptease has really wide midspread and low median which suggests it was controversial and not very popular. But some people even rated it with 5 stars. The movies Forbidden Planet, Short Cuts and The Adventures of Robin Hood were really popular among most users but are just a tithe from the top. And the rest of the movies were deemed as average in the eyes of the users.

3. Clustering the users [7pt]

Population and average age for each cluster:

	population	age
1	105	25.00952
2	33	54.33333
3	65	22.43077
4	96	32.36458
5	82	43.14634
6	15	57.60000
7	48	35.43750
8	142	28.52817
9	79	38.56962
10	65	47.70769
11	82	20.04878
12	1	7.00000
13	37	17.35135
14	46	50.69565
15	12	60.25000
16	14	14.07143
17	11	63.81818
18	2	10.50000
19	7	69.14286
20	1	73.00000

I have found no duplicates by comparing age,rating and cluster.
Not even when only comparing by rating and cluster.