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

Name: Gareth Lafay

School year: 2019/20

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

1. Conditional entropy **[1pt]**

---

entropy of H(OCCUPATION|RATING) ≈ 3.8

I used :

H(Y|X)=H(Y,X)-H(X)=H(Y)-I(X;Y)

I(X;Y)=H(X)+H(Y)-H(X,Y)

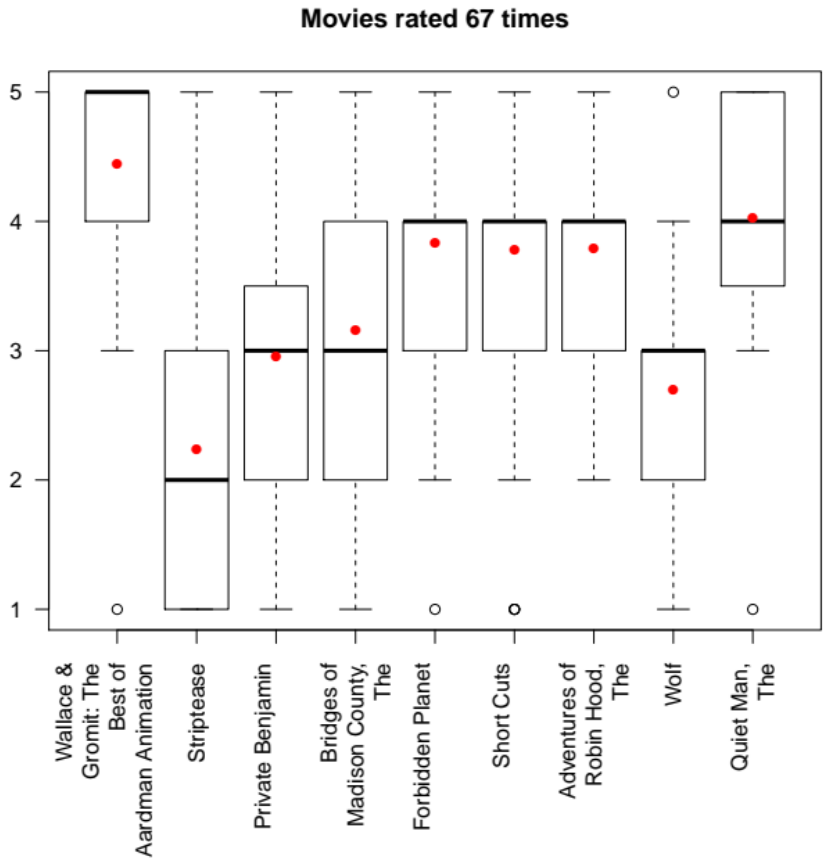
2. Boxplots of ratings of the movies rated 67 times [2pt]

Reasoning

I know that .N is a special variable that holds the number of rows in the current group, so using .(.N) and by=.(movie), I got the numbers of rows for each movie id. .() is a shorthand for list(). I then extracted the movies who had 67 rows meaning they were rated 67 times.

I merged the lists to get the data I wanted. Then I formatted the titles using regular expressions so that the labels looked more like the model.

par() and las() let me adjust the margins and align the labels vertically. box-plot() and points() are the functions that made the graph and displayed the means.



Interpretation of boxplots

We see that it usual to have outliers, for a lot of the movies the rating are all over the place, and the “whiskers” are long. 3 have extremely similar boxplots whith medians at the top of the box, with the average close by meaning a lot of people agreed on the rating which will make it easier to predict.

3. Clustering the users [7pt]

Head of dendogram (3.b)

```
[1] "entropy of H(OCCUPATION|RATING) = " "3.7599835206782"
 [1] "user"       "age"        "gender"     "occupation" "zip"
 [6] "ONE"        "TWO"        "THREE"      "FOUR"       "FIVE"
--[dendrogram w/ 2 branches and 943 members at h = 21.4]
  |--[dendrogram w/ 2 branches and 592 members at h = 16.9]
  |  |--[dendrogram w/ 2 branches and 3 members at h = 3.52]
  |  |  |--leaf 30
  |  |  `--[dendrogram w/ 2 branches and 2 members at h = 1.07] ..
  |  `--[dendrogram w/ 2 branches and 589 members at h = 9.48]
  |     |--[dendrogram w/ 2 branches and 198 members at h = 4.67] ..
  |     `--[dendrogram w/ 2 branches and 391 members at h = 6.37] ..
  `--[dendrogram w/ 2 branches and 351 members at h = 19.9]
     |--[dendrogram w/ 2 branches and 19 members at h = 5.82]
     |  |--[dendrogram w/ 2 branches and 11 members at h = 2.25] ..
     |  `--[dendrogram w/ 2 branches and 8 members at h = 3.87] ..
     `--[dendrogram w/ 2 branches and 332 members at h = 10.6]
        |--[dendrogram w/ 2 branches and 161 members at h = 4.59] ..
        `--[dendrogram w/ 2 branches and 171 members at h = 7.4] ..
etc...
```

Exploration of clusters

	cluster	nbr of users	avr age of users	nbr of duplicates
1	1	105	25.00952	0
2	2	33	54.33333	0
3	3	65	22.43077	0
4	4	96	32.36458	0
5	5	82	43.14634	0
6	6	15	57.60000	0
7	7	48	35.43750	0
8	8	142	28.52817	0
9	9	79	38.56962	0
10	10	65	47.70769	0
11	11	82	20.04878	0
12	12	1	7.00000	0
13	13	37	17.35135	0
14	14	46	50.69565	0
15	15	12	60.25000	0
16	16	14	14.07143	0
17	17	11	63.81818	0
18	18	2	10.50000	0
19	19	7	69.14286	0
20	20	1	73.00000	0