Data analysis Project 2

Correlation and Regression of Movie Ratings Data

Dataset description

This dataset features ratings data of 400 movies from 1097 research participants.

- 1st row: Headers (Movie titles/questions) – note that the indexing in this list is from 1
- Row 2-1098: Responses from individual participants
- Columns 1-400: These columns contain the ratings for the 400 movies (0 to 4, and missing)
- Columns 401-421: These columns contain self-assessments on sensation seeking behaviors (1-5)
- Columns 422-464: These columns contain responses to personality questions (1-5)
- Columns 465-474: These columns contain self-reported movie experience ratings (1-5)
- Column 475: Gender identity (1 = female, 2 = male, 3 = self-described)
- Column 476: Only child (1 = yes, 0 = no, -1 = no response)
- Column 477: Movies are best enjoyed alone (1 = yes, 0 = no, -1 = no response)

Note that we did most of the data munging for you already (e.g. Python interprets commas in a csv file as separators, so we removed all commas from movie titles), but you still need to handle missing data.

Q1:

Note: For all missing values in the data, use the average of the corresponding column so to fill in the missing data.

In this problem, under **the most correlated**, we consider the largest correlation in the absolute value.

1.1. For every user in the given data, find its most correlated user.

1.2. What is the pair of the most correlated users in the data?

1.3. What is the value of this highest correlation?

1.4. For users 0, 1, 2, \dots, 9, print their most correlated users.

Q2:

We want to find a model between the ratings and the personal part of the data. To do so, consider:

Part 1: the ratings of all users over columns 1-400:

-- Columns 1-400: These columns contain the ratings for the 400 movies (0 to 4, and missing);

call this part `df_rate`

and

Part 2: the part of the data which includes all users over columns 401-474

-- Columns 401-421: These columns contain self-assessments on sensation seeking behaviors (1-5)

-- Columns 422-464: These columns contain responses to personality questions (1-5)

-- Columns 465-474: These columns contain self-reported movie experience ratings (1-5)

call this part `df_pers` .

Our main task is to model:

```
df_pers = function(df_rate)
```

Note: Split the original data into training and testing as the ratio 0.80: 0.20.

2.1. Model `df_pers = function(df_rate)` by using the linear regression.

What are the errors on: (i) the training part; (ii) the testing part?

2.2. Model `df_pers = function(df_rate)` by using the ridge regression with hyperparamter values alpha from [0.0, 1e-8, 1e-5, 0.1, 1, 10].

For every of the previous values for alpha, what are the errors on: (i) the training part; (ii) the testing part?

What is a best choice for alpha?

2.3. Model `df_pers = function(df_rate)` by using the lasso regression with hyperparamter values alpha from [1e-3, 1e-2, 1e-1, 1].

For every of the previous values for alpha, what are the errors on: (i) the training part; (ii) the testing part?

What is a best choice for alpha?

Note: Ignore any `convergence warning` in case you may obtain in the Lasso regression.