Information

Exercise 1: Design experimental workflows for a given dataset

For this assignment, you will explore data collected by the MovieLens project, namely the MovieLens 100K Dataset (ml-100k). Download the dataset at https://grouplens.org/datasets/movielens/ from section “older datasets” and familiarize yourself with the contents. (Hint: the files u.data, u.item, and u.user contain the particularly relevant data.)

There are four questions in this Exercise:

- Question 1 deals with data exploration,
- Questions 2 and 3 with hypotheses, and
- Question 4 with data acquisition.

Each question is worth 25pts, the total sum is 100.

Question 1

Not yet answered

Marked out of 25.00

Question 1 - Dataset Exploration (25pts)

Explore the data with a tool of your choice (Python, R, Java, MATLAB, Excel, etc. -- whichever you are most comfortable with).

Report your findings on the type of data you found and interesting insights on data properties (like statistical moments, distributions, correlations etc.). Include figures where appropriate and justify why you deem these figures relevant and insightful. (Note that there is not one true answer expected here -- this exercise is about exploration and identifying potentially interesting research questions. Inspiration and creativity are important in this process!)

Report on **potential ethical issues** you find with the data and how you would address these issues.

Write up your findings in a **one-page PDF document** (a second page is allowed, but may only contain figures and tables to support the reported findings) and **submit this PDF as answer**. No text content exceeding the first page will be taken into consideration for grading!

Maximum file size: 20MB, maximum number of files: 1

Files

You can drag and drop files here to add them.

Accepted file types

PDF document .pdf

Marked out of 25.00

Question 2 - Hypothesis 1 (25pts)

Formulate a hypothesis that can be answered with the data available in the ml-100k data set. Chose a **classification task**, i.e. select a nominal variable as dependent variable.

- Clearly indicate what the **independent and dependent variable** is
- Describe the **control conditions**
- Which **performance indicator** (or which type of performance criteria) you deem effective to compare the different conditions
- Given the scale type of the dependent variable, what might be a **suitable strategy to build a prediction model**, i.e., what type of method/algorithm is appropriate to model data of this type
- How you would prepare the data and **design the experimental setup** to simulate the situation of evaluating on unseen data?

Do not repeat the example from the lecture! (Hint: there are many ways to phrase other hypotheses, e.g., by choosing alternative target variables and/or subsets of attributes)

Hypothesis:

Independent Variable:

Dependent Variable:

Control:

Performance Indicator:

Prediction Model:

Experimental Setup:

Question **3**

Not yet answered

Marked out of 25.00

Question 3 - Hypothesis 2 (25pts)

Formulate a hypothesis that can be answered with the data available in the ml-100k data set. Chose a **regression task**, i.e. select a variable with at least ordinal scale type as dependent variable.

- Clearly indicate what the **independent and dependent variable** is
- Describe the **control conditions**
- Which **performance indicator** (or which type of performance criteria) you deem effective to compare the different conditions
- Given the scale type of the dependent variable, what might be a **suitable strategy to build a prediction model**, i.e., what type of method/algorithm is appropriate to model data of this type
- How you would prepare the data and **design the experimental setup** to simulate the situation of evaluating on unseen data?

Do not repeat the example from the lecture or a trivial variation of Hypothesis 1!

Hypothesis:

Independent Variable:

Dependent Variable:

Control:

Performance Indicator:

Prediction Model:

Experimental Setup:

Marked out of 25.00

Question 4 - Data Collection (25pts)

Formulate a hypothesis that can ***not*** be answered with the data available in the ml-100k data set as is. Describe a possible strategy to obtain the data needed and possible problems (including ethical issues)

Hypothesis:

Independent Variable:

Dependent Variable:

Data Missing and Strategy for Data Collection:

Possible Issues:

◄ Experimental Designs II

Jump to...