Understanding Data Assessment

Set: Week 9

Due: 3rd May 2023 at noon UK time

Contents

Introduction

This assessment has two parts.

Part A: There are 10 Questions that you need to answer. There is **NO page** limit for this part

Part B: You are instructed to create a **3-page report**

A few things to remember:

- You need to include a screenshot of **all** code in the appendix of your report. You will not be marked on the code, but it will be used to check for plagiarism. If you fail to do this, you will get a **FAIL!!!**
- There is a 10 mark/per extra page penalty for part B
- The report must be a **single PDF** submission. See "Sample_template"

Overall marks breakdown:

1) Question 1: 40 marks
2) Question 2: 40 marks
3) Presentation of the report: 20 marks

Part A: Recommender system

You've been hired by Lu's Communications which is a UK based online company that offers the various electronic products. Lu's Communication would like to recommend their customers other products when they've purchased a product. For example:

> "If we know that Product A is similar to Product B, then we can recommend product B to all customers who've purchased product A."

Your task is to find/investigate this similarity between products by answering the following questions. All relevant data can be found in the folder "Part A data".

You have been given the "train.txt" file which consists of the following information:

(User 18073 , Product 5351 , Rating 1.0)

From this we can conclude that User "18073" gave the product "5351" a rating of 1.0 out of 10

Question 1) Import the data "train.txt" and identify the total number Users and the total number of Products.

Question 2) Create a data frame (or matrix, or array) $\mathbf{Y}$ which consists of the ratings, such that y_{nd} represents the rating given by User n to Product d.

- What are the dimensions of $\mathbf{Y}$?
- Plot the matrix $\mathbf{Y}$

Question 3) Find the top 5 Products. Explain your method and results.

Question 4) Import the data "test.txt" and identify the total number Users and the total number of Products.

Question 5) Using "test.txt", create a data frame (or matrix, or array) $\mathbf{X}$ which consists of the ratings, such that x_{nd} represents the rating given by User n to Product d.

- What are the dimensions of $\mathbf{X}$?

Question 6) For each Product in "test.txt", find the most similar product in "train.txt". To compute the distance between Product n (from $\mathbf{X}$) and Product m (from $\mathbf{Y}$) use the following

$$\text{For } i = 1, \dots, N_{user}$$

$$\text{If } x_{in} \text{ and } y_{im} \text{ exist}$$

$$d_{nm} = d_{nm} + (x_{in} - y_{im})^2$$

Where N_{user} is the total number of users, and d_{nm} represents the distance between Product n (from $\mathbf{X}$) and Product m (from $\mathbf{Y}$).

Question 7) Using the method from Question 6) Find the top 5 similar product in "train.txt", for each Product in "test.txt".

Question 8) Explain the **main** limitation of the method in Question 6)

Question 9) Propose an alternative to the distance method described in Question 6) by completing:

For $i = 1, \dots, N_{max}$

If #Answer here

$$d_{nm} = d_{nm} + (x_{in} - y_{im})^2$$

If #Answer here

$$d_{nm} = d_{nm} + \#Answer\ here$$

Question 10) Using your solution to Question 10), find the top 5 similar products in "train.txt", for each Product in "test.txt".

Marking scheme

Question Number	Mark
1	5
2	5
3	5
4	3
5	2
6	10
7	3
8	2
9	4
10	1
Total	40

Part B: University Module Data

You've been tasked to analyse the marks of students for two modules over a five-year period. The data is stored in the following "Part B data". You are instructed to create a short 3-page report which gives an overview of the performance of students over the two modules and predicts the average mark of each module for 2020 (assuming COVID did not take place).

Marking scheme

Report structure and presentation (out of 10)

- [8-10] Clearly presented, all main sections included. Tables, diagrams appropriate and explained.
- [4-7]: Clearly presented, all main sections included. Some tables and diagrams may not be appropriate or incompletely explained.
- [1-4]: Parts of report difficult to read or poorly structured. Missing or inappropriate tables, diagrams.
- [0]: Report incomplete or missing.

Content of report (out of 30)

- [25-30]: Clear and consistent motivation for the approach taken, explanation of the approach, arguments and analysis.
- [20-25]: Mainly clear and mainly consistent motivation for the approach taken, explanation of the approach, arguments and analysis.
- [15-20]: Attempt to describe the motivation, explanation, arguments and analysis, but unclear or inconsistent in parts.
- [5-15]: Incomplete description of the motivation for the approach taken, explanation. No or limited arguments and analysis.
- [0-5]: Report incomplete or missing significant parts.

Penalise

- 10 Marks per extra Page – Note only use 3 pages

Submission

You must submit the report (PDF format) on Aston Blackboard.

Deadline is 3rd May 2023 at noon UK time.