UNIVERSITY OF CAPE TOWN
DEPARTMENT OF STATISTICAL SCIENCES
STA5073Z – Data Science for Industry 2024

Assignment 1
Recommender Systems

Due Date: Wednesday 18 September 2024 @ 12:00

Please read the following instructions carefully:

- Doing your project in R is **highly recommended**. Python submissions will be accepted but then you are essentially "on your own" (and contact me about submission requirements).
- Your analysis must be completed in and the report compiled in **Quarto**.
- You should submit a single zip file using your **STUDENT ID as the filename** e.g. ABCXYZ001.zip
- The zip file (when extracted) should contain a folder housing both the source file(s) and the derived output i.e. a `.qmd` file, one or more data files (as well as any other supporting files required to render your Quarto document), and the output in the form of a `.pdf` report.
- The report should contain a description of the problem (don't just copy & paste from this file!), the approach you took, and your results.
- Your code should **not** be displayed in the final typeset document (use `echo = FALSE`).
- Reference any material you use, whether it be a blog post, a figure, code, etc.
- You may not share any coding or write-up with any other student. Please sign the plagiarism declaration provided on Vula and append it to your report.

Notes and hints:

1. This project must be done in `R` without the use of special purpose recommender systems packages (e.g. `recommenderlab`).
2. The factorization of large matrices is computationally intensive, and some of the methods used during lectures for demonstration purposes may take a very long time to run and/or run into memory problems. You may use the `recosystem` package for this step (recommended), or carve your own path by writing a custom gradient descent function or leveraging one or more packages to assist with this.

3. Given the high degree of sparsity in the full dataset it is expected that you will reduce it down to something more manageable and suited for the purposes of this assignment. This is somewhat subjective so there aren't necessarily clear rights and wrongs here, but this should be done thoughtfully and you should at least outline your considerations and offer some justification for the steps you take. If after reducing the data to a smaller set you find you are still struggling to make progress, then try using fewer users. For a given number of items, reducing rows/users will probably increase prediction error (as you have less data) but does not fundamentally alter the task.

4. Anyone else should be able to run the code in your Qmd to completion. Use `set.seed()` to set a random seed so your final results do not change.

5. You are expected to present a coherent report, i.e. a continuous story that anyone should be able to at least follow, and that will not leave someone knowledgeable in the field with questions as to why/how you did something, or what the results mean. This does not require great volume though, still be selective in what you provide! You should be able to say what you need to in about 10 pages, although this is only a rough guideline, it could even be less.

Problem

The aim of this project is to build an ensemble recommender system that recommends to users books they might enjoy, based on their past book evaluations. The data you will be working with is a partially preprocessed version of the "Book-Crossing" dataset. The data can be downloaded here from Kaggle.

It was collected by Cai-Nicolas Ziegler in a 4-week crawl (August / September 2004) from the Book-Crossing community with kind permission from Ron Hornbaker, CTO of Humankind Systems. It contains 278,858 users (anonymized but with demographic information) providing 1,149,780 ratings (explicit / implicit) about 271,379 books.

The recommender system should be able to provide recommendations both for existing users and new users. For a new user, you can assume that they will provide (explicit) ratings for a small number of books – say, 5 or fewer – when joining the platform, thereby avoiding the cold-start problem in this exercise.

Some EDA is expected – you must always explore the structure and scope of your data. This will also help inform the splitting of the data into training and testing sets.

Specific Objectives

1. Build recommender systems that predict the rating a user will give to a book based on each of item-based CF, user-based CF, and matrix factorization.

2. (a) Assess the accuracy of the matrix factorization recommender system. You may either use a single train/test sample or k-fold cross-validation for this.

(b) Assess the accuracy of the matrix factorization recommender system with or without regularization.

(c) Create a final model that ensembles the predictions from the three approaches. Assess the accuracy of the ensemble predictions.

If time allows, you may also explore how sensitive the model is to sampling variation and how accuracy changes with the number of users (and/or items) the model is based on. You are not expected to create a perfect recommender system, just a functional and sensible one. If you need to compromise or simplify at any point then that is fine – just be sure to communicate this.

Bonus Objective

1. You are encouraged to publish your assignment on GitHub, and use GitHub's pages feature to publish either an html version of your report, or a Quarto website. This could be the same content as your report or an alternative version more suitable for the web.

2. If you do this, please include a ReadMe file in your Vula submission (the zipped project folder) which provides a link to the GitHub repo and webpage for your assignment.

3. Note that the GitHub component of the assignment is not strictly required and will not be graded in detail, but a limited number of bonus marks will be allocated to those who complete this so it should be a relatively easy way to boost your grade for this assignment.
