Math 122a: Final Project

1 Project overview

For the final project, you will take on the role of a data scientist doing market research. You will analyze the (synthetic) dataset using the tools you have learned in class. You ultimate goal is to build a model to predict how users will rate certain products. You will work in a group to develop a strategy, explore the dataset, develop your model, and write a report describing your approach and what you have learned.

The modeling task can be accomplished by combining techniques covered in class, including SVD/PCA, clustering, regression, and neural networks. (You might not need to use all of these.)

2 Forming groups

You may work in groups of up to 5 students. Please email me the names of the students in your group by the evening of November 8, or email me to ask for help finding a group (in which case your group assignment will be announced on November 10).

3 Data files

There are two data files:

1. `user_history.csv`

   This file lists the historical engagement of 4500 users with a collection of 100 different websites. The first column gives the numerical ID of each user. The remaining columns correspond to the websites, whose (meaningless) names are listed in the header. The decimal values in each field represent the historical engagement of a user with a website — perhaps measured in average minutes per day. (So, e.g., User #100950 spends an average of 2.192897 minutes per day on the "alpine kimono" website. We might imagine this data was collected by tracking users' web surfing habits over time.)

2. `user_ratings.csv`

This file gives some ratings by 3000 of the users on 75 different products. Just over 33,000 ratings are provided. The first column gives the numerical ID of each user, the second column gives the (meaningless) name of the product, and the last column gives the user's rating of the product, on an integer scale from 0 to 10. (So, e.g., User #100953 gives a rating of 10 to the "Magnet Marvin" product. We might imagine this data was collected from user surveys.)

The User IDs are shared between the two files, so user #100953 refers to the same person in both files. But there is no explicit connection between the websites tracked in the first file and the products rated in the second. (Nevertheless, we might hope that the web surfing behavior of each user might tell us something about their interests and preferences in general, which might help us predict the ratings they will give to the products.)

4 Deliverables

Your ultimate goal is to predict the ratings each of the 4500 users would give to each of the 75 products referenced in the `user_ratings.csv` file. You should output these ratings in a CSV file named `predictions.csv`, using the same format as `user_ratings.csv` (three columns, for user ID, product name, and rating). (So the file will have 4500×75 lines, each with three comma-separated fields, plus a header line at the top.) Note that you are asked to produce predicted ratings for all users, including the 1500 users who appear in the `user_history.csv` file but not in the `user_ratings.csv` file.

In addition to your predictions CSV file, you should submit three additional documents as a group:

1. A report describing in detail your approach to modeling the user ratings. Describe what you have discovered about the dataset, and describe the choices you made in producing your model, and why you made them. Include figures to illustrate your findings, as appropriate.

2. All Python code used to create your predicted ratings and perform the analysis described in your report. You can include several Python files (scripts or Jupyter notebooks), but make sure that each file executes as expected (in the case of a notebook, when each cell is executed in sequence).

3. A README file describing the structure of your code and how to use it to reproduce your predictions

Finally, **each individual** member of your group must separately submit a single page document describing their own contributions to the project, as well as the contributions of the other group members. Each group member should create this report on their own.

Due date: 11:59pm Wednesday, December 8.

Email all components to `maunu@brandeis.edu`.

5 Grading

Your submission will be graded according to the following criteria:

- Have you submitted all required materials in the required format?
- Is your approach clearly described?
- Is your modeling approach conceptually sound, and well motivated by your analysis of the data?
- Did you demonstrate an understanding of methods you applied?
- Does your code behave correctly, and is it well-written, readable, and appropriately commented?
- Are your predictions accurate?
- Did you, the individual student, contribute significantly to the success of the group?

6 Getting started

This is a big project—in order to succeed, you'll need to break it into smaller tasks. For example,

- Can you read the data files into a more helpful format? If you had the desired rating predictions, could you output them in the desired format?
- Does the data have any structure? Is it possible to simplify the dataset?
- Can you produce predicted ratings for all 75 products for just the 3000 users who have rated some of the products? How could you check your work?
- Can you find a relationship between the product ratings and the website history data?
- Finally, can you extrapolate product ratings even for users who haven't provided any ratings?

It's not possible to make perfect predictions, but see how far you can push your predictions.