1 Overview

This exercise session is in two parts.
The goal of the first part is to gain more familiarity with recommender systems and in particular with collaborative filtering.
In the second part, you will train a simple logistic regression model to predict the usage of a bike sharing system.
Most of the exercises will be done within a group. Have fun!

2 Rate Movies

The first task is to provide ratings for some movies. Please do so according to the following instructions.

1. Form groups of 3 or 4.
2. Download "collaborative.zip" from Toledo. In "u.item", you'll find a list of movies. The first column of the file is the ID of the movie; the second is the name.
3. Select 10 of these movies in such a way that everybody has seen all or most of them. Ideally, the movies span the whole range from favorite to despised.
4. Each member of the group rates exactly 6 of the movies on a scale from 1 to 5.

3 Predict Rating for Unrated Movie based on few data

Now, you will use the collaborative filtering algorithm presented in class to make a prediction for an unrated movie.

1. Use your ratings from Exercise 2.
2. Pick 1 movie that you have not rated and that at least one other member of your group has rated.
3. Use Equation 1 to make a prediction for the movie you selected.

Here is a description of the algorithm that should be used for this exercise.

- $\hat{R}_{u,i}$ is the prediction for user u for unrated item i. It is defined in Equation 1.
- I_u is the set of all items rated by user u.
- $w(u,v)$ is the Pearson correlation between user u and user v. Note that the sums run over the items, $i \in I_u \cap I_v$, that are rated by both user u AND user v. It is defined in Equation 2.
- $\alpha_{u,i}$ is a normalization constant. It is defined in Equation 3.
- $\bar{R}_u$ is the average rating for user u. The average is taken over *all* items, I_u, that user u has rated. It is defined in Equation 4.

Tip: it is important to understand what exactly each summation is taken over.

$$\hat{R}_{u,i} = \bar{R}_u + \frac{1}{\alpha} \sum_{v:i \in I_v} w(u,v)(R_{v,i} - \bar{R}_v) \tag{1}$$

$$w(u,v) = \frac{\sum_{j \in I_u \cap I_v} (R_{u,j} - \bar{R}_u)(R_{v,j} - \bar{R}_v)}{\sqrt{\sum_{j \in I_u \cap I_v} (R_{u,j} - \bar{R}_u)^2 \sum_{j \in I_u \cap I_v} (R_{v,j} - \bar{R}_v)^2}} \tag{2}$$

$$\alpha = \sum_{v:i \in I_v} |w(u,v)| \tag{3}$$

$$\bar{R}_u = \frac{1}{|I_u|} \sum_{j \in I_u} R_{u,j} \tag{4}$$

4 Predict Rating based on many data

On Toledo, in Course Documents/Session_2, you will find a file named `collaborative.zip`. Download and unpack this file. It includes an Ipython notebook `cf.ipynb` and a data file named `u.data` which contains $100,000$ movie ratings by different users.
You will see that there are four empty functions in the notebook. Fill in the functions based on the instructions in the notebook and formulas from Exercise 3. Once you are done, use your program to give recommendations based on a set of ratings produced by yoursels.

5 Discussion Questions

Please discuss the following questions with your group.

1. Were you surpised with the quality of the predictions in exercises 3 and 4? Explain.
2. Why do you think the predictions where good or bad? Explain.
3. If you were doing exploratory data analysis on a set of movie rankings, what statistics about the data set would you look at? Why would you pick each of these statistics?

6 Bike Rental Usage Prediction

In this part of the exercise session, we will work again with the dataset from Capital Bikeshare system from Washington D.C., USA.
Download the data (`train.csv` and `test.csv`) and the notebook `bikes.ipynb` from Toledo.
Let us call the usage of the system high if the total number of the bikes rented is higher than on average at that time.

1. Introduce a new binary variable, `'is_busy'`, which indicates if the usage of the system is high at the moment.

   Tip: Keep in mind that the overall usage of the system in 2012 is higher compared to 2011 since more bikes became available. You may want to account for that when computing the averages.

2. Train a simple logistic regression model to predict whether the usage of the system is high for the records in the test set. Evaluate your model using the real counts from `test_solution.csv`.