Computer Science 6100/4100: Assignment 3

Due: December 7, 2007

General notes: This assignment is intended to give you some room to explore topics that you think are interesting. I've suggested a couple of ideas which I think are interesting projects to work on, but if you feel like working on something else, I encourage you to do so. However, I ask for two things: (1) make sure that what you are doing makes sense by doing a literature survey and writing about the relevant related research (2) after doing a preliminary literature survey please come and tell me what you're planning to do and what the state of the existing literature is. You have quite a lot of freedom even within the two ideas I've given below – these are not intended to be complete specifications of the problems. However, it is absolutely critical to justify the decisions that you take in your final writeup. Working in teams of two on the project is permitted.
A note on the writeup: You must submit a maximum 4-page paper that strictly adheres to the official ACM style file that can be found here (use the "tighter alternate style"):
`http://www.acm.org/sigs/publications/proceedings-templates`
This means you have to make significant editorial decisions about what to include and what not to, but that's part of the challenge of academic writing, at least in Computer Science!

1 Option 1: Netflix

Read about the Netflix contest and prize at:
`http://www.netflixprize.com`

Compare at least two approaches to this problem in terms of RMSE on the "probe set." The approaches need not be complicated, but keep in mind that the matrix is very sparse. A first, simple approach, is to predict the following for every (User, Movie) pair: $\alpha\overline{M} + \beta\overline{U}$, where $\overline{M}$ is the average rating for the movie and $\overline{U}$ is the average rating the user has given to all movies she has rated. Learn α and β from the data in some manner. Another approach is to find the k nearest neighbors that also rated the same movie, and then predict the average rating these users gave that movie. You will have to define a meaningful distance metric that does not degenerate because of the sparsity of data, and figure out what to do when no one who is "close" to this user has rated this movie.

One of the challenges in this problem will be dealing with the fact that the dataset is enormous. You should test your algorithms with a representative subset to figure out what is going on. Either find one someone has made or come up with a good way to construct a representative subset.

Feel free to experiment with other approaches, or variants of the approaches described above.

2 Option 2: Dating With Heterogeneous Preferences

The Das and Kamenica paper discussed in class considers a model of matching with learned preferences when the true preferences are in fact homogenous. What would happen if the true preferences were actually heterogeneous? Well, many things, but pick something you're interested in and explore it further. One possibility is to think about the concept of social welfare, which is just the sum of all utilities that everyone receives in a matching. Asymptotically speaking, how "good" is the final matching achieved compared to the social-welfare-maximizing matching under different matching mechanisms (you can consider just a couple of mechanisms – Gale-Shapley and simultaneous choice)? Keep in mind that any perfect matching under homogenous preferences achieves the same social welfare.

You can think about this in the general context or in the context of "pairwise homogenous" preferences, where each member of a couple likes the other one equally.

3 Option 3: Your idea here

If you want to pursure your own idea come and meet with me about it soon (the Monday or Tuesday before Thanksgiving would be best!). One interesting possibility that could be a fair amount of work is on trying to come up with a good prior and a Gittins index strategy that works well on the webpage data from the Vermorel and Mohri paper.