CIS606, Spring 2014:

Project 1 – Collaborative Filtering of Educational Data

Deadline: "End of day", 20th of February, 2014

All projects will be ***individual efforts*** and the bulk of the implementation should be in Octave. The main objective of this project is to (i) Get you writing code! (ii) Provide hands-on experience with applying machine learning algorithms to a very current and popular research problem.

Methods and data

Collaborative filtering (CF) is a domain of machine learning/data mining where the data is generated by the combined actions of a very large number of independent agents acting on a common set of objects or items. One of the most common applications of CF is in the development of recommendation systems, such as those deployed by Amazon and IMDB. CF problems are characterized by large, sparse and incomplete matrices. The elements of these matrices typically correlate with the degree to which the different agents are "connected" to the set of objects being acted on. These can be binary (e.g. "likes" on social networking sites), numerical (e.g. star ratings on IMDB) or some other similar arrangement.

In this project, you will be applying machine learning techniques to predict the performance of students on a set of problem steps. The data to be used was provided as part of the KDD 2010 data mining challenge, and can be obtained from the following site (registration required):

http://pslcdatashop.web.cmu.edu/KDDCup/

The data consists of a (very large) set of entries detailing the performance of students on various problem "steps", and the challenge is to predict their performance on previously unencountered steps. While a CF based approach is suggested, it is possible to produce the predictions using other methods as well, such as by framing the challenge as a classification problem.

The following are some notes to get you started:

1. On the website, five files are available for download:
   - i. algebra_2005_2006.zip
   - ii. algebra_2006_2007.zip
   - iii. bridge_to_algebra_2006_2007.zip
   - iv. algebra_2008_2009.zip
   - v. bridge_to_algebra_2008_2009.zip
2. The first three are "development data sets" while the last two are "challenge data sets". As the name suggests, the former is for testing and model development

(they are also smaller in size and easier to handle) while the latter were the sets which were actually used to score the contestants.

3. Each zipfile contains at a minimum a training file (with training data, ahem) and a test file, which are standard tab-delimited files. Development sets also have "master" files which provide the performance data for all entries while challenge sets come with a "submission" file which is a template to be used for submitting results to the competition site.
4. The students' performance which you are to predict is given by the respective student's success in completing a step at the first try, and is denoted by the attribute "***Correct First Attempt***". In the competition, the scoring is based on RMSE on predicted values of this attribute, and this is the score which you should report in your experiments.
5. You are permitted to adjust the scope of your experiments based on available time and computational resources (so, for example, you might want to start with the smaller development sets before moving to the larger challenge sets if time permits). However, it is still possible to submit predictions over the challenge test sets to the competition website, and if possible, you should try to do this and report your score and position on the leader board.
6. As a starting point for your research, a reading has been provided together with this project description:

   "Collaborative Filtering Applied to Educational Data Mining", 2010, (A Toscher and M Jahrer).

Requirements

To evaluate your performance in this project, you are required to prepare a submission package with the following elements:

i) A report which presents the aims and objectives, a very brief review of any required background information, a detailed description of your algorithm, justifications for any design choices, data and experimental methods used, and finally a discussion of your observations and results.

   (Length ~4pages, formatted according to the standard IEEE format:

   http://www.ieee.org/conferences_events/conferences/publishing/templates.html

   (You can use either LaTeX or Microsoft Word, but the final submission has to be in PDF format)

ii) Please retain all source code, data and executables (if applicable) for your system for at least the duration of the course. We may or may not request this from you but you should be prepared to demonstrate your system in the event any questions or uncertainties regarding your method arise.

Submission must be via e-mail, with the words "CIS606, Spring 2014, Project 1 report" in the subject.