Assignment 3: Netflix

Barbara Engelhardt, Princeton University out 03/27/2018; due 04/17/2018

Background

As the ability to collect data on consumer preferences has grown exponentially, recommendation algorithms that customize advertisements and suggestions to users' interests have become powerful tools for increasing engagement and revenue. From shopping online to choosing a TV show to binge, the offerings that are presented to us are informed by our past choices and ratings. In this homework you will try to understand how patterns can be identified in movie viewing data to inform recommendation systems (and why you can't tear yourself away from Netflix even when you have a problem set due).

As part of the Netflix Prize Competition, an attempt to improve recommendations algorithms, Netflix released large amounts of user reviews for a subset of movies and TV shows. Netflix ratings are on a 1 to 5 star scale with increasing stars corresponding to greater enjoyment. From these ratings data, we can gain insight into both the movies/TV shows available on Netflix and Netflix viewers.

Project definition

This is an open-ended homework with the goal of finding interesting and meaningful patterns in the Netflix data. A key idea to use in your analysis is that there exists some sort of *latent structure* among the movies or users. Latent structure that shed slight on which movies or users are similar can be exploited to inform predictions to new users or for new movies.

The data consist of the training set from the Netflix prize data set. A link will be provided on piazza to download from dropbox. Ratings on the 1 to 5 scale are provided for 17,770 movies from 2,649,430 users. The data set is sparse, meaning that most entries are zeros, indicating an absence of a review (not that the review is a 0). Only about $\sim 1\%$ of entries are non-zero. The data have been processed into a movies $\times$ users matrix. A separate file contains the movie titles, reference ID, and release year. The raw data also contain the year that a user reviewed each movie. A test set for the Netflix prize data is available with the raw data (link posted in piazza) if you are interested in making and evaluating recommendations.

The data are available for download in the course Piazza website under Resources.

Deliverables

Your deliverables for this project include:

- A five page (not including citations) summary of the project work, which should contain (as described in the Example project write up on Piazza):
  - A title, authors' names, and abstract for the project;
  - an introduction to the problem being addressed;
  - a brief description of the data;
  - a description of the methods developed and used, and how they were fitted using training data;
  - one page describing in detail a method that you applied to the data set;
  - a presentation of the results of the methods applied to the test data;
  - a discussion of the results, including the intuition you have gained on the behavior of the models;
  - a short summary and conclusion, including extensions that you believe would be particularly valuable based on the results;
  - a *complete* bibliography to support any related work or methods that are relevant to your project.

  A .zip file of your Python code that you developed for the project, with a README about how you ran the code.

Please put your PDF write up of the project, the Python code for the project, and a readme about how to run the code into
`https://dropbox.cs.princeton.edu/COS424_S2018/Assignment3`
by midnight on the assignment due date, with the file names
`<author1PUID>_<author2PUID>_hw3.pdf` (for write up)
`<author1PUID>_<author2PUID>_hw3.zip` (for code and readme).
Please only submit one PDF per pair of authors.

We strongly recommend *writing as you go*, which means starting to write the project report as you are downloading and analyzing the data. That said, you should avoid speculative writing, and only write results once you have them.

A simple analysis

We performed a rudimentary analysis to give a baseline of what we expect. We performed a singular value decomposition (SVD) on a subsample of this matrix with dimension 5 (selected by looking at a drop off in the eigenvalues). Note that, due to the size of the data, we had to use a truncated SVD routine designed for sparse data—-this is available in scikit-learn. Most modern computers can handle the full data set (e.g. 2017 Macbook can perform TruncatedSVD on the full data set). Older computers likely need to subsample the data. You can consider using Princeton computing resources and clusters to perform more computationally expensive analysis. We then examined the movies with largest absolute weights for each component to look for structure in the latent space.

Suggestions and Advice

This is an intentionally open-ended assignment. Some food for thought:

- *Sparsity*: The data are very big and also very sparse. The size means that you will have to be careful with how you handle the data. If you must work with the data in matrix form, sparse matrix representations will be the most efficient. Python (`scipy.sparse`) has support for sparse matrices.
- *Modifying the data type*: The data is provided with ratings, but one possible way to handle the data would be to simply turn each rating into binary values, indicating missing or not missing (i.e., intrinsic ratings). You should be careful about the possible loss of information in this case.
- *Sampling from the full matrix*: If you decide to use a stochastic optimization routine for learning, make sure that the underlying algorithm samples from the full matrix and not just from the non-zero entries. Otherwise, you will be making the missing-at-random assumption, which will result in a poor generalization performance. Carefully read the implementation details of any Python packages you use and make sure that the algorithm handles the sparse matrix data structure correctly.
- *Patterns in Feature Space*: These data reveal insights on both movies and viewers. Make sure you understand the assumptions for any model you choose and which category you are using as the features/samples. Careful feature selection can be a useful preprocessing step to find more meaningful signals. Adding additional features in the movie meta-data often helps. The latent variables in the models you fit to the data can also tell you what movies are most similar and what users are most similar. We also intuitively know that a user that hasn't rated a movie might not like that movie (so they haven't watched it) – can this idea be used in the model?
- *Latent variable models*: We have learned a lot of latent variable models in this course, and more are available elsewhere. These include PCA, latent Dirichlet allocation, clustering, non-negative matrix factorization, and Gaussian graphical models. You should think about which of these approaches is most applicable to these data and try them out.
- *Visualization*: High dimensional data is hard to grasp intuitively. Effective visualization of data in 2 or 3 dimensions can help inform further analysis. But projections can also hide potentially important information. Graphs of data exploration (histograms, violin plots, box-and-whiskers, etc.) are also helpful for selecting appropriate models and justifying assumptions.
- *Similarity Measures*: The range of possible outcomes, 0 to 5, is limited. Euclidean distances may not be the best measure of similarity as even "distant" reviews are not very far in Euclidean space. On the other hand, high dimensional data leads to increased distances between similar points. Think about what kinds of metrics make sense for comparing samples.

- *Missing Data*: A zero can represent either a user not having watched the video or a user declining to rate a movie. Some models are better suited for "not missing at random" (NMAR) data than others.
- *Prediction/Recommendation*: How can the features and patterns you find in the data be used for prediction in the test data? How can those predictions be translated into recommendations? What are the best metrics for recommendation system accuracy on the test set? How can you use your fitted models to predict whether a user rated a movie or not?

More information on the Netflix prize can be found on the website (https://www.netflixprize.com/).