---

# Visualization of latent factors from movies [Masoud Farivar]

## Overview

- The main goal of this mini-project is for you to create two-dimensional illustrations of the latent factor approach, similar to the one in Figure 2 of the reference [1], using the movielens dataset.
- This mini project is due Friday March 13th, 2015 at 2pm via Moodle. You can work in groups of up to three.
- You may implement this project in any programming language you choose, but you should provide a complete report with readable and commented codes.

## Data Format

The movielens dataset [2] consists of 100,000 ratings from 943 users on 1682 movies, where each user has rated at least 20 movies. Download and use the following files from the course website:

- **movies.txt** Each of the 1682 lines in this file contains comma separated list of the following fields for a movie:

  *movie id, movie title, unknown, Action, Adventure, Animation, Children's, Comedy, Crime, Documentary, Drama, Fantasy, Film-Noir, Horror, Musical, Mystery, Romance, Sci-Fi, Thriller, War, Western.*

  where the last 19 fields are the genres, a 1 indicates the movie is of that genre, a 0 indicates it is not; movies can be in several genres at once. The movie ids are the ones used in the data set.

- **data.txt** Each of the 100,000 lines in this file consists of

  *a user id, a movie id, a rating*

## Preliminaries

Let $m, n$ be the number of users and movies, respectively, and $Y$ be the $m \times n$ matrix of the movie ratings, where $y_{ij}$ corresponds to user $i$'s rating for the movie $j$. Note that most of the elements of the matrix are unknown and the goal of a recommender system is to predict these missing values.

# 1 Matrix Factorization (40 points)

In this part, the goal is to find matrices $U$ and $V$, such that $Y \simeq UV^T$ . The dimensions of $U$ are $m \times k$, are the dimensions of $V$ are $n \times k$, where $k$ is a parameter of choice. We will do this by solving the following minimization problem:

$$\underset{U,V}{\text{minimize}} \sum_{(i,j)\in ratings} \left(y_{ij} - u_i v_j^T\right)^2 + \frac{\lambda}{2}\left(\|U\|_{Fro}^2 + \|V\|_{Fro}^2\right) \quad (1)$$

Stochastic Gradient. Descent (SGD) and Alternating Least Squares (ALS) are two popular approaches to solve this problem. You may implement either of them for a full credit. Or alternatively, you can choose to get a partial credit (20 points) by just using any existing implementation, including the one from graphlab [3], to perform this factorization. In any case, you should document your approach. Choose $k = 20$, and justify your choices for other parameters and the stopping criterion in your solution method.

# 2 Visualization & Interpretation (60 points)

(a) In order to visualize the resulting latent vectors, apply SVD and use the first two components to project $U, V$ into a two-dimensional space.

(b) Now construct creative 2D-visualizations of the resulting latent vectors, similar to the one in Figure 2 of the reference [1] , and try to explain your plots and make interesting observations. Note that you do not have to place all movies and users in one plot.

# References

[1] Koren, Y., Bell, R., & Volinsky, C. (2009). Matrix factorization techniques for recommender systems. Computer, (8), 30-37.

[2] Herlocker, J. L., Konstan, J. A., Borchers, A., & Riedl, J. (1999, August). An algorithmic framework for performing collaborative filtering. In Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval (pp. 230-237). ACM.

[3] GraphLab's collaborative filtering library - available from
http://select.cs.cmu.edu/code/graphlab/pmf.html