

Homework 2

Lecturer: Inderjit Dhillon

Date Due: Feb 27, 2008

Keywords: *Netflix, Regression, Least Squares*

- **Million Dollar Question**

For this problem, you will work on the Netflix dataset, where one needs to predict missing (movie, user) ratings. The problem of predicting missing values for a recommender system is formally known as Collaborative Filtering. Read more about the Netflix challenge and the dataset at: <http://www.netflixprize.com>.

The complete Netflix dataset has 480,189 users, 17,770 movies, and 100,480,507 ratings. So, the scale of the problem is huge. For this assignment we will work on a much smaller subset of 2000 users, 5000 movies and 166,749 ratings. You can download the dataset and relevant files from:

<http://www.cs.utexas.edu/users/pjain/cs395T/hw2.tar.gz>.

We will take an approach based on Matrix Factorization which is also being used by the current leading team for Netflix challenge. Let us first introduce some notation. Let there be u users and m movies and let $R \in \mathbb{R}^{u \times m}$ be the ratings matrix where element R_{ij} is the rating given by user i for movie j . Note that the matrix R has many missing values - in the entire Netflix dataset, 98.82% of the entries are missing. Think of factoring the matrix $R \approx U^T M$, where $U \in \mathbb{R}^{k \times u}$ and $M \in \mathbb{R}^{k \times m}$ and $k \ll \min(u, m)$. Intuitively, you can think of column \mathbf{u}_i of U as a low-dimensional feature vector for the i -th user, and the column \mathbf{m}_j of M as a low-dimensional feature vector for the j -th movie.

Using the above notation, the rating R_{ij} for a movie j by user i is approximated as $R_{ij} \approx \mathbf{u}_i^T \mathbf{m}_j$. Thus, the Matrix Factorization approach involves finding U and M such that $U^T M$ is as close to R as possible. Formally the problem can be stated as:

$$\min_{U, M} \|R - U^T M\|_F^2 + \lambda(\|U\|_F^2 + \|M\|_F^2), \quad (1)$$

where $\|\cdot\|_F$ denotes the Frobenius norm, and where we have added regularization terms (as in ridge regression) and the regularization parameter $\lambda \in [0, 1]$

This problem is in general hard to solve as both U and M need to be inferred and no known method guarantees the optimal solution. Instead, we use a heuristic which works well in practice. We use the method of Alternating Minimization, where in each iteration we fix U to compute M , and subsequently fix M to compute U , and so on. Now we describe the method to compute M assuming U to be fixed.

Let U be fixed, then (1) reduces to:

$$\min_M \|R - U^T M\|_F^2 + \lambda \|M\|_F^2,$$

This problem can be decoupled in terms of columns of M , \mathbf{m}_j , i.e.,

$$\min_{\{\mathbf{m}_1, \mathbf{m}_2, \dots, \mathbf{m}_m\}} \sum_j \|\mathbf{r}_j - U^T \mathbf{m}_j\|_F^2 + \lambda \sum_j \|\mathbf{m}_j\|_2^2,$$

where \mathbf{r}_j denote the j -th column of R . Thus, we can solve a separate ridge regression problem for each j , i.e.,

$$\min_{\mathbf{m}_j} \|\mathbf{r}_j - U^T \mathbf{m}_j\|_F^2 + \lambda \|\mathbf{m}_j\|_2^2,$$

However, the vector \mathbf{r}_j , which represents ratings for movie j , may have many missing values. Let there be n_j known ratings for the movie j . Then, the ridge regression should be solved using only n_j ratings, i.e. let $\mathbf{r}_j^{(\mathcal{K}_j)} \in \mathbb{R}^{n_j}$ denote the known ratings for movie j , and $U^{(\mathcal{K}_j)} \in \mathbb{R}^{k \times n_j}$ denote the submatrix of U for the corresponding users who rated movie j . Thus the problem can be written as:

$$\min_{\mathbf{m}_j} \|\mathbf{r}_j^{(\mathcal{K}_j)} - (U^{(\mathcal{K}_j)})^T \mathbf{m}_j\|_F^2 + \lambda \|\mathbf{m}_j\|_2^2,$$

An analogous procedure can be used for computing U .

So, for each iteration compute U and M using the procedure given above. Repeat this procedure for a fixed number of iterations. Initialize U and M randomly, fix the number of iterations τ as 30 and fix $k = 10$.

1. Select $\lambda = 0$ and compute U and M . What are the problems you face?
2. Using ten-fold cross validation (validation sets are provided in the hw2.tar.gz file), arrive at the value of the regularization parameter λ . Report the optimal λ and also a plot showing Root Mean Squared Error (RMSE) averaged over all the validation sets vs λ .
3. For the optimal λ , find out the RMSE obtained for the Test Data.
4. What other initialization can be used? What is the problem with using the zero matrix as the initialization for U and M .
5. Now let us consider preprocessing the data in order to remove some inherent bias in the data. First center the complete data using global mean. Now, center each row of the ratings matrix A to mean 0, i.e. remove user mean for each row. Similarly center each column of the resulting matrix. Perform the above operations on known ratings only. After this preprocessing, repeat part (2) and (3) given above. Note that after factorization, add the global mean and the user/movie bias for prediction. Report RMSE obtained. How does RMSE change? Explain.