

CS532 Lab

Recommender Systems

Lokesh Jindal

Mehreen Ali

Urmish Thakker

Fall 2015

Contents

| | | |
|----------|---|-----------|
| 1 | Motivation | 2 |
| 2 | Approaches | 2 |
| 2.1 | Content-based | 2 |
| 2.2 | Collaborative Filtering | 2 |
| 3 | User Based Recommendation | 3 |
| 3.1 | Idea of Similarity | 3 |
| 3.1.1 | Similarity Weight Computation | 3 |
| 3.2 | Calculating Predicted Ratings | 4 |
| 3.3 | Rating Normalization | 5 |
| 3.3.1 | Mean-Centering | 5 |
| 3.3.2 | Z-score Normalization | 6 |
| 3.4 | How to Evaluate a Recommender System? | 6 |
| 3.4.1 | Root Mean Square Error (RMSE) | 6 |
| 3.4.2 | Top Five Rated Movies (TFRM) | 7 |
| 3.4.3 | Number of Good Movies Not Predicted (NMNP) | 7 |
| 3.4.4 | Number of Flipped Movies (NFP) | 7 |
| 4 | Model Based Systems | 9 |
| 4.1 | Singular Vector Decomposition | 10 |
| 4.2 | Generating Predictions | 10 |
| 4.2.1 | Latent Matrix Factorization | 10 |
| 4.2.2 | Average Value Based SVD Completion | 11 |
| 4.2.3 | Iterative SVD Completion | 11 |
| 4.3 | Nearest Neighbor | 12 |
| 4.3.1 | Using Predicted Matrix | 12 |
| 4.3.2 | Original Matrix in Lower Dimension | 12 |
| 4.3.3 | Which Nearest Neighbor approximation to Use? | 12 |
| 4.4 | Bringing it All Together | 12 |
| 5 | Miscellaneous Topics | 13 |
| 5.1 | Content Based Recommender Systems | 13 |
| 5.1.1 | Advantages and Drawbacks of Content-based Filtering | 13 |
| 5.2 | Serendipity | 14 |
| A | Latent Matrix Factorization | 15 |

1 Motivation

User modeling, adaptation and personalization methods have reached the mainstream. The tremendous growth of social networking websites and the computational power of mobile devices are generating extremely huge amounts of user data, and an increasing need of users to "personalize" their e-mail, phone etc.

The potential value of personalization is now clear both as a commodity for the benefit of end-users and as an enabler of better (or new) services a tactical opportunity to expand and improve businesses.

The main characteristic of recommender systems is that they attract the interest of industry and businesses while posing extremely interesting challenges.

In spite of noteworthy progress in the research community and the efforts of the industry to provide the end users the the benefits of new techniques, there are still many important gaps that make personalization and adaptation challenging for users. Research activities still focus on narrow problems, such as incremental accuracy improvements of current techniques or on a few applicative problems. This confines the range of other applications where personalization technologies might be useful as well.

Thus, we have come to a good point where we can take a step back to obtain a perspective in the research done in recommender systems.

2 Approaches

2.1 Content-based

The content-based systems are used to recommend items similar to those a user has liked in the past, i.e. recommendations are based on the content of items rather on other user's opinion. The content-based approach to recommendation has its roots in the information retrieval (IR) community, and employs many of the similar techniques. The system learns to recommend items that are similar to the ones that the user liked in the past. The similarity of items is calculated based on the features associated with the compared items. For example, if a user has positively rated a movie that belongs to the comedy genre, then the system can learn to recommend other movies from the same genre.

2.2 Collaborative Filtering

Collaborative filtering method finds a subset of users who have similar tastes and preferences to the target user and use this subset for offering recommendations. Rather than computing the similarity of the items, we compute the similarity of the users. Typically, for each user a set of "nearest neighbor" users is found with whose past ratings there is the strongest correlation. Scores for unseen items are predicted based on a combination of the scores known from the nearest neighbors.

Main approaches for collaborative filtering:

- User based
- Item based

| | Matrix | Titanic | DieHard | ForrestGump | Wall-E |
|-------|--------|---------|---------|-------------|--------|
| John | 5 | 1 | | 2 | 2 |
| Lucy | 1 | 5 | 2 | 5 | 5 |
| Eric | 2 | | 3 | 5 | 4 |
| Diane | 4 | 3 | 5 | 3 | |

Table 1: Example ratings

3 User Based Recommendation

3.1 Idea of Similarity

Let's say the users of a website like Imdb [TODO FIXME] rate the movies on a scale of 1 to 5, where a rating of 5 implies the user really liked the movie and a rating of 1 implies the opposite. This data can be stored in the form of a *utility matrix* where each row of the matrix corresponds to a user and each column corresponds to a movie. Consider an example rating set shown in Table 1 that can be represented in the form of *utility matrix* U as shown in equation 1. This 4x5 matrix stores the ratings given by 4 users to a set of 5 movies. As expected, not every user would have rated each of the 5 movies. The user-movie combination for which a rating does not exist is indicated by a 0.

$$U = \begin{bmatrix} 5 & 1 & 0 & 2 & 0 \\ 1 & 5 & 2 & 5 & 5 \\ 2 & 0 & 3 & 5 & 4 \\ 4 & 3 & 5 & 3 & 0 \end{bmatrix} \quad (1)$$



Exercise 1

Look at the ratings of user Eric in Table 1. Also look at the ratings given by other users in the table. Can you guess which user(s) is(are) similar to user Eric?

3.1.1 Similarity Weight Computation

The example data set in Table 1 is very small in size for which finding users that are similar to the user of interest might be trivial. For real data sets, we need to define a measure of similarity that can be used to find the items/users that are similar to each other. The computation of the similarity weights is one of the most critical aspects of building a *neighborhood-based* recommendation systems, since it can have a significant impact on the performance and accuracy of the system.

Cosine Similarity: We apply the traditional notion of Cosine Vector (CV) similarity to find the similarity between two users u and v . If \mathbf{x}_u is vector of ratings r_{ui} of a user u , then the cosine similarity of users u and v can be calculated using the equation 2.

$$CV(u, v) = \cos(\mathbf{x}_u, \mathbf{x}_v) = \frac{\sum_{i \in I_{uv}} r_{ui} r_{vi}}{\sqrt{\sum_{i \in I_u} r_{ui}^2 \sum_{j \in I_v} r_{vj}^2}} \quad (2)$$

where I_u and I_v represent the set of ratings of users u and v respectively, and I_{uv} is the set of ratings for the movies that have been rated by both users.

| | Matrix | Titanic | DieHard | ForrestGump | Wall-E |
|-------|--------|---------|---------|-------------|--------|
| Eric | -1.5 | 0 | -0.5 | 1.5 | 0.5 |
| Diane | 0.25 | -0.75 | 1.25 | -0.75 | 0 |

Table 2: Mean Centered Ratings



Exercise 2

In the example data set of Table 1, find the users that are most similar to user Eric. You can use the function *CosSimVecMatrix* (provided with the lab) to generate the similarity of user Eric with other users.

Hopefully that was easy and you have the answer - users that seem most similar to Eric are Lucy and Diane. That was easy. Let's look at the ratings of Diane who seems to be similar to Eric. If we calculate the means of ratings of Eric and Diane, 3.5 and 3.75, and subtract the means from their respective ratings, the new ratings look like as shown in Table 2.

Positive values in Table 2 represent a preference for the movie by the user since it's rating is more than the average, whereas a negative value implies a disliking for the movie. On analyzing these *mean-centered ratings*, Eric and Diane don't seem to be so similar anymore. In fact, they seem to have quite the opposite taste. What's going on?

Pearson Correlation Similarity: A major flaw in using *Cosine Similarity* to find the similarity between users is the fact that Cosine Similarity does not take into account the differences in the mean and variance of the ratings made by the users. Pearson Correlation (PC) similarity is a popular measure that can be used to compare the ratings of users since it removes the effects of mean and variance in ratings while calculating the similarity between users. The PC similarity between users u and v can be calculated using the equation 3.

$$PC(u, v) = PC(\mathbf{x}_u, \mathbf{x}_v) = \frac{\sum_{i \in I_{uv}} (r_{ui} - \bar{r}_u)(r_{vi} - \bar{r}_v)}{\sqrt{\sum_{i \in I_{uv}} (r_{ui} - \bar{r}_u)^2 \sum_{i \in I_{uv}} (r_{vi} - \bar{r}_v)^2}} \quad (3)$$

where \bar{r}_u represents the mean of the ratings given by user u . Note that PC accounts for variance in ratings only for the ratings of the movies I_{uv} that have been rated by both users u and v . The sign of similarity weight calculated using PC indicates whether the correlation between the two users is direct or inverse, and the magnitude of similarity weight (ranging from 0 to 1) represents the strength of correlation. These will be used to predict unknown ratings for a user using *neighborhood-based* methods later.



Exercise 3

Calculate the Pearson Correlation (PC) similarity of Eric with other users in Table 1. You can use the function *PCSimVecMatrix* provided with the lab. Which user(s) would you say is similar to Eric. You can do a similar analysis for comparing the Cosine similarities and Pearson Correlation similarities of other users.

3.2 Calculating Predicted Ratings

The Recommender Systems exist so that they can make recommendations to the user. As highlighted in section [TODO FIXME], collaborative filtering relies on finding users similar to a user of interest based

on the ratings given by them. We now have two important measures of finding similarity between users, namely Cosine similarity and Pearson Correlation similarity. It follows that these similarity measures can now be used to find the neighbors nearest to a user, and subsequently their ratings can be *combined* to give us predicted ratings for a number of movies that have actually not been rated by the user of interest. It is natural that the movies whose predicted ratings are higher are the potential candidates that the recommender system can recommend to the user.

Thus, the rating r_{ui} of a user u for a new item i , can be predicted using the ratings given to i by users most similar to u . Let w_{uv} denote the similarity weight between users u and v , and $N_i(u)$ be the set of k neighbors of u , then the predicted rating r_{ui} can be calculated as:

$$r_{ui} = \frac{\sum_{v \in N_i(u)} r_{vi}}{|N_i(u)|} \quad (4)$$

Note that the formula considers only the neighbors v of user u who have rated the movie i , $N_i(u)$ representing the set of such neighbors.



Exercise 4

Carefully look at the formula in the equation 4. Can you find a flaw in the formula? What does it use the similarity weights of user u with different users v for? Explain qualitatively, how the formula is not making full use of the similarity weights?

If you have an intuitive answer to Exercise 4, equation 5, which is more sophisticated in its calculation of the predicted rating, should make sense.

$$r_{ui} = \frac{\sum_{v \in N_i(u)} w_{uv} r_{vi}}{\sum_{v \in N_i(u)} |w_{uv}|} \quad (5)$$

We now take a small detour to look at the concept of *Rating Normalization* and will then come back to learn how to really generate a recommendation for a user.

3.3 Rating Normalization

Various users rate the movies differently, each having his/her own personal scale even when they have to give a rating on the same scale of 1 to 5. For example, while a rating of 4 by one user might imply a strong preference for a movie, the same rating by another user whose average rating across movies is 4, might not give us any interesting information to use. To tackle this, there are two popular normalization techniques: *Mean-Centering* and *Z-score Normalization*.

3.3.1 Mean-Centering

The idea of mean-centering is quite simple. It determines whether a rating is positive or negative by comparing it to the mean of the ratings for that user. Once we have the ratings for nearest neighbors, they can be *normalized* by subtracting their corresponding means and then used to generate the predicted rating for a user using the following relation.

$$r_{ui} = \bar{r}_u + \frac{\sum_{v \in N_i(u)} w_{uv} (r_{vi} - \bar{r}_v)}{\sum_{v \in N_i(u)} |w_{uv}|} \quad (6)$$

where \bar{r}_u is the mean of ratings of user u .

| | Matrix | Titanic | DieHard | ForrestGump | Wall-E |
|-------|--------|---------|---------|-------------|--------|
| John | 5 | 1 | | 2 | 2 |
| Lucy | 1 | 5 | 2 | 5 | 5 |
| Eric | 2 | | 3 | 5 | 4 |
| Diane | 4 | 3 | 5 | 3 | |
| Kim | 1 | 4 | | 5 | |

Table 3: Example ratings

3.3.2 Z-score Normalization

While mean-centering removes the user bias by removing their respective means from their ratings, it still does not consider the spread of ratings given by a user. For example, consider two users both of whom have an average rating of 2.8. Further, user X shows a lot of variation in his ratings from 1 to 5, whereas user Y has most ratings very close to 2.8. In this case, a rating of 4.5 by Y would imply an exceptional liking for the movie for user Y. The same cannot be said for user X. Z-score normalization takes care of this by normalizing the ratings so the ratings of a user have a mean of zero and a variance of 1. You can explore the MATLAB function *zscore*.



Exercise 5

Consider the Table 3 which has a new user Kim added to our earlier toy example. We are interested in generating a recommendation for Kim. Can you guess which movie we should recommend just by looking at the table? Now write a small code in MATLAB to generate those predictions following the steps mentioned below:

- Calculate the PC similarity of Kim with other users using the provided MATLAB function *PCSimVecMatrix*
- Find the nearest neighbors of Kim using a threshold for the similarity
- Use mean-centering to normalize the ratings of the neighbors
- Generate a recommendation for Kim using the normalized ratings of neighbors and similarity weights

Carefully read the specification of the provided MATLAB function *PCSimVecMatrix*. Can you use it to skip one of the steps mentioned in Exercise 5.

3.4 How to Evaluate a Recommender System?

We need some methodology to evaluate our recommendations. In this section we will discuss multiple ways to do so. A recommender system generally looks at a mix of these values to evaluate its results. We first separate our database into a test set and a train set. We compute our results based on the train set and make our predictions. We extract the predictions from our test set and look at the corresponding predictions we generated for them. In order to quantify the difference, we use the following techniques.

3.4.1 Root Mean Square Error (RMSE)

$$RMSE = \sqrt{\frac{\sum_{(x,i) \in T} (r_{xi} - r_{xi}^*)^2}{N}}$$

where T : The set of test data

$N = |T|$: The number of elements in the test data

r_{xi} : The actual rating of item 'i' by user 'x'

r_{xi}^* : The rating that our system predicts user 'x' will give item 'i'

We can try to design our algorithm to reduce this RMSE value, by minimizing the sum of squared errors (SSE) thereby giving us more accurate predictions of user ratings and hence improving our recommendations.

3.4.2 Top Five Rated Movies (TFRM)

Sort the movies in the test dataset in a decreasing order and look at the top five rated movies. For the movies in the test dataset, look at their corresponding predictions and sort them in a decreasing fashion. We find the intersection of the top 5 movies for both these sets to see how close we came to understanding user preference. let the top five movies in the test dataset for a user be represented by :

$$S_{test}$$

And the top five movies using the predictions for the movies rated in test dataset for the user be represented by:

$$S_{pred}$$

Thus,

$$Error = \frac{|S_{test} \cap S_{pred}|}{|S_{test}|}$$

3.4.3 Number of Good Movies Not Predicted (NMNP)

Another measure of error could be the number of movies that were rated by the user in the test dataset as good which the recommender system predicted as bad. For the notion of good, we use a threshold value. Any movie with a rating more than the threshold value above average is considered good. Thus, if S_{gtb} represent the number of good movies that the recommender system predicted as bad and S_g are the set of all good movies in the test data set, then the error is simply,

$$Error = 1 - \frac{|S_{gtb}|}{|S_g|}$$

3.4.4 Number of Flipped Movies (NFP)

This is an extension of the previous error metric. Here we look at all the movies that user rated good an bad. The definition of good remains the same as before. We define bad movies as movies which have ratings lower than $AverageValue - Threshold$. S_{flip} represent the sum of number of good movies that the recommender system predicted as bad and bad movies that the recommender system predicted as good, S_{gb} are the set of all good and bad movies in the test data set, then the error is simply,

$$Error = 1 - \frac{|S_{flip}|}{|S_{gb}|}$$



Exercise 6

In this exercise, you will extend the MATLAB code from the previous exercise [TODO FIXME] to generate recommendations using user-based method on a large data set. The training data is available in file `u1.base` and test data in `u1.test`. You can read these files into *utility matrix* form using the function *ConvertUDataToMatrix* that takes filename as input. You should pick a user from the test data, generate recommendations for him and calculate the *error* in the recommendation. You can repeat this for all the users in the test set and calculate the mean *error* across the users. This will give you a measure of how well the recommendation system performs for one *configuration*. A *configuration* is defined by the chosen values for a number of parameters:

- Choice of similarity criteria - Cosine similarity, PC similarity
- Choice of threshold value for selecting nearest neighbors based on their similarity weights
- Choice of threshold to limit the maximum number of neighbors chosen in the previous step
- Choice of normalization used - Mean Centering, Z-score Normalization, no normalization
- Choice of metric to calculate the error - RMSE, TFRM, NMNP, NFP

A set of values selected for these parameters will give you one configuration. Vary one or more of these parameters to run for different configurations and calculate the mean error for each. You DO NOT need to cover the entire space of configurations. Refer to Table [TODO FIXME] to decide on 4-5 configurations. Do some choices perform better than others? Try explaining any interesting observations.

| | Die Hard 1 | Die Hard 2 | Die Hard 3 | Die Hard 4 |
|-------|------------|------------|------------|------------|
| John | 4 | 4 | | |
| Susan | 4 | 2 | 3 | 3 |
| Jack | | | 4 | 4 |

Table 4: Truncate to Die Hard

4 Model Based Systems

The systems based on methods discussed earlier in the lab have been widely used and produce good results. But such algorithms have been shown to have several limitations.

- Sparsity - Nearest Neighbor (NN) Algorithms need exact matches. As a result algorithms sacrifice recommender system coverage and accuracy [TODO FIXME ref1]. To be more precise, correlation coefficient is only defined between customers who have products (movies in our database) in common. In an ecommerce environment where there are large number of items, one may find many customers who do not have any correlation with other customers [TODO FIXME ref2]. As a result Nearest Neighbor based algorithms are not able to recommend anything to these customers. This problem is known as the reduced coverage problem. The sparsity could also lead to a recommender system from missing certain obvious neighbors.

For example - John and Susan are highly correlated, while Susan and Jack are highly correlated. Conventional wisdom might suggest John and Jack should also have similar choices, however if John and Jack have very few ratings in common, such patterns could be easily missed

- Synonym Problem - In real life scenarios, different product names can refer to similar items. Correlation based recommender systems cannot find such latent association and thus end up treating these objects as two separate entities.

For example, let us consider two customers one of whom rates 10 different writing pad products as "high" and another customer rates 10 different notepads as "high". NN based recommender system will not capture their association.

One of the methods used to handle both the problems is the low rank approximation method. Let us do an exercise to understand the issues discussed above and see how low rank approximation can help us.



Exercise 7

Take a look at Table 4. The table illustrates the example we described while discussing *Sparsity*. So how do you ensure that John and Eric are considered similar to each other? Look at the code written in *Examples/ToyExampleModelSVDMotivation.m* and run it to see how low rank approximation helps us magically establish this relationship between John and Eric!! We will discuss what low rank approximation in the next section. For now, do you agree that this method is helping us with the sparsity problem?



Exercise 8

Look at *Examples/ToyExampleModelSVDMotivation2.m*. It explains a scenario similar to the talked about in Synonym Problem and looks at a potential solution. Run the code and observe the output? Do you think Low Rank approximation is finding some sort of latent association? What are the predictions for values not rated by the user?

4.1 Singular Vector Decomposition

The Singular Value Decomposition (SVD) is a well known matrix factorization method. Formally, the SVD of a $m \times n$ matrix A is a factorization of the form $A = U\Sigma V^T$ where U is a $m \times m$ orthogonal matrix, Σ is a $m \times n$ diagonal matrix with non-negative terms on the diagonal, and V is a $n \times n$ orthogonal matrix. The diagonal entries of Σ are known as the singular values of A . The m columns of U and the n column of V are known as the left and right singular vectors of A respectively.

The SVD gives the ‘best’ low-rank approximation of a matrix. To put this in a more formal notation, it minimizes the Frobenius form of the difference between the approximation and the original matrix.

Assume that matrix $A \in R^{m \times n}$ with rank $r > k$. The Frobenius norm approximation problem $\min \|A - Z\|_F$ where $\text{rank}(Z) = k$ has the solution:

$$Z = A_k = U_k \Sigma_k V_k^T$$

where U_k , Σ_k and V_k are the matrices obtained by truncating the SVD to contain only the first k singular vectors/values.

The SVD is implemented by the function *svd* in MATLAB. You can try the following code in matlab to get an idea:

```
A = magic(3)
[U,D,V] = svd(A)
```

You should observe that the columns of the matrices U and V are orthonormal. Also note that Σ is a diagonal matrix with non-negative and monotonically decreasing entries along the diagonal. These are the singular values of A .

4.2 Generating Predictions

We can use SVD in recommender system to capture certain latent relationship between customers and products/movies. We can use this relationship to capture the predicted likeliness of a certain product/movie by a consumer. *So the next big question is*, how do you calculate the SVD of an incomplete matrix? Or rather, how do you complete a matrix in general? In order to handle this, we will study various approximations. We will then run these methods and collect empirical evidence to validate their effectiveness.

4.2.1 Latent Matrix Factorization

Since we cannot generate an SVD of an incomplete matrix, we can try to find a low rank approximation to our matrix using factorization similar to SVD. To do this, we first fill the incomplete values with 0 or any random permissible number within the valid range of our ratings. Thus the factorization of a matrix R into two matrices P and Q could be viewed as a minimization problem where we are trying to reduce the error of our known values from their predicted values. In other words, our ideal factorization would be the one where after multiplyint P and Q , we get back the same value of ratings that the user had filled initially along with some predictions. In order to avoid overfitting and generate good predictions, we use regularized least squares approach to solve the factorization problem. Because Latent Matrix Factorization is difficult

to implement, we discuss this topics in a more formal context in the Appendix. We encourage the user to go through the tutorial in Appendix and run the code provided in the end to get an idea of Latent Matrix Factorization. We now discuss two other simple, yet effective approaches for Matrix Completion.

4.2.2 Average Value Based SVD Completion

Below is the description of steps used to generate predictions using Average Values -

1. Let R be the original sparse matrix where rows represent the user and columns represent the movies
2. Fill the empty cells in each column with average values of the product/movies in that column.
3. Calculate the average rating for each customer \overline{C} using the non zero values
4. Let $R_{norm} = R - \overline{C}$, i.e. mean center each row.
5. Factor R_{norm} using SVD and obtain U , S and V
6. Truncate U , S and V to U_k , S_k and V_k .
7. Compute the resultant matrix $Pred_{norm} = U_k * S_k * V_k$
8. Add the average customer value calculated in Step 3 to this matrix, i.e, $Pred = Pred_{norm} + \overline{C}$



Exercise 9

Implement a function for average value SVD completion. Use the steps described above and run the code on matrix generated from table 1. Compare the result with our implementation located in *Examples/-ToyExampleAverageValueSVD.m*. What do you think about the corresponding predictions? Do they make sense?

4.2.3 Iterative SVD Completion

Another simple approach to complete a matrix and generate predictions is to use iterative SVD.

1. Let R be the original sparse matrix where rows represent the user and columns represent the movies
2. Fill the empty cells in each row with 0. Let this new Matrix be R_{next}
3. Factorize R_{next} using SVD and obtain U , S and V
4. Truncate U , S and V to U_k , S_k and V_k .
5. Compute the resultant matrix $R_{next} = U_k * S_k * V_k$
6. Use the known values in R to replace values in R_{next}
7. Repeat the process from Step 3, T number of times. T is the number of times you want to repeat the process.



Exercise 10

Implement a function for iterative SVD completion. Use the steps described above and run your code on the matrix generated from table 1. Compare the result with our implementation located in *Examples/-ToyExampleIterativeSVD.m*. Look at the corresponding predictions. What do you think about them? How do they compare to the predictions using the previous methods?

4.3 Nearest Neighbor

By now you should have learnt to predict the ratings for unmarked items. So what do we do with these predictions? One thing we can do is recommend the top rated predictions for each user and that works perfectly fine. Another thing we can look into is the Nearest Neighbor approach that we discussed earlier in the lab.

4.3.1 Using Predicted Matrix

Once we have a prediction for all the movies/products of all the users, we can use nearest neighbors to generate item recommendation for a user. Why do we do this when we already have a prediction? Well, nearest neighbor method can lead to something called as "Serendipity". We will discuss the meaning of Serendipity in a later section. This method is based on the assumption - given a high accuracy of prediction, we might find better neighbors and thus better ratings.

4.3.2 Original Matrix in Lower Dimension

Another work around to find better neighbors is to look for neighbors in a lower dimension space of the original sparse matrix. The motivation here is that a lower dimension space will lead to a denser matrix and might help us find better neighbors.

4.3.3 Which Nearest Neighbor approximation to Use?

To answer this question we will resort to the golden answer to anything non-concrete - *"It Depends!!!"*. For the purpose of this lab we used Nearest Neighbor approach in a lower dimension space of predicted matrix. Empirically, it worked better. In general, one tries different approaches and sees what works best for their environment and their preference of error metrics.

4.4 Bringing it All Together

Now that you have all the concepts required to understand a Model Based Recommender System, we will encourage you to experiment with it to get an idea of the different trade-offs involved and the corresponding complexity in settling down on a method.



Exercise 11

Write a code to run Model Based Prediction on the entire movie lens dataset. Follow these steps -

1. Load `u1.base` and `u1.test`.
2. Create a Matrix of the form `user*movies` where uninitialized values to zero. We already have function calls to do both these steps so feel free to use them
3. Create two matrix completion for the `u1.base` dataset, one based on Average Value and one based on Incremental SVD
4. Look at the Top 5 predictions from both of them for any random customer.
5. Pick a customer from `u1.test` and look at the movies he has rated.
6. Look at the corresponding ratings in your predictions. Do they agree?
7. Calculate the various error metrics that we discussed earlier in the lab.

We have already implemented this code in *ModelBasedPrediction / EvaluateModelBasedSVD.m*. Play around with the parameters, look at the impact it has on different error rates. It is upto you to decide what factors are important and what constitutes a good system. We have provided you with outputs for many different parameters in the [table-reference-here](#)[TODO FIXME].

5 Miscellaneous Topics

5.1 Content Based Recommender Systems

Content-based recommender systems recommend an item to a user based upon a description of the item and the profile of the users interests. Systems implementing a content-based recommendation approach analyze a set of documents and/or descriptions of items previously rated by a user, and build a model or profile of user interests based on the features of the objects rated by that user. The profile is a structured representation of user interests, adopted to recommend new interesting items. The recommendation process basically consists in matching up the attributes of the user profile against the attributes of a content object. The result is a relevance judgment that represents the users level of interest in that object. If a profile accurately reflects user preferences, it is of tremendous advantage for the effectiveness of an information access process. For instance, it could be used to filter search results by deciding whether a user is interested in a specific Web page or not and, in the negative case, preventing it from being displayed.

5.1.1 Advantages and Drawbacks of Content-based Filtering

The adoption of the content-based recommendation paradigm has several advantages when compared to the collaborative one:

- **User Independence** - Content-based recommenders exploit solely ratings provided by the active user to build her own profile.
- **Transparency** - Explanations on how the recommender system works can be provided by explicitly listing content features or descriptions that caused an item to occur in the list of recommendations.
- **New Item** - Content-based recommenders are capable of recommending items not yet rated by any user i.e., they do not suffer from the first-rater problem.

Nonetheless, content-based systems have several shortcomings:

- **Limited Content Analysis** - Content-based techniques have a natural limit in the number and type of features that are associated with the objects they recommend. Domain knowledge is often needed, e.g., for movie recommendations the system needs to know the actors and directors, and sometimes, domain ontologies are also needed. No content-based recommendation system can provide suitable suggestions if the analyzed content does not contain enough information to discriminate items the user likes from items the user does not like.
- **Over-Specialization** - Content-based recommenders have no inherent method for finding something unexpected. The system suggests items whose scores are high when matched against the user profile, hence the user is going to be recommended items similar to those already rated. This drawback is also called serendipity problem to highlight the tendency of the content-based systems to produce recommendations with a limited degree of novelty.
- **New User** - Enough ratings have to be collected before a content-based recommender system can really understand user preferences and provide accurate recommendations. Therefore, when few ratings are available, as for a new user, the system will not be able to provide reliable recommendations.

5.2 Serendipity

Recommender Systems exist to help users discover an item that he or she would like among the large pool of items available. Most of the errors that we saw did not talk about the novelty of recommendation. For example, if I have seen *Star Wars: A New Hope*, *Star Wars: The Empire Strikes Back* and *Star Wars: Revenge of the Sith*, I will most likely find a lot of nearest neighbors whose collective votes leads to the recommendation for *The Phantom Menace*, *Attack of the Clones* and *Revenge of the Sith*. I will definitely end up liking that movie. Now consider a scenario where watch a movie of a completely different genre, say *Lincoln* and enjoy it a lot. An item based recommendation system would have never led me to discover this movie as all the movies I have rated fall into the Action/Saga/Fantasy category. Such discoveries are called serendipitous discoveries.

Serendipity is related to unexpectedness and involves a positive emotional response of the user about a previously unknown item. It is concerned with the novelty of recommendations and how far such recommendations may positively affect the user. Serendipity is generally not easy to measure and most systems consider a trade-off between various forms of accuracy and serendipity.

Appendix

A Latent Matrix Factorization

We talked about Latent Matrix Factorization in the document earlier. In this section, we will go through the basics of matrix factorization, and then we will look into a simple implementation in Python. Assume that we are dealing with user ratings (score in the range of 1 to 5) of items in a recommendation system.

Basic Idea

In recommendation system like Netflix or MovieLens, there are a group of users and a set of items. Given that each users have rated only some items in the system, we would like to predict how the users would rate the other not yet rated items, so that we can make appropriate recommendations to the users. We can represent all the information we have about the existing ratings in a matrix. Assume that we have 5 users and 10 items, and ratings range from 1 to 5 (integer values), the matrix may look something like this:

Table 5: User Ratings

| | D1 | D2 | D3 | D4 |
|----|----|----|----|----|
| U1 | 5 | 3 | - | 1 |
| U2 | 4 | - | - | 1 |
| U3 | 1 | 1 | - | 5 |
| U4 | 1 | - | - | 4 |
| U5 | - | 1 | - | 4 |

The task of predicting the missing ratings can be considered as filling in the blanks such that the values would be consistent with the existing ratings in the matrix.

The basic idea behind using matrix factorization is that there should be some latent features that can predict how a user rates an item. For example, two users would give high ratings to a certain movie if they both like the actors/actresses of the movie, or if the movie is an comedy movie, which is a genre preferred by both users. Hence, if we can discover these latent features, we should be able to determine a rating with respect to a certain user and a certain item, because the features associated with the user should match with the features associated with the item.

The Mathematics of Matrix Factorization

Having discussed the intuition behind matrix factorization, we can now go on to work on the mathematics. Firstly, we have a set U of users, and a set D of items. Let \mathbf{R} of size $|U| \times |D|$ be the matrix that contains all the ratings that the users have assigned to the items. Also, we assume that we would like to discover K latent features. Our task, then, is to find two matrices \mathbf{P} (a $|U| \times K$ matrix) and \mathbf{Q} (a $|D| \times K$ matrix) such that their product approximates \mathbf{R} :

$$R \approx P \times Q^T = \hat{R}$$

In this way, each row of \mathbf{P} would represent the strength of the associations between a user and the features. Similarly, each row of \mathbf{Q} would represent the strength of the associations between an item and the features. To get the prediction of a rating of an item d_j by u_i , we can calculate the dot product of the two vectors corresponding to u_i and d_j :

$$r_{ij} = p_i^T q_j = \sum_{k=1}^K p_{ik} q_{kj}$$

Now, we have to find a way to obtain \mathbf{P} and \mathbf{Q} . One way to approach this problem is the first initialize the two matrices with some values, calculate how 'different their product is to \mathbf{M} , and then try to minimize this difference iteratively. Such a method is called gradient descent, aiming at finding a local minimum of the difference.

Here we consider the squared error because the estimated rating can be either higher or lower than the real rating.

To minimize the error, we have to know in which direction we have to modify the values of p_{ik} and q_{kj} . In other words, we need to know the gradient at the current values, and therefore we differentiate the above equation with respect to these two variables separately:

$$\begin{aligned}\frac{\partial}{\partial p_{ik}} e_{ij}^2 &= -2(r_{ij} - \widehat{r_{ij}})(q_{kj}) = -2e_{ij}q_{kj} \\ \frac{\partial}{\partial q_{ik}} e_{ij}^2 &= -2(r_{ij} - \widehat{r_{ij}})(p_{ik}) = -2e_{ij}p_{ik}\end{aligned}$$

Having obtained the gradient, we can now formulate the update rules for both p_{ik} and q_{kj} :

$$\begin{aligned}p'_{ik} &= p_{ik} + \alpha \frac{\partial}{\partial p_{ik}} e_{ij}^2 = p_{ik} + 2\alpha e_{ij}q_{kj} \\ q'_{kj} &= q_{kj} + \alpha \frac{\partial}{\partial q_{kj}} e_{ij}^2 = q_{kj} + 2\alpha e_{ij}p_{ik}\end{aligned}$$

Here, α is a constant whose value determines the rate of approaching the minimum. Usually we will choose a small value for α , say 0.0002. This is because if we make too large a step towards the minimum we may run into the risk of missing the minimum and end up oscillating around the minimum.

A question might have come to your mind by now: if we find two matrices \mathbf{P} and \mathbf{Q} such that $\mathbf{P} \times \mathbf{Q}$ approximates \mathbf{R} , isn't that our predictions of all the unseen ratings will all be zeros? In fact, we are not really trying to come up with \mathbf{P} and \mathbf{Q} such that we can reproduce \mathbf{R} exactly. Instead, we will only try to minimize the errors of the observed user-item pairs. In other words, if we let T be a set of tuples, each of which is in the form of (u_i, d_j, r_{ij}) , such that T contains all the observed user-item pairs together with the associated ratings, we are only trying to minimize every e_{ij} for $(u_i, d_j, r_{ij}) \in T$. (In other words, T is our set of training data.) As for the rest of the unknowns, we will be able to determine their values once the associations between the users, items and features have been learnt.

Using the above update rules, we can then iteratively perform the operation until the error converges to its minimum. We can check the overall error as calculated using the following equation and determine when we should stop the process.

$$E = p_i^T q_j = \sum_{(u_i, d_j, r_{ij}) \in T} e_{ij} = \sum_{(u_i, d_j, r_{ij}) \in T} (r_{ij} - \sum_{k=1}^K p_{ik}q_{kj})^2$$

Regularization

The above algorithm is a very basic algorithm for factorizing a matrix. There are a lot of methods to make things look more complicated. A common extension to this basic algorithm is to introduce regularization to avoid over-fitting. This is done by adding a parameter β and modify the squared error as follows:

$$e_{ij}^2 = (r_{ij} - \sum_{k=1}^K p_{ik}q_{kj})^2 + \frac{\beta}{2} \sum_{k=1}^K (\|P\|^2 + \|Q\|^2)^2$$

In other words, the new parameter β is used to control the magnitudes of the user-feature and item-feature vectors such that P and Q would give a good approximation of R without having to contain large numbers.

In practice, β is set to some values in the range of 0.02. The new update rules for this squared error can be obtained by a procedure similar to the one described above. The new update rules are as follows.

$$\begin{aligned} \dot{p}_{ik} &= p_{ik} + \alpha \frac{\partial}{\partial p_{ik}} e_{ij}^2 = p_{ik} + \alpha(2e_{ij}q_{kj} - \beta p_{ik}) \\ \dot{q}_{kj} &= q_{kj} + \alpha \frac{\partial}{\partial q_{kj}} e_{ij}^2 = q_{kj} + \alpha(2e_{ij}p_{ik} - \beta q_{kj}) \end{aligned}$$

Implementation in Python

```
import numpy

def matrix_factorization(R, P, Q, K, steps=5000, alpha=0.0002, beta=0.02):
    Q = Q.T
    for step in xrange(steps):
        for i in xrange(len(R)):
            for j in xrange(len(R[i])):
                if R[i][j] > 0:
                    eij = R[i][j] - numpy.dot(P[i,:],Q[:,j])
                    for k in xrange(K):
                        P[i][k] = P[i][k] + alpha * (2 * eij * Q[k][j] - beta * P[i][k])
                        Q[k][j] = Q[k][j] + alpha * (2 * eij * P[i][k] - beta * Q[k][j])
    eR = numpy.dot(P,Q)
    e = 0
    for i in xrange(len(R)):
        for j in xrange(len(R[i])):
            if R[i][j] > 0:
                e = e + pow(R[i][j] - numpy.dot(P[i,:],Q[:,j]), 2)
                for k in xrange(K):
                    e = e + (beta/2) * (pow(P[i][k],2) + pow(Q[k][j],2))
    if e < 0.001:
        break
    return P, Q.T
```

We can try to apply it to our example mentioned above and see what we would get. Below is a code snippet in Python for running the example.

```
R = [
    [5,3,0,1],
    [4,0,0,1],
    [1,1,0,5],
    [1,0,0,4],
    [0,1,5,4],
]

R = numpy.array(R)

N = len(R)
M = len(R[0])
K = 2

P = numpy.random.rand(N,K)
Q = numpy.random.rand(M,K)

nP, nQ = matrix_factorization(R, P, Q, K)
nR = numpy.dot(nP, nQ.T)
```

And the matrix obtained from the above process would look something like this:

Table 6: Output of Matrix Factorization

| | D1 | D2 | D3 | D4 |
|----|------|------|------|------|
| U1 | 4.97 | 2.98 | 2.18 | 0.98 |
| U2 | 3.97 | 2.40 | 1.97 | 0.99 |
| U3 | 1.02 | 0.93 | 5.32 | 4.93 |
| U4 | 1.00 | 0.85 | 4.59 | 3.93 |
| U5 | 1.36 | 1.07 | 4.89 | 4.12 |

We can see that for existing ratings we have the approximations very close to the true values, and we also get some 'predictions' of the unknown values. In this simple example, we can easily see that U1 and U2 have similar taste and they both rated D1 and D2 high, while the rest of the users preferred D3, D4 and D5. When the number of features (K in the Python code) is 2, the algorithm is able to associate the users and items to two different features, and the predictions also follow these associations. For example, we can see that the predicted rating of U4 on D3 is 4.59, because U4 and U5 both rated D4 high.