

Matrix Factorization Methods for Recommender Systems

Data Mining
Prof. Dawn Woodard
School of ORIE
Cornell University

Recommender Systems

Recommender Systems:

- Achieve a similar goal to association rules, but use different techniques
- Have data on rating or purchase of products by various customers
- While association rules learn rules of the form “If A is purchased then B is likely to be purchased”, recommender systems identify products that have been rated highly/purchased by customers with similar tastes to yours
- Today’s lecture is drawn from Koren, Bell, and Volinsky (2009; “Matrix Factorization Techniques for Recommender Systems”)

Recommender Systems

From Amazon.com (Fig. from SPB text):



[See larger image](#)

[Share your own customer images](#)

Bound Away

Last Train Home

★★★★★ (2 customer reviews)

[More about this product](#)

List Price: \$16.98

Price: \$16.98 & eligible for **FREE Super Saver Shipping** on orders over \$25. [Details](#)

Availability: In Stock.

To ensure delivery by December 22, choose FREE Super Saver Shipping. [See more on holiday shipping.](#) Ships from and sold by **Amazon.com**. Gift-wrap available.

Want it delivered Tuesday, December 5? Order it in the next 9 hours and 5 minutes, and choose **One-Day Shipping** at checkout. [See details](#)

44 used & new available from \$8.99

Better Together

Buy this album with Time and Water ~ Last Train Home today!

Buy Together Today: \$33.96



+



Buy both now!

Recommender Systems

From Netflix:

The screenshot displays the Netflix homepage with a red header. The top navigation bar includes the Netflix logo, links for 'Buy / Redeem Gift', 'Jane Queue', 'Your Account & Help', and a search bar. Below this is a secondary navigation bar with tabs: 'Watch Instantly', 'Browse DVDs', 'Your Queue', and 'Movies You'll ♥'. A sub-navigation bar shows 'Suggestions (700)', 'Rate Movies', 'Taste Preferences', 'Movies You've Rated (651)', and 'Your Reviews'. The main content area features a 'Suggestions in All Genres' dropdown and a 'RATED MOVIES 651' badge. The 'Suggestions to Watch Instantly' section displays three movie cards: 'The Right Stuff', 'Paradise Now', and 'Chaplin: The Movie'. Each card includes a poster, a 'Because you enjoyed:' recommendation, a 'Play' button, and a star rating. The 'New Suggestions' section at the bottom shows 'The Secret in Their Eyes', 'To Kill a Mockingbird', and 'Beauty and the Beast: Special Edition'.

NETFLIX

Buy / Redeem Gift | Jane Queue | Your Account & Help

Watch Instantly | Browse DVDs | Your Queue | Movies You'll ♥

Movies, TV shows, actors, directors, genres

Suggestions (700) | Rate Movies | Taste Preferences | Movies You've Rated (651) | Your Reviews

Suggestions in All Genres

RATED MOVIES 651

Suggestions to Watch Instantly

See all

The Right Stuff
Because you enjoyed:
Apollo 13
Field of Dreams
Glory

Play

★★★★☆
Not Interested

Paradise Now
Because you enjoyed:
Dr. Strangelove
Amélie
30 Rock: Season 2

Play

★★★★☆
Not Interested

Chaplin: The Movie
Because you enjoyed:
The Last Emperor
La Vie en Rose

Play

★★★★☆
Not Interested

New Suggestions

See all

The Secret in Their Eyes
Because you enjoyed:
Amélie

To Kill a Mockingbird
Because you enjoyed:
Annie Hall

Beauty and the Beast: Special Edition
Because you enjoyed:

Share your Netflix movie ratings on Facebook.

Recommender Systems

Data takes the form of a **matrix of ratings** of different products by different users, where **many of the values are missing** because the user has not rated the product:

Users		Movies						
		101	102	103	104	105	106	107
		Rambo	Rocky	Garden State	Before Sunset	Training Day	Thor	Black Swan
1	Alice	5.0	3.0	2.5				
2	Bob	2.0	2.5	5.0	2.0			
3	Charlie	2.5			4.0	4.5		5.0
4	Damon	5.0		3.0	4.5		4.0	
5	Eddie	4.0	3.0	2.0	4.0	3.5	4.0	

Figure 1: User ratings for movies on a scale of 1-5

Recommender Systems

The goal is to fill in the missing data: if Alice had rated Black Swan, what rating would she have given it?

Users		Movies						
		101	102	103	104	105	106	107
		Rambo	Rocky	Garden State	Before Sunset	Training Day	Thor	Black Swan
1	Alice	5.0	3.0	2.5				
2	Bob	2.0	2.5	5.0	2.0			
3	Charlie	2.5			4.0	4.5		5.0
4	Damon	5.0		3.0	4.5		4.0	
5	Eddie	4.0	3.0	2.0	4.0	3.5	4.0	

Figure 1: User ratings for movies on a scale of 1-5

Matrix Factorization

Matrix Factorization:

- The method used to win the Netflix Prize in 2009
- Idea: Try to explain the ratings by identifying a moderate # of abstract “factors” that characterize the items.
- Learn these factors from the data
- For movies, these factors might capture:
 -
 -
 -
- If a movie has a high value for the factor capturing amount of action, for instance, the movie has a lot of action

Matrix Factorization

- The users each have values for the SAME set of factors.
- For users, each factor measures:

Recommender Systems

Illustration: 2 factors capture female- vs. male-oriented and serious vs. escapist:

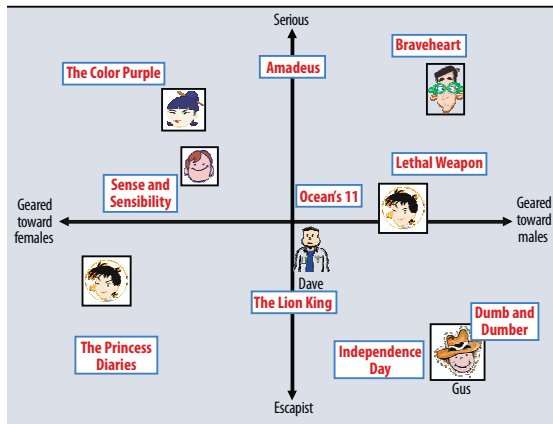


Figure 2 in Koren, Bell, and Volinsky (2009)

Recommender Systems

The factor values of some movies are shown, along with those of some hypothetical users:

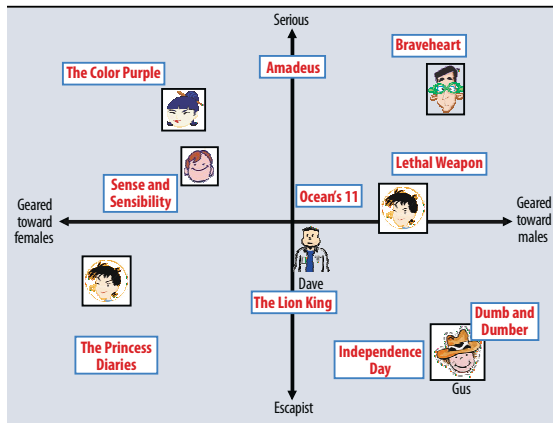


Figure 2 in Koren, Bell, and Volinsky (2009)

Matrix Factorization

- To calculate the estimated rating for a movie by a user, we would take the **inner product** of the movie's factor vector and the user's factor vector
- Example:

Recommender Systems

So we would expect Gus to love Independence Day, to hate The Color Purple, and to rate Braveheart about average

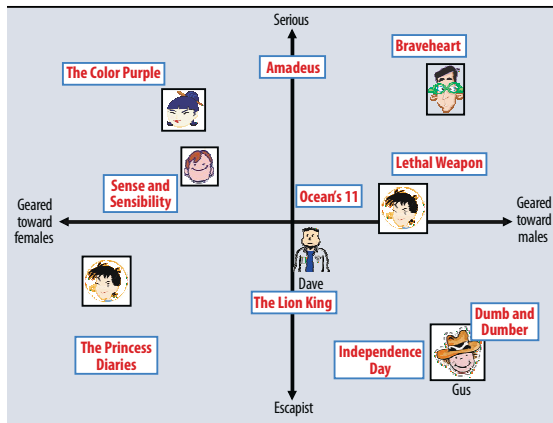


Figure 2 in Koren, Bell, and Volinsky (2009)

Matrix Factorization

Matrix Factorization in general:

- We want to identify some number K of abstract factors, and estimate the factor vectors for all the items and all the users.
- For each item i , write the unknown length- K factor (column) vector as q_i
- For each user u , write the unknown length- K factor (column) vector as p_u
- For the user-item pairs with known rating, write it r_{ui}
- The estimated rating is defined as

$$\hat{r}_{ui} = q_i^T p_u.$$

■

Matrix Factorization

How to estimate the q_i and p_u ?

- We want the difference between \hat{r}_{ui} and r_{ui} to be close for the user-item pairs for which we have observed a rating
- So perhaps we could choose q_i and p_u to minimize the sum of squared errors:

Matrix Factorization

- This approach tends to overfit to the data.
- The method works better if we create an incentive for the q_i and p_u to be close to the origin (when it does not increase the error much)
- So we instead minimize the objective

$$\sum_{(u,i) \in \mathcal{A}} (r_{ui} - q_i^T p_u)^2 + \lambda(\|q_i\|^2 + \|p_u\|^2)$$

where $\lambda > 0$ penalizes the length of q_i and p_u

- λ controls:
- We can choose λ by cross-validation (minimize RMSE of observed ratings on validation data)!

The Netflix Prize

- In 2006 Netflix announced a contest to improve its methods for movie recommendation
- Netflix released a training set of more than 100 million ratings from 500,000 anonymized customers, for more than 17,000 movies
- Each movie is rated on a scale 1-5
- Competing teams submitted predicted user ratings for a test set of 3 million ratings
- Then Netflix would calculate the RMSE on the test set
- The first team to improve over the Netflix algorithm by $> 10\%$ was to win \$1,000,000

The Netflix Prize

- The prize was won in 2009 by the team “BellKor’s Pragmatic Chaos” with 10.09% improvement over Netflix’s method
- Unfortunately Netflix was sued for privacy violations after it was discovered that some users could be identified by matching the anonymized ratings with film ratings on the Internet Movie Database

Recommender Systems

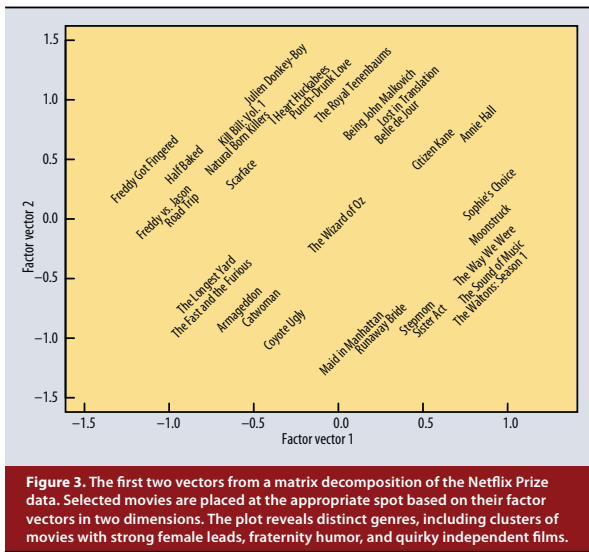


Figure 3. The first two vectors from a matrix decomposition of the Netflix Prize data. Selected movies are placed at the appropriate spot based on their factor vectors in two dimensions. The plot reveals distinct genres, including clusters of movies with strong female leads, fraternity humor, and quirky independent films.