User Preference Modeling： Film Ratings

This is unsupervised learning: Matrix factorization
Learn an underlying dot-product representation.

Theory：

This is the algorithm we use.

MAP inference coordinate ascent algorithm

Input: An incomplete ratings matrix M, as indexed by the set Ω. Rank d.

Output: N_1 user locations, $u_i \in \mathbb{R}^d$, and N_2 object locations, $v_j \in \mathbb{R}^d$.

Initialize each v_j. For example, generate $v_j \sim N(0, \lambda^{-1} I)$.

for each iteration **do**

- **for** $i = 1, \ldots, N_1$ **update user location**

$$u_i = \left(\lambda\sigma^2 I + \sum_{j \in \Omega_{u_i}} v_j v_j^T\right)^{-1} \left(\sum_{j \in \Omega_{u_i}} M_{ij} v_j\right)$$

- **for** $j = 1, \ldots, N_2$ **update object location**

$$v_j = \left(\lambda\sigma^2 I + \sum_{i \in \Omega_{v_j}} u_i u_i^T\right)^{-1} \left(\sum_{i \in \Omega_{v_j}} M_{ij} u_i\right)$$

Predict that user i rates object j as $u_i^T v_j$ rounded to closest rating option

We want to get the Matrix U and V to maximum the objective function.

Log joint likelihood and MAP

The MAP solution for U and V is the maximum of the log joint likelihood

$$U_{\text{MAP}}, V_{\text{MAP}} = \arg\max_{U,V} \sum_{(i,j)\in\Omega} \ln p(M_{ij}|u_i, v_j) + \sum_{i=1}^{N_1} \ln p(u_i) + \sum_{j=1}^{N_2} \ln p(v_j)$$

Calling the MAP objective function $\mathcal{L}$, we want to maximize

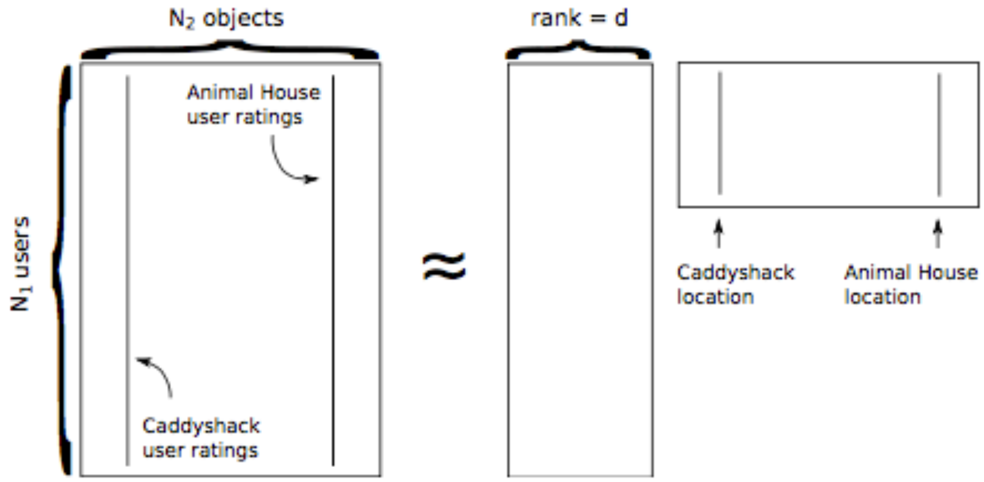
$$\mathcal{L} = -\sum_{(i,j)\in\Omega} \frac{1}{2\sigma^2}\|M_{ij} - u_i^T v_j\|^2 - \sum_{i=1}^{N_1} \frac{\lambda}{2}\|u_i\|^2 - \sum_{j=1}^{N_2} \frac{\lambda}{2}\|v_j\|^2 + \text{constant}$$

The squared terms appear because all distributions are Gaussian.

Imagine that the big matrix M on the left contains ratings from audiences for movies. Each row represents an audience, each column represents a movie. Each audience will rate only some of the movies, so there would be a lot of Nan values in the movie.

On the right side. The left small matrix represents audiences, we call it U, the right small matrix represents the movies, we call it V.
We would train the model using the algorithm for 100 iterations then we update U and V matrix per iteration. For the final iteration, we get the final U and V. We use this two matrix to calculate the predicted rating an audience for a movie.

For example, the rating from the ith audience for the jth movie, which is Mij is equal to the dot product of Ui and Vj, which is the ith row of U and jth column of V respectively.

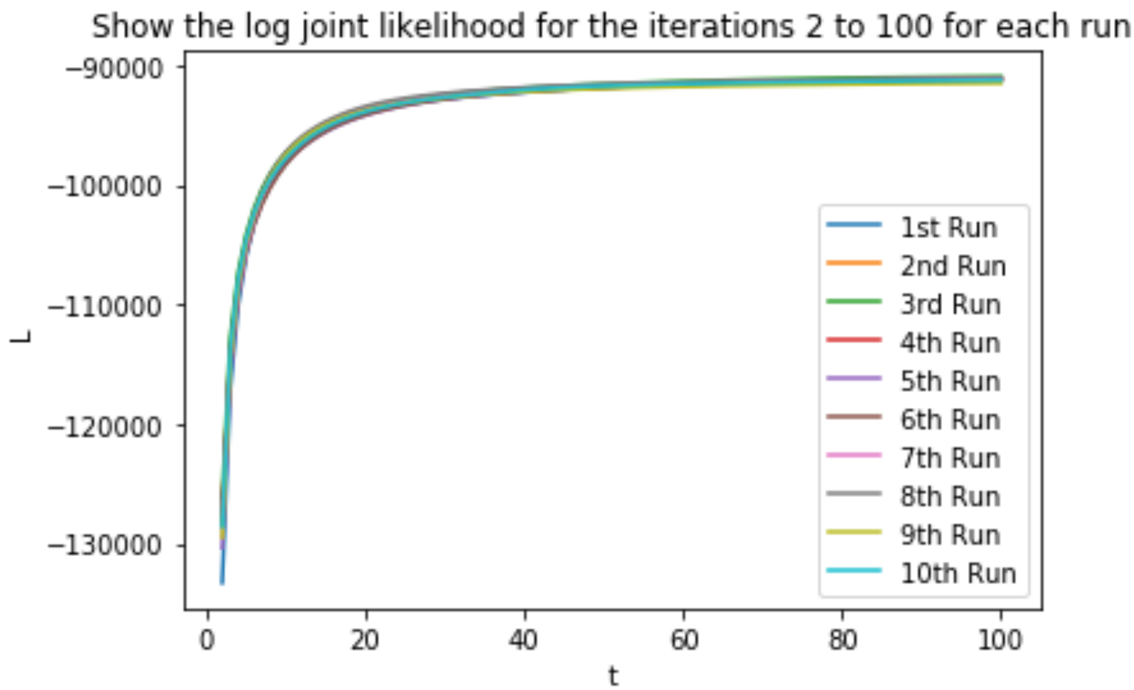


Data:

Training Data:
95000 pairs of user and movie rating pair.
943 users
1682 movies

Testing Data:
5000 pairs of user and movie rating pair

Implement the algorithm for ten times.
For each run, run the algorithm for 100 iterations.
Get the following graph showing the objective value vs. per iteration. The objective value is converging. Actually they all converge much ealier before the 100th iteration. So we could believe our model and result.



Calculate the RMSE on the testing data. Find the 3rd run has the highest objective value. The dataframe is sorted on the column "Value of Training Objective Function" in decreasing order.

	Number of Run	RMSE	Value of Training Objective Function
1	3	1.135890	-90926.773002
2	8	1.112526	-91071.487002
3	7	1.105095	-91122.178009
4	6	1.111487	-91130.091731
5	5	1.149956	-91139.434243
6	2	1.139418	-91158.631579
7	4	1.098039	-91184.861056
8	10	1.136748	-91225.683045
9	1	1.123147	-91256.867434
10	9	1.115471	-91499.521658

Use the U and V matrix we get in the 3rd run, get the following: For example, the ten nearest movie to "Lion King, The (1994)"", "Sleepless in Seattle (1993)", "Gone with the Wind (1939)", "Godfather: Part II, The (1974)", "101 Dalmatians (1996)" and "Bean (1997)" according to Euclidean distance.

10 Closest movies to "Lion King, The (1994)"

	Moive Name	Euclidean Distance with "The Lion King (1994)"
1	Aladdin (1992)	0.341041
2	Beauty and the Beast (1991)	0.458715
3	Toy Story (1995)	0.513229
4	Sleepless in Seattle (1993)	0.546174
5	Dave (1993)	0.672226
6	That Thing You Do! (1996)	0.679255
7	Apollo 13 (1995)	0.698691
8	When Harry Met Sally... (1989)	0.708046
9	Indiana Jones and the Last Crusade (1989)	0.713611
10	Back to the Future (1985)	0.728628

10 Closest movies to "Sleepless in Seattle (1993)"

	Moive Name	Euclidean Distance with "Sleepless in Seattle (1993)"
1	Mrs. Doubtfire (1993)	0.433582
2	American President, The (1995)	0.456897
3	Firm, The (1993)	0.482652
4	Mr. Holland's Opus (1995)	0.504514
5	Dave (1993)	0.526407
6	Lion King, The (1994)	0.546174
7	Aladdin (1992)	0.555524
8	Ransom (1996)	0.590753
9	Speed (1994)	0.597748
10	That Thing You Do! (1996)	0.639877

10 Closest movies to "Gone with the Wind (1939)"

	Moive Name	Euclidean Distance with "Gone with the Wind (1939)"
1	Mary Poppins (1964)	0.596637
2	Mr. Smith Goes to Washington (1939)	0.638874
3	It's a Wonderful Life (1946)	0.675656
4	Christmas Carol, A (1938)	0.743172
5	Snow White and the Seven Dwarfs (1937)	0.749223
6	Arsenic and Old Lace (1944)	0.753570
7	Old Yeller (1957)	0.763035
8	Miracle on 34th Street (1994)	0.782993
9	Wizard of Oz, The (1939)	0.783382
10	African Queen, The (1951)	0.795397

10 Closest movies to "Godfather: Part II, The (1974)"

	Moive Name	Euclidean Distance with "Godfather: Part II, The (1974)"
1	Godfather, The (1972)	0.352685
2	Cool Hand Luke (1967)	0.628378
3	Patton (1970)	0.643152
4	Lawrence of Arabia (1962)	0.743002
5	GoodFellas (1990)	0.758400
6	Unforgiven (1992)	0.762347
7	Taxi Driver (1976)	0.806701
8	2001: A Space Odyssey (1968)	0.818078
9	Chinatown (1974)	0.839072
10	Apocalypse Now (1979)	0.849496

10 Closest movies to "101 Dalmatians (1996)"

	Moive Name	Euclidean Distance with "101 Dalmatians (1996)"
1	Raw Deal (1948)	0.558573
2	Nightwatch (1997)	0.585806
3	Murder at 1600 (1997)	0.654093
4	Lassie (1994)	0.659407
5	Net, The (1995)	0.678053
6	Cool Runnings (1993)	0.706349
7	Kika (1993)	0.725907
8	Father of the Bride Part II (1995)	0.740802
9	Batman Returns (1992)	0.742844
10	Favor, The (1994)	0.754165

10 Closest movies to "Bean (1997)"

	Moive Name	Euclidean Distance with "Bean (1997)"
1	Disclosure (1994)	1.423766
2	Lost World: Jurassic Park, The (1997)	1.650735
3	Star Trek III: The Search for Spock (1984)	1.690907
4	Man in the Iron Mask, The (1998)	1.694907
5	Malice (1993)	1.748782
6	Bananas (1971)	1.775444
7	Madonna: Truth or Dare (1991)	1.781742
8	Highlander III: The Sorcerer (1994)	1.793188
9	Escape from L.A. (1996)	1.812162
10	That Darn Cat! (1965)	1.824453