Carnegie Mellon University
Tepper School of Business

46-886: Machine Learning Fundamentals

Amr Farahat

Recommender Systems:
Application to MovieLens Dataset

Much of this slide deck is derived/borrowed from course material I've co-taught at MIT

MovieLens Dataset

- We will work instead with the MovieLens dataset
  - 6,040 users
  - 3,900 movies
  - 1,000,209 ratings of movies
- On average, each user rated 166 of the 3,900 movies (4.2%)
  - The goal is to predict/infer/estimate ratings of the other 95.8% user-movie pairings
  - Need to estimate 22,555,791 missing user ratings

MovieLens Data

n: # of observations ($n = 1,000,209$)

	rating	userID	movieID	date
1	5	1	1193	2000-12-31 17:12:40
2	3	1	661	2000-12-31 17:35:09
3	3	1	914	2000-12-31 17:32:48
4	4	1	3408	2000-12-31 17:04:35
5	5	1	2355	2001-01-06 18:38:11
6	3	1	1197	2000-12-31 17:37:48
7	5	1	1287	2000-12-31 17:33:59
8	5	1	2804	2000-12-31 17:11:59
9	4	1	594	2000-12-31 17:37:48
10	4	1	919	2000-12-31 17:22:48
11	5	1	595	2001-01-06 18:37:48
12	4	1	938	2000-12-31 17:29:12
13	4	1	2398	2000-12-31 17:38:01
14	4	1	2918	2000-12-31 17:35:24
15	5	1	1035	2000-12-31 17:29:13
16	4	1	2791	2000-12-31 17:36:28
...	...	...	...	...
1000205	1	6040	1091	2000-04-25 22:35:41
1000206	5	6040	1094	2000-04-25 19:21:27
1000207	5	6040	562	2000-04-25 19:19:06
1000208	4	6040	1096	2000-04-25 22:20:48
1000209	4	6040	1097	2000-04-25 22:19:29

Implementation: Python's LightFm Package (Optional)

- Ideal for neighborhood-based Collaborative Filtering

- https://making.lyst.com/lightfm/docs/home.html
- https://github.com/lyst/lightfm/blob/master/examples/quickstart/quickstart.ipynb

Implementation: R's SoftImpute Package (Optional)

- Ideal for model-based Collaborative Filtering

```
library(softImpute)

# The training and test data have 3 columns: user id, movie id, ratings

mat <- Incomplete(train [,1], train [,2], train [,3])

fit <- softImpute(mat, rank.max=9, lambda=0, maxit=1000)

fit$u: gives the user-archetype combinations

fit$v * fit$d: gives the archetype-movie ratings

pred <- impute(fit, test[, 1], test[, 2])
```

library

matrix

model fitting

test-set predictions

- The following analysis is based on the softimpute package

Movie-Archetype Ratings

Archetype	Bird Box	Roma	Avengers: Infinity War	Black Panther	Solo: A Star Wars Story	Thor: Ragnarok	Zodiac	Incredibles 2	Fyre	…
1	3	5	2	5	5	3	3	4	1	…
2	5	4	5	2	3	1	5	4	5	…
…	…	…	…	…	…	…	…	…	…	…
…	…	…	…	…	…	…	…	…	…	…
9	2	5	3	5	4	5	1	4	3	…

	[,1]	[,2]	[,3]	[,4]	[,5]	[,6]	[,7]	[,8]	[,9]
[1,]	-0.010285232	6.956747e-03	-6.044632e-03	-1.289385e-02	1.574182e-02	2.227203e-02	1.707377e-02	1.977146e-02	-1.087540e-02
[2,]	-0.012344893	6.105412e-03	-6.572170e-03	5.094671e-03	2.240189e-02	-2.220966e-02	2.139380e-03	1.111205e-02	-5.284898e-03
[3,]	-0.010258076	2.449277e-02	1.835016e-03	1.336783e-02	6.813315e-03	-1.195719e-02	-1.190039e-02	8.282133e-03	6.439084e-03
[4,]	-0.006645143	1.238735e-02	-2.273594e-02	2.068818e-02	-2.364472e-03	-1.668273e-02	7.594584e-03	9.782163e-03	1.519675e-02
[5,]	-0.012251521	-1.145180e-02	-1.912628e-03	-1.364508e-02	-1.101168e-02	2.131751e-02	3.487680e-03	-6.096829e-03	1.380375e-02
[6,]	-0.012154679	5.379816e-03	2.531307e-03	6.138166e-03	4.464902e-02	3.451502e-03	7.005278e-03	1.567645e-04	-1.353828e-02
[7,]	-0.006511135	2.066783e-02	-3.972466e-03	1.285268e-02	2.398443e-03	-2.171474e-02	2.511292e-02	2.528996e-02	2.519582e-02
[8,]	-0.013592265	1.266743e-03	8.825011e-03	-3.575442e-02	-1.048578e-02	-5.233410e-03	1.850059e-02	6.996654e-03	1.801823e-02
[9,]	-0.011316615	1.005654e-02	-1.044946e-03	-2.322501e-02	-2.651104e-03	-8.045379e-05	-2.005455e-03	7.769253e-03	1.120376e-02
[10,]	-0.020120629	-9.499954e-03	2.092072e-02	2.156243e-04	1.281555e-02	7.191000e-03	2.399823e-03	3.773381e-03	6.424719e-03
[11,]	-0.011467069	5.962614e-03	2.508387e-03	-2.527401e-02	-6.866954e-03	6.121569e-03	-1.706545e-02	6.369188e-03	-1.531428e-02
[12,]	-0.008781337	7.993272e-03	-1.896017e-02	6.992801e-03	-1.207661e-02	-6.258294e-03	2.156462e-03	-4.593993e-03	-5.866125e-03
[13,]	-0.010064586	8.218105e-03	-1.266595e-03	1.456695e-02	5.331772e-03	-3.059282e-03	1.212330e-02	1.248064e-02	1.288025e-02
[14,]	-0.007630269	4.145655e-03	-2.127939e-02	-3.183541e-03	-5.485976e-03	-2.680324e-03	1.010888e-02	-1.539796e-02	2.457152e-03
[15,]	-0.010456894	1.755345e-02	1.430709e-02	-1.393102e-02	3.778003e-03	-1.148099e-02	5.170825e-03	3.189222e-03	-7.059932e-03
[16,]	-0.007264796	1.080755e-02	-6.762122e-03	-7.456529e-03	-2.870412e-03	2.272647e-02	1.607834e-02	-2.736593e-02	5.981621e-03
[17,]	-0.018684556	-1.382709e-02	6.696824e-03	1.781394e-03	-1.009749e-02	3.304828e-04	6.866780e-03	1.548684e-03	1.388230e-02
[18,]	-0.017681834	-8.977174e-03	5.932550e-04	1.641735e-03	1.354223e-02	8.720808e-03	-1.010710e-02	6.330836e-03	1.819685e-02
[19,]	-0.015832864	9.103966e-03	1.551252e-02	9.194664e-03	1.266139e-03	4.235855e-03	-1.851253e-02	-3.604855e-03	-7.856964e-03
[20,]	-0.007794994	2.018210e-02	-4.841692e-03	1.922852e-03	-1.397836e-02	-1.293446e-02	1.864254e-02	3.685368e-03	1.234879e-02
[21,]	-0.005758456	1.154898e-02	2.651461e-03	5.077005e-03	-1.244090e-02	2.343798e-02	6.655368e-03	-1.490443e-02	2.825629e-02
[22,]	-0.013520195	3.198393e-03	1.229410e-02	-3.984148e-03	-1.320959e-02	6.984560e-03	-1.844733e-02	7.871633e-03	-1.888646e-03

- The table $S_{a,m}$ has:
  - 9 rows corresponding to the archetypes created by the model
  - 3,900 columns corresponding to the movies
  - Non-integer values—positive or negative
- Hard to interpret results and define archetypes “in words”

Aggregation by Movie Genre

- For interpretability, we can classify movies in different genres and report the average rating per archetype-genre pair

Archetype	Action	Adventure	Comedy	Documentary	Fantasy	Horror	Musical	Romance	Sci-Fi	Thriller
Archetype 1	-4.46	-4.35	-3.92	-2.78	-4.66	-2.99	-4.63	-4.14	-4.44	-4.33
Archetype 2	3.48	2.38	-0.14	-5	2.95	-0.68	-1.23	-1.66	2.34	0.75
Archetype 3	-3.9	-3.32	-2.43	0.25	-2.8	-4.62	0.42	-1.11	-2.4	-2.41
Archetype 4	-2.24	-2.56	0.37	1.19	-1.55	-1.69	-1.64	1.64	-3.05	-0.54
Archetype 5	-0.86	0.51	0.95	-1.99	0.53	-5	3.13	2.76	-3.48	-1.43
Archetype 6	1.67	-0.7	-1.39	-1.01	-3.35	-1.24	-4.26	0.11	-0.5	1.92
Archetype 7	-3.58	-4.32	4.91	-3.3	-2.95	-0.15	-3.41	0.17	-4.64	-3.7
Archetype 8	0.38	1.48	-1.79	-0.71	2.08	2.6	1.54	0.17	2.4	-0.69
Archetype 9	-3.66	-3.14	0.2	-0.71	-5	1.82	0.89	-2.14	-3.84	-1.01

- Interpretation:
  - Archetype 2 seems oriented toward action/adventure/fantasy/SciFi
  - Archetype 7 seems focused on comedy exclusively
  - Etc.

Model Quality on the Test Set

- We use the observed user-movie ratings to assess model performance—leveraging the usual metrics of fit

1. R^2: how well the model fits the observed user-movie ratings
   - Defined as in linear regression
2. MAE: Mean Absolute Error

$$\boxed{\mathrm{MAE} = \frac{1}{N}\left(\sum_{\text{all test pairs }(u,m)} |w_{u,1}S_{1,m} + w_{u,2}S_{2,m} + \cdots + w_{u,9}S_{9,m} - \mathrm{OBSR}_{u,m}|\right)}$$

3. RMSE: Root-Mean-Square Error

$$\boxed{\mathrm{RMSE} = \sqrt{\frac{1}{N}\left(\sum_{\text{all test pairs }(u,m)} (w_{u,1}S_{1,m} + w_{u,2}S_{2,m} + \cdots + w_{u,9}S_{9,m} - \mathrm{OBSR}_{u,m})^2\right)}}$$

Results: MovieLens Ratings Prediction

- Results of our collaborative filtering model on the test set:
  - Fit of 32.4%, as measured by the R^2
  - Average error of 0.70–0.91 (on a ratings scale of 1 to 5), as measured by the MAE and RMSE

Model	R^2	MAE	RMSE
Collaborative Filtering Model	0.324	0.700	0.908

- As a point of reference, Netflix's Cinematch algorithm reported an RMSE of 0.952
  - If we were competing for the Netflix prize, these results would have achieved a 4.62% RMSE improvement

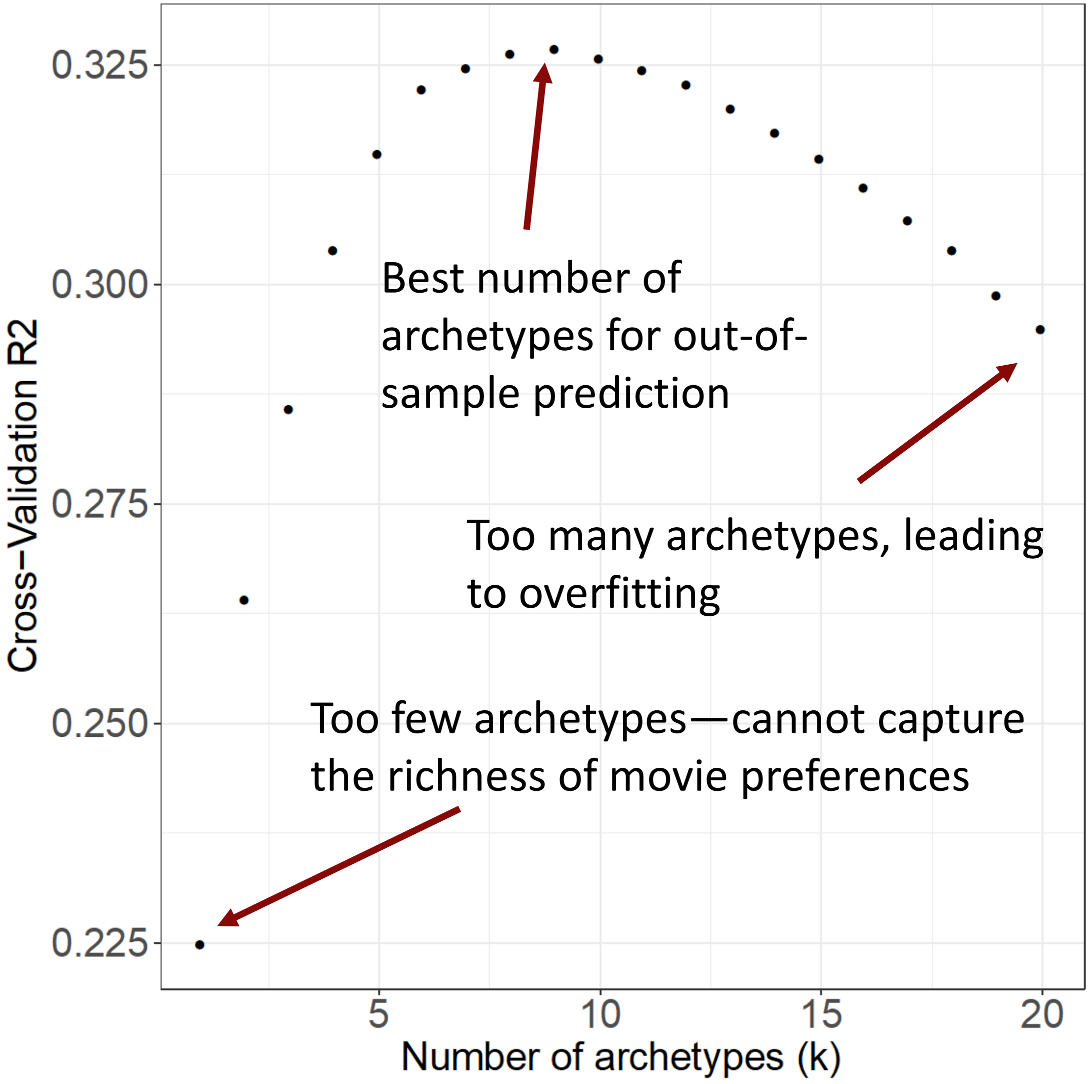
Number of Archetypal Users

- How many "archetypal users" should we use in the collaborative filtering model?
  - If we have too few archetypal users, the predictions will be too rough—we will not capture important patterns in the data
  - If we have too many archetypal users, the model will be overfit to the training set—resulting in poor out-of-sample performance
- We will use cross-validation to select the value of k that yields the best prediction on unseen data
  - For each value of k, perform 10-fold cross-validation*
  - Record the average R^2 using the current value of k. Call it **AvgR2**
  - Choose the value of k that yields the highest value of **AvgR2**

*Please don't confuse the *k* from *k*-fold cross-validation to the number of archetypes *k*

Cross-Validation Results

	k	r2	RMSE	MAE
1	1	0.2248233	0.9836520	0.7676702
2	2	0.2641345	0.9583857	0.7465296
3	3	0.2858431	0.9441433	0.7345625
4	4	0.3040083	0.9320585	0.7234563
5	5	0.3149789	0.9246835	0.7165823
6	6	0.3223250	0.9197120	0.7119121
7	7	0.3247095	0.9180925	0.7097197
8	8	0.3263409	0.9169829	0.7080140
9	9	0.3269581	0.9165627	0.7068340
10	10	0.3258245	0.9173342	0.7071070
11	11	0.3244393	0.9182761	0.7071731
12	12	0.3228265	0.9193716	0.7071403
13	13	0.3200789	0.9212349	0.7079009
14	14	0.3173621	0.9230736	0.7088670
15	15	0.3144179	0.9250621	0.7095384
16	16	0.3111153	0.9272875	0.7110537
17	17	0.3074181	0.9297725	0.7125528
18	18	0.3040299	0.9320440	0.7137350
19	19	0.2988428	0.9355108	0.7157445
20	20	0.2950465	0.9380400	0.7173114



An Ensemble Model

Additional Data

- Our Collaborative Filtering model has leveraged data on user-movie ratings to “reconstruct” missing data
- But we have access to additional data on the movies
  - Genre (out of 18 genres): Action, Adventure, …, Western
- And we also have access to additional data on the users
  - Gender
  - Age
  - Occupation (out of 21 categories): Administrator, Artist, …, Writer
  - Zip code—hence estimates of income, rural/urban living, etc.
- Last, recall that we know the date and time of each movie rating

Ensemble Learning

- How can we make use of this additional information?
- Approach 1: Forget about collaborative filtering, just use your favorite method (linear regression, CART, random forests, etc.)
- Approach 2: Combine the collaborative filtering (CF) model into another method (linear regression, CART, random forests, etc.)
  - Use the CF output as an additional independent variable!
  - For instance, let $CF_{u,m}$ be the predicted rating of user-movie pair (u,m) using our CF model. Then consider the following regression model:

$$\boxed{\mathrm{R}_{u,m} = \beta_0 + \beta_1 \cdot \mathrm{CF}_{u,m} + \begin{array}{c}\text{linear model}\\ \text{of movie data}\end{array} + \begin{array}{c}\text{linear model}\\ \text{of user data}\end{array}}$$

- An **ensemble model** combines different predictive methods

Extended Training Data

k: # of independent variables ($k = 51$); n: # of observations ($n = 990,206$)

	rating	wday	mon	year	hour	AgeRange	Jobacademic	...	Jobwriter	Male	MedianIncome	Urban	RegionMidwest	...	RegionWest	Action	...	Western	CF
1	5	0	12	2000	17	1	0	...	0	0	63015	1	1	...	0	0	...	0	4.390944
2	3	0	12	2000	17	1	0	...	0	0	63015	1	1	...	0	0	...	0	3.707804
3	3	0	12	2000	17	1	0	...	0	0	63015	1	1	...	0	0	...	0	4.757499
4	4	0	12	2000	17	1	0	...	0	0	63015	1	1	...	0	0	...	0	4.604990
5	5	6	1	2001	18	1	0	...	0	0	63015	1	1	...	0	0	...	0	4.100613
6	3	0	12	2000	17	1	0	...	0	0	63015	1	1	...	0	1	...	0	4.787918
7	5	0	12	2000	17	1	0	...	0	0	63015	1	1	...	0	1	...	0	2.715198
8	5	0	12	2000	17	1	0	...	0	0	63015	1	1	...	0	0	...	0	3.789038
9	4	0	12	2000	17	1	0	...	0	0	63015	1	1	...	0	0	...	0	4.363013
10	4	0	12	2000	17	1	0	...	0	0	63015	1	1	...	0	0	...	0	4.954122
11	5	6	1	2001	18	1	0	...	0	0	63015	1	1	...	0	0	...	0	4.544611
12	4	0	12	2000	17	1	0	...	0	0	63015	1	1	...	0	0	...	0	3.205334
13	4	0	12	2000	17	1	0	...	0	0	63015	1	1	...	0	0	...	0	3.547938
14	4	0	12	2000	17	1	0	...	0	0	63015	1	1	...	0	0	...	0	3.602538
15	5	0	12	2000	17	1	0	...	0	0	63015	1	1	...	0	0	...	0	4.933145
16	4	0	12	2000	17	1	0	...	0	0	63015	1	1	...	0	0	...	0	2.667993
...																			
1000205	1	2	4	2000	22	25	0	...	0	1	44031	1	0	...	0	0	...	0	1.000000
1000206	5	2	4	2000	19	25	0	...	0	1	44031	1	0	...	0	0	...	0	3.980936
1000207	5	2	4	2000	19	25	0	...	0	1	44031	1	0	...	0	0	...	0	3.509349
1000208	4	2	4	2000	22	25	0	...	0	1	44031	1	0	...	0	0	...	0	3.616338
1000209	4	2	4	2000	22	25	0	...	0	1	44031	1	0	...	0	0	...	0	3.436217

Dependent variable

Date/time data

User data

Movie data

CF prediction

Linear Regression: Results

```
                     Estimate Std. Error t value Pr(>|t|)
(Intercept)         8.403e-01  1.097e-02  76.635  < 2e-16 ***
wday1               1.254e-02  3.208e-03   3.910 9.25e-05 ***
...
wday6               1.236e-02  3.557e-03   3.475 0.000510 ***
mon2                1.130e-02  1.015e-02   1.114 0.265465
...
mon12              -3.734e-02  7.809e-03  -4.781 1.74e-06 ***
year2001           -7.206e-02  4.712e-03 -15.294  < 2e-16 ***
year2002           -8.974e-02  6.368e-03 -14.091  < 2e-16 ***
year2003           -1.625e-01  1.694e-02  -9.597  < 2e-16 ***
hour1               9.275e-03  5.992e-03   1.548 0.121661
...
hour23              8.898e-03  5.472e-03   1.626 0.103914
AgeRange            1.967e-03  9.164e-05  21.468  < 2e-16 ***
Jobacademic        -7.570e-03  3.602e-03  -2.102 0.035575 *
...
Jobwriter          -2.685e-02  4.081e-03  -6.579 4.75e-11 ***
Male               -2.214e-02  2.195e-03 -10.090  < 2e-16 ***
Urban               1.837e-02  3.655e-03   5.025 5.04e-07 ***
Action             -5.835e-02  2.553e-03 -22.850  < 2e-16 ***
...
War                 4.015e-02  3.715e-03  10.809  < 2e-16 ***
CF                  7.741e-01  1.101e-03 703.120  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.8935 on 990126 degrees of freedom
Multiple R-squared:  0.3604,        Adjusted R-squared:  0.3604
F-statistic:  7062 on 79 and 990126 DF,  p-value: < 2.2e-16
```

- We knew we would do at least as well as CF on the training set
- In fact, we observe an improvement in the model fit
  - Increase of in-sample R^2 from 0.33 to 0.36
- What about out-of-sample performance?

Updated Prediction Results

- The ensemble model enhances predictive performance, as compared to "just" the collaborative filtering model

Model	R^2	MAE	RMSE
Collaborative Filtering Model	0.324	0.700	0.908
Ensemble Model	0.358	0.698	0.885

- Recall that Netflix's Cinematch reported an RMSE of 0.952
  - These results would have achieved a 7.04% RMSE improvement
- Ensemble models are often winners in machine learning contests
  - For instance, the winners of the Netflix Prize merged their teams and used their different models to create an ensemble model