

A student in a dark blue shirt is standing in a classroom, drawing network graphs on a chalkboard. The chalkboard is filled with several star-like and interconnected node graphs drawn in white chalk. The student is holding a piece of chalk and looking at the board. In the foreground, there is a red bucket containing a black eraser and a red-handled tool. The background shows a brick wall and a chalkboard with some faint writing.

STATS 507 Project

MovieLens—Predicting and Analysing User Ratings of Movies

Trong Dat Do and Simon Fontaine
dodat@umich.edu, simfont@umich.edu



April 17th, 2020

Contents

- 1 Context
- 2 Datasets description
- 3 Predictive models
- 4 Results
- 5 Exploratory analysis
- 6 References

A close-up, front-facing view of Darth Vader's helmeted head and shoulders. He is in a meditative pose with his hands clasped in front of his chest. The lighting is dramatic, highlighting the metallic textures of his armor and the iconic breathing apparatus. The background is a dark, gradient grey.

MOTIVATION

Prediction

- Build a **predictive** model for user ratings
- Inspired from the **Netflix prize** (Bennett, Lanning, et al. 2007)
- Incorporate **external information**
- Actionable: could be part of a **recommender system**

Exploratory analysis

- Analyze how user and movie information **relate** to model predictions
- **Cluster** users and movies and identify relationships with ratings
- Actionable: **improve** production and marketing **decisions**

A still from the movie 'Boyz n the City' featuring Laurence Fishburne and Ice Cube. Both are wearing dark suits, white shirts, and dark ties. They are standing in a room with wood-paneled walls. The text 'MOVIELENS DATASETS' is overlaid in the center in a bold, blue, sans-serif font.

MOVIELENS DATASETS

MovieLens Datasets

User ratings

Harper and Konstan 2015

- Collected in 1997-1998
- 100K user ratings of movies (94692 after tidying)
- Ratings 1-5
- 1682 movies (935 after tidying)
- 943 users
- Extra information ...

	user_id	movie_id	rating
0	196	242	3
1	186	302	3
2	22	377	1
3	244	51	2
4	166	346	1
...
99995	880	476	3
99996	716	204	5
99997	276	1090	1
99998	13	225	2
99999	12	203	3

MovieLens Datasets

Movie information—Movie genres

movie_id	movie_title	Genre(19)		
		Action	...	Western
1	Toy Story (1995)	0	...	0
2	GoldenEye (1995)	1	...	0
3	Four Rooms (1995)	0	...	0
4	Get Shorty (1995)	1	...	0
5	Copycat (1995)	0	...	0
...
1678	Mat' i syn (1997)	0	...	0
1679	B. Monkey (1998)	0	...	0
1680	Sliding Doors (1998)	0	...	0
1681	You So Crazy (1994)	0	...	0
1682	Scream of Stone (Schrei aus Stein) (1991)	0	...	0

MovieLens Datasets

Movie information—Movie tags

movie_id	tag_id	tag_relevance
1	0	0.032
	1	0.035
	2	0.070
	3	0.114
	4	0.105
...
108932	1123	0.327
	1124	0.030
	1125	0.006
	1126	0.161
	1127	0.028

tag_id	tag
0	007
1	007 (series)
2	18th century
3	1920s
4	1930s
...	...
1123	writing
1124	wuxia
1125	wwii
1126	zombie
1127	zombies

MovieLens Datasets

User information

user_id	age	gender	occupation
1	24	M	technician
2	53	F	other
3	23	M	writer
4	24	M	technician
5	33	F	other
...
939	26	F	student
940	32	M	administrator
941	20	M	student
942	48	F	librarian
943	22	M	student



PREDICTIVE MODELS

Predictive models

K -Nearest-Neighbors

- Euclidean distance between features
- Standardize features
- Weight groups of features

$$\mathbf{x} = (\mathbf{x}_{\text{genres}}, \mathbf{x}_{\text{tags}}, \mathbf{x}_{\text{user}})$$

$$\tilde{\mathbf{x}} := (\mathbf{x}_{\text{genres}}, \alpha_{\text{tags}}\mathbf{x}_{\text{tags}}, \alpha_{\text{user}}\mathbf{x}_{\text{user}})$$

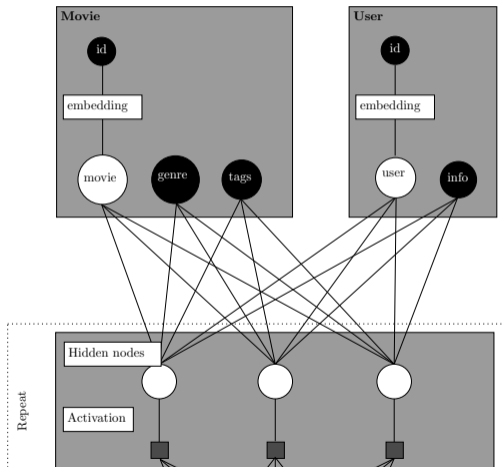
- Aggregate neighboring ratings
 - Average (regression)
 - Majority vote (classification)

Tuning

- Number of neighbors K
- Aggregation function
- Feature weights

Predictive models

Neural Networks



Tuning

- Size of user and movie embeddings
- Whether to use features
- Number and size of hidden layers
- Response transformation and loss function
- Weight decay parameter

Predictive models

Matrix completion using SVD

user_id	movie_id	Ratings
1	1	5
1	2	4
1	5	2
2	2	3
2	5	5
3	1	4
3	3	2
3	4	3
4	1	5
4	5	2

Ratings dataframe



	Movie_id				
user_id	5	4	?	?	2
?	3	?	?	5	
4	?	2	3	?	
5	?	?	?	2	

Pivotal matrix

Philosophy: Lots of people have same taste in movies

Therefore, the pivotal matrix is **Low-rank**
Use **SVD** to approximate ?

Very famous recommendation algorithm

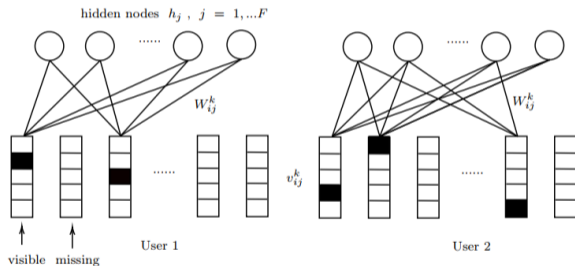
- Bennett, Lanning, et al. 2007: Netflix prize
- eHarmony: Dating website

Tuning

- Truncated dimension of SVD
- Number of iterations

Predictive models

Restricted Boltzmann Machine



- Neural network with hidden input layer.
- Implemented from Salakhutdinov, Mnih, and Hinton 2007

Tuning

- Number of hidden nodes
- Number of Gibbs sampling iterations
- Learning rate

A man in a light-colored suit and a blue and white checkered shirt is sitting on a wooden park bench. To his right is a brown suitcase with a white gift box on top, tied with a red and white ribbon. The background shows a park setting with trees and a stone wall.

RESULTS

K-NN

- + K around 50
- + **Averaging ratings**
- + $\alpha_{\text{tags}} \approx 0.5, \alpha_{\text{user}} \approx 200$

NN

- + **Regression transformations**
- + Embedding: user(128), movies(128)
- + Two layers (1024, 128), ReLU
- + Using weight decay
- + **Using features**

SVD

- **Truncated dimension: 4**
- Number of iterations around 100

RBM

- Number of hidden nodes around 15
- Gibbs sampling: 1 (good enough and fast)
- Learning rate 1.0

- The best model of each four methods:

Model	CV MSE	CV Acc.	Test MSE	Test Acc.
<i>K</i> -NN	0.9898	0.3741	0.9737	0.3767
NN	0.9474	0.3886	0.9092	0.3968
SVD	0.9533	0.4116	0.8810	0.4333
RBM	1.0362	0.3609	1.0308	0.3682

- Matrix completion (SVD) exhibits great generalization performance both in terms of MSE and prediction accuracy
- Fast, simple (almost no tuning) and interpretable model

A cinematic scene featuring two Vikings in the foreground, both with blue face paint. The Viking on the left has a long beard and is wearing a leather tunic with a curved blade. The Viking on the right has long hair and is also in a leather tunic. In the background, a large group of Vikings is visible, many holding spears. The scene is set outdoors under a clear sky.

EXPLORATORY ANALYSIS

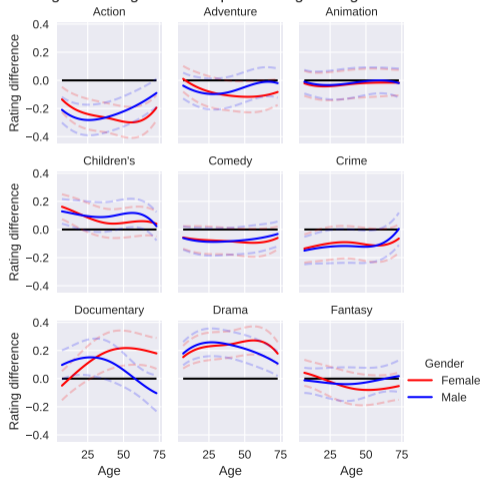
Exploratory analysis

Correlating predictions with features

Using the Matrix completion (SVD) model:

- Make every user rate every movie
- Compute difference between prediction and the average prediction per user
- Subset by movie genre
- Fit a GAM on age for each gender

Movie genre rating difference per user age and gender



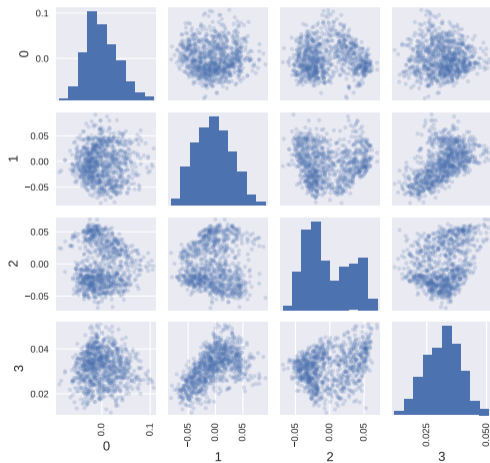
Exploratory analysis

Exploring the SVD

Interpreting U and V in the decomposition

- U and V : principle components for users and movies
- Clustering patterns are visible when plotting U and V
- We can use clustering methods to get the clusters and see how do they correlate to the features

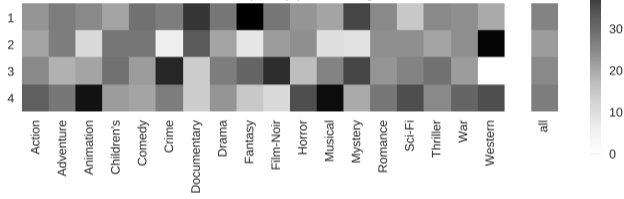
Movie clusters



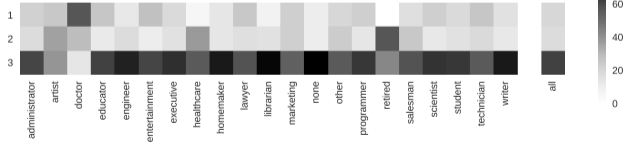
Exploratory analysis

Bi-clustering

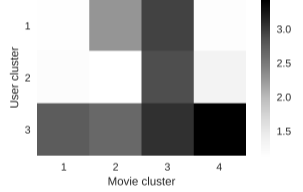
Cluster frequency per movie genre



Cluster frequency per user occupation



Ratings for user-movie cluster combinations





THANK YOU!

References

- James Bennett, Stan Lanning, et al. (2007). “The netflix prize”. In: *Proceedings of KDD cup and workshop*. Vol. 2007. Citeseer, p. 35.
- F. Maxwell Harper and Joseph A. Konstan (Dec. 2015). “The MovieLens Datasets: History and Context”. en. In: *ACM Transactions on Interactive Intelligent Systems* 5.4, pp. 1–19. ISSN: 21606455. DOI: 10.1145/2827872.
- Ruslan Salakhutdinov, Andriy Mnih, and Geoffrey Hinton (2007). “Restricted Boltzmann Machines for Collaborative Filtering”. In: *Proceedings of the 24th International Conference on Machine Learning*. ICML '07. Corvallis, Oregon, USA: Association for Computing Machinery, pp. 791–798. ISBN: 9781595937933. DOI: 10.1145/1273496.1273596. URL: <https://doi.org/10.1145/1273496.1273596>.

A close-up photograph of a man's face peering through a narrow vertical slit in a light-colored wooden door. The man has a wide, toothy grin, showing his teeth, and his eyes are looking slightly to the right. The lighting is bright, highlighting his skin and the texture of the wood. The word "QUESTIONS?" is overlaid in the center of the image in a bold, dark blue font.

QUESTIONS?