



SVD Application: Categorizing Movie Data

Laura Lyman Margot Gerritsen

Stanford University

ICME Summer Workshops
Fundamentals of Data Science

August 2, 2021

Simple example with movie data

Let A be a rectangular data matrix, whose

- rows correspond to users,
- columns correspond to movies, and
- entries are ratings that users have given movies.

Specifically, $A_{ij} = \text{rating user } i \text{ has given movie } j$. For example,

$$A = \begin{bmatrix} & \text{The Matrix} & \text{Alien} & \text{Serenity} & \text{Casablanca} & \text{Amelie} \\ 1 & 1 & 1 & 0 & 0 & 0 \\ 2 & 2 & 2 & 0 & 0 & 0 \\ 1 & 1 & 1 & 0 & 0 & 0 \\ 5 & 5 & 5 & 0 & 0 & 0 \\ 0 & 0 & 0 & 2 & 2 & 2 \\ 0 & 0 & 0 & 3 & 3 & 3 \\ 0 & 0 & 0 & 1 & 1 & 1 \end{bmatrix}.$$

Simple example with movie data

Let A be a rectangular data matrix, whose

- rows correspond to users,
- columns correspond to movies, and
- entries are ratings that users have given movies.

Specifically, $A_{ij} = \text{rating user } i \text{ has given movie } j$. For example,

User 3 rated "Alien" 1 out of 5

$$A = \begin{bmatrix} & \text{The Matrix} & \text{Alien} & \text{Serenity} & \text{Casablanca} & \text{Amelie} \\ 1 & 1 & 1 & 0 & 0 & 0 \\ 2 & 2 & 2 & 0 & 0 & 0 \\ 1 & 1 & 1 & 0 & 0 & 0 \\ 5 & 5 & 5 & 0 & 0 & 0 \\ 0 & 0 & 0 & 2 & 2 & 0 \\ 0 & 0 & 0 & 3 & 3 & 0 \\ 0 & 0 & 0 & 1 & 1 & 0 \end{bmatrix}$$

User 5 saw Casablanca and Amelie and rated them 2/5

Simple example with movie data

Let A be a rectangular data matrix, whose

- rows correspond to users,
- columns correspond to movies, and
- entries are ratings that users have given movies.

Specifically, $A_{ij} = \text{rating user } i \text{ has given movie } j$. For example,

$A_{ij} = 0 \iff$
user i has not
rated movie j

$$A = \begin{bmatrix} & \text{The Matrix} & \text{Alien} & \text{Serenity} & \text{Casablanca} & \text{Amelie} \\ 1 & 1 & 1 & 0 & 0 & 0 \\ 2 & 2 & 2 & 0 & 0 & 0 \\ 1 & 1 & 1 & 0 & 0 & 0 \\ 5 & 5 & 5 & 0 & 0 & 0 \\ 0 & 0 & 0 & 2 & 2 & 0 \\ 0 & 0 & 0 & 3 & 3 & 0 \\ 0 & 0 & 0 & 1 & 1 & 0 \end{bmatrix} \quad \left\{ \begin{array}{l} \text{e.g. first 4 users} \\ \text{didn't rate} \\ \text{Casablanca or} \\ \text{Amelie} \end{array} \right. \quad \left\{ \begin{array}{l} \text{last 3 users} \\ \text{only} \end{array} \right.$$

Simple example with movie data

Let A be a rectangular data matrix, whose

- rows correspond to users,
- columns correspond to movies, and
- entries are ratings that users have given movies.

Specifically, $A_{ij} =$ rating user i has given movie j . For example,

	The Matrix	Alien	Serenity	Casablanca	Amelie
1	1	1	1	0	0
2	2	2	2	0	0
1	1	1	1	0	0
5	5	5	5	0	0
	<hr/>				
0	0	0	0	2	2
0	0	0	0	3	3
0	0	0	0	1	1

Sci Fi category {

ROMANCE ♥ category }

SVD for movie data

Using the compact SVD,

$$\left[\begin{array}{ccc|cc} 1 & 1 & 1 & 0 & 0 \\ 2 & 2 & 2 & 0 & 0 \\ 1 & 1 & 1 & 0 & 0 \\ 5 & 5 & 5 & 0 & 0 \\ \hline 0 & 0 & 0 & 2 & 2 \\ 0 & 0 & 0 & 3 & 3 \\ 0 & 0 & 0 & 1 & 1 \end{array} \right] = \underbrace{\begin{bmatrix} 0.18 & 0 \\ 0.36 & 0 \\ 0.18 & 0 \\ 0.90 & 0 \\ 0 & 0.53 \\ 0 & 0.80 \\ 0 & 0.27 \end{bmatrix}}_U \times \underbrace{\begin{bmatrix} 9.64 & 0 \\ 0 & 5.29 \end{bmatrix}}_{\Sigma} \times \underbrace{\begin{bmatrix} 0.58 & 0 \\ 0.58 & 0 \\ 0.58 & 0 \\ 0 & 0.71 \\ 0 & 0.71 \end{bmatrix}}_{V^T}^T.$$

Romance SCI FI

- Σ : # of categories and their relative strength in the data set

SVD for movie data

Using the compact SVD,

if user 3
rates
higher

Romance	Sci Fi	1	1	1	0	0
		2	2	2	0	0
Romance	Sci Fi	1	1	1	0	0
		5	5	5	0	0
Romance	Sci Fi	0	0	0	2	2
		0	0	0	3	3
Romance	Sci Fi	0	0	0	1	1

1	1	1	0	0
2	2	2	0	0
4	4	4	0	0
5	5	5	0	0
0	0	0	2	2
0	0	0	3	3
0	0	0	1	1

$$\Sigma = \begin{bmatrix} 11.75 & 0 \\ 0 & 5.29 \end{bmatrix}$$

- Σ : # of categories and their relative strength in the data set

SVD for movie data

Using the compact SVD,

Romance			Sci Fi	
1	1	1	0	0
2	2	2	0	0
<u>1</u>	<u>1</u>	<u>1</u>	0	0
5	5	5	0	0
0	0	0	<u>2</u>	<u>2</u>
0	0	0	3	3
0	0	0	1	1

if user 3
rates
higher

1	1	1	0	0
2	2	2	0	0
4	4	4	0	0
5	5	5	0	0
0	0	0	2	2
0	0	0	3	3
0	0	0	1	1

if user 5
changes to
Sci Fi

1	1	1	0	0
2	2	2	0	0
1	1	1	0	0
5	5	5	0	0
2	2	2	0	0
0	0	0	3	3
0	0	0	1	1

$$\Sigma = \begin{bmatrix} 11.75 & 0 \\ 0 & 5.29 \end{bmatrix}$$

increased

$$\Sigma = \begin{bmatrix} 10.25 & 0 \\ 0 & 4.47 \end{bmatrix}$$

increased

decreased

- Σ : # of categories and their relative strength in the data set

SVD for movie data

Using the compact SVD,

$$\left[\begin{array}{ccc|cc} 1 & 1 & 1 & 0 & 0 \\ 2 & 2 & 2 & 0 & 0 \\ 1 & 1 & 1 & 0 & 0 \\ 5 & 5 & 5 & 0 & 0 \\ \hline 0 & 0 & 0 & 2 & 2 \\ 0 & 0 & 0 & 3 & 3 \\ 0 & 0 & 0 & 1 & 1 \end{array} \right] = \underbrace{\begin{bmatrix} 0.18 & 0 \\ 0.36 & 0 \\ 0.18 & 0 \\ 0.90 & 0 \\ 0 & 0.53 \\ 0 & 0.80 \\ 0 & 0.27 \end{bmatrix}}_U \times \underbrace{\begin{bmatrix} 9.64 & 0 \\ 0 & 5.29 \end{bmatrix}}_{\Sigma} \times \underbrace{\begin{bmatrix} 0.58 & 0 \\ 0.58 & 0 \\ 0.58 & 0 \\ 0 & 0.71 \\ 0 & 0.71 \end{bmatrix}}_{V^T}^T.$$

- Σ : # of categories and their relative strength in the data set
- U : “user to category (genre)” similarity matrix

SVD for movie data

Using the compact SVD,

$$\begin{bmatrix} 1 & 1 & 1 & 0 & 0 \\ 2 & 2 & 2 & 0 & 0 \\ 1 & 1 & 1 & 0 & 0 \\ \underline{5} & \underline{5} & \underline{5} & 0 & 0 \\ \hline 0 & 0 & 0 & 2 & 2 \\ 0 & 0 & 0 & 3 & 3 \\ 0 & 0 & 0 & 1 & 1 \end{bmatrix} = \underbrace{\begin{bmatrix} 0.18 & 0 \\ 0.36 & 0 \\ 0.18 & 0 \\ 0.90 & 0 \\ 0 & 0.53 \\ 0 & 0.80 \\ 0 & 0.27 \end{bmatrix}}_{U} \times \underbrace{\begin{bmatrix} 9.64 & 0 \\ 0 & 5.29 \end{bmatrix}}_{\Sigma} \times \underbrace{\begin{bmatrix} 0.58 & 0 \\ 0.58 & 0 \\ 0.58 & 0 \\ 0 & 0.71 \\ 0 & 0.71 \end{bmatrix}}_{V^T}.$$

Strongest SciFi preference (highest ratings)

- Σ : # of categories and their relative strength in the data set
- U : “user to category (genre)” similarity matrix

SVD for movie data

Using the compact SVD,

$$\begin{bmatrix} 1 & 1 & 1 & 0 & 0 \\ 2 & 2 & 2 & 0 & 0 \\ 1 & 1 & 1 & 0 & 0 \\ \underline{5} & \underline{5} & \underline{5} & 0 & 0 \\ \hline 0 & 0 & 0 & 2 & 2 \\ 0 & 0 & 0 & \underline{3} & \underline{3} \\ 0 & 0 & 0 & 1 & 1 \end{bmatrix} = \underbrace{\begin{bmatrix} 0.18 & 0 \\ 0.36 & 0 \\ 0.18 & 0 \\ 0.90 & 0 \\ 0 & 0.53 \\ 0 & 0.80 \\ 0 & 0.27 \end{bmatrix}}_U \times \underbrace{\begin{bmatrix} 9.64 & 0 \\ 0 & 5.29 \end{bmatrix}}_{\Sigma} \times \underbrace{\begin{bmatrix} 0.58 & 0 \\ 0.58 & 0 \\ 0.58 & 0 \\ 0 & 0.71 \\ 0 & 0.71 \end{bmatrix}}_{V^T}^T.$$

Annotations:

- Strongest SciFi preference (highest ratings) points to the circled value 0.90 in the matrix U .
- User 6 fits best w/ romance category (highest ratings) points to the circled value 0.80 in the matrix U .

- Σ : # of categories and their relative strength in the data set
- U : “user to category (genre)” similarity matrix

SVD for movie data

Using the compact SVD,

each user gives identical ratings

$$\left\{ \begin{array}{ccc|cc} 1 & 1 & 1 & 0 & 0 \\ 2 & 2 & 2 & 0 & 0 \\ \hline 1 & 1 & 1 & 0 & 0 \\ 5 & 5 & 5 & 0 & 0 \\ \hline 0 & 0 & 0 & 2 & 2 \\ 0 & 0 & 0 & 3 & 3 \\ 0 & 0 & 0 & 1 & 1 \end{array} \right] = \underbrace{\begin{bmatrix} 0.18 & 0 \\ 0.36 & 0 \\ 0.18 & 0 \\ 0.90 & 0 \\ 0 & 0.53 \\ 0 & 0.80 \\ 0 & 0.27 \end{bmatrix}}_U \times \underbrace{\begin{bmatrix} 9.64 & 0 \\ 0 & 5.29 \end{bmatrix}}_{\Sigma} \times \underbrace{\begin{bmatrix} 0.58 & 0 \\ 0.58 & 0 \\ 0.58 & 0 \\ 0 & 0.71 \\ 0 & 0.71 \end{bmatrix}}_{V^T}.$$

first 3 movies belong to SciFi to the same extent (and are separate from romance)

- Σ : # of categories and their relative strength in the data set
- U : “user to category (genre)” similarity matrix
- V : “category (genre) to movie” similarity matrix

SVD for movie data

Using the compact SVD,

each user gives identical ratings

$$\left\{ \begin{array}{ccc|cc} 1 & 1 & 1 & 0 & 0 \\ 2 & 2 & 2 & 0 & 0 \\ \hline 1 & 1 & 1 & 0 & 0 \\ 5 & 5 & 5 & 0 & 0 \\ \hline 0 & 0 & 0 & 2 & 2 \\ 0 & 0 & 0 & 3 & 3 \\ 0 & 0 & 0 & 1 & 1 \end{array} \right] = \underbrace{\begin{bmatrix} 0.18 & 0 \\ 0.36 & 0 \\ 0.18 & 0 \\ 0.90 & 0 \\ 0 & 0.53 \\ 0 & 0.80 \\ 0 & 0.27 \end{bmatrix}}_U \times \underbrace{\begin{bmatrix} 9.64 & 0 \\ 0 & 5.29 \end{bmatrix}}_{\Sigma} \times \underbrace{\begin{bmatrix} 0.58 & 0 \\ 0.58 & 0 \\ 0.58 & 0 \\ 0 & 0.71 \\ 0 & 0.71 \end{bmatrix}}_{V^T}^T.$$

↑ all rate Casablanca and Amelie equally

- Σ : # of categories and their relative strength in the data set
- U : “user to category (genre)” similarity matrix
- V : “category (genre) to movie” similarity matrix

Diversifying within a category

Suppose instead the second user — let's call him Pete — has non-homogenous ratings; he really prefers Alien to the other Sci Fi movies and rates it a 5.

Diversifying within a category

Suppose instead the second user — let's call him Pete — has non-homogenous ratings; he really prefers Alien to the other Sci Fi movies and rates it a 5.

- Now 3 nonzero singular values
- In compact SVD, we now need an extra column in U and V^T to span the Sci-fi columns of A (which are no longer identical)

$$\left[\begin{array}{ccc|cc} 1 & 1 & 1 & 0 & 0 \\ 2 & 5 & 2 & 0 & 0 \\ 1 & 1 & 1 & 0 & 0 \\ 5 & 5 & 5 & 0 & 0 \\ \hline 0 & 0 & 0 & 2 & 2 \\ 0 & 0 & 0 & 3 & 3 \\ 0 & 0 & 0 & 1 & 1 \end{array} \right] = \underbrace{\begin{bmatrix} \mathbf{0.16} & 0 & \mathbf{0.10} \\ \mathbf{0.52} & 0 & -\mathbf{0.85} \\ \mathbf{0.16} & 0 & \mathbf{0.10} \\ \mathbf{0.82} & 0 & \mathbf{0.50} \\ \hline 0 & 0.53 & 0 \\ 0 & 0.80 & 0 \\ 0 & 0.27 & 0 \end{bmatrix}}_U \underbrace{\begin{bmatrix} \mathbf{10.47} & 0 & 0 \\ 0 & 5.29 & 0 \\ 0 & 0 & \mathbf{2.11} \end{bmatrix}}_{\Sigma} \underbrace{\begin{bmatrix} \mathbf{0.52} & \mathbf{0.67} & \mathbf{0.52} & 0 & 0 \\ 0 & 0 & 0 & 0.71 & 0.71 \\ \mathbf{0.48} & -\mathbf{0.74} & \mathbf{0.48} & 0 & 0 \end{bmatrix}}_{V^T}$$

Diversifying within a category

Suppose instead the second user — let's call him Pete — has non-homogenous ratings; he really prefers Alien to the other Sci Fi movies and rates it a 5.

- Now 3 nonzero singular values
- In compact SVD, we now need an extra column in U and V^T to span the Sci-fi columns of A (which are no longer identical)

$$\left[\begin{array}{ccc|cc} 1 & 1 & 1 & 0 & 0 \\ 2 & 5 & 2 & 0 & 0 \\ 1 & 1 & 1 & 0 & 0 \\ 5 & 5 & 5 & 0 & 0 \\ \hline 0 & 0 & 0 & 2 & 2 \\ 0 & 0 & 0 & 3 & 3 \\ 0 & 0 & 0 & 1 & 1 \end{array} \right] = \underbrace{\begin{bmatrix} \mathbf{0.16} & 0 & \mathbf{0.10} \\ \mathbf{0.52} & 0 & -\mathbf{0.85} \\ \mathbf{0.16} & 0 & \mathbf{0.10} \\ \mathbf{0.82} & 0 & \mathbf{0.50} \\ \hline 0 & \mathbf{0.53} & 0 \\ 0 & \mathbf{0.80} & 0 \\ 0 & \mathbf{0.27} & 0 \end{bmatrix}}_U \underbrace{\begin{bmatrix} \mathbf{10.47} & 0 & 0 \\ 0 & \mathbf{5.29} & 0 \\ 0 & 0 & \mathbf{2.11} \end{bmatrix}}_{\Sigma} \underbrace{\begin{bmatrix} \mathbf{0.52} & \mathbf{0.67} & \mathbf{0.52} & 0 & 0 \\ 0 & 0 & 0 & \mathbf{0.71} & \mathbf{0.71} \\ \mathbf{0.48} & -\mathbf{0.74} & \mathbf{0.48} & 0 & 0 \end{bmatrix}}_{V^T}$$

 = unchanged values (romance category)

Mixing categories

Instead of updating ratings within a category, what if our second user (Pete) is interested in Sci-fi *and* romance? In fact, he *loves* Amelie.

Mixing categories

Instead of updating ratings within a category, what if our second user (Pete) is interested in Sci-fi *and* romance? In fact, he *loves* Amelie.

- Now there is “fill-in” where there were previously 0s when diversified ratings still fell into separate categories

Mixing categories

Instead of updating ratings within a category, what if our second user (Pete) is interested in Sci-fi *and* romance? In fact, he *loves* Amelie.

- Now there is “fill-in” where there were previously 0s when diversified ratings still fell into separate categories
- The magnitude of the “fill-in” corresponds to the strength of the new connection i.e. the mixed Sci-fi-romance grouping

Mixing categories

Instead of updating ratings within a category, what if our second user (Pete) is interested in Sci-fi *and* romance? In fact, he *loves* Amelie.

- Now there is “fill-in” where there were previously 0s when diversified ratings still fell into separate categories
- The magnitude of the “fill-in” corresponds to the strength of the new connection i.e. the mixed Sci-fi-romance grouping
- A weaker connection (say $A_{2,5} = 2$ instead of 4) gives a smaller corresponding Σ_{33} , along with fill-in entries smaller in magnitude

$$\begin{bmatrix} 1 & 1 & 1 & 0 & 0 \\ 2 & 2 & 2 & 0 & \textcolor{blue}{4} \\ 1 & 1 & 1 & 0 & 0 \\ 5 & 5 & 5 & 0 & 0 \\ 0 & 0 & 0 & 2 & 2 \\ 0 & 0 & 0 & 3 & 3 \\ 0 & 0 & 0 & 1 & 1 \end{bmatrix} = \begin{bmatrix} 0.17 & \textcolor{red}{0.06} & -0.06 \\ 0.43 & \textcolor{red}{-0.44} & 0.79 \\ 0.17 & \textcolor{red}{0.06} & -0.06 \\ 0.86 & \textcolor{red}{0.29} & -0.31 \\ \textcolor{red}{0.05} & -0.45 & \textcolor{red}{-0.28} \\ \textcolor{red}{0.08} & -0.68 & \textcolor{red}{-0.42} \\ \textcolor{red}{0.03} & -0.23 & \textcolor{red}{-0.14} \end{bmatrix} \begin{bmatrix} 9.80 & 0 & 0 \\ 0 & 5.97 & 0 \\ 0 & 0 & 2.30 \end{bmatrix}^T$$

= fill in
from mixing
categories