Recommender Systems and Dimensionality Reduction

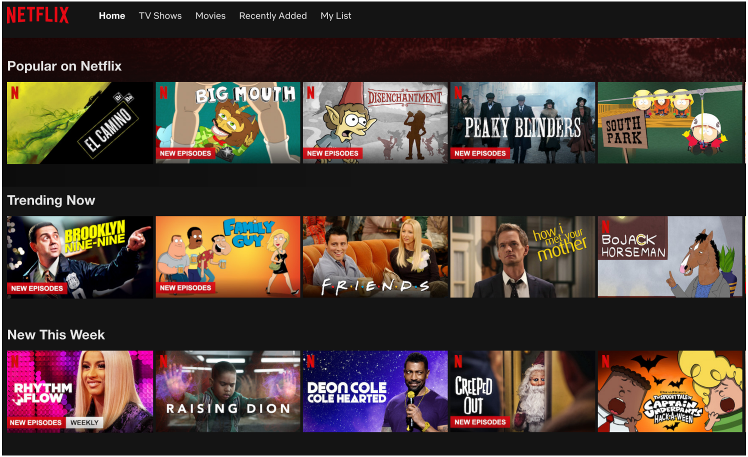
Sibylle Hess

Technische Universiteit Eindhoven
University of Technology

1

Informal Problem Description

Recommending Movies like Netflix does



Who Would You Recommend What and Why?

	Star Wars	Interstellar	Blade Runner	Tron	2001: Space O.	Mars Attacks	Dune	Matrix	Robo Cop	Aliens	Terminator	Solaris	Avatar	12 Monkeys
Grace	🤩	🤩		🙈			🤩		🤩	🤩	🤩		🤩	
Carol		🤩	🤩	🤩				🤩				🤩		🤩
Alice	🤩	🤩	🤩	🤩	🤩							🤩	🙈	
Bob	🤩					🤩		🤩	🤩	🤩	🤩		🤩	
Eve	🤩				🙈	🤩		🤩	🤩				🤩	
Chuck	🤩	🤩		🤩	🤩	🤩	🤩	🤩		🤩	🤩		🙈	🤩

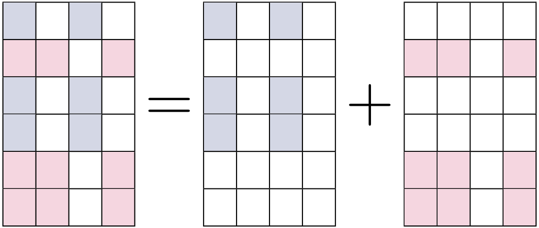
🤩: Yeeey 🙈: Naaay

Who Would You Recommend What and Why?

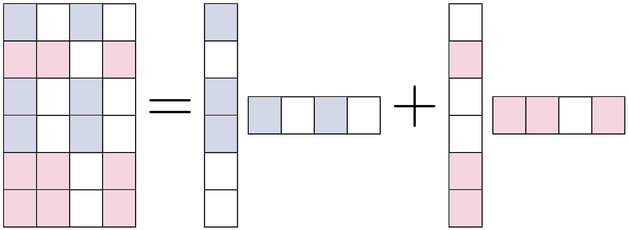
	Star Wars	Interstellar	Blade Runner	Tron	2001: Space O.	Mars Attacks	Dune	Matrix	Robo Cop	Aliens	Terminator	Solaris	Avatar	12 Monkeys
Grace	🤩	🤩		🙈			🤩		🤩	🤩	🤩		🤩	
Carol		🤩	🤩	🤩				🤩				🤩		🤩
Alice	🤩	🤩	🤩	🤩	🤩							🤩	🙈	
Bob	🤩					🤩		🤩	🤩	🤩	🤩		🤩	
Eve	🤩				🙈	🤩		🤩	🤩				🤩	
Chuck	🤩	🤩		🤩	🤩	🤩	🤩	🤩		🤩	🤩		🙈	🤩

🤩: Yeeey 🙈: Naaay

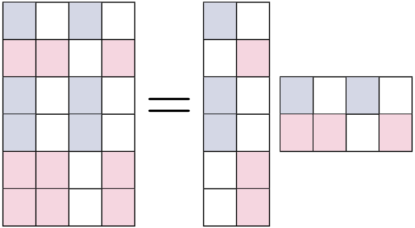
What is this Color Scheme in Math?



What is this Color Scheme in Math?



What is this Color Scheme in Math? A Matrix Product!



2

Derive the Formal Problem Definition

The Rank-r Matrix Factorization Problem

Given: a data matrix $D \in \mathbb{R}^{n \times d}$ and a rank $r < \min\{n, d\}$.

Find: matrices $X \in \mathbb{R}^{d \times r}$ and $Y \in \mathbb{R}^{n \times r}$ whose product approximates the data matrix:

$$\min_{X,Y} \|D - YX^\top\|^2 \qquad \text{s.t. } X \in \mathbb{R}^{d \times r}, Y \in \mathbb{R}^{n \times r}$$

The Rank-r MF Problem is Nonconvex

Theorem (MF is Nonconvex)

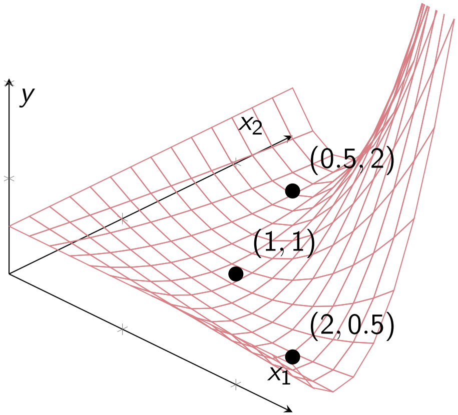
The rank-r matrix factorization problem, defined for a matrix $D \in \mathbb{R}^{n\times d} \neq 0$ and a rank $1 \leq r < \min\{n, d\}$ as

$$\min_{X,Y} RSS(X, Y) = \|D - YX^\top\|^2 \quad s.t.\ X \in \mathbb{R}^{d\times r}, Y \in \mathbb{R}^{n\times r}$$

is a nonconvex optimization problem.

Proof: follows from the fact that the set of global minimizers is not a convex set.

Example: One-dimensional Matrix Factorization



$$f(x_1, x_2) = (1 - x_1 x_2)^2$$

The rank-r MF problem is nonconvex. Does that mean that we can only determine local minimizers?

No, the global minimum is given by truncated SVD.

3 Optimization

Singular Value Decomposition

Theorem (SVD)

For every matrix $D \in \mathbb{R}^{n \times d}$ *there exist orthogonal matrices* $U \in \mathbb{R}^{n \times n}$, $V \in \mathbb{R}^{d \times d}$ *and* $\Sigma \in \mathbb{R}^{n \times d}$ *such that*

$$D = U \Sigma V^\top, \text{ where}$$

- $U^\top U = UU^\top = I_n, V^\top V = VV^\top = I_d$
- Σ *is a rectangular diagonal matrix,* $\Sigma_{11} \geq \ldots \geq \Sigma_{ll}$ *where* $l = \min\{n, d\}$

The column vectors $U_{\cdot s}$ and $V_{\cdot s}$ are called left and right singular vectors and the values $\sigma_i = \Sigma_{ii}$ are called singular values $(1 \leq i \leq l)$.

Solutions to the Rank-r Matrix Factorization Problem

Theorem (Truncated SVD)

Let $D = U\Sigma V^\top \in \mathbb{R}^{n\times d}$ be the singular decomposition of D. Then the global minimizers X and Y of the rank-r MF problem

$$\min_{X,Y} \|D - YX^\top\|^2 s.t.\ X \in \mathbb{R}^{d\times r}, Y \in \mathbb{R}^{n\times r}.$$

satisfy

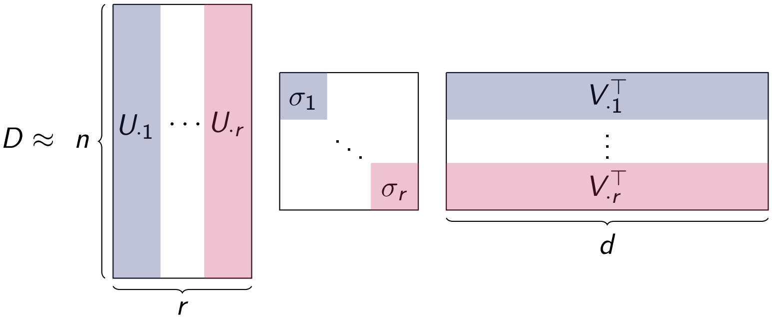
$$YX^\top = U_{\cdot\mathcal{R}}\Sigma_{\mathcal{R}\mathcal{R}}V_{\cdot\mathcal{R}}^\top,\ \text{where } \mathcal{R} = \{1,\ldots,r\}.$$

The proof follows from the orthogonal invariance of the Frobenius norm, yielding:

$$\min_{X,Y} \|D - YX^\top\|^2 = \|\Sigma - U^\top YX^\top V\|^2$$

Truncated SVD

The approximation $D \approx U_{\cdot\mathcal{R}}\Sigma_{\mathcal{R}\mathcal{R}}V_{\cdot\mathcal{R}}^{\top}$ is called truncated SVD.



Ok, so the truncated SVD solves the task to determine a low-rank approximation of my data.

How can we apply the low-rank approximation to provide recommendations?

Fill missing values with the mean value and compute the truncated SVD.

Matrix Completion for Recommender Systems

Movies

Users		A	B	C	D
	1	★★★★★	?	★★☆☆☆	★☆☆☆☆
	2	?	★☆☆☆☆	★★★★★	?
	3	★★★★★	★☆☆☆☆	★★★★★	★★☆☆☆
	4	★★★★★	?	★★★★★	★★★☆☆
	5	★★★★★	★★★★★	?	?
	6	?	★★★★☆	★★★★★	★★★☆☆

Can we fill the ? with the rating which would be given by the user if (s)he had seen the movie?

Matrix Completion by SVD

Quick hack: replace the ? with the mean rating $\mu = 3$.

Movies

Users		A	B	C	D
	1	5	μ	2	1
	2	μ	1	5	μ
	3	5	1	5	2
	4	5	μ	5	3
	5	5	5	μ	μ
	6	μ	4	5	3

The Low-Rank Matrix Approximation Provides Recommendations

$$\begin{pmatrix} 5 & \mu & 1 & 1 \\ \mu & 1 & 5 & \mu \\ 2 & 1 & 5 & 3 \\ 4 & \mu & 4 & 2 \\ 5 & 5 & \mu & 1 \\ \mu & 1 & 5 & 3 \end{pmatrix} \approx \begin{pmatrix} 4.3 & 3.7 & 1.4 & 0.6 \\ 2.8 & 1.2 & 5.1 & 3.0 \\ 2.2 & 0.7 & 5.0 & 2.9 \\ 4.2 & 2.8 & 3.9 & 2.1 \\ 5.5 & 4.5 & 2.7 & 1.3 \\ 2.8 & 1.2 & 5.1 & 3.0 \end{pmatrix}$$

$$= \begin{pmatrix} -0.3 & 0.5 \\ -0.4 & -0.4 \\ -0.4 & -0.4 \\ -0.4 & 0.1 \\ -0.5 & 0.5 \\ -0.4 & -0.4 \end{pmatrix} \begin{pmatrix} -9.0 & -5.8 & -9.5 & -5.3 \\ 2.6 & 3.3 & -3.3 & -2.2 \end{pmatrix}$$

Interpretation of MF for Recommender Systems

$$\begin{pmatrix} 5 & \mu & 1 & 1 \\ \mu & 1 & 5 & \mu \\ 2 & 1 & 5 & 3 \\ 4 & \mu & 4 & 2 \\ 5 & 5 & \mu & 1 \\ \mu & 1 & 5 & 3 \end{pmatrix} \approx \begin{pmatrix} -0.3 & 0.5 \\ -0.4 & -0.4 \\ -0.4 & -0.4 \\ -0.4 & 0.1 \\ -0.5 & 0.5 \\ -0.4 & -0.4 \end{pmatrix} \begin{pmatrix} -9.0 & -5.8 & -9.5 & -5.3 \\ 2.6 & 3.3 & -3.3 & -2.2 \end{pmatrix}$$

Every user's preferences are approximated by a linear combination of the rows in the second matrix:

$$\begin{aligned} \begin{pmatrix} 5 & \mu & 1 & 1 \end{pmatrix} \approx & -0.3 \cdot \begin{pmatrix} -9.0 & -5.8 & -9.5 & -5.3 \end{pmatrix} \\ & + 0.5 \cdot \begin{pmatrix} 2.6 & 3.3 & -3.3 & -2.2 \end{pmatrix} \end{aligned}$$

Matrix Completion by SVD

$$
\begin{pmatrix}
5 & \mu & 1 & 1 \\
\mu & 1 & 5 & \mu \\
2 & 1 & 5 & 3 \\
4 & \mu & 4 & 2 \\
5 & 5 & \mu & 1 \\
\mu & 1 & 5 & 3
\end{pmatrix}
\approx
\begin{pmatrix}
4.3 & 3.7 & 1.4 & 0.6 \\
2.8 & 1.2 & 5.1 & 3.0 \\
2.2 & 0.7 & 5.0 & 2.9 \\
4.2 & 2.8 & 3.9 & 2.1 \\
5.5 & 4.5 & 2.7 & 1.3 \\
2.8 & 1.2 & 5.1 & 3.0
\end{pmatrix}
$$

$$
=
\begin{pmatrix}
-0.3 & 0.5 \\
-0.4 & -0.4 \\
-0.4 & -0.4 \\
-0.4 & 0.1 \\
-0.5 & 0.5 \\
-0.4 & -0.4
\end{pmatrix}
\begin{pmatrix}
-9.0 & -5.8 & -9.5 & -5.3 \\
2.6 & 3.3 & -3.3 & -2.2
\end{pmatrix}
$$

Question: What happens if observations are sparse?

How can we prevent the approximation to the inserted mean values?

Adapt the objective to approximate only observed entries.

Making 3rd place in the Netflix Price 2009

Given: a data matrix $D \in \mathbb{R}^{n \times d}$ having observed entries D_{ik} for $(i,k) \in \mathcal{O} \subseteq \{1,\ldots,n\} \times \{1,\ldots d\}$ the set of observed matrix entries, and a rank $r < \min\{n,d\}$.

Find: matrices $X \in \mathbb{R}^{d \times r}$ and $Y \in \mathbb{R}^{n \times r}$ whose product approximates the data matrix only on observed entries, indicated by $\mathbb{1}_{\mathcal{O}}$:

$$\min_{X,Y} \|\mathbb{1}_{\mathcal{O}} \circ (D - YX^\top)\|^2 = \sum_{(i,k)\in\mathcal{O}} (D_{ik} - Y_{i\cdot}X_{k\cdot}^\top)^2$$

$$\text{s.t. } X \in \mathbb{R}^{d \times r}, Y \in \mathbb{R}^{n \times r}$$

Optimization: Coordinate Descent

Truncated SVD solves the

Rank-r Matrix Factorization Problem

Now something different:

Finding low-dimensional representations of the data by truncated SVD

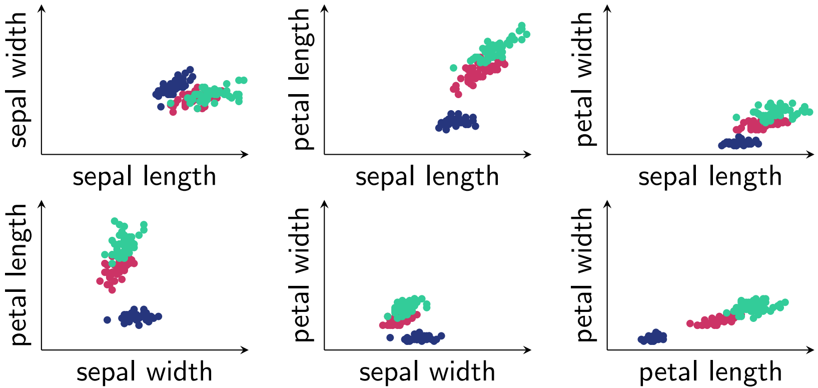
1

Informal Problem Description

Exploring the Iris Dataset

sepal length	sepal width	petal length	petal width	class
5.1	3.5	1.4	0.2	setosa
6.4	3.5	4.5	1.2	versicolor
5.9	3.0	5.0	1.8	virginica
⋮	⋮	⋮	⋮	⋮

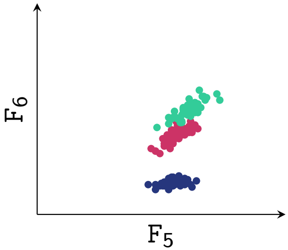
The First Step of Data Analysis: Visualization



Which views are good?

We can also Generate our Own Features

$$F_5 = F_1 + F_2$$
$$F_6 = F_3 + F_4$$



- F_1: sepal length
- F_2: sepal width
- F_3: petal length
- F_4: petal width

How do we find good low-dimensional views on our data? How to create good new features?

Find the linear combination of features with highest variance.

2

Derive the Formal Problem Definition

Defining a new Feature by a Linear Combination

Given the $n \times d$ data matrix D gathering n observations of d features $\mathtt{F}_1, \ldots, \mathtt{F}_d$, we define a new feature:

$$\mathtt{F}_{d+1} = \sum_{k=1}^{d} \alpha_k \mathtt{F}_k.$$

We have n observations of this new feature, given by

$$D_{\cdot d+1} = \sum_{k=1}^{d} \alpha_k D_{\cdot k} = D\boldsymbol{\alpha} \in \mathbb{R}^n$$

The Sample Mean of the new Feature

Given observations $D_{\cdot d+1} = D\boldsymbol{\alpha}$ of the new feature $\mathtt{F}_{d+1} = \sum_{k=1}^{d} \alpha_k \mathtt{F}_k$, we compute the sample mean as

$$\mu_{\mathtt{F}_{d+1}} = \frac{1}{n}\sum_{i=1}^{n} D_{id+1} = \boldsymbol{\mu}_{\mathtt{F}}^\top \alpha, \qquad \text{where } \boldsymbol{\mu}_{\mathtt{F}} = \begin{pmatrix} \mu_{\mathtt{F}_1} \\ \vdots \\ \mu_{\mathtt{F}_d} \end{pmatrix}$$

is the vector gathering all sample means for the d features.

The Sample Variance of the new Feature

Given observations $D_{\cdot d+1} = D\boldsymbol{\alpha}$ of the new feature

$$\mathtt{F}_{d+1} = \sum_{k=1}^{d} \alpha_k \mathtt{F}_k, \qquad \text{with sample mean} \qquad \mu_{\mathtt{F}_{d+1}} = \boldsymbol{\mu}_{\mathtt{F}}^\top \boldsymbol{\alpha},$$

we compute the sample variance as

$$\sigma^2_{\mathtt{F}_{d+1}} = \frac{1}{n}\sum_{i=1}^{n}(D_{id+1} - \mu_{\mathtt{F}_{d+1}})^2 = \frac{1}{n}\left\| \left(D - \mathbf{1}\boldsymbol{\mu}_{\mathtt{F}}^\top\right)\boldsymbol{\alpha} \right\|^2$$

Sample Statistics of the new Feature

Given observations $D_{\cdot d+1} = D\boldsymbol{\alpha}$ of the new feature

$$\mathtt{F}_{d+1} = \sum_{k=1}^{d} \alpha_k \mathtt{F}_k,$$

the sample mean and variance is given by

$$\mu_{\mathtt{F}_{d+1}} = \boldsymbol{\mu}_{\mathtt{F}}^{\top}\boldsymbol{\alpha}, \qquad \sigma^2_{\mathtt{F}_{d+1}} = \frac{1}{n}\left\|\left(D - \mathbf{1}\boldsymbol{\mu}_{\mathtt{F}}^{\top}\right)\boldsymbol{\alpha}\right\|^2.$$

We are interested in the direction of maximal variance, so we can restrict the length of vector $\boldsymbol{\alpha}$: $\|\boldsymbol{\alpha}\| = 1$

Finding the Direction of Maximal Sample Variance

The direction of largest variance $\boldsymbol{\alpha}$ is the solution to the following optimization problem:

$$\begin{aligned}
\max_{\|\boldsymbol{\alpha}\|=1} \sigma_{d+1}^2 &= \max_{\|\boldsymbol{\alpha}\|=1} \frac{1}{n} \left\| \left(D - 1\boldsymbol{\mu}_{\mathrm{F}}^\top \right) \boldsymbol{\alpha} \right\|^2 \\
&= \max_{\|\boldsymbol{\alpha}\|=1} \frac{1}{n} \boldsymbol{\alpha}^\top \left(D - 1\boldsymbol{\mu}_{\mathrm{F}}^\top \right)^\top \left(D - 1\boldsymbol{\mu}_{\mathrm{F}}^\top \right) \boldsymbol{\alpha} \\
&= \max_{\|\boldsymbol{\alpha}\|=1} \frac{\boldsymbol{\alpha}^\top C^\top C \boldsymbol{\alpha}}{n},
\end{aligned}$$

where $C = D - 1\boldsymbol{\mu}_{\mathrm{F}}^\top$ is the centered data matrix.

So, the direction of largest variance is given by the operator norm of the centered data matrix.

How can we derive a low-dimensional representation of the data?

Find the r orthogonal directions of largest variance.

The Principal Components Analysis Task

Given: a data matrix $D \in \mathbb{R}^{n \times d}$ and a rank r.

Find: the r orthogonal direction of largest variance, given by the columns $Z_{\cdot s}$ which are the solution to the following optimization problem:

$$\max_{Z} \operatorname{tr}(Z^\top C^\top C Z) \qquad \text{s.t. } Z \in \mathbb{R}^{n \times r},\ Z^\top Z = I$$

where $C = D - 1\mu_{\mathrm{F}}^\top$ is the centered data matrix.

3 Optimization

What is the solution Z of the objective of PCA?

The right singular vectors of C.

SVD Solves the Objective of PCA

Theorem (Value of the Operator Norm)

Let $C = U\Sigma V^\top \in \mathbb{R}^{n\times d}$ be the SVD of the matrix C. The solution of the optimization problem

$$\max_Z \operatorname{tr}(Z^\top C^\top CZ) \qquad s.t.\ Z \in \mathbb{R}^{n\times r},\ Z^\top Z = I$$

is given by $Z = V_{\cdot\mathcal{R}}$ for $\mathcal{R} = \{1, \ldots r\}$.

Proof (sketch): Show that the objective above is equivalent to

$$\min_Z \|C^\top C - Z\Sigma^2_{\mathcal{R}\mathcal{R}} Z^\top\|^2 \qquad \text{s.t. } Z \in \mathbb{R}^{n\times r},\ Z^\top Z = I.$$

Principal Components Analysis

function $\mathrm{PCA}(D, r)$

$C \leftarrow D - \mathbf{1}\boldsymbol{\mu}_{\mathrm{F}}^{\top}$ ▷ Center the data matrix

$(U_{\cdot\mathcal{R}}, \Sigma_{\mathcal{R}\mathcal{R}}, V_{\cdot\mathcal{R}}) \leftarrow \textsc{TruncatedSVD}(C, r)$

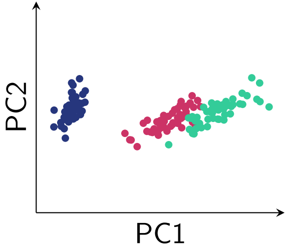
return $CV_{\cdot\mathcal{R}}$ ▷ the low-dimensional view on the data

end function

PCA can be implemented such that the novel data representation is centered (returning $CV_{\cdot\mathcal{R}}$) or not (returning $DV_{\cdot\mathcal{R}}$).

Two-Dimensional PCA on the Iris Dataset

$$\mathrm{PC1} = 0.36\mathrm{F}_1 - 0.08\mathrm{F}_2 + 0.85\mathrm{F}_3 + 0.36\mathrm{F}_4$$
$$\mathrm{PC2} = 0.66\mathrm{F}_1 + 0.73\mathrm{F}_2 - 0.17\mathrm{F}_3 - 0.07\mathrm{F}_4$$



- F_1: sepal length
- F_2: sepal width
- F_3: petal length
- F_4: petal width

PCA enables

Dimensionality Reduction

Onto the Directions with

Maximal Variance