annotated version

Machine Learning Course - CS-433

Matrix Factorizations

Dec 8, 2020

changes by Martin Jaggi 2020,2019, 2018, 2017,
©Martin Jaggi and Mohammad Emtiyaz Khan 2016

Last updated on: December 8, 2020

EPFL

Motivation

In the Netflix prize, the goal was to predict ratings of users for movies, given the existing ratings of those users for other movies. We are going to study the method that achieved the best error (for a single method).

movie user rating

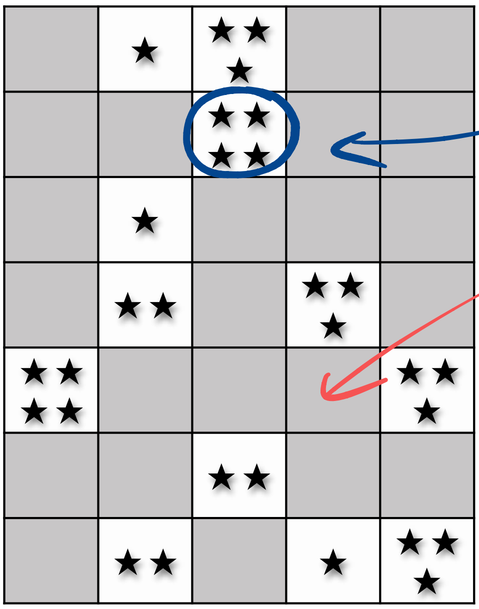
((2,3), 4)

n

users N

d

movies



D

to predict

matrix 99.99% unobserved

D = 20k

N = 500k

The Movie Ratings Data

Given movies $d = 1, 2, \dots, D$ and users $n = 1, 2, \dots, N$, we define $\mathbf{X}$ to be the $D \times N$ matrix containing all rating entries. That is, x_{dn} is the rating of n-th user for d-th movie.

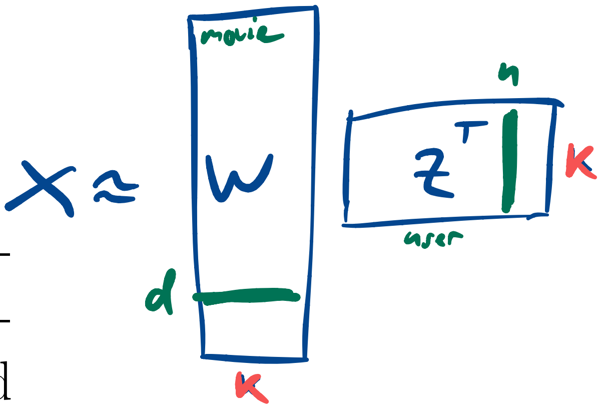
Note that most ratings x_{dn} are missing and our task is to predict those missing ratings accurately.

$\min_{w,z} \left(\mathcal{L}(w,z) = f(wz^\top) \right)$

Prediction Using a Matrix Factorization

We will aim to find $\mathbf{W}, \mathbf{Z}$ s.t.

$$\mathbf{X} \approx \mathbf{W}\mathbf{Z}^\top .$$



So we hope to 'explain' each rating x_{dn} by a numerical representation of the corresponding movie and user
- in fact by the inner product of a movie feature vector with the user feature vector.

model = $w_{d:}$ $z_{n:}$

stars

Loss function

$$\min_{\mathbf{W},\mathbf{Z}} \left(\mathcal{L}(\mathbf{W},\mathbf{Z}) := \frac{1}{2} \sum_{(d,n)\in\Omega} \left[x_{dn} - (\mathbf{W}\mathbf{Z}^\top)_{dn} \right]^2 \right)$$

where $\mathbf{W} \in \mathbb{R}^{D\times K}$ and $\mathbf{Z} \in \mathbb{R}^{N\times K}$ are tall matrices, having only $K \ll D, N$ columns.

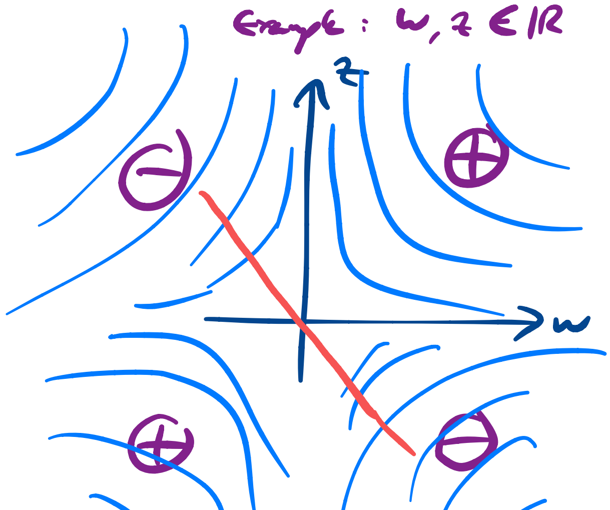
The set $\Omega \subseteq [D] \times [N]$ collects the indices of the observed ratings of the input matrix $\mathbf{X}$.

Each row of those matrices is the feature representation of a movie (rows of $\mathbf{W}$) or a user (rows of $\mathbf{Z}$) respectively.

① convex ? no!

$\mathcal{L}(w,z) = f(wz^\top) = wz^\top$

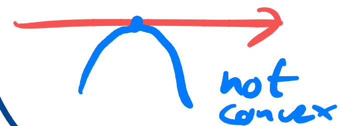
Example : $w, z \in \mathbb{R}$



Is this cost jointly convex w.r.t. $\mathbf{W}$ and $\mathbf{Z}$? ① Is the model identifiable? ② no!

given w^*, z^*

↳ $\beta w^*, \frac{1}{\beta} z^*$ also optimal



Choosing K

w $\leftarrow d$

K is the number of *latent* features.

$K=2$ example

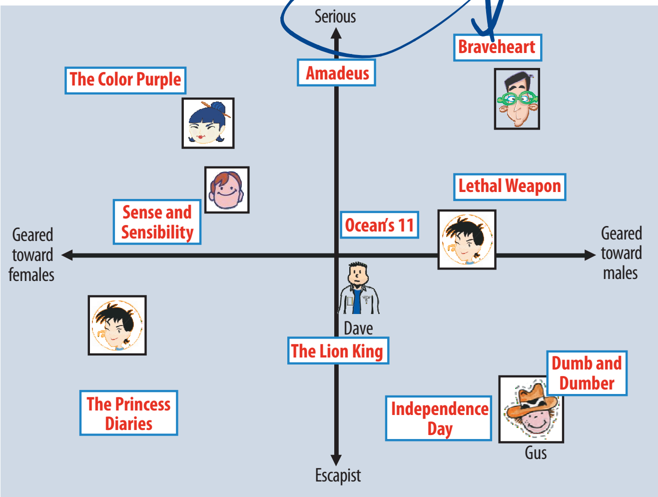


Figure 2. A simplified illustration of the latent factor approach, which characterizes both users and movies using two axes—male versus female and serious versus escapist.

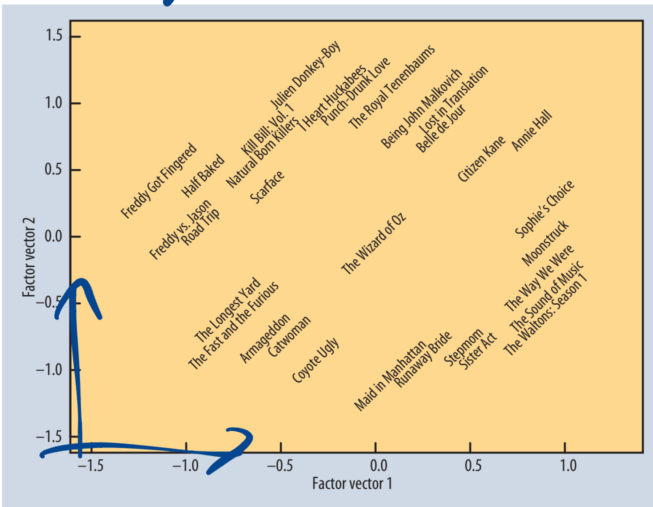


Figure 3. The first two vectors from a matrix decomposition of the Netflix Prize data. Selected movies are placed at the appropriate spot based on their factor vectors in two dimensions. The plot reveals distinct genres, including clusters of movies with strong female leads, fraternity humor, and quirky independent films.

Recall that for K-means, K was the number of clusters. (Similarly for GMMs, K was the number of latent variable dimensions).

W	Z^T
1 ·	x
x ·	1

if $K \geq \max(D, N)$

Large K facilitates overfitting.

Regularization

We can add a regularizer and minimize the following cost:

$$\mathcal{L}(W,Z) = \underbrace{\frac{1}{2}\sum_{(d,n)\in\Omega}\left[x_{dn} - (\mathbf{WZ}^\top)_{dn}\right]^2}_{f(WZ^T)} + \frac{\lambda_w}{2}\|\mathbf{W}\|^2_{\text{Frob}} + \frac{\lambda_z}{2}\|\mathbf{Z}\|^2_{\text{Frob}}$$

where $\lambda_w, \lambda_z > 0$ are scalars.

Ref: sifter.org 2006 Blog post

Stochastic Gradient Descent (SGD)

The training objective is a sum over $|\Omega|$ terms (one per rating):

rating

$f_{dn}(w,z) \to \mathbb{R}$

$$\mathcal{L}(w,z) = \frac{1}{|\Omega|} \sum_{(d,n)\in\Omega} \underbrace{\tfrac{1}{2}\left[x_{dn} - (\mathbf{W}\mathbf{Z}^\top)_{dn}\right]^2}_{f_{d,n}}$$

$$\boxed{\mathcal{L} = \frac{1}{|\Omega|} \sum_{n,d\in\Omega} f_{dn}(w,z)}$$

Derive the stochastic gradient for $\mathbf{W}, \mathbf{Z}$, given one observed rating $(d, n) \in \Omega$.

$\nabla_w f_{dn}(w,z) \in \mathbb{R}^{D\times K}$

$\nabla_z f_{dn}(w,z) \in \mathbb{R}^{N\times K}$

user n
movie d

For one fixed element (d, n) of the sum, we derive the gradient entry (d', k) for $\mathbf{W}$, that is $\frac{\partial}{\partial w_{d',k}} f_{d,n}(\mathbf{W}, \mathbf{Z})$, and analogously entry (n', k) of the $\mathbf{Z}$ part:

cost : $O(K)$

∇_w

$$\frac{\partial}{\partial w_{d',k}} f_{d,n}(\mathbf{W}, \mathbf{Z})$$

prediction error

$3 - 3.7 = -0.7$

$$= \begin{cases} -\left[x_{dn} - \overbrace{(\mathbf{W}\mathbf{Z}^\top)_{dn}}\right] z_{n,k} & \text{if } d' = d \\ 0 & \text{otherwise} \end{cases}$$

$d \to$ $w_{d:}$ $z_{n:}$

∇_z

$$\frac{\partial}{\partial z_{n',k}} f_{d,n}(\mathbf{W}, \mathbf{Z})$$

$$= \begin{cases} -\left[x_{dn} - (\mathbf{W}\mathbf{Z}^\top)_{dn}\right] w_{d,k} & \text{if } n' = n \\ 0 & \text{otherwise} \end{cases}$$

model updates :

$w^{(t+1)} := w^{(t)} - \gamma \nabla_w f_{dn}(w,z)$

$z^{(t+1)} := z^{(t)} - \gamma \nabla_z f_{dn}(w,z)$

cost $O(K)$

Alternating Least-Squares (ALS)

For simplicity, let us first assume that there are no missing ratings, that is $\Omega = [D] \times [N]$. Then

optional regularizer

$+ \frac{\lambda_w}{2}\|W\|_F + \frac{\lambda_z}{2}\|Z\|_F^2$

$$\mathcal{L} = \frac{1}{2}\sum_{d=1}^{D}\sum_{n=1}^{N}\left[x_{dn} - (\mathbf{WZ}^\top)_{dn}\right]^2$$

$$= \frac{1}{2}\|\underbrace{\mathbf{X}}_{\text{ratings}} - \mathbf{WZ}^\top\|^2_{\text{Frob}} \; .$$

$\min_Z \mathcal{L}(W, Z)$

$= \frac{1}{2}\|X - W Z^T\|^2_{Frob}$

$\nabla_Z \mathcal{L}(\cdot,\cdot) \stackrel{!}{=} 0$

as in Least squares

We can use coordinate descent to minimize the cost plus regularizer: We first minimize w.r.t. $\mathbf{Z}$ for fixed $\mathbf{W}$ and then minimize $\mathbf{W}$ given $\mathbf{Z}$.

$$\mathbf{Z}^{\top *} := (\mathbf{W}^\top\mathbf{W} + \lambda_z\mathbf{I}_K)^{-1}\mathbf{W}^\top\mathbf{X}$$
$$\mathbf{W}^{\top *} := (\mathbf{Z}^\top\mathbf{Z} + \lambda_w\mathbf{I}_K)^{-1}\mathbf{Z}^\top\mathbf{X}^\top$$

$\min_W \mathcal{L}(W, Z)$

What is the computational complexity? How can you decrease the cost when N and D are large?

if using regularizer.

as in Ridge regression

ALS with Missing Entries

Can you derive the ALS updates for the more general setting, when only the ratings $(d, n) \in \Omega$ contribute to the cost, i.e.

partially observed

$$\mathcal{L} = \frac{1}{2} \sum_{(d,n)\in\Omega} \left[x_{dn} - (\mathbf{W}\mathbf{Z}^\top)_{dn}\right]^2$$

$$\boxed{\begin{aligned} \nabla_w \mathcal{L}(w,z) &\overset{!}{=} 0 \\ \nabla_z \mathcal{L}(w,z) &\overset{!}{=} 0 \end{aligned}}$$

Hint: Compute the gradient with respect to each group of variables, and set to zero.