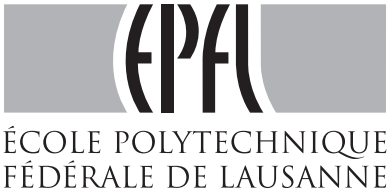
Matrix Factorization

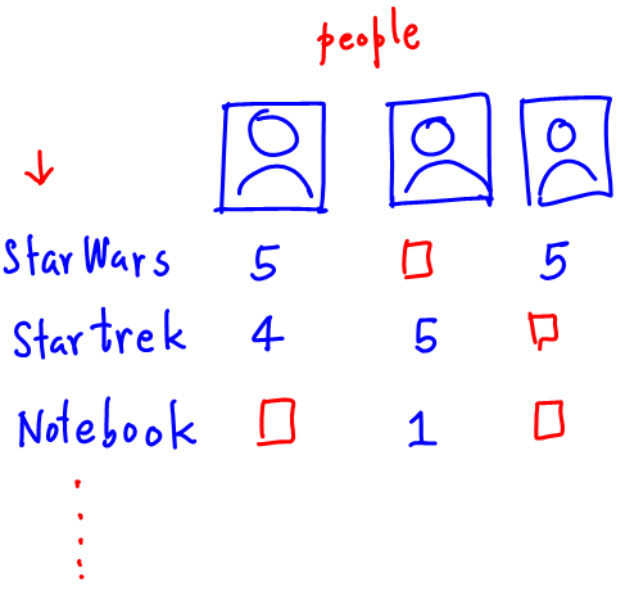
Mohammad Emtiyaz Khan
EPFL

Nov 19, 2015



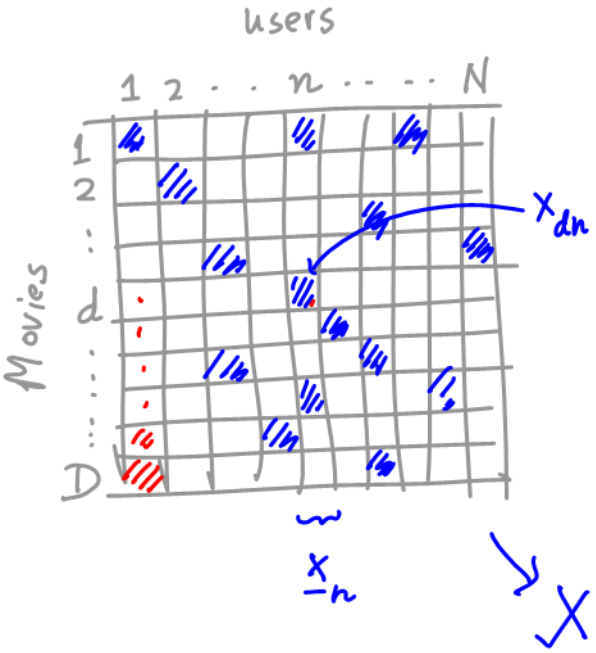
Motivation

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).



The movie-rating data

We index D movies with $d = 1, 2, \ldots, D$ and N users by $n = 1, 2, \ldots, N$. We denote the rating of n'th user for d'th movie by x_{dn}. Define $\mathbf{X}$ to be a $D \times N$ rating matrix with columns $\mathbf{x}_n \in \mathbb{R}^D$ containing ratings of the n'th user.



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

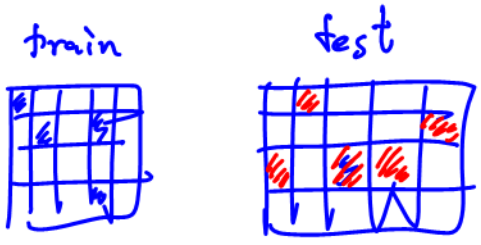
Q What is the generalization error in this case?

Let $\mathcal{O}_n = \{$ all movies rated by the n'th user $\}$

Train error: $\sum_{n=1}^{N} \sum_{d \in \mathcal{O}_n} (x_{dn} - \hat{x}_{dn})^2$

Generalization ~~Test~~ error: $\sum_{n=1}^{N} \sum_{d \notin \mathcal{O}_n} (x_{dn} - \hat{x}_{dn})^2$

* Simulate the reality: Create train & test pairs by "punching holes".



* Read about A/B testing in Wikipedia

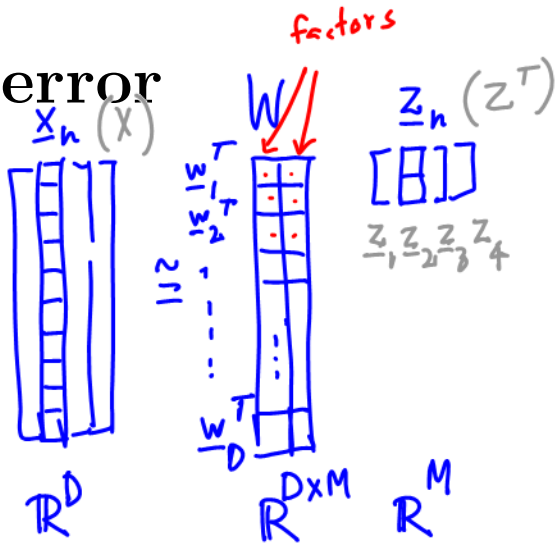
Minimizing reconstruction error

To simplify the problem, let us assume that there are no missing ratings.

Project data vectors $\mathbf{x}_n$ to a smaller dimension $\mathbf{z}_n \in \mathbb{R}^M$ with $M < D$, such that the *reconstruction error* is minimized (assume M is known):

$$\min_{\mathbf{W},\mathbf{Z}} \frac{1}{2}\sum_{n=1}^{N}\sum_{d=1}^{D}(x_{dn} - \mathbf{w}_d^T\mathbf{z}_n)^2$$

where $\mathbf{W}$ is a $D \times M$ matrix and $\mathbf{Z}$ is an $N \times M$ matrix, and $\mathbf{w}_d^T$ and $\mathbf{z}_n^T$ are rows of these matrices.



$$\underline{x}_n \cong W\underline{z}_n$$
$$X \cong WZ^T$$
$$D\times N \quad D\times M \quad M\times N$$
$$x_{dn} \cong \underline{w}_d^T\underline{z}_n$$
$$= w_{d1}z_{n1} + w_{d2}z_{n2} + \cdots + w_{dM}z_{nM}$$

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

Given W, convex wrt Z, and vice-versa.

$$0 \neq \begin{bmatrix} \frac{\partial^2}{\partial w^2} & \frac{\partial^2}{\partial w\,\partial z} \\ \times & \frac{\partial^2}{\partial z^2} \end{bmatrix} (x - wz)^2$$

$\theta \to P_\theta(x)$ one to one

No since WQQ^TZ → orthogonal

Regularization

We can add a regularizer and minimize the following cost:

$$\mathcal{L}(W,Z) = \frac{1}{2}\sum_{n=1}^{N}\sum_{d=1}^{D}(x_{dn} - \mathbf{w}_d^T\mathbf{z}_n)^2 + \frac{\lambda_w}{2}\sum_{d=1}^{D}\mathbf{w}_d^T\mathbf{w}_d + \frac{\lambda_z}{2}\sum_{n=1}^{N}\mathbf{z}_n^T\mathbf{z}_n$$

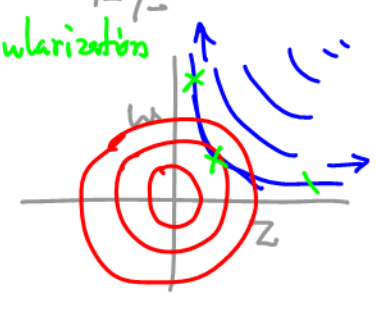
$(y - X^T\beta)^2$ $\quad \beta^T\beta$

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

Regularization

$$\mathcal{L}(Z|W) = \frac{1}{2}\sum_{n=1}^{N}\left[(\underline{x}_n - W\underline{z}_n)^T(\underline{x}_n - W\underline{z}_n) + \lambda_z\underline{z}_n^T\underline{z}_n\right] + \text{const}$$

$\underline{y} - X\beta$



$\beta^* = (X^TX + \lambda I)^{-1} X^T y \Rightarrow z_n^* = (W^TW + \lambda_z I)^{-1} W^T x_n$

MxM MxD Dx1

$DM^2 + M^3$

Alternating least-squares (ALS)

We can use coordinate descent algorithm, by first minimizing w.r.t. $\mathbf{Z}$ given $\mathbf{W}$ and then minimizing $\mathbf{W}$ given $\mathbf{Z}$.

$$\mathbf{Z}^T \leftarrow (\mathbf{W}^T\mathbf{W} + \lambda_z \mathbf{I}_M)^{-1}\mathbf{W}^T\mathbf{X}$$
$$\mathbf{W}^T \leftarrow (\mathbf{Z}^T\mathbf{Z} + \lambda_w \mathbf{I}_M)^{-1}\mathbf{Z}^T\mathbf{X}^T$$

$w_d^* = (Z^TZ + \lambda_w I_M)^{-1} Z^T x_d$ ↑ rows of X

X DxN
W DxM
Z NxM

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

$(DM^2 + M^3)N$
but since $D \gg M$, $\approx O(DNM^2)$ Almost

Extensions

Dealing with missing entries

Denote the set of movies not rated by n'th user by $\mathbb{O}_n$, then we can modify the cost as follows:

$$\frac{1}{2}\sum_{n=1}^{N}\sum_{d\in\mathbb{O}_n}(x_{dn} - \mathbf{w}_d^T\mathbf{z}_n)^2$$

Derive the modified ALS algorithm.

Q: Write the update for z_n given W & x_{O_n}

x_n x_{O_n}

$x_{O_n} \simeq W_{O_n,:} z_n$

2x1 2xM Mx1

$z_n^{(k+1)} = (W_{O_n,:}^{(k)T} W_{O_n,:}^{(k)} + \lambda_z I_M)^{-1} W_{O_n,:}^T x_{O_n}$

Algorithm

* start with some $W^{(0)}$
* Iterate

$Z^{(k+1)} \leftarrow W^{(k)}$

$W^{(k+1)} \leftarrow Z^{(k+1)}$

Adding an offset

Since many ratings are missing, we cannot normalize the data. A solution is to add an offset term in the cost.

$$\frac{1}{2}\sum_{n=1}^{N}\sum_{d\in\mathbb{O}_n}(x_{dn}-\mathbf{w}_d^T\mathbf{z}_n-w_{0d}-z_{0n}-\mu)^2$$

where w_{0d} and z_{0n} are the offset terms for d'th movie and n'th user while μ is the global offset.

Derive the modified ALS algorithm.

A Probabilistic Model

Write the probabilistic model associated with the regularized cost.

Generalization to Exponential Family

Using generalized linear model, we can extend matrix factorization to different types of datasets. See the following paper for details: "A Generalization of Principal Component Analysis to the Exponential Family" by Collins et. al., 2001.

Choosing number of factors M, λ_z, λ_w