

Matrix Factorization

Mohammad Emtiyaz Khan
EPFL

Nov 12, 2015



ÉCOLE POLYTECHNIQUE
FÉDÉRALE DE LAUSANNE

©Mohammad Emtiyaz Khan 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 $d = 1, 2, \dots, D$ and N users by $n = 1, 2, \dots, N$. 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.

What is the generalization error in this case?

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.

Is this cost jointly convex w.r.t. \mathbf{W} and \mathbf{Z} ? Is the model identifiable?

Regularization

We can add a regularizer and minimize the following cost:

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

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

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} .

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

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

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

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.

Adding an offset

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

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

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