CSE 8803RS: Recommendation Systems

Matrix Factorization: Bayesian Formulations

Steven P. Crain

School of Computational Science & Engineering
College of Computing
Georgia Institute of Technology

February 8, 2011

Probabilistic Matrix Factorization

R Salakhurutdinov and A Mnih 2008
University of Toronto
Problem:

- Existing CF methods could not scale to large datasets.
- Existing CF methods had bad prediction accuracy for users with few ratings.

Key idea: Bayesian methods provide natural and scalable regularization to matrix factorization methods that improves accuracy for users with few ratings.

Probabilistic Matrix Factorization (PMF)

Map the possibles ratings onto $[0, 1]$.

We can then model the rating based on latent user and item factors:

$$r_{ui} = g(U_u^T V_i) + \epsilon \tag{1}$$

$g(x)$ is the logistic function $1/(1 + \exp(-x))$.
ϵ is Gaussian noise with variance σ^2.

Advantages:

- Natural probabilistic extension of conventional matrix factorization.
- Logistic function simulates the human tendency to reserve extreme ratings.
- This formulation leads to efficient training algorithms.

Log Likelihood

$$P(U,V|R,\sigma^2,\sigma_U^2,\sigma_V^2) = \frac{P(R|U,V,\sigma^2)P(U|\sigma_U^2)P(V|\sigma_V^2)}{P(R)} \tag{2}$$

$$= \frac{\prod_u \prod_i \left[\mathcal{N}(R_{ui}|g(U_u^T V_i),\sigma^2)\right]_{ui}^{I} \prod_u \mathcal{N}(U_u|0,\sigma_U^2\mathbb{I}) \prod_i \mathcal{N}(V_i|0,\sigma_V^2\mathbb{I})}{P(R)}$$

$$\ln P(U,V|R,\sigma^2,\sigma_U^2,\sigma_V^2) = \tag{3}$$

$$-\frac{1}{2\sigma^2}\sum_u\sum_i I_{ui}\left(R_{ui} - g(U_u^T V_i)\right)^2 - \frac{1}{2}\sum_u\sum_i I_{ui}\ln\sigma^2$$

$$-\frac{1}{2\sigma_U^2}\sum_u U_u^T U_u - \frac{ND}{2}\sum_u \ln\sigma_U^2$$

$$-\frac{1}{2\sigma_V^2}\sum_i V_i^T V_i - \frac{MD}{2}\sum_u \ln\sigma_V^2 - \ln P(R)$$

Rearranging gives the same equation as derived by Rennie and Srebo (2005):

$$Minimize \frac{1}{2} \sum_u \sum_i I_{ui} \left(R_{ui} - g(U_u^T V_i) \right)^2 + \frac{\lambda_U}{2} \|U\|_{\mathcal{F}}^2 + \frac{\lambda_V}{2} \|V\|_{\mathcal{F}}^2 \qquad (4)$$

- Provides a probabilistic interpretation of fast maximum margin matrix factorization.
- The λ parameters can be autotuned by use of an appropriate prior in an EM framework.

Constrained PMF

Recall the method of writing U in terms of V:

$$U_u = \sum_i I_{ui} V_i \tag{5}$$

- Constrained PMF generalizes this technique.

Constrained PMF User Model

$$U_u = Y_u + \frac{\sum_i I_{ui} W_i}{\sum_i I_{ui}} \tag{6}$$

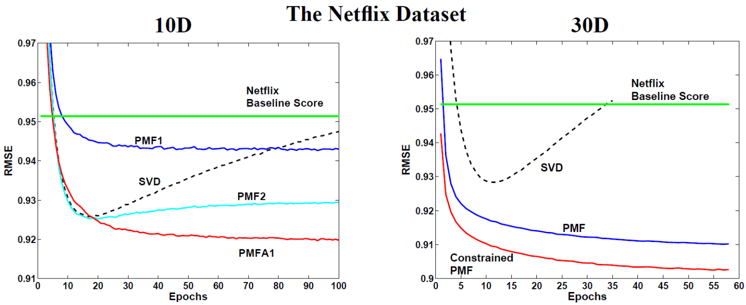
- Motivation is to separate what we can accurately model from peculiarities.
  - The fact that the user rated an item contributes to a prior for the user's latent profile.
  - Y_u captures the residual peculiarities of the user's profile.
- Compare with subtracting user rating bias.

Constrained PMF Optimization

$$Minimize \frac{1}{2} \sum_u \sum_i I_{ui} \left(R_{ui} - g \left(\left[Y_u + \frac{\sum_i I_{uk} W_k}{\sum_k I_{uk}} \right]^T V_i \right) \right)^2 + \frac{\lambda_Y}{2} \|Y\|_{\mathcal{F}}^2 + \frac{\lambda_W}{2} \|W\|_{\mathcal{F}}^2 + \frac{\lambda_V}{2} \|V\|_{\mathcal{F}}^2 \quad (7)$$

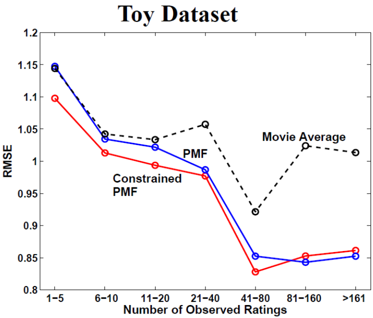
- Provides a probabilistic interpretation of fast maximum margin matrix factorization.
- The λ parameters can be autotuned by use of an appropriate prior in an EM framework.

Results: Overall



- PMF1: $\lambda_U = 0.01, \lambda_V = 0.001$; PMF2: $\lambda_U = 0.001, \lambda_V = 0.0001$
- PMFA1: Adaptive spherical priors, diagonal covariance similar

Results: Cold Start



Conclusions

- PMF provides higher accuracy than SVD.
- Autotuning performs reasonably well, though not compared to thorough parameter search.
- Cold start accuracy is not much better than using the movie average ratings.

Global Analytic Solution for Variational Bayesian Matrix Factorization

S Nakajima, M Sugiyama and R Tomioka 2010
Nikon, Tokyo Institute of Technology and University of Tokyo
Problem:

- Matrix factorization is normally expensive because of non-convexity.

Key idea: A variational Bayesian matrix factorization can be solved analytically.
Important limitation: The approach requires full observations, so it cannot be applied to partially observed ratings matrices.

Setup

Suppose there is a rank-H $L \times M (L \leq M)$ matrix $U = BA^T$. Now, we get n observations V^i of U subject to Gaussian noise with variance σ^2.

$$P(V^i|A,B) = \mathcal{N}(V^i|BA^T, \sigma^2) \tag{8}$$

$$P(V|A,B) \propto \exp\left(-\frac{1}{2\sigma^2}\sum_i \|V^i - BA^T\|_{\mathcal{F}}^2\right) \tag{9}$$

- We need to add some constraint to make the problem identifiable.
- Since we are taking a probabilistic approach, natural constraints are priors on A, B.
  - We place spherical Gaussian priors with variance c_{ah} and c_{bh} on columns A_h and B_h.
- Even so, the dependency between A and B makes exact inference impossible and approximate inference is expensive.

Variational Bayesian (VB) Approach

- The VB approach bounds the real optimization problem with an approximation that is tractable.
- A current approximation yields a probability distribution over possible models, which is used to find a better approximation.
- In this case, we will see that the globally best approximation can be found analytically.

Variational Model

We suppose that all columns of A and B are independently sampled from Gaussian distributions with arbitrary centers and covariance matrices.

$$r(A, B|V) = \prod_h \mathcal{N}(A_h|\mu_{ah}, \Sigma_{ah})\mathcal{N}(B_h|\mu_{bh}, \Sigma_{bh}) \tag{10}$$

This equation can be iteratively optimized for A, B given μ, Σ and vice versa, which is the conventional VBMF approach.
The variations σ^2, c^2 can also be estimated in the process for empirical VBMF.

A Closer Look

Explicit function to be optimized:

$$\frac{nLM}{2}\log\sigma^2 + \frac{1}{2}\sum_h \left(M \log c_{ah}^2 - \log|\Sigma_{ah}| + \frac{\alpha_h}{c_{ah}^2} + L\log c_{bh}^2 - \log|\Sigma_{bh}| + \frac{\beta_h}{c_{bh}^2} \right) + \frac{1}{2\sigma^2}\sum_i \left\| V^i - \sum_h \mu_{bh}\mu_{ah}^T \right\|_{\mathcal{F}}^2 + \frac{n}{2\sigma^2}\sum_h \left(\alpha_h\beta_h - \|\mu_{ah}\|^2 \|\mu_{bh}\|^2 \right) \quad (11)$$

Notation

$$\bar{V} = \frac{1}{n} \sum_i V^i \tag{12}$$

$$\bar{V} = \sum_h \gamma_h \omega_{bh} \omega_{ah}^T (\text{by SVD}) \tag{13}$$

Let $\hat{\gamma}_h$ be the second largest real root of

$$t^4 + \xi_3 t^3 + \xi_2 t^2 + \xi_1 t + \xi_0 \tag{14}$$

ξ are analytic functions of $L, M, n, \gamma_h, \sigma, c_{ah}$ and c_{bh}. $\tilde{\gamma}_h$ is also defined by another analytic function of those variables.

Globally Optimal VB Solution

Theorem: The global VB solution can be expressed as

$$\hat{U}^{VB} = \sum_h \hat{\gamma}_h^{VB} \omega_{bh} \omega_{ah}^T \tag{15}$$

$$\hat{\gamma}_h^{VB} = \begin{cases} \hat{\gamma}_h & \text{if } \gamma_h > \tilde{\gamma}_h, \\ 0 & \text{otherwise.} \end{cases} \tag{16}$$

Experiments

- Artificial data
- Concrete slump test data from UCI
- Compared against iterative VBMF.
- In all cases, immediately arrived at a better solution than VBMF.
- Greatly improved computational cost.
- Only applicable when matrices V^i are fully observed.