CSE 8803RS: Recommendation Systems

Lecture 10: CF Incorporating User and Item Features

Hongyuan Zha

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

General Setting

- Users: $u, v \in \mathcal{U}$; Items: $i, j \in \mathcal{I}$
- Ratings: r_{ui} indicating degree of preference of user u for item j: *dyadic* data, (u, i) dyad
- **Problem.** Ratings are not defined over all $\mathcal{U} \times \mathcal{I}$, need to predict those missing ratings
  — prediction based users' past interactions with items
- New twists: user and item features
  — User features: age, gender, geo-location
  — Item features: genre, title words, cast, release date of the movie
  — User-item features: Is the user's favorite actor playing a lead role in the movie?

Regression-based Latent Factor Model (RLFM)

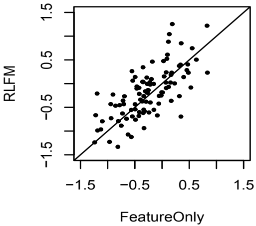
- Improve prediction for warm-start problem by simultaneously incorporating features and past interactions
- Prediction for cold-start problem through features
- Regularized SVD

$$\min_{U,V} \; E(U,V) = \frac{1}{2} \sum_{(i,j)\in O} (A_{ij} - u_i v_j^T)^2 + \frac{\lambda}{2}(\|U\|_F^2 + \|V\|_F^2)$$

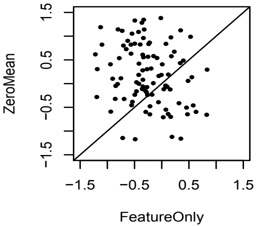
- Regularization: shrink to zero, *ZeroMean*
  — Equivalent to Gaussian prior $u_i \sim \mathcal{N}(0, \sigma^2 I)$ and $v_j \sim \mathcal{N}(0, \sigma^2 I)$
- New prior: replace the zero mean with a feature-based regression

First User Latent Factor

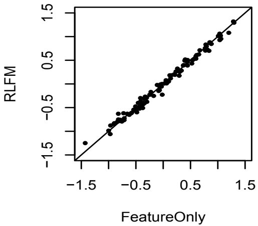
First user latent feature for heavy users and light users



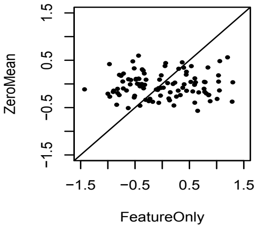
(a) RLFM for heavy users



(b) ZeroMean for heavy users



(c) RLFM for light users



(d) ZeroMean for light users

First User Latent Factor

- Light users
  — ZeroMean: shrink to zero, leading to bias
  — RLFM: closely anchored around known feature
- Heavy users
  — ZeroMean: all over the place, may lead to overfitting
  — RLFM: deviate from feature in a smooth way

Notations

- User features: $w_i \in R^p$; Item features: $z_j \in R^q$; Dyad features: $x_{ij} \in R^s$. $X_{ij} = [x_{ij}, w_i, z_j]$. Rating y_{ij}
- Observed data $\{X_{ij}, y_{ij}\}$. Need to build a probabilistic model $P(\{y_{ij}\} \mid \{X_{ij}\})$
- Remove systematic effects by x_{ij}, considering $A_{ij} - b^T x_{ij}$

Models in Matrix Notations

W user features, Z item features

- ZeroMean: $A \approx UV^T$
- Interaction: $A \approx WXZ^T$
- FeatureOnly: $A \approx WS^T + TZ^T + (WG)(ZD)^T$
  — X be low-rank
  — in the paper: $T = ed_0^T$, and $S = eg_0^T$
- RLFM: $A \approx WS^T + TZ^T + (U + WG)(V + ZD)^T$
  — matrices marked with red are unknown parameters

RLFM and Regression-based Prior

- Rather than using shrinking to zero, RLFM can be interpreted as using regression-based prior
- Ignore the $WS^T + TZ^T$, and define

$$\hat{U} = U + WG, \quad \hat{V} = V + ZD$$

Then $A \approx \hat{U}\hat{V}^T$

- If we use regularization on U and V, then

$$\min_{\hat{U},\hat{V},G,D} \|\mathcal{S} \odot (A - \hat{U}\hat{V}^T)\|_F^2 + \lambda(\|\hat{U} - WG\|_F^2 + \|\hat{V} - ZD\|_F^2)$$

Probabilistic Model for RLFM

W user features, Z item features

- First stage: Generalized linear model for ratings,

$$y_{ij} \sim N(m_{ij}, \sigma^2), \quad m_{ij} = b^T x_{ij} + \alpha_i + \beta_j + u_i v_j^T$$

- Second stage:

$$\alpha_i = w_i g_0^T + \epsilon_i^\alpha, \quad \epsilon_i^\alpha \sim N(0, a_\alpha)$$

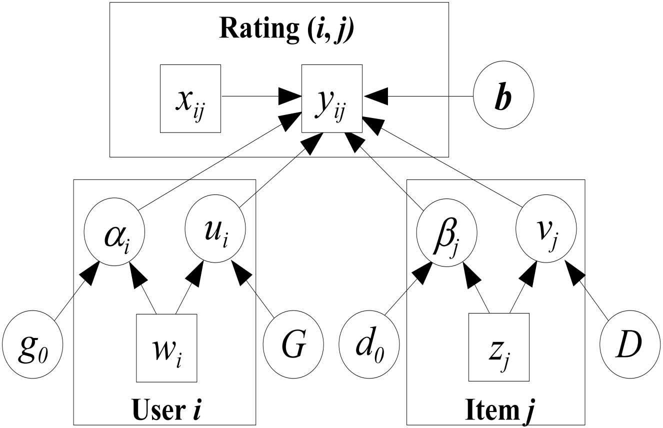
$$\beta_j = z_j d_0^T + \epsilon_i^\beta, \quad \epsilon_i^\beta \sim N(0, a_\beta)$$

$$u_i = w_i G + \epsilon_i^u, \quad \epsilon_i^u \sim N(0, A_u)$$

$$v_j = z_j D + \epsilon_i^v, \quad \epsilon_i^v \sim N(0, A_v)$$

- Parameters: $\Theta = \{b, g_0, d_0, G, D, a_\alpha, a_\beta, A_u, A_v\}$

RLFM Graphical Model representation



MLE and Prediction

- Two sets of parameters:
  — $\delta_{ij} = (\alpha_i, \beta_j, u_i, v_j)$: (latent) parameters
  — $\Theta = \{b, g_0, d_0, G, D, a_\alpha, a_\beta, A_u, A_v\}$: hyperparamters
- Given data $\{X_{ij}\}$ and $\{y_{ij}\}$
- For a fixed Θ, posterior distribution of the latent factors,

$$p(\delta_{ij} | \{y_{ij}\}, \{X_{ij}\}, \Theta)$$

- Prediction distribution for y_{ij}^{new}, given X_{ij}^{new},

$$p(y_{ij}^{new} | X_{ij}^{new}, \{y_{ij}\}, \{X_{ij}\}, \Theta) = \int p(y_{ij}^{new} | X_{ij}^{new}, \delta_{ij}) dp(\delta_{ij} | \{y_{ij}\}, \{X_{ij}\}, \Theta)$$

Bayesian and Empirical Bayesian

- For full Bayesian, we need prior on Θ and another set of hyperparameters to go with it
- Empirical Bayesian (Type 2 MLE, GMLE, evidence approximation), settles on a specific Θ obtained by

$$\max_{\Theta} p(\{y_{ij}\} \mid \{X_{ij}\}, \Theta) \Rightarrow \Theta_{MLE}$$

- As in ordinary regression, we assume $\{X_{ij}\}$ *fixed*
— and it will be removed from subsequent formula,

$$\max_{\Theta} p(\{y_{ij}\} \mid \Theta) \Rightarrow \Theta_{MLE}$$

Approximate Prediction

- Settle on a specific δ_{ij},

$$y_{ij}^{new} = \mathcal{E}(y_{ij}^{new} | X_{ij}^{new}, \hat{\delta}_{ij}) = \hat{b}^T x_{ij}^{new} + \hat{\alpha}_i + \hat{\beta}_j + \hat{u}_i \hat{v}_j^T$$

  where for $\phi \in \{\alpha_i, \beta_j, u_i, v_j\}$,

$$\hat{\phi} = \mathcal{E}(\phi | \{y_{ij}\}, \Theta_{MLE}) \approx \frac{1}{L} \sum_{\ell=1}^{L} \phi^{(\ell)}$$

- Recall $p(\{\delta_{ij}\} | \{y_{ij}\}, \Theta_{MLE})$
  — will have L samples $\phi^{(\ell)}, \ell = 1, \ldots, L$ drawn from the above using Gibbs sampling

A Brief Review of EM

Given distribution $p(\{y_{ij}\}, \{\delta_{ij}\}|\Theta)$, want to maximize $p(\{y_{ij}\}|\Theta)$ w.r.t. Θ

1. Choose initial Θ^{old}
2. **E Step.** Evaluate $p(\{\delta_{ij}\}|\{y_{ij}\}, \Theta)$
3. **M Step.** Find Θ^{new}

$$\max_{\Theta} \sum_{\{\delta_{ij}\}} p(\{\delta_{ij}\}|\{y_{ij}\}, \Theta^{old}) \log p(\{y_{ij}\}, \{\delta_{ij}\}|\Theta) \Rightarrow \Theta^{new}$$

4. $\Theta^{old} \leftarrow \Theta^{new}$, **goto** Step 2

Monte Carlo EM (MCEM)

For the **E Step**,

1. Rather than evaluating $p(\{\delta_{ij}\}|\{y_{ij}\}, \Theta^{old})$
2. We draw L samples from $p(\{\delta_{ij}\}|\{y_{ij}\}, \Theta^{old}) \Rightarrow p^*(\{\delta_{ij}\}|\{y_{ij}\}, \Theta^{old})$
3. Find Θ^{new}

$$\max_{\Theta} \sum_{\{\delta_{ij}\}} p^*(\{\delta_{ij}\}|\{y_{ij}\}, \Theta^{old}) \log p(\{y_{ij}\}, \{\delta_{ij}\}|\Theta) \Rightarrow \Theta^{new}$$

A Brief Review of Gibbs Sampling

Sample from $p(z_1, z_2, z_3)$. From some initial value (z_1^0, z_2^0, z_3^0)

1. $z_1^{k+1} \sim p(z_1 | z_2^k, z_3^k)$
2. $z_2^{k+1} \sim p(z_2 | z_1^{k+1}, z_3^k)$
3. $z_3^{k+1} \sim p(z_3 | z_1^{k+1}, z_2^{k+1})$

Negative complete data log-likelihood $-\log p(\{y_{ij}\}, \{\delta_{ij}\} | \Theta)$,

$$
\begin{aligned}
&\text{Constant} + \tfrac{1}{2} \textstyle\sum_{ij} \left(\frac{1}{\sigma_{ij}^2} (y_{ij} - \alpha_i - \beta_j - x_{ij}' \boldsymbol{b} - u_i' v_j)^2 + \log \sigma_{ij}^2 \right) \\
&+ \tfrac{1}{2} \textstyle\sum_i \left(\frac{1}{a_\alpha} (\alpha_i - g_0' w_i)^2 + \log a_\alpha + (u_i - G w_i)' A_u^{-1} (u_i - G w_i) + \log(\det A_u) \right) \\
&+ \tfrac{1}{2} \textstyle\sum_j \left(\frac{1}{a_\beta} (\beta_j - d_0' z_i)^2 + \log a_\beta + (v_j - D z_j)' A_v^{-1} (v_j - D z_j) + \log(\det A_v) \right)
\end{aligned}
$$

Conditional Distribution Used in Gibbs Sampling

Use completing the squares, all the conditionals are Gaussian,

Latent factors $\{\alpha_i\}$

$$\text{Let } o_{ij} = y_{ij} - \beta_j - x'_{ij}\boldsymbol{b} - u'_i v_j$$

$$\frac{\partial}{\partial \alpha_i}\Delta = \left(\frac{1}{a_\alpha} + \sum_{j\in\mathcal{J}_i}\frac{1}{\sigma_{ij}^2}\right)\alpha_i - \left(\frac{g'_0 w_i}{a_\alpha} + \sum_{j\in\mathcal{J}_i}\frac{o_{ij}}{\sigma_{ij}^2}\right)$$

$$Var[\alpha_i|\text{Rest}] = \left(\frac{1}{a_\alpha} + \sum_{j\in\mathcal{J}_i}\frac{1}{\sigma_{ij}^2}\right)^{-1}$$

$$E[\alpha_i|\text{Rest}] = Var[\alpha_i|\text{Rest}]\left(\frac{g'_0 w_i}{a_\alpha} + \sum_{j\in\mathcal{J}_i}\frac{o_{ij}}{\sigma_{ij}^2}\right)$$

Latent factors $\{u_i\}$

$$\text{Let } o_{ij} = y_{ij} - \alpha_i - \beta_j - x'_{ij}\boldsymbol{b}$$

$$\frac{\partial}{\partial u_i}\Delta = \left(A_u^{-1} + \sum_{j\in\mathcal{J}_i}\frac{v_j v'_j}{\sigma_{ij}^2}\right)u_i - \left(A_u^{-1}Gw_i + \sum_{j\in\mathcal{J}_i}\frac{o_{ij}v_j}{\sigma_{ij}^2}\right)$$

$$Var[u_i|\text{Rest}] = \left(A_u^{-1} + \sum_{j\in\mathcal{J}_i}\frac{v_j v'_j}{\sigma_{ij}^2}\right)^{-1}$$

$$E[u_i|\text{Rest}] = Var[u_i|\text{Rest}]\left(A_u^{-1}Gw_i + \sum_{j\in\mathcal{J}_i}\frac{o_{ij}v_j}{\sigma_{ij}^2}\right)$$

Monte Carlo EM (MCEM)

For the **M Step**,

1. Five separate regression problem to compute each pair in $\Theta = \{b, \sigma, g_0, a_\alpha, d_0, a_\beta, G, A_u, D, A_v\}$