CSE 8803RS: Recommendation Systems

Lecture 11: Bayesian Probabilistic Matrix Factorization

Hongyuan Zha

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

Probabilistic Models

- Rating matrix $A \approx UV^T$, let I_{ij} be the indicator,

$$p(A|U,V,\alpha) = \prod_{i=1}^{M}\prod_{j=1}^{N}\left(\mathcal{N}(A_{ij}|U_i V_j^T, \alpha^{-1})\right)^{I_{ij}}$$

$$p(U|\alpha_U) = \prod_{i=1}^{M}\mathcal{N}(U_i|0, \alpha_U^{-1} I)$$

$$p(V|\alpha_V) = \prod_{j=1}^{N}\mathcal{N}(V_j|0, \alpha_V^{-1} I)$$

- Log-posterior,

$$\log p(U,V|A,\alpha,\alpha_U,\alpha_V) = \log p(A|U,V,\alpha) + \log p(U|\alpha_U) + \log p(V|\alpha_V)$$

Probabilistic Models

- The negative log-posterior is equivalent to

$$E(U,V) = \frac{1}{2}\sum_{i=1}^{M}\sum_{j=1}^{N} I_{ij}(A_{ij} - U_i V_j^T)^2 + \frac{\lambda_U}{2}\|U\|_F^2 + \frac{\lambda_V}{2}\|V\|_F^2$$

  and $\lambda_U = \alpha_U/\alpha$, and $\lambda_V = \alpha_V/\alpha$
- Complexity control done by selecting λ_U and λ_V using cross-validation for example
- Use prior of the hyper-parameters $\alpha_U, \alpha_V, \alpha$ and jointly optimize all parameters
- User evidence function $p(R|\alpha_U, \alpha_V, \alpha)$

Full Bayesian Models

- Models for user and item profiles,

$$p(U|\mu_U, \Lambda_U) = \prod_{i=1}^{M} \mathcal{N}(U_i|\mu_U, \Lambda_U^{-1})$$

$$p(V|\mu_V, \Lambda_V) = \prod_{j=1}^{N} \mathcal{N}(V_j|\mu_V, \Lambda_V^{-1})$$

- Gassian-Wishart priors on hyper-parameters $\Theta_U = (\mu_U, \Lambda_U)$ and $\Theta_V = (\mu_V, \Lambda_V)$,

$$p(\Theta_U|\Theta_0) = p(\mu_U|\Lambda_U)p(\Lambda_U) = \mathcal{N}(\mu_U|\mu_0, (\beta_0\Lambda_U)^1)\mathcal{W}(\Lambda_U|W_0, \nu_0)$$

with $\Theta_0 = (\mu_0, \beta_0, W_0, \nu_0)$

Predictive Distribution

- Integrate out parameters and hyper-parameters,

$$p(A_{ij}|A, \Theta_0) = \int p(A_{ij}|U, V) dp(U, V, \Theta_U, \Theta_V|A, \Theta_0)$$

- Monte Carlo methods,

$$p(A_{ij}|R, \Theta_0) = \frac{1}{n} \sum_{k=1}^{n} p(A_{ij}|U^k, V^k)$$

$$(U^k, V^k, \Theta_U^k, \Theta_V^k) \sim p(U, V, \Theta_U, \Theta_V|A, \Theta_0), \quad k = 1, \ldots, n$$

Gibbs Sampling

- Cycle through the variables, sampling each one from its distribution conditional on the current values of all other variables

$$p(U, V, \Theta_U, \Theta_V | A, \Theta_0) \sim p(A|U, V, \alpha) p(U|\Theta_U) p(V|\Theta_V) \times$$

$$\times p(\Theta_U|\Theta_0) p(\Theta_V|\Theta_0)$$

- Conditional probability,

$$p(U_i|A, V, \Theta_U, \alpha) \sim \prod_{j=1}^{N} \left(\mathcal{N}(A_{ij}|U_i V_j^T, \alpha^{-1})\right)^{I_{ij}} p(U_i|\mu_U, \Lambda_U)$$

$$= \mathcal{N}(U_i|\mu_i^*, (\Lambda_{Ui}^*)^{-1})$$

Using the technique of completing the squares,

$$\Lambda_{Ui}^* = \Lambda_U + \alpha \sum_{j=1}^{N} (V_j V_j^T)^{I_{ij}}$$

Gibbs Sampling

- Using the technique of completing the squares,

$$\Lambda_{Ui}^{*} = \Lambda_U + \alpha \sum_{j=1}^{N} (V_j V_j^T)^{I_{ij}}$$

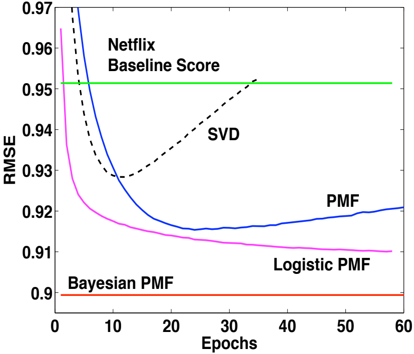
$$\mu_i^{*} = (\Lambda_{Ui}^{*})^{-1} \left(\alpha \sum_{j=1}^{N} (R_{ij} V_j)^{I_{ij}} + \Lambda_U \mu_U \right)$$

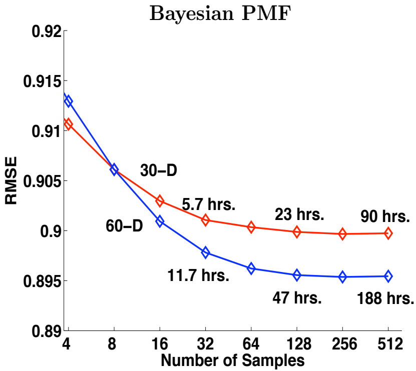
- $p(\Theta_U | U, \Theta_0)$

Gibbs sampling for Bayesian PMF

1. Initialize model parameters $\{U^1, V^1\}$
2. For t=1,...,T
   - Sample the hyperparameters (Eq. 14):
     $$\Theta_U^t \sim p(\Theta_U | U^t, \Theta_0)$$
     $$\Theta_V^t \sim p(\Theta_V | V^t, \Theta_0)$$
   - For each $i = 1, ..., N$ sample user features in parallel (Eq. 11):
     $$U_i^{t+1} \sim p(U_i | R, V^t, \Theta_U^t)$$
   - For each $i = 1, ..., M$ sample movie features in parallel:
     $$V_i^{t+1} \sim p(V_i | R, U^{t+1}, \Theta_V^t)$$

Netflix Results





Netflix Results