Case Study 4: Collaborative Filtering

Probabilistic Models for Matrix Factorization

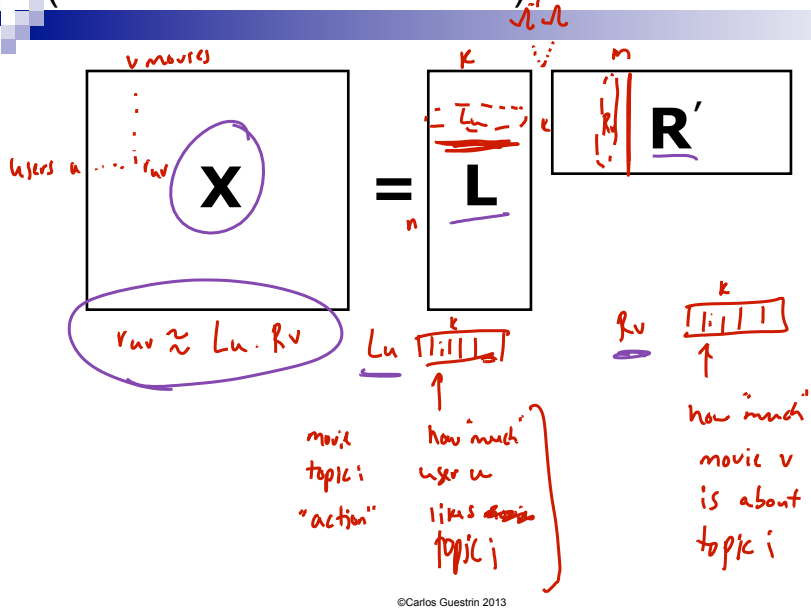
Cold-Start Problem

Machine Learning/Statistics for Big Data
CSE599C1/STAT592, University of Washington

Carlos Guestrin

March 12th, 2013

Interpreting Low-Rank Matrix Completion (aka Matrix Factorization)



Stochastic Gradient Descent

$$\min_{L,R} \frac{1}{2} \sum_{r_{uv}} (L_u \cdot R_v - r_{uv})^2 + \frac{\lambda_u}{2} ||L||_F^2 + \frac{\lambda_v}{2} ||R||_F^2$$

- Observe one rating at a time r_{uv}^t
- Gradient observing r_{uv}:

$$\frac{\partial F}{\partial L_u} = \varepsilon_t R_v + \lambda_u L_u \qquad \frac{\partial F}{\partial R_v} = \varepsilon_t L_u + \lambda_v R_v \qquad \nabla F_t = \begin{bmatrix} \varepsilon_t R_v + \lambda_u L_u \\ \varepsilon_t L_u + \lambda_v R_v \end{bmatrix}$$

- Updates:

$$\begin{bmatrix} L_u^{(t+1)} \\ R_v^{(t+1)} \end{bmatrix} \leftarrow \begin{bmatrix} (1-\eta_t \lambda_u) L_u^{(t)} - \eta_t \varepsilon_t R_v^{(t)} \\ (1-\eta_t \lambda_v) R_v^{(t)} - \eta_t \varepsilon_t L_u^{(t)} \end{bmatrix}$$

step size η_t : $\begin{bmatrix} L \\ R \end{bmatrix} \leftarrow \begin{bmatrix} L \\ R \end{bmatrix} - \eta_t \nabla F_t$

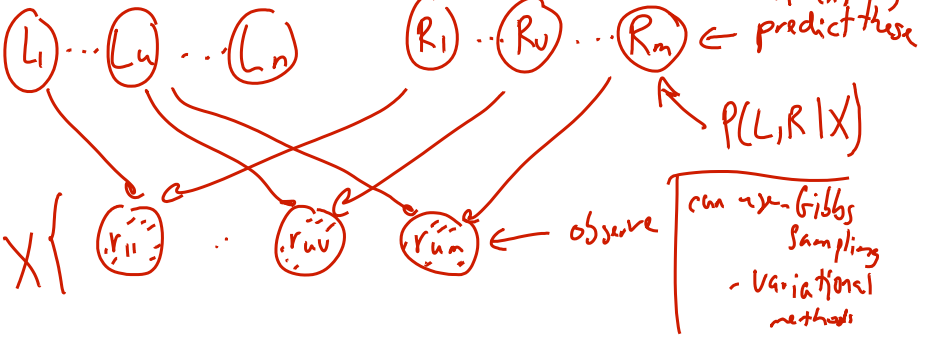
fast & easy to implement

What's Matrix Factorization Optimizing???

$P(L,R|X) \propto P(L,R,X) = P(L)\,P(R)\,P(X|L,R)$

- A generative process:
  - Pick user factors $L_{u1}, \ldots, L_{un} \Rightarrow L_{ui} \overset{iid}{\sim} N(0, \sigma_u^2)$ ← $P(L)$
  - Pick movie factors $R_{v1}, \ldots, R_{vk} \Rightarrow R_{vi} \overset{iid}{\sim} N(0, \sigma_v^2)$ ← $P(R)$
  - For each (user,movie) pair observed:
    - Pick rating as $L_u R_v$ + noise ← $N(0, \sigma_r^2) \Rightarrow r_{uv} | L,R \sim N(L_u \cdot R_v, \sigma_r^2)$

Latent Factor Model



Maximum A Posteriori for Matrix Completion

$$P(L,R|X) \propto P(L,R,X) = P(L)\,P(R)\,P(X|L,R)$$

$$\propto e^{\frac{-1}{2\sigma_u^2}\sum_{u=1}^{n}\sum_{i=1}^{k}L_{ui}^2} e^{\frac{-1}{2\sigma_v^2}\sum_{v=1}^{m}\sum_{i=1}^{k}R_{vi}^2} e^{\frac{-1}{2\sigma_r^2}\sum_{r_{uv}}(L_u\cdot R_v - r_{uv})^2}$$

$$\max_{L,R} \ell(L,R|X) = \log P(L,R|X)$$

$$= -\frac{1}{2\sigma_u^2}\sum_u\sum_i L_{ui}^2 - \frac{1}{2\sigma_v^2}\sum_v\sum_i R_{vi}^2 - \frac{1}{2\sigma_r^2}\sum_{r_{uv}}(L_u\cdot R_v - r_{uv})^2 + \text{const}$$

$$\min_{L,R} \frac{\lambda_u}{2}\|L\|_F^2 + \frac{\lambda_v}{2}\|R\|_F^2 + \frac{1}{2}\sum_{r_{uv}}(L_u\cdot R_v - r_{uv})^2$$

$$\lambda_u = \frac{\sigma_r^2}{\sigma_u^2} \qquad \lambda_v = \frac{\sigma_r^2}{\sigma_v^2}$$

MAP versus Regularized Least-Squares for Matrix Completion

- MAP under Gaussian Model:

$$P(L,R|X) \propto P(L,R,X)$$

$$\propto e^{\frac{-1}{2\sigma_u^2}\sum_{u=1}^{n}\sum_{i=1}^{k}L_{ui}^2} e^{\frac{-1}{2\sigma_v^2}\sum_{v=1}^{m}\sum_{i=1}^{k}R_{vi}^2} e^{\frac{-1}{2\sigma_r^2}\sum_{r_{uv}}(L_u\cdot R_v - r_{uv})^2}$$

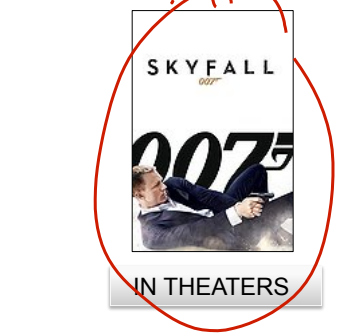
- Least-squares matrix completion with L_2 regularization:

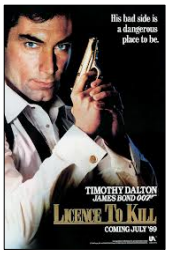
$$\min_{L,R} \frac{1}{2}\sum_{r_{uv}}(L_u\cdot R_v - r_{uv})^2 + \frac{\lambda_u}{2}||L||_F^2 + \frac{\lambda_v}{2}||R||_F^2$$

- Understanding as a probabilistic models is very useful! E.g.
  - Change priors
  - Incorporate other sources of information or dependencies

Cold-Start Problem

- **Challenge:** Cold-start problem (new movie or user)
- **Methods:** use features of movie/user





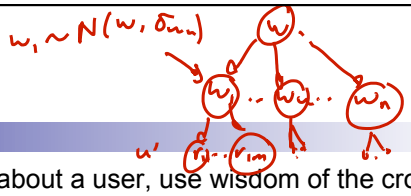
Cold-Start More Formally

- No observations about a particular user:

$$\min_{L,R} \frac{1}{2} \sum_{r_{uv}} (L_u \cdot R_v - r_{uv})^2 + \frac{\lambda_u}{2} ||L||_F^2 + \frac{\lambda_v}{2} ||R||_F^2$$

- A simpler model for collaborative filtering:
  - Observe ratings:
  - Given features of a movie:
  - Fit linear model:
  - Minimize:

Personalization



- If we don't have any observations about a user, use wisdom of the crowd
  - **Address cold-start problem**

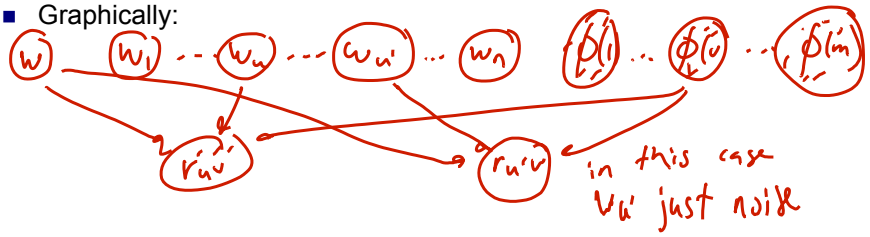
For user u', predict $r_{u'v} = w \cdot \phi(v)$

- But, as we gain more information about the user, forget the crowd:

parameter w_u per user, represents deviation from crowd prefs

$r_{uv} \sim N((w + w_u) \cdot \phi(v), \sigma_r^2)$

- Graphically:



in this case $w_{u'}$ just noise

User Features…

- In addition to movie features, may have information user:

$\phi(u) = ($ 25 (age), F (gender), MSc (education), A+ (grade in Big Data class) $)$

- Combine with features of movie:

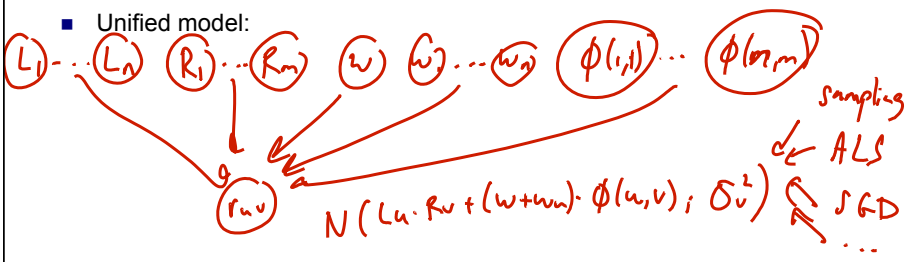
$\phi(u,v) = (\ldots \phi(u) \ldots, \ldots \phi(v) \ldots, \ldots$ cross features $\ldots)$

- Unified linear model:

$r_{uv} \sim N((w + w_u) \cdot \phi(u,v), \sigma_r^2)$

Feature-based Approach versus Matrix Factorization

- Feature-based approach:
  - Feature representation of user and movies fixed
  - Can address cold-start
- Matrix factorization approach:
  - Suffers from cold-start problem
  - User & movie features are learned from data
- Unified model:



MAP for Unified Collaborative Filtering via SGD

$$\min_{L,R,w,\{w_u\}_u} \frac{1}{2}\sum_{r_{uv}}(L_u \cdot R_v + (w+w_u)\cdot\phi(u,v) - r_{uv})^2 + \frac{\lambda_u}{2}||L||_F^2 + \frac{\lambda_v}{2}||R||_F^2 + \frac{\lambda_w}{2}||w||_2^2 + \frac{\lambda_{wu}}{2}\sum_u ||w_u||_2^2$$

- Gradient step observing $r_{uv}^{(t)}$
  - For L,R

$$\begin{bmatrix} L_u^{(t+1)} \\ R_v^{(t+1)} \end{bmatrix} \leftarrow \begin{bmatrix} (1-\eta_t\lambda_u)L_u^{(t)} - \eta_t\epsilon_t R_v^{(t)} \\ (1-\eta_t\lambda_v)R_v^{(t)} - \eta_t\epsilon_t L_u^{(t)} \end{bmatrix}$$

  - For w and w_u:

$$\begin{bmatrix} w^{(t+1)} \\ w_u^{(t)} \end{bmatrix} \leftarrow \begin{bmatrix} (1-\eta_t\lambda_w)w^{(t)} - \eta_t\epsilon_t\phi(u,v) \\ (1-\eta_t\lambda_{w_u})w_u^{(t)} - \eta_t\epsilon_t\phi(u,v) \end{bmatrix}$$

only update w_u for user u in $r_{uv}^{(t)}$

What you need to know…

- Probabilistic model for collaborative filtering
  - Models, choice of priors
  - MAP equivalent to optimization for matrix completion
- Cold-start problem
- Feature-based methods for collaborative filtering
  - Help address cold-start problem
- Unified approach