Case Study 4: Collaborative Filtering

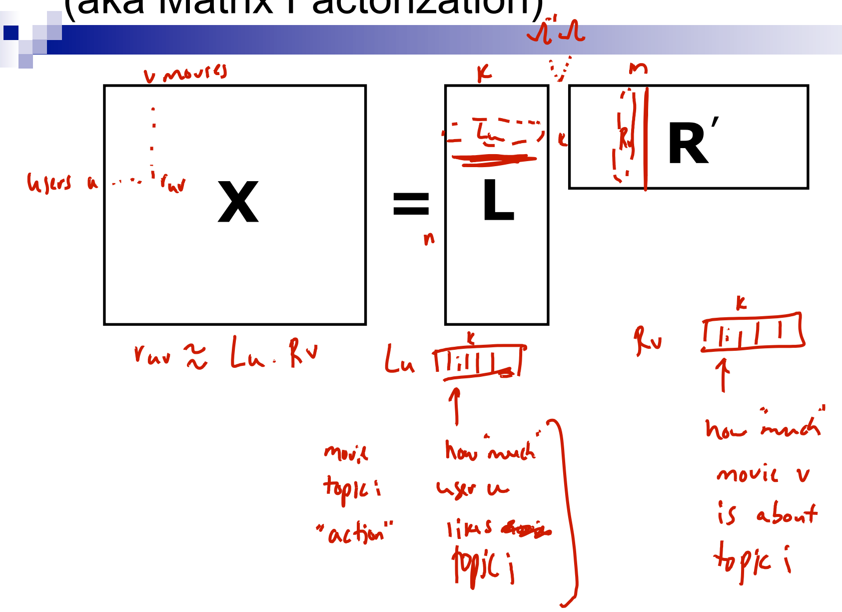
Probabilistic Models for Matrix Factorization

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

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

$\sum_u L_u \cdot L_u$

- Observe one rating at a time r_{uv}^t $\quad \varepsilon_t = L_u^{(t)} \cdot R_v^{(t)} - r_{uv}$
- Gradient observing r_{uv}:

$$\frac{\partial F}{\partial L_u} = \varepsilon_t R_v + \lambda_u L_u$$

$$\frac{\partial F}{\partial R_v} = \varepsilon_t L_u + \lambda_v R_v$$

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

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

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

✓ fast & easy to implement

What's Matrix Factorization Optimizing???

- A generative process:
  - Pick user factors
  - Pick movie factors
  - For each (user,movie) pair observed:
    - Pick rating as $L_u R_v$ + noise

Maximum A Posteriori for Matrix Completion

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

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:

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

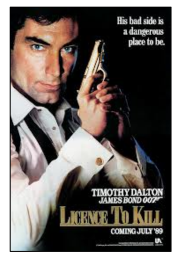
- Understanding as a probabilistic models is very useful! E.g.,
  - Change priors
  - Incorporate other sources of information or dependencies

Cold-Start Problem

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

$\phi(\text{Skyfall}) = \begin{array}{l}\text{action}\\ \text{romance}\\ \\ \vdots\end{array}\begin{pmatrix}7\\2\\0\\\vdots\end{pmatrix}$

$\phi(\text{FRWL}) = \begin{pmatrix}8\\1\\6\\\vdots\end{pmatrix}$



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

User Features…

- In addition to movie features, may have information user:
- Combine with features of movie:
- Unified linear model:

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}
  - 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:

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