Case Study 4: Collaborative Filtering

Review:
Cold Start Problem

Machine Learning for Big Data
CSE547/STAT548, University of Washington
Emily Fox

February 18th, 2014

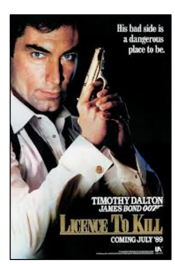
Cold-Start Problem

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

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

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

IN THEATERS



Cold-Start Problem More Formally

- Consider a new user u' and predicting that user's ratings
  - No previous observations

    $r_{u'v} = ?\quad \forall v$ want $L_{u'}$ to predict ratings for this user
  - Objective considered so far:

    doesn't depend on $L_{u'}$ (no obs. $r_{u'v}$)

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

all r_{uv}? (only obs. values)

only term appearing in ALS step for u'

  - Optimal user factor:

    $L_{u'} = 0$ only penalty term in opt. $L_{u'}$ $(\min ||L_{u'}||_2^2)$
  - Predicted user ratings: always predict $r_{u'v} = 0\ \forall v$ ... problem

An Alternative Formulation

- A simpler model for collaborative filtering
  - We would not have this issue if we assumed all users were identical

    - If all users shared a feature vector w
    - w informed by all ratings + can use it for new user u'

    shared
  - What about for new movies? What if we had side information?

    Create a movie feature vector: ↳ synopsis / info

    fixed movie features $\phi(v) = (\text{'action'}, 1994, \text{Tarantino}, \ldots)$ — genre, year, director — dim k
  - What dimension should w be? same length as movie feature vector
  - Fit linear model:

    For all users, u, $\boxed{r_{uv} \approx w \cdot \phi(v)}$ want to learn this
  - Minimize:

    only 1 param $\min_w \sum_{r_{uv}} (w \cdot \phi(v) - r_{uv})^2 + \lambda_w ||w||$ ← LS, LASSO

Personalization

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

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

- Clearly, not all users are the same ... shared w is strong assumption
- Just as in personalized click prediction, consider model with global and user-specific parameters

Consider user-specific deviations w_u from the crowd w

init. to 0

$$r_{uv} \approx \underbrace{(w + w_u)}_{\text{global preference vector}} \cdot \phi(v)$$

user deviation from the crowd

- As we gain more information about the user, forget the crowd

w_u more informed

User Features…

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

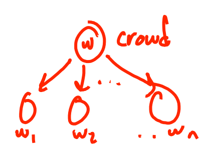
$$\phi(u) = (\underset{\text{age}}{25}, \underset{\text{gender}}{F}, \underset{\text{education}}{MSc}, \underset{\text{grade in Big Data}}{A^+}, \ldots)$$

- Combine with features of movie:

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

- Unified linear model:

$$r_{uv} \approx (w + w_u) \cdot \phi(u,v)$$



Feature-based Approach versus Matrix Factorization

- Feature-based approach:
  - Feature representation of user and movies fixed
  - Can address cold-start problem

informed by side info + chosen representation

- Matrix factorization approach:
  - Suffers from cold-start problem
  - User & movie features are learned from data L_u, R_v
- A unified model: combine both ideas

$$r_{uv} = L_u \cdot R_v + (w + w_u) \cdot \phi(u,v)$$

solve via:
ALS
SGD
⋮

Unified Collaborative Filtering via SGD

$$F = \min_{L,R,w,\{w_u\}_u} \frac{1}{2} \sum_{r_{uv}} (L_u \cdot R_v + (w + w_u) \cdot \phi(u,v) - r_{uv})^2$$

$$+ \frac{\lambda_u}{2}||L||_F^2 + \frac{\lambda_v}{2}||R||_F^2 + \frac{\lambda_w}{2}||w||_2^2 + \frac{\lambda_{wu}}{2}\sum_u ||w_u||_2^2$$

← unified model

- Gradient step observing r_{uv} $\quad \epsilon_t = L_u^{(t)} \cdot R_v^{(t)} + (w^{(t)} + w_u^{(t)}) \cdot \phi(u,v) - r_{uv}^{(t)}$
  - For L,R
$$\begin{bmatrix} L_u^{(t+1)} \\ R_v^{(t+1)} \end{bmatrix} \leftarrow \begin{bmatrix} (1-\eta_t\lambda_u)L_u^{(t)} - \eta_t\epsilon_t R_v^{(t)} \\ (1-\eta_t\lambda_v)R_v^{(t)} - \eta_t\epsilon_t L_u^{(t)} \end{bmatrix}$$
same as before w/ new defn of ϵ_t
  - For w and w_u: $\nabla_w F^{(t)} = \epsilon_t \phi(u,v) + \lambda_w w^{(t)}$

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

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

What you need to know…

- Cold-start problem
- Feature-based methods for collaborative filtering
  - Help address cold-start problem
- Unified approach

Case Study 4: Collaborative Filtering

Connections with Probabilistic Matrix Factorization

Machine Learning for Big Data
CSE547/STAT548, University of Washington
Emily Fox
February 18th, 2014

Matrix Completion Problem

$\mathbf{X} =$ 

X_{ij} known for black cells

X_{ij} unknown for white cells

Rows index users

Columns index movies

- Filling missing data?

well-approx. by rank k matrix

$X_{n \times m} \approx L_{n \times k} \, R^T_{k \times m}$

sparse

$n \cdot m$ params

filled in matrix θ

Soln for pred. θ: learn low-dim feature vecs for each user + movie

$\underbrace{k \cdot n + k \cdot m}_{\text{learn this}} \ll \underbrace{n \cdot m}_{\text{predict this}}$ (low-rank approx)

Coordinate Descent for Matrix Factorization: Alternating Least-Squares

$$\boxed{\min_{L,R} \sum_{(u,v):r_{uv}\neq ?} (L_u \cdot R_v - r_{uv})^2 + \lambda_u \|L\| + \lambda_v \|R\|}$$

- Fix movie factors, optimize for user factors
  - Independent least-squares over users

$$\min_{L_u} \sum_{v \in V_u} (L_u \cdot R_v - r_{uv})^2 + \lambda_u \|L\|$$

- Fix user factors, optimize for movie factors
  - Independent least-squares over movies

$$\min_{R_v} \sum_{u \in U_v} (L_u \cdot R_v - r_{uv})^2 + \lambda_v \|R\|$$

- System may be underdetermined: use regularization
- Converges to local optima

Probabilistic Matrix Factorization (PMF)

- A generative process:
  - Pick user factors $L_{u_1}, \ldots, L_{u_k}$ $\quad L_{u_i} \overset{iid}{\sim} N(0, \sigma_u^2) \leftarrow P(L)$ prior on user factors
  - Pick movie factors $R_{v_1}, \ldots, R_{v_k}$ $\quad R_{v_i} \overset{iid}{\sim} N(0, \sigma_v^2) \leftarrow P(R)$
  - For each (user,movie) pair observed:
    - Pick rating as $L_u \cdot R_v$ + noise $\rightarrow r_{uv} \mid L_u, R_v \sim N(L_u \cdot R_v, \sigma_r^2)$

      noise: $N(0, \sigma_r^2)$ $\qquad$ $P(X \mid L, R)$ likelihood
- Joint probability:

$$P(L,R,X) = P(L)\,P(R)\,P(X \mid L,R)$$

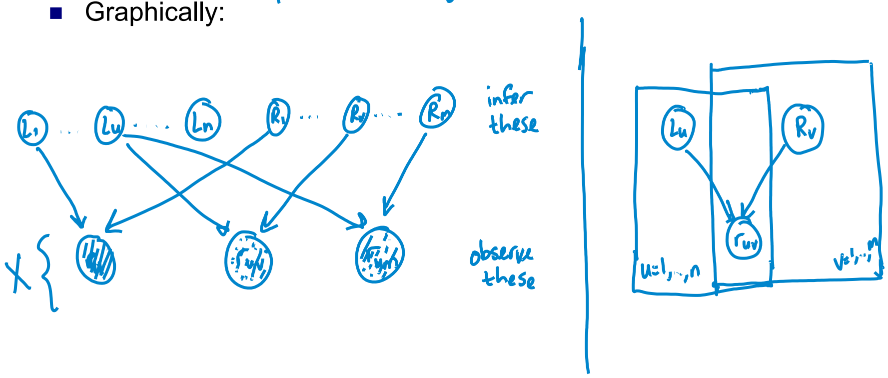
PMF Graphical Model

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

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

posterior $\propto$ joint prob.

- Graphically:



Maximum A Posteriori for Matrix Completion

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

var. mean obs

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

$$\max_{L,R} \log p(L,R|X) = -\frac{1}{2\sigma_u^2}\underbrace{\sum_u\sum_i L_{ui}^2}_{\|L\|_F^2} - \frac{1}{2\sigma_v^2}\underbrace{\sum_v\sum_i R_{vi}^2}_{\|R\|_F^2} - \frac{1}{2\sigma_r^2}\sum_{r_{uv}}(L_u \cdot R_v - r_{uv})^2 + \text{const}$$

$\lambda_u = \frac{\sigma_r^2}{\sigma_u^2}$ $\lambda_v = \frac{\sigma_r^2}{\sigma_v^2}$ $\Updownarrow$ multiply above by $-\sigma_r^2$

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

MAP versus Regularized Least-Squares for Matrix Completion

- MAP under Gaussian Model:

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

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

- Least-squares matrix completion with L_2 regularization:

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

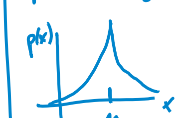
objectives are equivalent!

- Understanding as a probabilistic model is very useful! E.g.,
  - Change priors

$L_{u_i} \overset{iid}{\sim} N(0,\sigma_u^2)$, $R_{v_i} \overset{iid}{\sim} N(0,\sigma_v^2)$ } L_2 reg | $L_{u_i} \overset{iid}{\sim}$ Laplace$(0,\sigma_u)$, $R_{v_i} \overset{iid}{\sim}$ Laplace$(0,\sigma_v)$ } L_1 reg

Laplace(μ,b)

$p(x) \propto e^{\frac{-|x-\mu|}{b}}$



  - Incorporate other sources of information or dependencies

very key! many extensions of PMF

What you need to know…

- Probabilistic model for collaborative filtering
  - Models, choice of priors
  - MAP equivalent to optimization for matrix completion

Case Study 4: Collaborative Filtering

Gibbs Sampling for Bayesian Inference

Machine Learning for Big Data
CSE547/STAT548, University of Washington
Emily Fox
February 18th, 2014

Posterior Computations

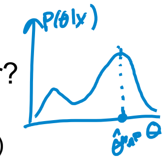
- MAP estimation focuses on point estimation:

$$\hat{\theta}^{MAP} = \arg\max_{\theta} p(\theta \mid x)$$

data

parameters

- What if we want a full characterization of the posterior?
  - Maintain a measure of uncertainty
  - Estimators other than posterior mode (different loss functions)
  - Predictive distributions for future observations



$$p(x^{N+1} \mid x^1,\dots,x^N) = \int p(x^{N+1} \mid \theta)\, p(\theta \mid x^1,\dots,x^N)\, d\theta$$ ← assuming x^i iid given θ (exch.)

belief about θ having seen obs. $x^1,\dots,x^N$

integrate over uncertainty in model params

contrast with:

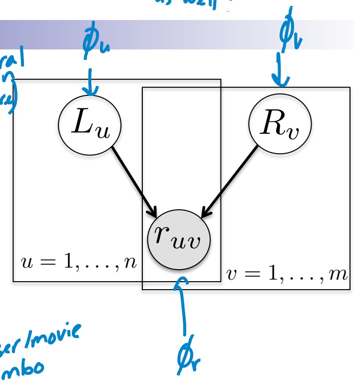
$p(x^{N+1} \mid \hat{\theta}^{MAP}(x^1,\dots,x^N))$ ← make pred w/ $\hat{\theta}^{MAP}$ after N obs.

- Often ~~no closed-form characterization~~ (e.g., mixture models, PMF, etc.)

Bayesian PMF Example

Full Bayesian approach: place priors on ϕ as well!

- Latent user and movie factors: (more general than before)
  - $L_u \sim N(\mu_u, \Sigma_u) \quad u=1,\dots,n$
  - $R_v \sim N(\mu_v, \Sigma_v) \quad v=1,\dots,m$



- Observations $r_{uv} \sim N(L_u' R_v, \sigma_r^2)$
- Hyperparameters:

$$\phi = \{\underbrace{\mu_u, \Sigma_u}_{\phi_u}, \underbrace{\mu_v, \Sigma_v}_{\phi_v}, \underbrace{\sigma_r^2}_{\phi_r}\}$$

- Want to predict new movie rating:

new user/movie combo

posterior given obs. so far

$$p(r_{uv}^* \mid X, \phi) = \int p(r_{uv}^* \mid L_u, R_v)\, p(L, R \mid X, \phi)\, dL\, dR$$

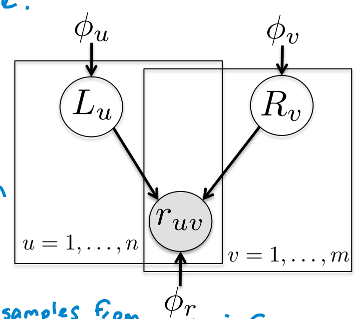
new rating; obs. ratings

Bayesian PMF Example

$$p(r_{uv}^* \mid X, \phi) = \int p(r_{uv}^* \mid L_u, R_v) p(L, R \mid X, \phi) dLdR$$

ANALYTICALLY INTRACTABLE!

- Monte Carlo methods:



Approx. as

$$p(r_{uv}^* \mid X, \phi) \approx \frac{1}{N} \sum_{k=1}^{N} p(r_{uv}^* \mid L_u^{(k)}, R_v^{(k)})$$

samples from posterior ... how?

- Ideally: $(L^{(k)}, R^{(k)}) \overset{iid}{\sim} p(L, R \mid X, \phi)$ ← ind samples from posterior

$$p(L, R \mid X) = \frac{p(X \mid L, R) p(L) p(R)}{p(X) = \int p(X \mid L, R) p(L) p(R) \, dL dR}$$

← issue!

Again, intractable :(

Bayesian PMF Example

- Want posterior samples $(L^{(k)}, R^{(k)}) \sim p(L, R \mid X, \phi)$
- What can we sample from?
  - Hint: Same reasoning as behind ALS, but sampling rather than maximization

What if we condition on R? Can we sample L?

Yes! And decomposes over users:

$$p(L \mid X, R, \phi) \propto p(X \mid L, R, \phi_r) \, p(L \mid \phi_u)$$

← Cond. on R

$$= \prod_{r_{uv} \neq ?} p(r_{uv} \mid L_u, R_v, \phi_r) \prod_{u=1}^{n} p(L_u \mid \phi_u)$$

$$= \prod_{u=1}^{n} \Big[\underbrace{p(L_u \mid \phi_u) \prod_{v \in V_u} p(r_{uv} \mid L_u, R_v, \phi_r)}_{\propto\, p(L_u \mid X, R, \phi)} \Big]$$

← breaks down over users

↑ all movies rated by user u

Bayesian PMF Example

- For user u:

prior ↙ likelihood for user u

$$p(L_u \mid X, R, \phi_u) \propto p(L_u \mid \phi_u) \prod_{v \in V_u} p(r_{uv} \mid L_u, R_v, \phi_r)$$

$$\propto N(L_u \mid \mu_u, \Sigma_u) \prod_{v \in V_u} N(r_{uv} \mid L_u \cdot R_v, \sigma_r^2)$$

$$= N(L_u \mid \tilde{\mu}_u, \tilde{\Sigma}_u)$$ ← via conjugacy

↖ posterior is in the same family as prior

where $\tilde{\Sigma}_u^{-1} = \Sigma_u^{-1} + \sigma_r^{-2} \sum_{v \in V_u} R_v R_v^T$

$\tilde{\mu}_u = \tilde{\Sigma}_u \left(\sigma_r^{-2} \sum_{v \in V_u} r_{uv} R_v + \Sigma_u \mu_u \right)$

- Symmetrically for R_v conditioned on L (breaks down over movies)
- Luckily, we can use this to get our desired posterior samples

Gibbs Sampling

← Type of Markov chain Monte Carlo (MCMC) approach

- Want draws: (generically for n params $\underline{\theta} = (\theta_1, \ldots, \theta_n)$)

$(\theta_1, \ldots \theta_n) \sim \pi(\underline{\theta})$ can't sample directly from π

e.g. $(L_1, \ldots, L_n, R_1, \ldots, R_m \mid X) \sim p(L, R \mid X)$

- Construct Markov chain whose steady state distribution is π
- Then, asymptotically correct … eventually samples from the Markov chain are samples from desired π
- Simplest case: (Gibbs)

For $k = 1, \ldots, N_{iter}$

for $i = 1, \ldots, n$ ← can use random order

$$\theta_i^{(k)} \sim p(\theta_i \mid \underbrace{\theta_1^{(k)}, \ldots, \theta_{i-1}^{(k)}, \theta_{i+1}^{(k-1)}, \ldots, \theta_n^{(k-1)}}_{\text{cond. on everything else}})$$

↑ Gibbs assumes this "full conditional" has a closed-form that we can sample from

Bayesian PMF Gibbs Sampler

- Outline of Bayesian PMF sampler

1. Init $L^{(1)}, R^{(1)}$
2. For $k=1,\dots,N_{iter}$
   (i) Sample hyperparams $\phi^{(k)} = \{\phi_u^{(k)}, \phi_v^{(k)}, \phi_r^{(k)}\}$
   (ii) For each user $u=1,\dots,n$ sample in parallel
   $$L_u^{(k+1)} \sim p(\underline{L_u} \mid X, R^{(k)}, \phi^{(k)})$$
   (iii) For each movie $v=1,\dots,m$ sample in parallel
   $$R_v^{(k+1)} \sim p(R_v \mid X, L^{(k+1)}, \phi^{(k)})$$

Very similar to ideas of ALS (systematically)

Bayesian PMF Results

From Salakhutdinov and Mnih, ICML 2008

- Netflix data with:
  - Training set = 100,480,507 ratings from 480,189 users on 17,770 movie titles
  - Validation set = 1,408,395 ratings.
  - Test set = 2,817,131 user/movie pairs with the ratings withheld.

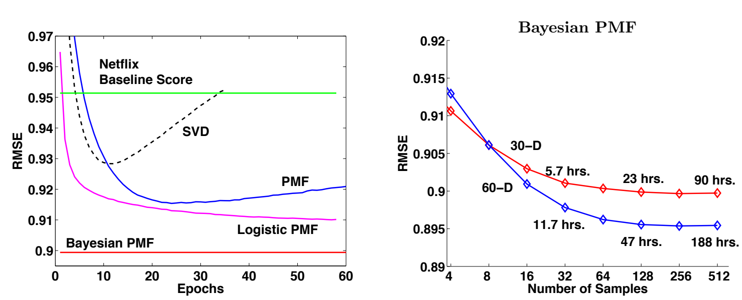


Figure 2. Left panel: Performance of SVD, PMF, logistic PMF, and Bayesian PMF using 30D feature vectors, on the Netflix validation data. The y-axis displays RMSE (root mean squared error), and the x-axis shows the number of epochs, or passes, through the entire training set. Right panel: RMSE for the Bayesian PMF models on the validation set as a function of the number of samples generated. The two curves are for the models with 30D and 60D feature vectors.

Bayesian PMF Results

From Salakhutdinov and Mnih, ICML 2008

- Bayesian model better controls for overfitting by averaging over possible parameters (instead of committing to one)

D	Valid. RMSE PMF	Valid. RMSE BPMF	% Inc.	Test RMSE PMF	Test RMSE BPMF	% Inc.
30	0.9154	0.8994	1.74	0.9188	0.9029	1.73
40	0.9135	0.8968	1.83	0.9170	0.9002	1.83
60	0.9150	0.8954	2.14	0.9185	0.8989	2.13
150	0.9178	0.8931	2.69	0.9211	0.8965	2.67
300	0.9231	0.8920	3.37	0.9265	0.8954	3.36

Table 1. Performance of Bayesian PMF (BPMF) and linear PMF on Netflix validation and test sets.

What you need to know…

- Idea of full posterior inference vs. MAP estimation
- Gibbs sampling as an MCMC approach
- Example of inference in Bayesian probabilistic matrix factorization model