Case Study 4: Collaborative Filtering

GraphLab Review,
ALS, SGD, Gibbs Sampling
for MF in GraphLab

Machine Learning for Big Data
CSE547/STAT548, University of Washington
Emily Fox
February 27th, 2014

©Emily Fox 2014

The GraphLab Framework

Graph Based
Data Representation

Update Functions
User Computation

Scheduler
Consistency Model

©Emily Fox 2014
What do I recommend???

Interpreting Low-Rank Matrix Completion (aka Matrix Factorization)

\[
X = LR'
\]

- \( X \): movies x users
- \( L \): movies x topics
- \( R' \): topics x users
- \( r_{uv} \): movie topic i
- \( Lu \): how much user u likes topic i
- \( R'v \): how much movie v is about topic i

©Emily Fox 2014
Matrix Completion as a Graph

\[ X = \]

\( X_{ij} \) known for black cells
\( X_{ij} \) unknown for white cells
Rows index movies
Columns index movies

Coordinate Descent for Matrix Factorization: Alternating Least-Squares

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

- 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
Alternating Least Squares Update Function

\[
\min_{L_u} \sum_{v \in V_u} (L_u \cdot R_v - r_{uv})^2 \quad \min_{R_v} \sum_{u \in U_u} (L_u \cdot R_v - r_{uv})^2
\]

SGD for Matrix Factorization in GraphLab

\[
\epsilon_t = L^{(t)}_u \cdot R^{(t)}_v - r_{uv} \\
L^{(t+1)}_u \\
R^{(t+1)}_v \\
\left[ \begin{array}{c} 
(1 - \eta_t \lambda_u) L^{(t)}_u - \eta_t \epsilon_t R^{(t)}_v \\
(1 - \eta_t \lambda_v) R^{(t)}_v - \eta_t \epsilon_t L^{(t)}_u 
\end{array} \right]
\]
Bayesian PMF Example

- Latent user and movie factors:
  \[ L_u \sim N(M_u, \Sigma_u), \quad u=1,\ldots,n \]
  \[ R_v \sim N(M_v, \Sigma_v), \quad v=1,\ldots,m \]

- Observations
  \[ r_{uv} \sim N(L_u R_v, \sigma^2) \]

- Hyperparameters:
  \[ \phi = \left\{ M_u, \Sigma_u, M_v, \Sigma_v, \sigma^2 \right\} \]

- Want to predict new movie rating:
  \[ p(r_{uv}^* | X, \phi) = \int p(r_{uv}^* | L_u, R_v) p(L_u, R_v | X, \phi) \, dL_u dR_v \]

Outline of Bayesian PMF Gibbs Sampler

- Outline of Bayesian PMF sampler
  1. Init. \( L^{(0)}, R^{(0)} \)
  2. For \( k=1,\ldots,N_{iter} \)
     (i) Sample hyperparams \( \phi^{(k)} \)
     (ii) For each user \( u=1,\ldots,n \) sample in parallel
         \[ L_u^{(k+1)} \sim p(L_u | X, R^{(k)}, \phi^{(k)}) \]
     (iii) For each movie \( v=1,\ldots,m \) sample in parallel
         \[ R_v^{(k+1)} \sim p(R_v | X, L^{(k+1)}, \phi^{(k)}) \]

Very similar to ideas of ALS (systematically)
Bayesian PMF Example

- For user u:
  \[ p(L_u \mid X, R, \phi_u) \propto p(L_u \mid \phi_u) \prod_{v \in V_u} p(r_{uv} \mid L_u, R_v, \phi_r) \]
  \[ = N(L_u \mid \tilde{\mu}_u, \tilde{\Sigma}_u) \prod_{v \in V_u} N(r_{uv} \mid L_u R_v, \sigma_v^2) \]
  where \( \tilde{\Sigma}_u = \Sigma_u^{-1} + \sigma_v^2 \Sigma \)
  \[ \tilde{\mu}_u = \tilde{\Sigma}_u \left( \sigma_v^2 \sum_{v \in V_u} r_{uv} R_v + \Sigma \mu_u \right) \]

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

PMF Gibbs Sampling in GraphLab

\[ p(L_u \mid X, R, \phi_u) = N(\mu_u, \Sigma_u) \]
\[ \Sigma_u = \Sigma_u^{-1} + \sigma_v^2 \sum_{v \in V_u} R_v^T R_v \]
\[ \mu_u = \Sigma_u \left( \sigma_v^2 \sum_{v \in V_u} r_{uv} R_v + \Sigma \mu_u \right) \]
GraphLab

Release 2.2 available now
http://graphlab.org

Documentation... Code... Tutorials... (more on the way)

GraphChi 0.1 available now
http://graphchi.org

What you need to know...

- Data-parallel versus graph-parallel computation
- Bulk synchronous processing versus asynchronous processing
- GraphLab system for graph-parallel computation
  - Data representation
  - Update functions
  - Scheduling
  - Consistency model
- ALS, SGD and Gibbs for matrix factorization/PMF in GraphLab
Reading

- Papers under “Case Study IV: Parallel Learning with GraphLab”

- Optional:
  - Parallel Splash BP

Acknowledgements

- Slides based on Carlos Guestrin’s GraphLab talk
Case Study 5: Mixed Membership Modeling

Clustering Documents Revisited, Latent Dirichlet Allocation

Machine Learning for Big Data
CSE547/STAT548, University of Washington
Emily Fox
February 27th, 2014

©Emily Fox 2014

Document Retrieval

- **Goal:** Retrieve documents of interest
- **Challenges:**
  - Tons of articles out there
  - How should we measure similarity?
Task 1: Find Similar Documents

- First considered:
  - **Input:** Query article
  - **Output:** Set of k similar articles

Task 2: Cluster Documents

- Then examined:
  - Cluster documents based on topic
**Document Representation**

- Bag of words model

**Bag of words model**

- Previously: vector of word counts (e.g. tf-idf)
- Performed operations on this vector

- Now: $x^d = \{w_1, \ldots, w_{N_d}\}$, unordered set of $N_d$ words
- $w_i \in \text{vocab}$

---

**A Generative Model**

- Documents: $x^1, \ldots, x^D$ with $x^d = \{w_1^d, \ldots, w_{N_d}^d\}$
- Associated topics: $z^1, \ldots, z^D$ with $z^d \in \{1, \ldots, K\}$
- Parameters: $\theta = \{\pi, \beta\}$

As before:

- $\Pi = [\pi_1, \ldots, \pi_K]$ topic probabilities
- $p_r(z^d = k) = \pi_k$

- $\beta = \{\beta_1, \ldots, \beta_K\}$
- $A_k$ prob. of word $v$

- $A_k \rightarrow \text{prob. of word } v$
A Generative Model

- Documents: $x_1, \ldots, x_D$
- Associated topics: $z_1, \ldots, z_D$
- Parameters: $\theta = \{\pi, \beta\}$
- Generative model:

\[
\begin{align*}
    z^d &\sim \pi \quad \text{generate topic} \\
    w_i^d &\mid z^d \sim \beta^z_k \quad i=1, \ldots, N_d \\
\end{align*}
\]

Given topic $z^d:k$ for doc $d$, draw each word from $\beta^z_k$

Model In Pictures

- Mixture weights (on topics)

\[\pi\]

- Topic distributions (on words)

\[\beta_1, \ldots, \beta_K\]

- For each document,

\[
\begin{align*}
    z^d &\sim \pi \\
    w_i^d &\mid z^d \sim \beta^z_k \\
\end{align*}
\]

©Emily Fox 2014
Bayesian Document Model

- Model parameters $\pi, \{\beta_k\}$ unknown
- Bayesian approach

- Need distribution on pmf’s

The Simplex in 3D

- The simplex defines the hyperplane of vectors that sum to 1

$$0 \leq \theta_k \leq 1 \quad \sum_{k=1}^{3} \theta_k = 1$$
Dirichlet Distributions

The Dirichlet distribution is defined on the simplex

\[ p(\pi | \alpha) = \frac{\Gamma(\sum_k \alpha_k)}{\prod_k \Gamma(\alpha_k)} \prod_{k=1}^{K} \pi_k^{\alpha_k-1} \]

Moments:

\[ E_{\alpha_k}[\pi_k] = \frac{\alpha_k}{\alpha_0} \]

\[ Var_{\alpha_k}[\pi_k] = \frac{K - 1}{K^2(\alpha_0 + 1)} \]

Dirichlet Probability Densities

\[ \pi \sim \text{Dir}(\alpha_1, \ldots, \alpha_K) \]

\[ \pi \sim \text{Dir}(1,1,1) \]

\[ \pi \sim \text{Dir}(4,4,4) \]

\[ \pi \sim \text{Dir}(4,9,7) \]

\[ \pi \sim \text{Dir}(0.2,0.2,0.2) \]

\[ \pi \sim \text{Dir}(0.1,0.1,0.1) \]

\[ \pi \sim \text{Dir}(1/2,1/2,0) \]

\[ \pi \sim \text{Dir}(1/4,1/4,1/2) \]

\[ \pi \sim \text{Dir}(1/2,1/2,0) \]
Dirichlet Samples

- Samples are **sparse** for small values of $\alpha_i$

\[
\mathbb{E}_{\alpha}[\pi_k] = \frac{\alpha_k}{\alpha_0}
\]

### Model Summary

- **Prior on model parameters**
  - E.g., symmetric Dirichlet for $\pi$

- Dirichlet prior for topic parameters $\beta_k$

- **Sample observations as**
  \[
  z^d \sim \pi \\
  w^d_i \mid z^d \sim \beta_{z^d}
  \]
Posterior Inference via Sampling

- Iterate between sampling

What form do these complete conditionals take?
- First a look at statements of conditional independence in directed graphical models

Conditional Independence in Bayes Nets

- Consider 4 different junction configurations

Conditional versus unconditional independence:
Bayes Ball Algorithm

- Consider 4 different junction configurations

A node is conditionally independent of all other nodes in the graph given its Markov blanket

- Gibbs sampling iterates between full conditionals

→ simplify to
Recall that the plate notation is really indicating

**Complete Conditional for** $\pi$

- Recall conjugate Dirichlet prior
  
  $\pi \sim \text{Dir}(\alpha_1, \ldots, \alpha_K)$
  
  $p(\pi | \alpha) = \frac{\Gamma(\sum_k \alpha_k)}{\prod_k \Gamma(\alpha_k)} \prod_k \alpha_k^{\alpha_k - 1}$

- Likelihood:
  - Dirichlet posterior
    - Count occurrences of
    - Then,

  - Conjugacy: **Posterior** has same form as **prior**
Complete Conditional for $\beta_k$:

- Again, Dirichlet prior
- Consider docs $d$ such that
  - For these observations,
  - Do any other docs depend on $\beta_k$?
- Then,

  

Complete Conditional for $z^d$:

- We have $z^d \sim \pi$
- $w_i^d \mid z^d, \{\beta_k\} \sim \beta z^d$
- Calculate the posterior for each value of $z^d$ ("responsibility" of each topic to the doc):

  $$r_{dk} = p(z^d = k \mid \{w_i^d\}, \pi, \beta) = \frac{\pi_k p(\{w_i^d\} \mid \beta_k)}{\sum_j \pi_j p(\{w_i^d\} \mid \beta_j)}$$
- Sample each cluster indicator as
Task 3: Mixed Membership Models

Now: Document may belong to multiple clusters

Latent Dirichlet Allocation (LDA)
But we only observe the documents; the other structure is hidden.

We compute the posterior $p$.
LDA Generative Model

- Observations: \( w^d_1, \ldots, w^d_{N_d} \)
- Associated topics: \( z^d_1, \ldots, z^d_{N_d} \)
- Parameters: \( \theta = \{\{\pi^d\}, \{\beta_k\}\} \)
- Generative model:

\[
p(\cdot) = \prod_{k=1}^{K} p(\beta_k | \lambda) \prod_{d=1}^{D} p(\pi^d | \alpha) \left( \prod_{i=1}^{N_d} p(z^d_i | \pi^d) p(w^d_i | z^d_i, \beta) \right)
\]
Example Inference – Topic Weights

- **Data:** The OCR’ed collection of *Science* from 1990-2000
  - 17K documents
  - 11M words
  - 20K unique terms (stop words and rare words removed)
- **Model:** 100-topic LDA model

Example Inference – Topic Words

- human
- genome
- dna
- genetic
- genes
- sequence
- gene
- molecular
- sequencing
- map
- information
- genetics
- mapping
- project
- sequences
- evolution
- evolutionary
- species
- organisms
- life
- origin
- biology
- groups
- phylogenetic
- living
- diversity
- group
- new
- united
- common
- disease
- host
- bacteria
- diseases
- resistance
- bacterial
- new
- infectious
- malaria
- parasite
- parasites
- two
- united
- united
- tuberculosis
- simulations
- computer
- models
- information
- data
- computers
- network
- systems
- model
- methods
- networks
- software
- new

*Sweeping down: Computer analysis plots are so rich, they make many of the advanced theories and ancient genetics obsolete.*

(C) Emily Fox 2014
What you need to know…

- Bayesian specification of document clustering model
- Rules of conditional and unconditional independence in directed graphical models (Bayes nets)
  - Bayes’ ball
  - Markov blanket
- Gibbs sampling for Bayesian document model
- Latent Dirichlet allocation (LDA) motivation and generative model specification

Reading

- Mixed Membership Models: KM Sec. 27.3
Acknowledgements

- Thanks to Dave Blei for some of the LDA material