Missing Data Problems and Collaborative Filtering

CSC2535

Richard Zemel & Ben Marlin
Department of Computer Science
University of Toronto
March 23, 2011
Contents:

Introduction

Notation

Theory Of Missing Data
  Factorizations
  MCAR
  MAR
  NMAR
  Inference and Learning

Multinomial Models
  Multinomial Mixture
  Multinomial Mixture/CPT-v
  Demo

Collaborative Filtering
  Collaborative Prediction
  Yahoo! Data
  Jester Data
  Results
Introduction
Introduction

Collaborative filtering – users assign ratings to items → system uses information from all users to recommend previously unseen items that a user might like

One approach to recommendation: predict ratings for all unrated items, recommend highest predicted ratings

Critical assumption: missing ratings are missing at random

One way to violate: value of variable affects probability value will be missing – bias in observed ratings, and hence learned parameters

Also complementary bias in standard testing procedure – distribution of observed data different from distribution of complete data, so estimated error on observed test data poor estimate of complete data error
Introduction: Survey Sampling Example

Data Variables

R
Did the respondent answer the question?

X
Answers to questions.
**Introduction:** Medical Diagnosis Example

Latent Variables

![Image of a doctor] (Z) Diseases

Data Variables

Symptoms

Response Variables

Was a test performed to check for the symptom?
Introduction: Recommender Systems Example

Latent Variables

Data Variables

Response Variables

Z  Preferences and Tastes

X  Ratings or Purchase History

R  Did the user rate or buy the item?
## Introduction: Basic Notation

<table>
<thead>
<tr>
<th>Symbol</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>$N$</td>
<td>Number of data cases.</td>
</tr>
<tr>
<td>$D$</td>
<td>Number of data dimensions.</td>
</tr>
<tr>
<td>$C$</td>
<td>Number of classes.</td>
</tr>
<tr>
<td>$V$</td>
<td>Number of multinomial values.</td>
</tr>
<tr>
<td>$K$</td>
<td>Number of clusters or hidden units.</td>
</tr>
</tbody>
</table>
## Introduction: Notation for Missing Data

<p>| | | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>$x_n$</td>
<td>0.1</td>
<td>0.9</td>
<td>0.2</td>
<td>0.7</td>
<td>0.3</td>
</tr>
<tr>
<td>$r_n$</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>$o_n$</td>
<td>1</td>
<td>4</td>
<td>5</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$m_n$</td>
<td>2</td>
<td>3</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$x_n^o$, $x_n^o$</td>
<td>0.1</td>
<td>0.7</td>
<td>0.3</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$x_n^m$, $x_n^m$</td>
<td>0.9</td>
<td>0.2</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Data Vector</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Response Vector</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Observed Dimensions</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Missing Dimensions</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Observed Data</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Missing Data</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Contents:

Introduction

Notation

Theory Of Missing Data
  Factorizations
  MCAR
  MAR
  NMAR
  Inference and Learning

Multinomial Models
  Multinomial Mixture
  Multinomial Mixture/CPT-v
  Demo

Collaborative Filtering
  Collaborative Prediction
  Yahoo! Data
  Jester Data
  Results
Theory of Missing Data: Generative Process

Latent Variables

Data Variables

Response Variables

Data Model

Missing Data Model
Theory of Missing Data: Factorizations

Data/Selection Model Factorization:

\[ P(x, r, z|\theta, \mu) = P(r|x, z, \mu)P(x, z|\theta) \]

• The probability of selection depends on the true values of the data variables and latent variables.

Pattern Mixture Model Factorization:

\[ P(x, r, z|\vartheta, \nu) = P(x, z|r, \vartheta)P(r|\nu) \]

• Each response vector defines a different pattern, and each pattern has a different distribution over the data.
Theory of Missing Data: Classification

Missing Completely at Random:

- Response probability is independent of data variables and latent variables.

\[ P(r | x, z, \mu) = P(r | \mu) \]

MCAR Examples:

Send questionnaires to a random subset of the population or use random digit dialing.
Theory of Missing Data: Classification

Missing at Random:

• Typically written in a short-hand form that looks like a statement of probabilistic independence:

\[ P(r|x, z, \mu) = P(r|x^o, \mu) \]

• MAR is actually a different type of condition that requires a particular set of symmetries hold in \( P(r|x,z,\mu) \):

\[ P(r|x^o(r), x^m(r), Z, \mu) = f(r, x^o(r), \mu) \quad \text{... for all } x^m \]
Theory of Missing Data: Classification

Missing at Random Examples:

Respondents are not required to provide information about their employer if they are not currently employed.

Doctor only orders test B if the result of test A was negative. If result of test A is positive, result for test B is missing.
**Theory of Missing Data: Classification**

What Does it mean to be Missing at Random?

- MAR is *not* a statement of independence between random variables. MAR requires that particular symmetries hold so that $P(R=r|X=x)$ can be determined from observed data only.

<table>
<thead>
<tr>
<th>$X \backslash R$</th>
<th>0 0</th>
<th>0 1</th>
<th>1 0</th>
<th>1 1</th>
</tr>
</thead>
<tbody>
<tr>
<td>0 0</td>
<td>$\alpha$</td>
<td>$\beta$</td>
<td>$\gamma$</td>
<td>$1 - \alpha - \beta - \gamma$</td>
</tr>
<tr>
<td>0 1</td>
<td>$\alpha$</td>
<td>$\delta$</td>
<td>$\gamma$</td>
<td>$1 - \alpha - \delta - \gamma$</td>
</tr>
<tr>
<td>1 0</td>
<td>$\alpha$</td>
<td>$\beta$</td>
<td>$\lambda$</td>
<td>$1 - \alpha - \beta - \lambda$</td>
</tr>
<tr>
<td>1 1</td>
<td>$\alpha$</td>
<td>$\delta$</td>
<td>$\lambda$</td>
<td>$1 - \alpha - \delta - \lambda$</td>
</tr>
</tbody>
</table>
Theory of Missing Data: Classification

Not Missing at Random:

• Allows for arbitrary dependence of response probabilities on missing data values and latent variables:

\[ P(r | x, z, \mu) \] No Simplifications

An Easy Way to Violate MAR:

• Let the probability that a data variable is observed depend on the value of that data variable.
Theory of Missing Data: Classification

Not Missing at Random Examples:

Snowfall reading is likely to be missing if weather station is covered with snow.

Participants in a longitudinal health study for a heart medication may die of a heart attack during the study.

Users are more likely to rate or buy items they like than items they don’t like.
Theory of Missing Data: Inference

MCAR/MAR Posterior:

\[
P(\theta|x^o, r) \propto \int \int \int P(x, z|\theta)P(r|x, z, \mu)P(\theta|\omega)P(\mu|\eta)d\mu dZ dx^m
\]

\[
\propto \int f(r, x, \mu)P(\mu|\eta)d\mu \cdot \int \int P(x, z|\theta)P(\theta|\omega)dZ dx^m
\]

\[
\propto P(x^o|\theta)P(\theta|\omega)
\]

• When MCAR or MAR holds, the posterior can be greatly simplified. Inference for \(\theta\) does not depend on \(r, \mu,\) or \(\eta\). The missing data can be ignored.
Theory of Missing Data: Inference

NMAR Posterior:

\[ P(\theta|x^o, r) \propto \int \int \int P(x, z|\theta)P(r|x, z, \mu)P(\theta|\omega)P(\mu|\eta) d\mu dZ dx^m \]

- When MAR fails to hold, the posterior does not simplify.

- Basing inference on the observed data posterior and ignoring the missing data model leads to provably biased inference for data model parameters.
Contents:

Introduction

Notation

Theory Of Missing Data
  Factorizations
  MCAR
  MAR
  NMAR
  Inference and Learning

Multinomial Models
  Multinomial Mixture
  Multinomial Mixture/CPT-v
  Demo

Collaborative Filtering
  Collaborative Prediction
  Yahoo! Data
  Jester Data
  Results
Multinomial Models: Mixture

Probability Model:

\[
P(Z_n = k | \theta) = \theta_k
\]

\[
P(X_n = x_n | Z_n = k, \beta) = P(x_n | \beta_k)
\]

\[
P(\theta, \beta | \alpha, \phi) = P(\theta | \alpha) \prod_k P(\beta_k | \phi)
\]

Properties:

• Allows for a fixed, finite number of clusters.

• In the multinomial mixture, \( P(x_n | \beta_k) \) is a product of discrete distributions. The prior on \( \beta \) and \( \theta \) is Dirichlet.
Multinomial Models: Mixture

Dirichlet Distribution:

Bayesian mixture modeling becomes much easier when conjugate priors are used for the model parameters. The conjugate prior for the mixture proportions \( \theta \) is the Dirichlet distribution.

\[
P(\theta | \alpha) = \frac{\Gamma \left( \sum_k \alpha_k \right)}{\prod_k \Gamma (\alpha_k)} \prod_k \theta_k^{\alpha_k - 1}
\]

\[
E[\theta_k | \alpha] = \frac{\alpha_k}{\sum_{k=1}^K \alpha_k}
\]

\[
P(\theta | \alpha, z) = \frac{\Gamma (N + \sum_k \alpha_k)}{\prod_k \Gamma (C_k + \alpha_k)} \prod_k \theta_k^{C_k + \alpha_k - 1}
\]
Multinomial Models: Mixture

MAP EM Algorithm:

E-Step: \( q_n(k) \leftarrow \frac{\theta_k \prod_{d=1}^{D} \prod_{v=1}^{V} \beta_{vdk}[r_{dn}=1][x_{dn}=v]}{\sum_{k'=1}^{K} \theta_{k'} \prod_{d=1}^{D} \prod_{v=1}^{V} \beta_{vdk'}} \)

M-Step: \( \theta_k \leftarrow \frac{\alpha_k - 1 + \sum_{n=1}^{N} q_n(k)}{N - K + \sum_{k=1}^{K} \alpha_k} \)

\( \beta_{vdk} \leftarrow \frac{\phi_{vdk} - 1 + \sum_{n=1}^{N} q_n(k)[r_{dn}=1][x_{dn}=v]}{\sum_{n=1}^{N} q_n(k)[r_{dn}=1] - V + \sum_{v=1}^{V} \phi_{vdk}} \)
Multinomial Models: Mixture/CPT-v

Probability Model:

\[ P(\theta | \alpha) = \mathcal{D}(\theta | \alpha) \]

\[ P(\beta | \phi) = \prod_{k=1}^{K} \prod_{d=1}^{D} \mathcal{D}(\beta_{dk} | \phi_{dk}) \]

\[ P(Z_n = k | \theta) = \theta_k \]

\[ P(\mathbf{X} = \mathbf{x}_n | Z_n = k, \beta) = \prod_{d=1}^{D} \prod_{v=1}^{V} \beta_{vdk}^{[x_{dn} = v]} \]

\[ P(\mu | \xi) = \prod_{v} \mathcal{B}(\mu_v | \xi_v) \]

\[ P(\mathbf{R} = \mathbf{r}_n | \mathbf{X} = \mathbf{x}_n, \mu) = \prod_{d=1}^{D} \prod_{v=1}^{V} \mu_v^{[r_{dn} = 1][x_{dn} = v]} (1 - \mu_v)^{[r_{dn} = 0][x_{dn} = v]} \]
Multinomial Models: Mixture/CPT-v

MAP EM Algorithm (E-Step):

\[ q_n(k) = P(z_n = k|x_n^o, r_n, \theta, \beta, \mu) \]
\[ = \frac{\theta_k \prod_{d=1}^D (\prod_{v=1}^V (\beta_{vd} \mu_v)[x_{dn}=v])^{[r_{dn}=1]} \left( \sum_{v=1}^V \beta_{vd} (1 - \mu_v) \right)^{[r_{dn}=0]}}{\sum_{k=1}^K \theta_k \prod_{d=1}^D (\prod_{v=1}^V (\beta_{vd} \mu_v)[x_{dn}=v])^{[r_{dn}=1]} \left( \sum_{v=1}^V \beta_{vd} (1 - \mu_v) \right)^{[r_{dn}=0]}} \]

\[ q_n(k, v, d) = P(z_n = k, x_{dn} = v|x_n^o, r_n, \theta, \beta, \mu) \]
\[ = q_n(k) \left( \frac{\mu_v \beta_{vd}}{\sum_{v'=1}^V \mu_{v'} \beta_{v'd}} \right)^{[r_{dn}=1]} \left( \frac{(1 - \mu_v) \beta_{vd}}{\sum_{v'=1}^V (1 - \mu_{v'}) \beta_{v'd}} \right)^{[r_{dn}=0]} \]
**Multinomial Models: Mixture/CPT-v**

**MAP EM Algorithm (M-Step):**

\[
\theta_k = \frac{\alpha_k - 1 + \sum_{n=1}^{N} q_n(k)}{N - K + \sum_{k=1}^{K} \alpha_k}
\]

\[
\beta_{vdk} = \frac{\phi_{vdk} - 1 + \sum_{n=1}^{N} q_n(k) [r_{dn} = 1][x_{dn} = v] + q_n(k, v, d)[r_{dn} = 0]}{\sum_{n=1}^{N} q_n(k) - V + \sum_{v=1}^{V} \phi_{vdk}}
\]

\[
\mu_v = \frac{\xi_{1v} - 1 + \sum_{n=1}^{N} \sum_{d=1}^{D} [r_{dn} = 1][x_{dn} = v]}{\xi_{1v} + \xi_{0v} - 2 + \sum_{n=1}^{N} \sum_{d=1}^{D} [r_{dn} = 1][x_{dn} = v] + q_n(v, d)[r_{dn} = 0]}
\]
Other Models for Missing Data:

- K-Nearest Neighbors
- Probabilistic Principal Components Analysis
- Factor Analysis
- Mixtures of Gaussians
- Mixtures of PPCA/FA
- Probabilistic Matrix Factorization
- Maximum Margin Matrix Factorization
- Conditional Restricted Boltzmann Machines
Contents:

Introduction

Notation

Theory Of Missing Data
  Factorizations
  MCAR
  MAR
  NMAR
  Inference and Learning

Multinomial Models
  Multinomial Mixture
  Multinomial Mixture/CPT-v
  Demo

Collaborative Filtering
  Collaborative Prediction
  Yahoo! Data
  Jester Data
  Results
Collaborative Filtering: Collaborative Prediction Problem

<p>| | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td><img src="star1.png" alt="Star Rating" /> 🌟🌟🌟🌟</td>
<td>?</td>
<td>?</td>
<td>🌟🌟🌟🌟🌟</td>
<td>🌟🌟🌟🌟🌟</td>
</tr>
<tr>
<td><img src="star1.png" alt="Star Rating" /> ?</td>
<td>🌟🌟🌟🌟🌟</td>
<td>?</td>
<td>?</td>
<td>?</td>
</tr>
<tr>
<td><img src="star1.png" alt="Star Rating" /> 🌟🌟🌟🌟</td>
<td>?</td>
<td>🌟🌟🌟🌟🌟</td>
<td>?</td>
<td>?</td>
</tr>
<tr>
<td><img src="star1.png" alt="Star Rating" /> 🌟🌟🌟🌟</td>
<td>?</td>
<td>🌟🌟🌟🌟🌟</td>
<td>🌟🌟🌟🌟🌟</td>
<td>?</td>
</tr>
</tbody>
</table>
Collaborative Filtering: Yahoo!

Data was collected through an online survey of Yahoo! Music LaunchCast radio users.

- 1000 songs selected at random.
- Users rate 10 songs selected at random from 1000 songs.
- Answer 16 questions.
- Collected data from 35,000+ users.

Image copyright Yahoo! Inc. 2006. Used with permission.
Collaborative Filtering: Yahoo!

\[ P_{i} P_{E}^{U}(x = v | r = 1) \cdot \frac{\sum_{i} \sum_{m} \left[ \underbrace{P_{E}^{R}(x = v | r = 1)}_{\text{Randomly Selected}} \right] - 1}{\sum_{i} \sum_{m} \left[ \underbrace{P_{i} m - 1}_{\text{User Selected}} \right]} \]
More Empirical Distributions
Collaborative Filtering: Jester

Jester gauge set of 10 jokes used as complete data. Synthetic missing data was added.

- 15,000 users randomly selected
- Missing data model: $\mu_v(s) = s(v-3) + 0.5$
Experimental Protocol

Randomly partition users into 5 blocks of 1080 users

Three sets of ratings:

1. Observed ratings – all but one of original ratings
2. Test ratings for user-selected – remaining one
3. Test ratings for randomly-selected – ten survey responses

User-selected items – same distribution as observed
Randomly selected test items -- MCAR
Experimental Protocol

Weak Generalization

• Learn on training user observed ratings
• Evaluate on training user test ratings

Strong Generalization

• Learn on training user observed ratings
• Evaluate on test user test ratings
Data Sets: User Splits

- **User Selected Ratings**
  - **Test Users**: 1, 5, 4
  - **Train Users**: 2, 3, 1, 4
  - **Holdout Users**: 4, 3, 1

- **Randomly Selected Ratings**
  - **Test Users**: 3
  - **Train Users**: 4, 1
  - **Holdout Users**: 5
Data Sets: User Splits

User Selected Ratings

Randomly Selected Ratings

Test Ratings for User Selected Items

Observed Ratings

Test Ratings for Randomly Selected items
Collaborative Filtering: Results

Jester Results: MM vs MM/CPT-v

Weak MAE vs Effect Strength

Strong MAE vs Effect Strength
Collaborative Filtering: Results

Yahoo! Results: MM vs MM/CPT-v+
Collaborative Filtering: Results
Comparison of Results on Yahoo! Data

<table>
<thead>
<tr>
<th>Method</th>
<th>Complexity</th>
<th>Rand MAE</th>
<th>User MAE</th>
</tr>
</thead>
<tbody>
<tr>
<td>EM MM</td>
<td>1</td>
<td>0.7725 ± 0.0024</td>
<td>0.7626 ± 0.0077</td>
</tr>
<tr>
<td>EM MM/CPT-v</td>
<td>20</td>
<td>0.5431 ± 0.0012</td>
<td>0.6631 ± 0.0026</td>
</tr>
<tr>
<td>EM MM/Logit</td>
<td>5</td>
<td>0.5038 ± 0.0030</td>
<td>0.7029 ± 0.0186</td>
</tr>
<tr>
<td>EM MM/CPT-v+</td>
<td>5</td>
<td>0.4456 ± 0.0033</td>
<td>0.7235 ± 0.0059</td>
</tr>
<tr>
<td>MCMC DP</td>
<td>N/A</td>
<td>0.7624 ± 0.0063</td>
<td>0.5767 ± 0.0077</td>
</tr>
<tr>
<td>MCMC DP/CPT-v</td>
<td>N/A</td>
<td>0.5549 ± 0.0026</td>
<td>0.6670 ± 0.0071</td>
</tr>
<tr>
<td>MCMC DP/CPT-v+</td>
<td>N/A</td>
<td>0.4428 ± 0.0027</td>
<td>0.7537 ± 0.0026</td>
</tr>
<tr>
<td>CD cRBM</td>
<td>20</td>
<td>0.7179 ± 0.0025</td>
<td>0.5421 ± 0.0081</td>
</tr>
<tr>
<td>CD cRBM-v</td>
<td>1</td>
<td>0.4553 ± 0.0031</td>
<td>0.7501 ± 0.0066</td>
</tr>
</tbody>
</table>
Application to Ranking

\[
NDCG(n) = \frac{\sum_{i=1}^{T} \frac{2^{x_{ni}^t} - 1}{\log(1 + \pi(i, n))}}{\sum_{i=1}^{T} \frac{2^{x_{ni}^t} - 1}{\log(1 + \pi(i, n))}}
\]

- \(\hat{x}_{ni}^t\) : mean of posterior predictive distribution for test item i.
- \(\pi(i, n)\) : rank of test item i according to \(\hat{x}_{ni}^t\).
- \(\pi(i, n)\) : rank of test item i according to \(x_{ni}^t\).
Ranking Results

Higher Accuracy

Longer Recommendation Lists