Last Time

MLE for incomplete data.
MLE no longer factors

Parameter Estimation Optimization
Gradient ascent
EM algorithm

This Time

Promised, but will not deliver a PS 5 discussion
(Forgot to take KBs home)

Incomplete Data
EM Algorithm II
Exact solution for multinomial missing data
Bayesian inference

Finding the best structure  (Maybe)
Scoring metrics
Structure optimization

These slides draw heavily on Friedman+Goldszmidt’s 1998 tutorial
Last Time: EM

Presented the Neal+Hinton [93] formulation.\[ F(Q,\Theta) = \int Q(X) \log P(D, X \mid \Theta) dX - \int Q(X) \log Q(X) dX \]

Makes convergence obvious.

**E-step** \[ Q_{k+1} \leftarrow \arg \max_Q F(Q, \Theta_k) \]

**M-step** \[ \Theta_{k+1} \leftarrow \arg \max_{\Theta} F(Q_{k+1}, \Theta) \]

---

Last Time: Prove the M-Step of EM

\[ \Theta_{k+1} \leftarrow \arg \max_{\Theta} \int_X P(X \mid D, \Theta_k) \log P(X \mid D, \Theta) dX \]

\[ = \int_X P(X \mid D, \Theta_k) \sum_i \log L_j[\Theta_j \mid S_i(D, X)] dX \]

\[ = \sum_i \left( \int_X P(X \mid D, \Theta_k) \log L_j[\Theta_j \mid S_i(D, X)] dX \right) \]

\[ = \sum_i E_{X_{i\mid D, \Theta_k}} \log L_j[\Theta_j \mid S_i(D, X)] \]

\[ \leq \sum_i \log E_{X_{i\mid D, \Theta_k}} [S_i(X)] \]

**Implication:** Select \( \Theta \) that optimizes \( L[\Theta \mid D, X] \) for *expected sufficient statistics*.
Exact Solution

Single incomplete case.
Dirichlet prior.
“Parameter independence” assumption.
H is unobserved, Y is observed.

$$P\{\Theta, | Y\} = \sum_{Z} P\{Z | Y\} P\{\Theta, | Y, Z\}$$

$$= (1 - P\{pa_i | Y\}) P\{\Theta\} + \sum_{x} P\{x, pa_i | Y\} P\{\Theta, | x, pa_i\}$$

$$= (1 - P\{pa_i | Y\}) D(N_i) + \sum_{x} P\{x, pa_i | Y\} D(N_i(x, pa_i))$$

$$= \sum_{j} w_j D(N_{j_i})$$

Answer is a mixture of Dirichlet distributions.

Bayesian Inference

Determining $\Theta$

Complete Data: exact results for conjugate distributions
Incomplete Data: No closed form for Bayesian prediction

Today: Simple story
Primary interest is in finding the best $\Theta$.

Later:
Primary interest will be Bayesian model selection w/incomplete data
Interested in finding an approximation to $P(D|\Theta)$ and $\Theta$.
Approximations:
- MAP/Laplace (today)
- MCMC (today)
- BIC
- Cheeseman/Stutz
Bayesian Inference with Incomplete Data

"Parameter independence"

\[
P\{x[M + 1] \mid D\} = \int p\{x[M + 1] \mid \Theta_x, \Theta_{\Theta x}\} p\{\Theta_x, \Theta_{\Theta x} \mid D\}\theta\]

Problems:
No closed form solution for Bayesian prediction

Common Approximations:
MCMC
Maximum A-posteriori (MAP)

MCMC Solution

for (i = 1 to M)
MAP Approximation

Approximate

\[ P\{X[M+1]|D\} = P\{X[M+1]|\tilde{\Theta}\} \]

\[ \tilde{\Theta} = \text{arg max} \ P\{\Theta|D\} \]

Later: We will show that MAP arises from Gaussian approximation.

Limitations

\( \tilde{\Theta} \) cannot lie on the boundary of \( \Theta \)

Given D, there is a unique MAP

**Aliasing:** Interchange labels on a hidden variable and \( P\{\Theta|D\} \) is unchanged.

**Reduced dimensionality:** The likelihood can be encoded with a smaller set of parameters than \( \Theta \). Leads to an infinite number of MAP values.

Gaussian (MAP) Approximation

Finding MAP:

Same as ML, save Maximizing \( P(\Theta|D) \) instead of \( L(\Theta:D) \)

Include “prior” statistics in addition to statistics from data.

Given structure, the MAP assumption means:

Select

\[ \tilde{\Theta} = \text{arg max}_\Theta \ P\{\Theta|D\} \]
Where are we?

<table>
<thead>
<tr>
<th>Complete Data</th>
<th>Known Structure</th>
<th>Unknown Structure</th>
</tr>
</thead>
<tbody>
<tr>
<td>Complete Data</td>
<td>Statistical parameter estimation</td>
<td>Optimization over structures</td>
</tr>
<tr>
<td></td>
<td>Presented Wednesday</td>
<td>Right now.</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Incomplete Data</th>
<th>Statistical parameter estimation</th>
<th>Optimization over structures and parameters</th>
</tr>
</thead>
<tbody>
<tr>
<td>Incomplete Data</td>
<td>Parameter optimization</td>
<td>Presented Friday</td>
</tr>
<tr>
<td></td>
<td>Presented Friday</td>
<td>Bayesian estimation covered today.</td>
</tr>
</tbody>
</table>

Structure Optimization (Complete Data)

Select graph that maximizes a score

A

B

C

D

E

P(A)
P(B|A)
P(C)
P(D|A,C)
P(D|A)
P(E|D)
Structure-Based Learning Techniques

Goal: find the best structure to explain the data

(Controversial benefit): Identify “causal” relations via the V-structures of the corresponding graph.

Two approaches

Constraint-based
- Test for CISs (Just like PS 4)
- Search for network consistent with the CIS statements.
- Sensitive to errors in independence tests.
- No statistical foundation

Score-based (what we will talk about)
- Define a score that describes how well structure models observations
- Search for structure that maximizes the score

Focus on score-based approaches

Define score

I. Likelihood
II. MDL
III. BDe

Find structure that maximizes score

Special case: Trees
General DAGs

Issues

Model averaging
I. Likelihood score

Use likelihood function

\[
L[G, \Theta_G : D] = \prod_{m=1}^{M} P\{X[m] | G, \Theta_G \}
\]

\[
= \prod_{m=1}^{M} \prod_{i=1}^{N} P\{X[m] | P_{a_{i}}^{G}[m]; G, \Theta_G \}
\]


I. Likelihood Score

Rewrite likelihood score:

\[
I(G : D) = \log L[G, D]
\]

\[
= M \sum \left( I(X_i : P_{a_i}) - H(X_i) \right)
\]

where

\[
H(X) = -\int P(X) \log P(X) dX
\]

\[
I(G : D) = H(X) + H(Y) - H(X, Y)
\]  

**mutual information**
I. Likelihood: Pluses and Minuses

\[ I(G : D) = M \sum_i (I(X_i : Pa_i) - H(X_i)) \]

Pluses: \[ I(G : D) = M \sum_i (I(X_i : Pa_i) - H(X_i)) \]
- The larger the dependency of each variable on its parents, the higher the score.

Minuses:
- Adding arcs always improves the score.
- \[ I(X : Y) \leq I(X : Y, Z) \]
- Maximally connected network overfits data.

I. Preventing Overfitting

Restrict hypotheses
- Restricted # of parents or # of parameters.

Minimum description length (MDL)
- Description length measures complexity of model or data.
- Select model that minimizes \( DL(\text{model}) + DL(\text{data|model}) \)
  - Recall class on information theory...

Bayesian score
- Average over parameter values… Low likelihood if there is little data per fitted parameter.
II: Minimum Description Length

[Reassanen, 1987]

Prefer graphs that provide maximum compression of the data.
Compression means that the network summarizes the data.
Pick network that minimizes
\[ DL(\text{network}) + DL(\text{data|network}) \]

II. MDL

Description length of the data + structure:
\[
DL(D : G) = DL(G) + \frac{\log M}{2} \dim(G) - I(G : D)
\]

- bits to describe graph
- accuracy required to represent each parameter (M different possible values)
- # of parameters
- bits required to represent data given graph.

MDL is defined to be:
\[
MDL = -DL(D : G)
\]
II. MDL

MDL

\[ \text{MDL}(D : G) = I(G : D) - \frac{\log M}{2} \text{dim}(G) - DL(G) \]

Likelihood is linear in M

\[ I(G : D) = \sum_M \log P\{X[m]|G\} \]
\[ = M \cdot E[\log P\{X[m]|G\}] \]

Generally use only those terms that increase in M

\[ \text{MDL}(D : G) = I(G : D) - \frac{\text{dim}(G) \log M}{2} \]

Note: As M increases, the relative weight of the penalty decreases

II. MDL Properties

As \( M \to \infty \), the “true” structure \( G_{\text{max}} \) has the maximum score.

For sufficiently large \( M \), the maximal scoring structures are \textbf{independence equivalent} to \( G_{\text{max}} \)

Recall: Two DAGs, G and G', represent equivalent sets of CISs whenever G and G' have the same undirected version AND the identical V-structures.
III. Bayesian inference

\[ P\{x[M + 1] \mid D\} = \sum_G P\{x[M + 1] \mid D,G\}P\{G \mid D\} \]

where

\[ P\{G \mid D\} \propto P\{D \mid G\}P\{G\} \]

\[ = \left( \int \left( P\{D \mid G, \Theta\}P\{\Theta \mid G, \Theta\} \right)P\{G\} \right) \]

III. Marginal Likelihood for Binomials

Observe M coin tosses

Chain rule:

\[ P\{X[1], \ldots, X[M]\} \]

\[ = P\{X[1]\}P\{X[2] \mid X[1]\} \cdots P\{X[M] \mid X[1], \ldots, X[M - 1]\} \]

We showed that:

\[ P\{X[m + 1] \mid X[1], \ldots, X[m]\} = \frac{N^m_{\mu} + \alpha_{\mu}}{m + \alpha_{\mu} + \alpha_{\tau}} \]

SO...
III. Marginal Likelihood for Binomials (cont’d)

\[ p(x[1], \ldots, x[M]) = \frac{\alpha_h \cdots N_h + 1 + \alpha_h}{\alpha_h + \alpha_T \cdot N_h + 1 + \alpha_h + \alpha_T} \cdot \frac{\alpha_r \cdots N_r + 1 + \alpha_r}{N_h + \alpha_h + \alpha_T \cdot N_h + N_r + 1 + \alpha_h + \alpha_T} \]

\[ p(D) = \frac{\Gamma(\alpha_h + \alpha_r)}{\Gamma(N_h + N_r + \alpha_h + \alpha_r)} \cdot \frac{\Gamma(N_h + \alpha_h)}{\Gamma(\alpha_h)} \cdot \frac{\Gamma(N_r + \alpha_r)}{\Gamma(\alpha_r)} \]

III. Marginal Likelihood for Multinomials

\[ p(D) = \frac{\Gamma\left(\sum \alpha_k\right)}{\Gamma\left(\sum (\alpha_k + N_k)\right)} \cdot \prod_k \frac{\Gamma(\alpha_k + N_k)}{\Gamma(\alpha_k)} \]

In general networks:

\[ p(D|G) = \prod_{i=1}^{N} \left( \prod_{p_i} \frac{\Gamma(\alpha(p_i^G))}{\Gamma(\alpha(p_i^G) + N(p_i^G))} \cdot \prod_{x_i} \frac{\Gamma(\alpha(x, p_i^G) + N(x, p_i^G))}{\Gamma(\alpha(x, p_i^G))} \right) \]

\( \alpha(\ldots) \): parameters for each family given G

\( N(\ldots) \): counts from data
III. BDe Score

Problem:
Need prior counts $\alpha(\cdot)$ for each $G$

BDe prior:
Use prior of the form $M_0, B_0 = (G_0, \Theta_0)$
$M_0$ examples distributed according to $B_0$

Use $\alpha(x_i, pa_i^G) = M_0 P\{x_i, pa_i^G \mid G_0, \Theta_0\}$
In general, $pa_i^G \neq pa_i^{G_0}$

Nice property:
Equivalent networks are assigned the same scores.

III. BDe Score Properties

Seems different from MDL, BUT
Scores are asymptotically equivalent

Can show:
$$\log P\{D \mid G\} = I(G : D) - \frac{\dim(G)}{2} \log M + O(1)$$

Bayes score is asymptotically equivalent to MDL score
Constants ($P(G)$ and $DL(G)$) are negligible when $M$ is large
Scoring metrics: Summary

Likelihood, MDL and (log) BDe all have the form

$$Score(G : D) = \sum_i Score(X_i \mid \text{Pa}^G_i : N(X_i, \text{Pa}^G_i))$$

MDL and BDe are asymptotically equivalent
All three scoring metrics assign the same score to equivalent networks.