This Time

Continuous variables
  Gaussian belief networks
Learning problems for BNs
  Basic parameter estimation

Next Time

Parameter estimation in belief networks
Plates
Conjugate distributions
Incomplete data

Continuous variables

Principal problem:
  Closed form solutions don’t exist for most distribution classes
  and queries on general belief networks.
Exceptions:
  Multinomial (all semester)
  Conditional gaussian distributions (today)
  Mixtures of gaussian and discrete variables (later)

Solution techniques for non-gaussian continuous variables
  Approximation
  Query and structure-specific algorithms
  Mixtures
Conditional gaussian distributions

Gaussian

\[ P(x) = N(x; \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp \left( -\frac{(x-\mu)^2}{2\sigma^2} \right) \]

Conditional Gaussian

Variance fixed, mean is a linear function of the values for parents

\[ P(y \mid X) = N(y; \mu + B^T X, \sigma^2) \]

Need a representation for joint distributions.

JPDs: Multivariate gaussian distributions

“Normal” form: Specified in terms of covariance and mean.

\[ P(X) = N(X; \mu, \Sigma) = p \exp \left( -\frac{1}{2} (X - \mu)^\top \Sigma^{-1} (X - \mu) \right) \]

where \( p = \frac{1}{\sqrt{(2\pi)^d |\Sigma|}} \)
Canonical form [Lauritzen]

Standard form:
\[ P(X) = N(X; \mu, \Sigma) = \exp\left( -\frac{1}{2} (X - \mu)^T \Sigma^{-1} (X - \mu) \right) \]
where \( p = \frac{1}{\sqrt{(2\pi)^r |\Sigma|}} \)

Canonical form:
\[ P(X) = N(X; g, h, K) = \exp\left( g + X^T h - \frac{1}{2} X^T K X \right) \]
\[ K = \Sigma^{-1} \quad \text{"precision"} \]
\[ \mu = \Sigma h \]
\[ p = \exp\left( g + \frac{1}{2} \mu^T \Sigma^{-1} \mu \right) \]

Why?
Easy multiplication/division
Uniform representation for conditional and joint distributions.

Multiplication and Division with Canonical Form

\[ \phi_1(X) = N(X; g_1, h_1, K_1) \quad \phi_2(Y) = N(Y; g_2, h_2, K_2) \]

Extend to same domain \( Z = X \cup Y \) by adding zeros to appropriate dimensions
\[ \phi_1(Z) = N(Z; g_1, h_1, K_1) \quad \phi_2(Z) = N(Z; g_2, h_2, K_2) \]
\[ \phi_1(Z) \phi_2(Z) = N(Z; g_1 + g_2, h_1 + h_2, K_1 + K_2) \]
\[ \frac{\phi_1(Z)}{\phi_2(Z)} = N(Z; g_1 - g_2, h_1 - h_2, K_1 - K_2) \]

Compare to “normal form” multiplication and division:
\[ \phi_1(Z) \phi_2(Z) = N\left( Z; \left( \mu_1 \Sigma_1^{-1} + \mu_2 \Sigma_2^{-1} \right) \left( \Sigma_1^{-1} + \Sigma_2^{-1} \right)^{-1} \left( \Sigma_1^{-1} + \Sigma_2^{-1} \right)^{-1} \right) \]
\[ \frac{\phi_1(Z)}{\phi_2(Z)} = N\left( Z; \left( \mu_1 \Sigma_1^{-1} - \mu_2 \Sigma_2^{-1} \right) \left( \Sigma_1^{-1} - \Sigma_2^{-1} \right)^{-1} \left( \Sigma_1^{-1} - \Sigma_2^{-1} \right)^{-1} \right) \]
Inference

Objective:
Support the join tree algorithm.

Algorithmic Needs:
1. Create potentials from CPDs
2. Multiply and divide potentials (done)
3. Marginalize potentials
4. Enter evidence
5. Absorption

Creating potentials from CPDs

\[
P(y | X) = N\left(y; \mu + b^T X, \sigma^2 \right)
\]

\[
p(y | X) = \exp\left[-\frac{(y - \mu - b^T X)^2}{2\sigma^2}\right]
\]

\[
= \exp\left[-\frac{1}{2\sigma^2}(X^T y - b b^T - b y^T X + (X^T y) \frac{1}{\sigma^2} (\frac{-b \mu}{\mu}^T - \frac{1}{2} \frac{\mu^2}{\sigma^2} + \log p)\right]
\]

Conditional distribution in canonical form:

\[
g = -\frac{1}{2} \frac{\mu^2}{\sigma^2} - \frac{1}{2} \log(2\pi\sigma^2) \quad h = \frac{\mu}{\sigma^2} \begin{pmatrix} -b \\ 1 \end{pmatrix} \quad K = \frac{1}{\sigma^2} \begin{pmatrix} b b^T & -b \\ -b^T & 1 \end{pmatrix}
\]
Vector-valued gaussian variables

Convenient to have vector-valued variables.

Digression: Vector-valued nodes

Say that a vector-valued node $Y$ has parents $X$

CPD is

$$ P(Y \mid X) = N(Y; \mu + B^T X, \Sigma) $$

$$ P(Y \mid X) = p \exp \left[ -\frac{1}{2} (y - \mu - B^T X)^T \Sigma^{-1} (y - \mu - B^T X) \right] $$

$$ = \exp \left[ -\frac{1}{2} (X^T Y \begin{pmatrix} B \Sigma^{-1} B^T \\ -\Sigma^{-1} \end{pmatrix} - \Sigma^{-1} Y^T X - B \Sigma^{-1} \mu \right] - \frac{1}{2} \mu^T \Sigma^{-1} \mu + \log p $$

Canonical form:

$$ r = -\frac{1}{2} \mu^T \Sigma^{-1} \mu - \frac{n}{2} \log(2\pi) - \frac{1}{2} \log |\Sigma| $$

$$ h = \begin{pmatrix} -B \Sigma^{-1} \mu \\ \Sigma^{-1} \mu \end{pmatrix} $$

$$ K = \begin{pmatrix} B \Sigma^{-1} B^T & -B \Sigma^{-1} \\ -\Sigma^{-1} B^T & \Sigma^{-1} \end{pmatrix} $$
Marginalize CF

A bit more complicated than for normal form

\[
Y = \begin{pmatrix} Y_1 \\ Y_2 \end{pmatrix}, \quad h = \begin{pmatrix} h_1 \\ h_2 \end{pmatrix}, \quad K = \begin{pmatrix} K_{11} & K_{12} \\ K_{21} & K_{22} \end{pmatrix}
\]

\[
\phi(Y_2) = \int \phi(y) \nu_{Y_1} = N(y_2; \hat{g}, \hat{h}, \hat{K})
\]

\[
\hat{g} = g + \frac{1}{2} \left( \left[ \log(2\pi) - \log|K_{11}| \right] + h_1^T K_{11}^{-1} h_1 \right)
\]

\[
\hat{h} = h_2 - K_{21} K_{11}^{-1} h_1
\]

\[
\hat{K} = K_{22} - K_{21} K_{11}^{-1} K_{12}
\]

Need \( K_{11} \) to have full rank.

Compare to:

\[
N(y_2; \mu_2, \hat{\Sigma}) = N(y_2; \mu_2, \hat{K}^{-1})
\]

Adding Evidence

Since \( K \) is a precision, we cannot represent the evidence directly in each clique.

Instead, we need to remove the evidence from each separator and clique potential

Say that we observe \( Y = y \):

Start with: \( \phi(x) = \phi((x \ y)) \)

\[
\phi(x, y) = \exp \left[ g + \begin{pmatrix} X^T \\ y^T \end{pmatrix} \begin{pmatrix} h_1 \\ h_2 \end{pmatrix} - \frac{1}{2} \begin{pmatrix} X^T & y^T \end{pmatrix} \begin{pmatrix} K_{xx} & K_{xy} \\ K_{yx} & K_{yy} \end{pmatrix} \begin{pmatrix} X \\ y \end{pmatrix} \right]
\]

Rearrange:

\[
\phi(x, y) = \exp \left[ \left( g + h_2^T y - \frac{1}{2} y^T K_{yy} y \right) + X^T (h_2 - K_{xy} y) - \frac{1}{2} X^T K_{xx} X \right]
\]
Absorption

On first pass:
1. Initialize all g, h, K’s to zero.
2. Multiply conditional probabilities into cliques.
3. Note that we cannot marginalize X out of a clique if $K_{XX}$ is not of full rank (call these “complete” cliques).
4. Absorb from C1 to C2 only if we have received a message from a clique containing the CPD for X for each X in C1\C2.

Some properties...

1. Variance is a function of the number of observations, not the value of those observations.
2. There is a compact representation for the FULL joint distribution
   - For binary: $O(2^k)$ where $k$ is the size of the largest clique.
   - For gaussian: $O(n^2)$
3. This means that there is an efficient algorithm on the full joint distribution.
   - Marginalization: $O(M(n))$
   - $M(n) = n^3$ Naïve inversion algorithm
   - $M(n) = n^{3.81} = n^{2.81}$ Best known inversion algorithm
Learning

Why do we care?

If the number of parents is limited to \( k \), need to learn only parameters \( o(2^k) \).

If we are learning the full joint distribution, we need to learn parameters \( o(n^n) \).

Example: ALARM network (pulmonary function)

- 37 binary variables.
- Belief net: 509 independent parameters.
- Joint distribution: \( 2^{37} \approx 6.9 \times 10^{10} \) independent parameters.

The learning problem

<table>
<thead>
<tr>
<th></th>
<th>Known Structure</th>
<th>Unknown Structure</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Complete Data</strong></td>
<td>Statistical parameter estimation</td>
<td>Optimization over structures</td>
</tr>
<tr>
<td><strong>Incomplete Data</strong></td>
<td>Parameter optimization</td>
<td>Optimization over structures and parameters</td>
</tr>
</tbody>
</table>

**Learning Tasks:**

- **JPD Modeling:** best model for \( P(X) \)
- **Classification:** model for \( P(X,C) \) that provides “best” \( P(C|X) \)
Parameter Estimation (Complete Data)

\[
\begin{array}{c}
A \\
B \\
E
\end{array} \hspace{1cm}
\begin{array}{c}
C \\
D
\end{array} \quad \mathbb{P}(A) \\
\mathbb{P}(B|A) \\
\mathbb{P}(C) \\
\mathbb{P}(D|A,C) \\
\mathbb{P}(D|A) \\
\mathbb{P}(E|D)
\]

Parameter Estimation (Incomplete Data)

\[
\begin{array}{c}
A \\
B \\
E
\end{array} \hspace{1cm}
\begin{array}{c}
C \\
D
\end{array} \quad \mathbb{P}(A|h) \\
\mathbb{P}(B|A) \\
\mathbb{P}(C|h) \\
\mathbb{P}(D|A,C) \\
\mathbb{P}(D|A) \\
\mathbb{P}(E|D) \\
\mathbb{P}(h)
\]

Common: occurs whenever people are selective about tests or observations to collect.
Example: Medical data
Optimization over Structure (Complete Data)

Optimization over Structure and Parameters (Incomplete Data)
Statistics 101

Binomial experiment

\[ \text{Tails} \quad \text{Heads} \]

Infer the unknown probability \( \theta \) of heads from experiments \( X[1], \ldots, X[M] \)

Statistical parameter fitting

I.I.D.

Assume \( X[1] \ldots X[M] \) are sampled from the same distribution
Each sample is independent of the rest.

Likelihood:

\[
L(\theta : D) = P(D | \theta) = \prod_{m=1}^{M} P(X[m] | \theta)
\]

Likelihood for sequence H,T,T,H,H,H is

\[
L(\theta : D) = \theta \cdot (1-\theta) \cdot (1-\theta) \cdot \theta \cdot \theta \cdot \theta \cdot \theta
\]
Sufficient Statistics

In order to compute the likelihood, all we need are the number of heads and tails, $N_h$ and $N_t$.

$$L(\theta : D) = \theta^{N_h} \cdot (1 - \theta)^{N_t}$$

A **sufficient statistic** is a function on the data that summarizes all of the information relevant for computing the likelihood.

If $s(D) = s(D')$ then $L(\theta : D) = L(\theta : D')$

---

Maximum Likelihood Estimation (MLE)

Learn parameters that maximize the likelihood function

$$\frac{\partial}{\partial \theta} \theta^{N_h} \cdot (1 - \theta)^{N_t} = N_h \theta^{N_h-1} \cdot (1 - \theta)^{N_t} - N_t \theta^{N_h} \cdot (1 - \theta)^{N_t-1} = 0$$

$$N_h \cdot (1 - \theta) = N_t \theta$$

$$N_h = (N_t + N_h) \theta$$

$$\theta = \frac{N_h}{N_t + N_h}$$

In our example $N=(2,5)$, so $\theta = \frac{2}{7}$
Next Time

Parameter estimation in belief networks
Plates
Conjugate distributions
Parameter optimization with incomplete data