Today

Bayesian parameter estimation
Parameter estimation in Bayes networks
Plates
Conjugate distributions
Incomplete data

Last Time

Thumbtack example
Maximum Likelihood Estimation (MLE):
Likelihood for M Bernoulli trials with our thumbtack:
\[
L(\theta : D) = P(D | \theta) = \prod_{i=1}^{m} P(X[i] | \theta) = \theta^{N_h} \cdot (1 - \theta)^{N_t}
\]

Likelihood for sequence H,T,T,H,H,H is
\[
L(\theta : D) = 6 \cdot (1 - \theta) \cdot (1 - \theta) \cdot 6 \cdot \theta \cdot \theta \cdot \theta
\]

Select parameter \( \hat{\theta} \) that yields the maximum likelihood
\[
\hat{\theta} = \arg \max_{\theta} \theta^{N_h} \cdot (1 - \theta)^{N_t} = \frac{N_h}{N_h + N_t} = \frac{5}{7}
\]
MLE in Bayes Nets

Bayes net with variables \( X = \{x_1, \cdots, x_n\} \)
and parameters \( \Theta = \{\Theta_1, \cdots, \Theta_n\} \)

where \( \{\Theta_1, \cdots, \Theta_n\} \) are mutually exclusive sets of parameters characterizing
\( P\{x_1 | pa_1\}, \cdots, P\{x_n | pa_n\} \)

Likelihood for \( M \) i.i.d. samples, \( D = \{X[1], \cdots, X[M]\} \):

\[
L[\Theta : D] = P(D | \Theta) \\
= \prod_{n=1}^{M} P\{X[m] | \Theta\} \quad \text{i.i.d. assumption}
\]
\[
= \prod_{n=1}^{M} \prod_{m=1}^{N} P\{x_n[m] | pa_n[m] \Theta_n\} \quad \text{factoring in bayes nets}
\]
\[
= \prod_{n=1}^{N} \prod_{m=1}^{M} P\{x_n[m] | pa_n[m] \Theta_n\}
\]
\[
= \prod_{n=1}^{N} L_n(\Theta_n, D)
\]

MLE in Bayes Nets, cont’d

Since \( L(\Theta : D) = \prod_{n=1}^{N} L_n(\Theta_n : D) \) we can maximize the overall
likelihood by maximizing the individual factors.

Let \( \text{family}_n = \{x_n\} \cup \text{pa}_n(x_n) \)

If the parameters \( \{\Theta, \cdots, \Theta_n\} \) are mutually exclusive, then they can be estimated independently of one another using \( D_n = \{\text{family}_n[1], \cdots, \text{family}_n[M]\} \)
Likelihood for Multinomials

Multinomial variable \( x \) with values 1, \ldots, \( K \)

\[
P\{x = k\} = \theta_k, \quad \sum_{k=1}^{K} \theta_k = 1
\]

Sufficient statistic is \( (N_1, \ldots, N_k) \) the set of counts for each possible outcome

Likelihood is

\[
L(\Theta, D) = \prod_{k=1}^{K} \theta_k^{N_k}
\]

MLE is

\[
\hat{\theta}_k = \frac{N_k}{\sum_i N_i}
\]

Conditional Multinomial Distributions

\( P\{x_n \mid pa_n\} \) consists of a distinct multinomial for each value of \( pa_n \):

\[
\begin{align*}
P\{x_n \mid pa_n = (0,0,0,0)\} \\
P\{x_n \mid pa_n = (0,0,0,1)\} \\
P\{x_n \mid pa_n = (1,1,1,1)\}
\end{align*}
\]
Likelihood for Conditional Multinomial Distributions

$$L_\theta(\Theta : D) = \prod_{m=1}^{M} P\{x_m \mid pa_m[m]\}$$

$$= \prod_{\theta_{m}} \prod_{x_{m}}^{x_{m}} \prod_{s}^{s} P\{x_{m} \mid pa_{m}\}$$

$$= \prod_{\theta_{m}} \prod_{x_{m}}^{x_{m}} \theta_{s_{m}}^{N(x_{m}, pa_{m})}$$

where $N(x_{m}, pa_{m})$ is the number of times that $x_{m}$ and $pa_{m}$ occur in the data and $\theta_{s_{m}}$ is the parameter for $P\{x_{m} \mid pa_{m}\}$

MLE: $\hat{\theta}_{s_{m}} = \frac{N(x_{m}, pa_{m})}{N(pa_{m})}$

So...

OK, now we can determine the MLE for a multinomial Bayes net with independent parameters and complete data.

Let’s revisit our original thumbtack problem

What if the thumbtack were really a penny? …based on the sample (8 heads, 2 tails), we will conclude that $\hat{\theta} = \frac{8}{10} = 0.8$

Is this a good estimate for the probability of heads for a penny?
Bayesian Inference

Bayesian answer: MLE is not the best answer if we have a lot of prior knowledge.

We have a lot of experience with coins that we can use to argue that the probability ought to be close to 0.5.

Solution is to use a prior distribution over the parameter to reflect our prior beliefs and determine a posterior distribution over belief given both the prior and the likelihood.

The Bayes net for sampling:

i.i.d.: $x[m]$ are independent given $\theta$

Two interesting problems:
  - determine the distribution over $\theta$
  - determine the probability of $x$ given the sample and prior.
Bayesian Inference

Distribution over $\theta$:

$$P[\theta \mid X[1],...,X[M]] = \frac{P\{X[1],...,X[M] \mid \theta\}P[\theta]}{P\{X[1],...,X[M]\}} \propto P[\theta] \prod_{m} P\{X[m] \mid \theta\}$$

Distribution over $X$:

$$P\{X \mid X[1],...,X[M]\} = \int P\{X \mid \theta\}P[\theta \mid X[1],...,X[M]\}d\theta$$

$$= \int \theta \cdot P[\theta \mid X[1],...,X[M]\}d\theta$$

$$= E[\theta \mid X[1],...,X[M]]$$

Plates

For $i = 1$ to $M$
BUGS Example

Very likely to crash this presentation...

M.L. sez: BUGS is called BUGS because it has BUGS

My experience:
OLE stuff is extremely unreliable.
Compilation is also unreliable.

None-the-less, it is way cool.

Comparison between MLE and Bayesian Prediction

Say that we choose a uniform prior on $\theta$
for the thumbtack problem. Originally, we concluded
that the MLE for 5 heads and 2 tails was

$$\hat{\theta} = \max_{\theta} \theta^{N_h} (1 - \theta)^{N_i} = \frac{N_h}{N_h + N_i} = \frac{5}{7}$$

The Bayesian prediction is

$$\bar{\theta} = \int \theta \cdot \theta^{N_h} (1 - \theta)^{N_i} d\theta = \frac{N_h + 1}{N_h + N_i + 2} = \frac{6}{9} = \frac{2}{3}$$
MLE and Bayesian Inference

Thm: If \( P(\theta) > 0 \) whenever \( \theta = \arg \max_{\theta} P(D | \theta) \) then
\[
\lim_{n \to \infty} \left( \arg \max_{\theta} P(D | \theta) \right) = \lim_{n \to \infty} ( P(\theta | D) )
\]
Moral, the bayesian prediction equals the MLE if we observe enough data and select a prior that is positive for all feasible likelihoods.

“Sensible” Prior Distributions

For a binomial experiment, the likelihood function is:
\[
L(\theta : D) = \theta^N \cdot (1 - \theta)^V
\]
Say that we use a beta prior:
\[
P(\theta) = \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} \theta^{\alpha-1} \cdot (1 - \theta)^{\beta-1} \propto \theta^{\alpha-1} \cdot (1 - \theta)^{\beta-1}
\]
The posterior distribution is:
\[
P(\theta | D) = \frac{\Gamma(\alpha + N + \beta + N)}{\Gamma(\alpha + N)\Gamma(\beta + N)} \theta^{\alpha + N - 1} \cdot (1 - \theta)^{\beta + N - 1} \propto \theta^{\alpha + N - 1} \cdot (1 - \theta)^{\beta + N - 1}
\]
Same functional form!
Conjugate Families

(fixed dimension sufficient statistic) It is desirable to reason about the likelihood of data using a sufficient statistic of fixed dimension, regardless of the size or values in a sample.

(fixed dimension sufficient statistic) implies that there must exist families, \( \Psi \), of distributions, such that if the prior distribution \( P\{ \Theta \} \) belongs to family \( \Psi \) then the posterior distribution \( P\{ \Theta | D \} \) is also in family \( \Psi \).

These families of distributions are called conjugate families.

<table>
<thead>
<tr>
<th>Prior</th>
<th>Likelihood</th>
<th>Posterior</th>
</tr>
</thead>
<tbody>
<tr>
<td>Beta</td>
<td>( \theta^\alpha \cdot (1 - \theta)^\beta )</td>
<td>( \theta^{\alpha + N_\theta} \cdot (1 - \theta)^{\beta + N_\beta} )</td>
</tr>
<tr>
<td>Dirichlet</td>
<td>( K(A) \prod_{i} \theta^{\gamma_i} )</td>
<td>( K(A + N) \prod_{i} \theta^{\gamma_i + N_i} )</td>
</tr>
<tr>
<td>Gaussian</td>
<td>( K(\mu, \tau) \exp \left( -\frac{\tau}{2}(\theta - \mu)^2 \right) )</td>
<td>( K(\mu, \tau + n) \exp \left( -\frac{\tau}{2}(\theta - \mu)^2 \right) )</td>
</tr>
</tbody>
</table>

Some conjugate families
Next Time

Parameter estimation with incomplete data
EM

Monday

P.S. 5 Discussion
Beginnings of Structure Optimization
Read: The Heckerman tutorial (course reader)

Sufficient Statistics

S is a sufficient statistic for f(x|w) if L(x1|w) = L(x2|w) whenever S(x1) = S(x2).

THM: A statistic S is sufficient for a family of pdf’s f(.|θ) iff f(x|θ) can be factored as follows for all values x and θ:

\[ f(x|θ) = u(x)v[S(x),θ] \]