

# Today

Bayesian parameter estimation

Parameter estimation in Bayes networks

Plates

Conjugate distributions

Incomplete data

# Last Time

Thumbtack example

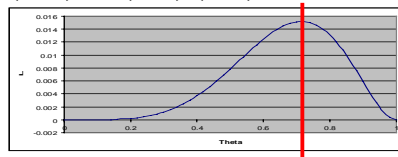
Maximum Likelihood Estimation (MLE):

Likelihood for M Bernoulli trials with our thumbtack:

$$L(\theta : D) = P(D | \theta) = \prod_{m \in M} P(X[m] | \theta) = \theta^{N_h} \cdot (1 - \theta)^{N_t}$$

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

$$L(\theta : D) = \theta \cdot (1 - \theta) \cdot (1 - \theta) \cdot \theta \cdot \theta \cdot \theta \cdot \theta$$



Select parameter  $\hat{\theta}$  that yields the maximum likelihood

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

## MLE in Bayes Nets

Bayes net with variables  $X = \{x_1, \dots, x_n\}$

and parameters  $\Theta = \{\Theta_1, \dots, \Theta_n\}$

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

Likelihood for M i.i.d. samples,  $\mathbf{D} = \{X[1], \dots, X[M]\}$ :

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

## MLE in Bayes Nets, cont'd

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

Let  $family_n = \{x_n\} \cup pa_n(x_n)$

If the parameters  $\{\Theta_1, \dots, \Theta_n\}$  are mutually exclusive, then they can be estimated independently of one another using  $D_n = \{family_n[1], \dots, family_n[M]\}$

## Likelihood for Multinomials

Multinomial variable  $x$  with values  $1, \dots, K$

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

Sufficient statistic is  $(N_1, \dots, N_K)$  the set of counts for each possible outcome

Likelihood is  $L(\Theta, \mathbf{D}) = \prod_{k=1}^K \theta_k^{N_k}$

$$\text{MLE is } \hat{\theta}_k = \frac{N_k}{\sum_i N_i}$$

## Conditional Multinomial Distributions

$P\{x_n | pa_n\}$  consists of a distinct multinomial for each value of  $pa_n$ :

$$P\{x_n | pa_n = (0,0,0,0,0)\}$$

$$P\{x_n | pa_n = (0,0,0,0,1)\}$$

$$P\{x_n | pa_n = (1,1,1,1,1)\}$$

## Likelihood for Conditional Multinomial Distributions

$$\begin{aligned}
 L_n(\Theta_n : \mathbf{D}) &= \prod_{m=1}^M P\{x_n[m] | pa_n[m]\} \\
 &= \prod_{pa_n} \prod_{m \in 1, \dots, M \text{ s.t. } pa_n = pa_n[m]} P\{x_n[m] | pa_n\} \\
 &= \prod_{pa_n} \prod_{x_n} \theta_{x_n|pa_n}^{N(x_n, pa_n)}
 \end{aligned}$$

where  $N(x_n, pa_n)$  is the number of times that  $x_n$  and  $pa_n$  occur in the data

and  $\theta_{x_n|pa_n}$  is the parameter for  $P\{x_n | pa_n\}$

**MLE:** 
$$\hat{\theta}_{x_n|pa_n} = \frac{N(x_n, pa_n)}{N(pa_n)}$$

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

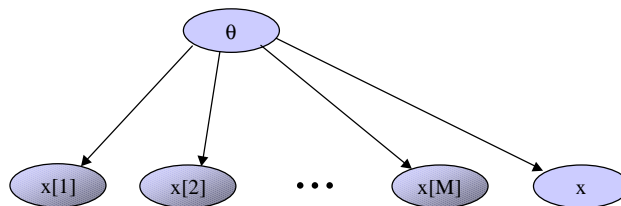
Bayesian answer: MLE is not the best answer if we have a lot of prior knowledge.

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.

## Bayesian Inference

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

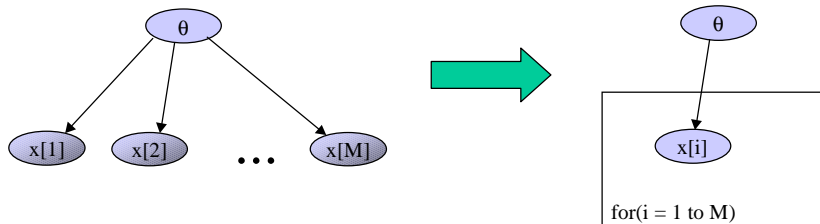
Distribution over  $\theta$ :

$$P\{\theta | X[1], \dots, X[M]\} = \frac{P\{X[1], \dots, X[M] | \theta\} P\{\theta\}}{P\{X[1], \dots, X[M]\}} \propto P\{\theta\} \prod_m P\{X[m] | \theta\}$$

Distribution over  $X$ :

$$\begin{aligned} P\{X | X[1], \dots, X[M]\} &= \int P\{X | \theta\} P\{\theta | X[1], \dots, X[M]\} d\theta \\ &= \int \theta \cdot P\{\theta | X[1], \dots, X[M]\} d\theta \\ &= E[\theta | X[1], \dots, X[M]] \end{aligned}$$

## Plates



## 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.

way cool: [Californian] nifty, neat.

## 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

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

The Bayesian prediction is

$$\hat{\theta} = \int \theta \cdot \theta^{N_h} (1-\theta)^{N_t} d\theta = \frac{N_h + 1}{N_h + N_t + 2} = \frac{6}{9} = \frac{2}{3}$$

## MLE and Bayesian Inference

Thm: If  $P\{\theta\} > 0$  whenever  $\theta = \arg \max_{\theta'} P\{\mathbf{D} | \theta'\}$

$$\text{then } \lim_{n \rightarrow \infty} \left( \arg \max_{\theta'} P\{\mathbf{D} | \theta'\} \right) = \lim_{n \rightarrow \infty} (P\{\theta | \mathbf{D}\})$$

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:

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

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:

$$P(\theta | D) = \frac{\Gamma(\alpha + N_h + \beta + N_t)}{\Gamma(\alpha + N_h)\Gamma(\beta + N_t)} \theta^{\alpha + N_h - 1} \cdot (1 - \theta)^{\beta + N_t - 1} \propto \theta^{\alpha + N_h - 1} \cdot (1 - \theta)^{\beta + N_t - 1}$$

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*.

## Some conjugate families

Prior	Likelihood	Posterior
$K(\alpha, \beta)\theta^{\alpha-1} \cdot (1-\theta)^{\beta-1}$ Beta	$\theta^{N_h} \cdot (1-\theta)^{N_t}$ Binomial	$K(\alpha + N_h, \beta + N_t)\theta^{\alpha+N_h-1} \cdot (1-\theta)^{\beta+N_t-1}$
$K(\mathbf{A})\prod_{\alpha_i \in \mathbf{A}} \theta^{\alpha_i-1}$ Dirichlet	$\prod_{N_k \in \mathbf{N}} \theta^{N_k}$ Multinomial	$K(\mathbf{A} + \mathbf{N})\prod_k \theta^{\alpha_i+N_k-1}$
$K(\mu, \tau)\exp\left(-\frac{\tau}{2}(\theta-\mu)^2\right)$ Gaussian	$\propto \prod_m \exp\left(-\frac{1}{2}(x-\theta)^2 r^2\right)$ Gaussian with unknown mean	$K(\mu', \tau + nr)\exp\left(-\frac{\tau}{2}(\theta-\mu')^2\right)$ $\mu' = \frac{\tau\mu + nr\bar{x}}{\tau + nr}$

## 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  $L(x_1|w) = L(x_2|w)$   
whenever  $S(x_1) = S(x_2)$

*THM:* A statistic  $S$  is sufficient for a family of pdf's  $f(\cdot|\theta)$   
iff  $f(x|\theta)$  can be factored as follows for all values  $x$  and  
 $\theta$ :

$$f(x|\theta) = u(x)v[S(x),\theta]$$