Bayesian networks: basic parameter learning
Machine Intelligence

Thomas D. Nielsen

September 2008
Example: Physical Measurements

Mass of an atomic particle is measured in repeated experiments. Measurement results = true mass + random error.
Estimation

Example: Physical Measurements

Mass of an atomic particle is measured in repeated experiments. Measurement results = true mass + random error.
Example: Physical Measurements

Mass of an atomic particle is measured in repeated experiments. Measurement results = true mass + random error.
Estimation

Example: Physical Measurements

Mass of an atomic particle is measured in repeated experiments. Measurement results = true mass + random error.

Estimate of true mass: mean value of normal distribution that best “fits” the data.
Example: Coin Tossing

Is the Euro fair?
Example: Coin Tossing

Is the Euro fair?

Toss Euro 1000 times and count number of heads and tails:
Example: Coin Tossing

Is the Euro fair?

Result: heads: 521, tails: 479.

Probability of Euro falling heads (estimate): value that best “fits” the data: 521/1000.
Estimation: Classical

Structure of Estimation Problem

**Given:** Data produced by some random process that is characterized by one or several numerical parameters.

**Wanted:** Infer value of (some) parameters.

**(Classical) Method:** Obtain estimate for parameter by a function that maps possible data sets into the parameter space.
Parametric Family

Let $W$ be a set, $\Theta \subseteq \mathbb{R}^k$ for some $k \geq 1$. For every $\theta \in \Theta$ let $P_\theta$ be a probability distribution on $W$. Then $\{P_\theta \mid \theta \in \Theta\}$ is called a parametric family (of distributions).

Example 1: $W = \{h, t\}$, $\Theta = [0, 1]$. $P_\theta$: distribution with $P(h) = \theta$ (and $P(t) = 1 - \theta$).
**Parametric Family**

Let $W$ be a set, $\Theta \subseteq \mathbb{R}^k$ for some $k \geq 1$. For every $\theta \in \Theta$ let $P_\theta$ be a probability distribution on $W$. Then $\{P_\theta \mid \theta \in \Theta\}$ is called a parametric family (of distributions).

**Example 1:** $W = \{h, t\}$, $\Theta = [0, 1]$. $P_\theta$: distribution with $P(h) = \theta$ (and $P(t) = 1 - \theta$).

**Example 2:** $W = \{w_1, \ldots, w_k\}$,

$$\Theta = \{\theta = (p_1, \ldots, p_k) \in [0, 1]^k \mid \sum p_i = 1\}.$$  

$P_\theta$: distribution with $P(w_i) = p_i$. 

---

*Basic parameter learning*  
*September 2008*
Parametric Family

Let $W$ be a set, $\Theta \subseteq \mathbb{R}^k$ for some $k \geq 1$. For every $\theta \in \Theta$ let $P_\theta$ be a probability distribution on $W$. Then $\{P_\theta \mid \theta \in \Theta\}$ is called a parametric family (of distributions).

Example 1: $W = \{h, t\}$, $\Theta = [0, 1]$. $P_\theta$: distribution with $P(h) = \theta$ (and $P(t) = 1 - \theta$).

Example 2: $W = \{w_1, \ldots, w_k\}$,

$$\Theta = \{\theta = (p_1, \ldots, p_k) \in [0, 1]^k \mid \sum p_i = 1\}.$$

$P_\theta$: distribution with $P(w_i) = p_i$.

Example 3: $W = \mathbb{R}$, $\Theta = \mathbb{R} \times \mathbb{R}^+$. For $\theta = (\mu, \sigma) \in \Theta$: $P_\theta$ normal distribution with mean $\mu$ and standard deviation $\sigma$. 

Basic parameter learning September 2008
Sample

A family $X_1, \ldots, X_N$ of random variables is called \textit{independent identically distributed (iid)} if the family is independent, and $P(X_i) = P(X_j)$ for all $i, j$. 
A family $X_1, \ldots, X_N$ of random variables is called *independent identically distributed (iid)* if the family is independent, and $P(X_i) = P(X_j)$ for all $i, j$. A *sample* $s_1, \ldots, s_N \in W$ of observations (or data items) is interpreted as the observed values of an iid family of random variables with distribution $P(X_i) = P_\theta$. 
Sample

A family $X_1, \ldots, X_N$ of random variables is called **independent identically distributed (iid)** if the family is independent, and $P(X_i) = P(X_j)$ for all $i, j$. A sample $s_1, \ldots, s_N \in W$ of observations (or data items) is interpreted as the observed values of an iid family of random variables with distribution $P(X_i) = P_\theta$.

Likelihood Function

Given a parametric family $\{P_\theta \mid \theta \in \Theta\}$ of distributions on $W$, and a sample $s = (s_1, \ldots, s_N) \in W^N$. The function

$$\theta \mapsto P_\theta(s) := \prod_{i=1}^{N} P_\theta(s_i), \quad \text{resp.} \quad \theta \mapsto \log P_\theta(s) = \sum_{i=1}^{N} \log P_\theta(s_i),$$

is called the **likelihood function** (resp. **log-likelihood function**) for $\theta$ given $s$. 
Estimation: Classical

Maximum Likelihood Estimator

Given: parametric family and sample $s$. Every $\theta^* \in \Theta$ with

$$\theta^* = \arg \max_{\theta \in \Theta} P_\theta(s)$$

is called a maximum likelihood estimate for $\theta$ (given $s$).
Maximum Likelihood Estimator

Given: parametric family and sample $s$. Every $\theta^* \in \Theta$ with

$$\theta^* = \arg \max_{\theta \in \Theta} P_{\theta}(s)$$

is called a *maximum likelihood estimate* for $\theta$ (given $s$).

Since the logarithm is a strictly monotone function, maximum likelihood estimates are also obtained by maximizing the log-likelihood:

$$\theta^* = \arg \max_{\theta \in \Theta} \log P_{\theta}(s)$$
Estimation: classical

Thumbtack example

We have tossed a thumb tack 100 times. It has landed pin up 80 times, and we now look for the model that best fits the observations/data:

Structure

\[
\begin{align*}
T & \quad 0.1 \\
T & \quad 0.2 \\
T & \quad 0.3 \\
\end{align*}
\]

Probability, \( P(pinup) = \)

\[
\begin{align*}
M_{0.1} & \\
M_{0.2} & \\
M_{0.3} & \\
\end{align*}
\]
Estimation: classical

Thumbtack example

We have tossed a thumb tack 100 times. It has landed pin up 80 times, and we now look for the model that best fits the observations/data:

<table>
<thead>
<tr>
<th>Structure</th>
<th>T</th>
<th>T</th>
<th>T</th>
</tr>
</thead>
<tbody>
<tr>
<td>Probability, $P(\text{pinup}) =$</td>
<td>0.1</td>
<td>0.2</td>
<td>0.3</td>
</tr>
<tr>
<td>Model</td>
<td>$M_{0.1}$</td>
<td>$M_{0.2}$</td>
<td>$M_{0.3}$</td>
</tr>
</tbody>
</table>

We can measure how well a model fits the data using:

$$P(\mathcal{D}|M_{\theta}) = P(\text{pinup}, \text{pinup}, \text{pindown}, \ldots, \text{pinup}|M_{\theta})$$

$$= P(\text{pinup}|M_{\theta})P(\text{pinup}|M_{\theta})P(\text{pindown}|M_{\theta}) \cdot \ldots \cdot P(\text{pinup}|M_{\theta})$$

This is also called the likelihood of $M_{\theta}$ given $\mathcal{D}$. 
Estimation: classical

Thumbtack example

We have tossed a thumb tack 100 times. It has landed pin up 80 times, and we now look for the model that best fits the observations/data:

<table>
<thead>
<tr>
<th>Structure</th>
<th>Probability, $P(pinup)$ =</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>0.1</td>
</tr>
<tr>
<td></td>
<td>0.2</td>
</tr>
<tr>
<td></td>
<td>0.3</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Model</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$M_{0.1}$</td>
</tr>
<tr>
<td></td>
<td>$M_{0.2}$</td>
</tr>
<tr>
<td></td>
<td>$M_{0.3}$</td>
</tr>
</tbody>
</table>

We select the parameter $\hat{\theta}$ that maximizes:

$$\hat{\theta} = \arg \max_\theta P(D|M_\theta)$$

$$= \arg \max_\theta \prod_{i=1}^{100} P(d_i|M_\theta)$$

$$= \arg \max_\theta \mu \cdot \theta^{80} (1 - \theta)^{20}.$$
Estimation: classical

**Thumbtack example**

We have tossed a thumb tack 100 times. It has landed pin up 80 times, and we now look for the model that best fits the observations/data:

<table>
<thead>
<tr>
<th>Structure</th>
<th>T</th>
<th>T</th>
<th>T</th>
</tr>
</thead>
<tbody>
<tr>
<td>Probability, $P(pinup) =$</td>
<td>0.1</td>
<td>0.2</td>
<td>0.3</td>
</tr>
</tbody>
</table>

Model

By setting:

$$\frac{d}{d\theta} \mu \cdot \theta^{80} (1 - \theta)^{20} = 0$$

we get the maximum likelihood estimate:

$$\hat{\theta} = 0.8.$$
Maximum Likelihood Estimates for Multinomial Distribution

Consider the family of multinomial distribution defined as $W = \{w_1, \ldots, w_k\}$,

$$\Theta = \{\theta = (p_1, \ldots, p_k) \in [0, 1]^k \mid \sum p_i = 1\}.$$ $\Theta$: distribution with $P(w_i) = p_i$.

For $\{P_\theta \mid \theta \in \Theta\}$ and $s \in W^N$: there exists exactly one maximum likelihood estimate $\theta^* = (p_1^*, \ldots, p_k^*)$ given by

$$p_i^* = \frac{1}{N} |\{j \in \{1, \ldots, N\} \mid s_j = w_i\}|$$

[i.e. $\theta^*$ is just the empirical distribution defined by the data on $W$]
Proof: (for \( W = \{w_1, w_2\} \)):

\[
\begin{align*}
p_1^* &= \frac{1}{N} |\{j \in \{1, \ldots, N\} \mid s_j = w_1\}| \\
p_2^* &= \frac{1}{N} |\{j \in \{1, \ldots, N\} \mid s_j = w_2\}| \quad (= 1 - p_1^*)
\end{align*}
\]

\[
\log P_\theta(s) = \sum_{j=1}^{N} \log P_\theta(s_j) = N \cdot (p_1^* \log(p_1) + p_2^* \log(p_2))
\]

\[
= N \cdot (p_1^* \log(p_1) + (1 - p_1^*) \log(1 - p_1))
\]

Differentiated w.r.t. \( p_1 \):

\[
N \cdot \left(\frac{p_1^*}{p_1} - \frac{(1 - p_1^*)}{(1 - p_1)}\right)
\]

Only root: \( p_1 = p_1^* \).
Consistency

Let \( W = \{w_1, \ldots, w_k\} \), and the data \( s_1, s_2, \ldots, s_N \) be generated by the distribution \( P_\theta \) with parameters \( \theta = (p_1, \ldots, p_k) \).

Then for all \( \epsilon > 0 \) and \( i = 1, \ldots, k \):

\[
\lim_{N \to \infty} P_\theta(|p^*_i - p_i| \geq \epsilon) = 0
\]

Note: \( p^* \) is a function of \( s \). The probability \( P_\theta(|p^*_i - p_i| \geq \epsilon) \) is the probability that by sampling from \( P_\theta \) a sample \( s \) will be obtained, so that for the \( p^* \) computed from \( s \) the inequality \(|p^*_i - p_i| \geq \epsilon\) holds.

Similar consistency properties hold for many other types of maximum likelihood estimates.
Chebyshev’s Inequality

A quantitative bound:

\[ P_{\theta}(|p_i^* - p_i| \geq \epsilon) \leq \frac{1}{\epsilon^2 N} p_i (1 - p_i) \]
Prior Beliefs

Classical approach: toss Euro 3 times and always obtain heads
⇒ infer $P(\text{heads}) = 1$.

Inference solely based on observed data – no use of prior knowledge, beliefs, etc.

Bayesian methods: start with an encoding of prior beliefs, and show how to modify them on the basis of observed data.
Estimation: Bayesian

Approach

Model a-priori assumptions on parameter by probability distribution on $\Theta$. Based on observed sample $s$ update to a-posteriori distribution.

**Example 1:** Apparently symmetric coin. $W = \{\text{heads}, \text{tails}\}$. $\theta = P(\text{heads})$.

Observation: sample with 3 *heads* and 7 *tails*
Example 2: Result of first clinical tests of new type of drug. $W = \{\text{pos}, \text{neg}\}$. $\theta = P(\text{pos})$.

Observation: sample with 3 pos and 7 neg
Example 3:
Consider a thumbtack with \( P(pin \ up \mid \theta) = \theta \). If we have no idea about \( \theta \), then we set \( f_{\text{prior}}(\theta) = 1 \). Assume we perform one experiment and get \( pin \ up \):

\[
f_{\text{post}}(\theta \mid pin \ up) = \frac{P(pin \ up \mid \theta)f_{\text{prior}}(\theta)}{P(pin \ up)}
\]
Example 3:
Consider a thumbtack with \( P(\text{pin up} \mid \theta) = \theta \). If we have no idea about \( \theta \), then we set \( f_{\text{prior}}(\theta) = 1 \). Assume we perform one experiment and get \text{pin up}:

\[
f_{\text{post}}(\theta \mid \text{pin up}) = \frac{P(\text{pin up} \mid \theta) f_{\text{prior}}(\theta)}{P(\text{pin up})} = \frac{\theta}{P(\text{pin up})}
\]
Example 3:
Consider a thumbtack with \( P(pin \ up \mid \theta) = \theta \). If we have no idea about \( \theta \), then we set \( f_{prior}(\theta) = 1 \). Assume we perform one experiment and get \( pin \ up \):

\[
\begin{align*}
   f_{post}(\theta \mid pin \ up) &= \frac{P(pin \ up \mid \theta)f_{prior}(\theta)}{P(pin \ up)} \\
   &= \frac{\theta}{P(pin \ up)} \\
   \end{align*}
\]

\[
\begin{align*}
P(pin \ up) &= \int_{0}^{1} \theta \, d\theta = \frac{1}{2}.
\end{align*}
\]
Example 3:
Consider a thumbtack with \( P(pin \ up | \theta) = \theta \). If we have no idea about \( \theta \), then we set \( f_{prior}(\theta) = 1 \). Assume we perform one experiment and get \textit{pin up}:

\[
\begin{align*}
    f_{post}(\theta | pin \ up) &= \frac{P(pin \ up | \theta)f_{prior}(\theta)}{P(pin \ up)} \\
    &= \frac{\theta}{P(pin \ up)} \\
    &= 2\theta \\

    P(pin \ up) &= \int_{0}^{1} \theta \, d\theta = \frac{1}{2}.
\end{align*}
\]
Example 3:
Consider a thumbtack with $P(\text{pin up} | \theta) = \theta$. If we have no idea about $\theta$, then we set $f_{\text{prior}}(\theta) = 1$. Assume we perform one experiment and get $\text{pin up}$:

$$f_{\text{post}}(\theta | \text{pin up}) = 2\theta$$

Assume that we now get a $\text{pin down}$

$$f_{\text{post}_2}(\theta | \text{pin up, pin down}) = \frac{P(\text{pin down, pin up} | \theta)f(\theta)}{P(\text{pin down, pin up})}$$
Example 3: 
Consider a thumbtack with $P\text{(pin up} \mid \theta) = \theta$. If we have no idea about $\theta$, then we set $f_{\text{prior}}(\theta) = 1$. Assume we perform one experiment and get pin up: 

$$f_{\text{post}}(\theta \mid \text{pin up}) = 2\theta$$

Assume that we now get a pin down

$$f_{\text{post}_2}(\theta \mid \text{pin up, pin down}) = \frac{P\text{(pin down, pin up} \mid \theta)f(\theta)}{P\text{(pin down, pin up})}$$

$$= \frac{P\text{(pin down} \mid \theta)P\text{(pin up} \mid \theta)f(\theta)}{P\text{(pin down, pin up})}$$
Example 3:
Consider a thumbtack with \( P(pin\ up\ |\ \theta) = \theta \). If we have no idea about \( \theta \), then we set \( f_{prior}(\theta) = 1 \). Assume we perform one experiment and get \( pin\ up \):

\[
f_{post}(\theta \mid pin\ up) = 2\theta
\]

Assume that we now get a \( pin\ down \)

\[
f_{post_2}(\theta \mid pin\ up, pin\ down) = \frac{P(pin\ down, pin\ up \mid \theta)f(\theta)}{P(pin\ down, pin\ up)}
= \frac{P(pin\ down \mid \theta)P(pin\ up \mid \theta)f(\theta)}{P(pin\ down, pin\ up)}
= \frac{P(pin\ down \mid \theta)\theta}{P(pin\ down, pin\ up)}
\]
**Example 3:**
Consider a thumbtack with \( P(pin \ up \mid \theta) = \theta \). If we have no idea about \( \theta \), then we set \( f_{prior}(\theta) = 1 \).
Assume we perform one experiment and get *pin up*:

\[
 f_{post}(\theta \mid pin \ up) = 2\theta
\]

Assume that we now get a *pin down*

\[
 f_{post_2}(\theta \mid pin \ up, pin \ down) = \frac{P(pin \ down, pin \ up \mid \theta)f(\theta)}{P(pin \ down, pin \ up)} = \frac{P(pin \ down \mid \theta)P(pin \ up \mid \theta)f(\theta)}{P(pin \ down, pin \ up)} = \frac{P(pin \ down \mid \theta)\theta1}{P(pin \ down, pin \ up)} = \frac{(1 - \theta)\theta}{P(pin \ down, pin \ up)}
\]

Note: \( P(pin \ down, pin \ up) = \int_0^1 (1 - \theta)\theta d\theta = \frac{1}{6} \)
Example 3:
Consider a thumbtack with $P(\text{pin up} \mid \theta) = \theta$. If we have no idea about $\theta$, then we set $f_{\text{prior}}(\theta) = 1$. Assume we perform one experiment and get $\text{pin up}$:

$$f_{\text{post}}(\theta \mid \text{pin up}) = 2\theta$$

Assume that we now get a $\text{pin down}$

$$f_{\text{post}_2}(\theta \mid \text{pin up}, \text{pin down}) = \frac{P(\text{pin down, pin up} \mid \theta)f(\theta)}{P(\text{pin down, pin up})}$$

$$= \frac{P(\text{pin down} \mid \theta)P(\text{pin up} \mid \theta)f(\theta)}{P(\text{pin down, pin up})}$$

$$= \frac{P(\text{pin down} \mid \theta)\theta \cdot 1}{P(\text{pin down, pin up})}$$

$$= \frac{(1 - \theta)\theta}{P(\text{pin down, pin up})}$$

$$= 6(1 - \theta)\theta$$

Note: $P(\text{pin down, pin up}) = \int_0^1 (1 - \theta)\theta \, d\theta = \frac{1}{6}$
Update a Prior Distribution

Let \( \{ P_\theta \mid \theta \in \Theta \} \) be a parametric family, \( f_{prior} \) a density on \( \Theta \).

Let \( s_1, \ldots, s_N \in W^N \) be a sample. The posterior density \( f_{post} \) on \( \Theta \) is defined as

\[
f_{post}(\theta) = c \cdot f_{prior}(\theta) P_\theta(s),
\]

where \( c \in \mathbb{R} \) is a normalization constant such that

\[
\int_0^1 f_{post}(\theta) d\theta = 1.
\]
Bayesian Estimator

The Bayesian estimate of the parameter $\theta$ is the mean value of the current distribution on $\Theta$:

$$\theta^* := \int_0^1 \theta f_{\text{current}}(\theta) d\theta$$

Example

In Example 3 we had $f_{\text{post}_2}(\theta \mid \text{pin up, pin down}) = 6(1 - \theta)\theta$, hence

$$\theta^* := \int_0^1 \theta 6(1 - \theta)\theta = \frac{1}{2}.$$
Bayesian Estimator

The Bayesian estimate of the parameter \( \theta \) is the mean value of the current distribution on \( \Theta \):

\[
\theta^* := \int_{0}^{1} \theta f_{\text{current}}(\theta) d\theta
\]

Alternative: MAP (maximum a-posteriori) estimate:

\[
\theta^{\text{map}} := \arg \max_{\theta} f_{\text{current}}(\theta)
\]
The Beta Distribution

For \( W = \{0, 1\} \), \( \Theta = [0, 1] \), \( P_\theta(1) = \theta \) usually use Beta distributions defined by densities:

\[
f_{a,b}(\theta) := \frac{\Gamma(a + b)}{\Gamma(a) \cdot \Gamma(b)} \theta^{a-1}(1 - \theta)^{b-1} \quad (a, b \in \mathbb{R})
\]

where \( \Gamma(x) := \int_0^\infty t^{x-1} e^{-t} dt \). For \( n \in \mathbb{N} \): \( \Gamma(n) = (n - 1)! \).

For positive integer parameters \( a, b \):

\[
f_{a,b}(\theta) = \frac{(a+b-1)!}{(a-1)!(b-1)!} \theta^{a-1}(1 - \theta)^{b-1} = (a + b - 1)\binom{a+b-2}{a-1} \theta^{a-1}(1 - \theta)^{b-1}
\]

i.e. the product of the likelihood of \( \theta \) given a sample of \( a - 1 \) 1’s and \( b - 1 \) 0’s, and a normalization factor \( (a + b - 1) \).
Estimation: Bayesian

Examples

- Beta(1,1)
- Beta(5,5)
- Beta(50,50)
- Beta(4,8)
- Beta(8,12)
- Beta(53,57)
Updating a Beta prior

Updating a Beta prior $f_{a,b}$ with a sample containing $k$ 1's and $l$ 0's leads to a posterior equal to $f_{a+k,b+l}$.

The mean of the beta distribution is

$$\theta^* = \int_0^1 \theta f_{a,b}(\theta) d\theta = \frac{a}{a+b}$$
Estimation: Bayesian

Asymptotic Indepdence from prior

beta(1,1)  beta(6,6)  beta(51,51)
beta(3,7)  beta(8,12)  beta(53,57)
Dirichlet Distribution

For $W = \{w_1, \ldots, w_k\}$ the Beta distribution is generalized by the Dirichlet distribution:

$$f_{a_1, \ldots, a_k}(p_1, \ldots, p_k) := \frac{\Gamma(a_1 + \ldots + a_k)}{\Gamma(a_1) \cdots \Gamma(a_k)} p_1^{a_1} \cdots p_k^{a_k} \quad (a_i \in \mathbb{R})$$
Estimation

Literature

- Any other textbook on statistics.