Advanced AI Techniques

I. Bayesian Networks / 2. Parameter Learning

Wolfram Burgard, Luc de Raedt, Bernhard Nebel, Lars Schmidt-Thieme

Institute for Computer Science
University of Freiburg
http://www.informatik.uni-freiburg.de/

Tutorials

• we will hand out every Thursday 1 exercise sheet with approx. 3 exercises.

• your solutions are due Thursday one week later (hand it in to your tutor)

• every Thursday there will be (in this room)
  – 13–14 tutorial group I (alternating Niels Landwehr and Sunna Torge),
  – 14–15 lecture (for everybody)
  – 15–16 tutorial group II (Alexander Scivos)

• solutions will be corrected by your tutor

• co-operation in groups of \( \leq 3 \) students is encouraged (1 solution / group).

• in the tutorial solutions will be discussed you submitted the week before.

Please create student groups now and enter your names on one of the registration forms.
1. Maximum Likelihood Parameter Estimates

2. Bayesian Parameter Estimates / One Variable

3. Bayesian Parameter Estimates / Several Variables

Given

- a bayesian network structure $G := (V, E)$ on a set of variables $V$ and

- a data set $D \in \text{dom}(V)^*$ of cases.

Learning the parameters of the bayesian network means to find vertex potentials

$$(p_v)_{v \in V}$$

s.t. some optimality criterion w.r.t. $G$ and $D$ holds.
The simplest criterion is the maximum likelihood criterion, i.e., the probability of the data given the bayesian network is maximal:

$$\text{find } (p_v)_{v \in V} \text{ s.t. } p(D) \text{ is maximal},$$

where $p$ denotes the JPD build from $(p_v)_{v \in V}$.

$$p(D) = \prod_{d \in D} p(d) = \prod_{d \in D} \prod_{v \in V} p_v(d|\text{fam}(v))$$

$(p_v)_{v \in V}$ with maximal $p(D)$ are called maximum likelihood estimates. $p$ is also called likelihood.

Data:

<table>
<thead>
<tr>
<th>X</th>
<th>Y</th>
<th>$p_1$(case)</th>
<th>$p_2$(case)</th>
</tr>
</thead>
<tbody>
<tr>
<td>H</td>
<td>T</td>
<td>0.25</td>
<td>0.25</td>
</tr>
<tr>
<td>H</td>
<td>H</td>
<td>0.25</td>
<td>0.2</td>
</tr>
<tr>
<td>T</td>
<td>H</td>
<td>0.25</td>
<td>0.3</td>
</tr>
<tr>
<td>T</td>
<td>T</td>
<td>0.25</td>
<td>0.25</td>
</tr>
<tr>
<td>T</td>
<td>H</td>
<td>0.25</td>
<td>0.3</td>
</tr>
<tr>
<td>T</td>
<td>T</td>
<td>0.25</td>
<td>0.3</td>
</tr>
<tr>
<td>H</td>
<td>T</td>
<td>0.25</td>
<td>0.25</td>
</tr>
<tr>
<td>T</td>
<td>T</td>
<td>0.25</td>
<td>0.25</td>
</tr>
</tbody>
</table>

$p(D) = 1.5259 \cdot 10^{-5} \quad 2.1093 \cdot 10^{-9}$

$\text{JPD}_1$:  

$$\begin{array}{c|cc}
Y & H & T \\
\hline
X = H & .25 & .25 \\
T & .25 & .25 \\
\end{array}$$

$\text{JPD}_2$:  

$$\begin{array}{c|cc}
Y & H & T \\
\hline
X = H & .2 & .25 \\
T & .3 & .25 \\
\end{array}$$
Lemma 1. \( p(D) \) is maximal iff

\[
p_v(x|y) := \frac{\left| \{ d \in D \mid d|_v = x, d|_{\text{pa}(v)} = y \} \right|}{\left| \{ d \in D \mid d|_{\text{pa}(v)} = y \} \right|}
\]

(if there are \( d \in D \) with \( d|_{\text{pa}(v)} = y \), otherwise \( p_v(x|y) \) can be choosen arbitrarily – \( p(D) \) does not depend on it).

Instead of the likelihood \( p \) often \( \log p \) is used, called log-likelihood.

Proof. Due to independence of the cases and the factorization of the JPD in bayesian networks, \( p(D) \) factors as

\[
p(D) = \prod_{d \in D} p(d) = \prod_{d \in D} \prod_{v \in V} p_v(d|_{\text{fam}(v)}) = \prod_{v \in V} \prod_{d \in D|_{\text{fam}(v)}} p_v(d)
\]

which is maximal if for all \( v \in V \)

\[
p_v(D) := \prod_{d \in D|_{\text{fam}(v)}} p_v(d) = \prod_{x \in \text{dom}(v)} \prod_{y \in \text{dom}(\text{pa}(v))} p_v(x|y)^{n_D(x,y)}
\]

is maximal, with count data

\[
n_D(x, y) := \left| \{ d \in D \mid d|_v = x, d|_{\text{pa}(v)} = y \} \right|
\]

\( p_v \) in turn is maximal if for all \( x \in \text{dom}(v) \)

\[
\prod_{y \in \text{dom}(\text{pa}(v))} p_v(x|y)^{n_D(x,y)}
\]

is maximal. As beneath \( p_v(x|y) \in [0, 1] \) the only constraint to \( p_v(x|y) \) is

\[
\sum_{y \in \text{dom}(\text{pa}(v))} p_v(x|y) = 1
\]

we have with an arbitrary \( y_0 \in \text{dom}(\text{pa}(v)) \)

\[
\prod_{y \in \text{dom}(\text{pa}(v))} p_v(x|y)^{n_D(x,y)} \cdot (1 - \sum_{y \in \text{dom}(\text{pa}(v))} p_v(x|y))^{n_D(x,y_0)}
\]

Taking logarithms we get

\[
\sum_{y \in \text{dom}(\text{pa}(v))} n_D(x, y) \log p_v(x|y) + n_D(x, y_0) \log (1 - \sum_{y \in \text{dom}(\text{pa}(v))} p_v(x|y))
\]
Let $\text{dom}(\text{pa}(v)) := \{y_0, \ldots, y_n\}$ an enumeration and write $x_i := (x, y_i)$, $p_i := p_v(x|y_i)$ and $n_i := n_D(x, y_i)$, then we can simplify notation to

$$L(p) := \sum_{i=1}^{n} n_i \log p_i + n_0 \log(1 - \sum_{i=1}^{n} p_i)$$

To be minimal, derivative has to vanish:

$$\frac{\partial L}{\partial p_j} = n_j \frac{1}{p_j} - n_0 \frac{1}{1 - \sum_{i=1}^{n} p_i} = 0$$

which yields

$$p_j = \frac{n_j}{n_0} (1 - \sum_{i=1}^{n} p_i)$$

Summing over $j = 1, \ldots, n$ we get

$$\sum_{i=1}^{n} p_i = \left( \sum_{i=1}^{n} \frac{n_i}{n_0} \right) (1 - \sum_{i=1}^{n} p_i)$$

Solving for $\sum_{i=1}^{n} p_i$ yields

$$\sum_{i=1}^{n} p_i = \frac{\sum_{i=1}^{n} n_i}{n_0 + \sum_{i=1}^{n} n_i}$$

and substituting this in the equations for $p_j$ finally yields

$$p_j = \frac{n_j}{n_0} \left( 1 - \sum_{i=1}^{n} p_i \right) = \frac{n_j}{n_0 + \sum_{i=1}^{n} n_i}$$
Simplest case:

- one variable $X$,
- variable is binary (= 2 states, H and T)

$\Rightarrow$ 1 parameter $\theta$.

### Example I: flip a coin

We flip a coin with possible outcomes head (H) or tail (T).

**actual sample:**

| actual | H | T | H | H | H |

**parameter estimation:**

\[
p(X = H) = \frac{4}{5} = 0.8
\]
1. Be able to combine
   - prior / background knowledge with
   - actual observations / new data

Example II: "flip a cat"

We observe cats falling from window seats with possible outcomes
   - lands on its paws (P) or
   - does not land on its paws (¬P)

actual sample:

<table>
<thead>
<tr>
<th>actual</th>
<th>¬P</th>
<th>P</th>
<th>¬P</th>
<th>¬P</th>
</tr>
</thead>
</table>

parameter estimation:

\[ \hat{p}(X = \neg P) = \frac{4}{5} = 0.8 \]
2. Be able to express different prior probabilities:
   - all events have same prior probability (e.g., coin, dice)
   - each event has a specific prior probability (e.g., cats landing on paws vs. not).

3. Be able to express different strengths of prior believes:

   **strong prior believes:**
   
   Many contradicting actual observations are necessary to overwrite prior believes.
   
   "I am quite sure in advance."

   **weak prior believes:**
   
   Already a few contradicting actual observations are sufficient to overwrite prior believes.
   
   "I guess, but really do not know in advance."
A Simple Model for Prior Believes

We model

• prior probabilities by a probability distribution $p_{\text{prior}}$.

• the strength of the prior believes by a prior sample size $n_{\text{prior}}$.

$$\hat{p} := \frac{n_{\text{prior}}}{n_{\text{prior}} + n_{\text{actual}}} p_{\text{prior}} + \frac{n_{\text{actual}}}{n_{\text{prior}} + n_{\text{actual}}} \hat{p}_{\text{actual}}$$

Prior sample size quantifies, how many actual observations we need s.t. prior and actual estimates have the same influence on our final estimates.

---

actual sample:

<table>
<thead>
<tr>
<th>actual</th>
<th>H</th>
<th>T</th>
<th>H</th>
<th>H</th>
<th>H</th>
</tr>
</thead>
</table>

$n_{\text{actual}} = 5$

<table>
<thead>
<tr>
<th></th>
<th>H</th>
<th>T</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\hat{p}_{\text{actual}}$</td>
<td>0.8</td>
<td>0.2</td>
</tr>
</tbody>
</table>

combined estimate:

$$\hat{p} := \frac{n_{\text{prior}}}{n_{\text{prior}} + n_{\text{actual}}} p_{\text{prior}} + \frac{n_{\text{actual}}}{n_{\text{prior}} + n_{\text{actual}}} \hat{p}_{\text{actual}}$$

$$= \frac{10}{15} \cdot 0.5 + \frac{5}{15} \cdot 0.8 = 0.6$$

<table>
<thead>
<tr>
<th>$\hat{p}$</th>
<th>H</th>
<th>T</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0.6</td>
<td>0.4</td>
</tr>
</tbody>
</table>
Prior Sample Size

Prior sample size can be understood literally as the size of a prior sample that we combine with the actual sample for our estimations.

actual sample:

\[
\begin{array}{cc}
\text{actual} & H & T & H & H & H \\
\end{array}
\]

\[n_{\text{actual}} = 5\]

\[
\begin{array}{cc}
\hat{p}_{\text{actual}} & 0.8 & 0.2 \\
\end{array}
\]

prior sample:

\[
\begin{array}{ccccccccccc}
\text{prior} & H & H & H & H & T & T & T & T & T \\
\end{array}
\]

combined sample:

\[
\begin{array}{ccccccccccc}
\text{combined} & H & H & H & H & H & T & T & T & T & T & H & H \\
\end{array}
\]

\[
\begin{array}{cc}
\hat{p} & 0.6 & 0.4 \\
\end{array}
\]

But,

- not all prior probabilities and prior sample sizes can be expressed equivalently as prior samples, e.g.,

\[
\begin{array}{cc}
\hat{p}_{\text{prior}} & 0.5 & 0.5 \\
\end{array}
\]

\[n_{\text{prior}} = 10\]

- prior sample size also can be choosen as fractional value, e.g.,

\[n_{\text{prior}} = 0.1\]
So far, if we specify

\[
\begin{array}{c|cc}
    & H & T \\
\hline
p_{\text{prior}} & 0.5 & 0.5 \\
\end{array}
\]

then . . .

. . . for the discrete attribute \( X \) in our model

we specify its prior distribution

\[ p_{\text{prior}}(X) \]

consisting of

\[
\begin{align*}
p_{\text{prior}}(X = H) &= \theta = 0.5 \\
p_{\text{prior}}(X = T) &= 1 - \theta = 0.5
\end{align*}
\]

. . . for the parameter \( \Theta := p_{\text{prior}}(X = H) \)

we specify its expected value \( \hat{\theta} = 0.5 \).
Figure 1: $p(\Theta) = \beta_{5,5}(\Theta)$: we expect the true parameter to be at 0.5.

$\hat{\theta} = 0.5$

$\int_{0.4}^{0.6} p(\Theta)d\Theta = .467$

Figure 2: $p(\Theta) = \beta_{5,5}(\Theta)$: we expect the true parameter to be at 0 or at 1.

$\hat{\theta} = 0.5$

$\int_{0.4}^{0.6} p(\Theta)d\Theta = .128$
Compute the expected value $\theta$ of its a posterior distribution

$$\hat{\theta}_{\text{MAP}} := E(p(\theta \mid d))$$

called **maximal a posterior estimator (MAP)** of $\Theta$.

Use Bayes’ formula:

$$p(\theta \mid d) = \frac{p(d \mid \theta) p(\theta)}{p(d)}$$

$$p(d \mid \theta) = \prod_{x \in d} \theta^\delta_{x=H}(1 - \theta)^{\delta_{x=T}}$$

$$= \theta^{|\{x \in d \mid x=H\}|}(1 - \theta)^{|\{x \in d \mid x=T\}|}$$

actual sample:

<table>
<thead>
<tr>
<th></th>
<th>H</th>
<th>T</th>
<th>H</th>
<th>H</th>
<th>H</th>
</tr>
</thead>
</table>

$$p(d \mid \theta) = \theta^{|\{x \in d \mid x=H\}|}(1 - \theta)^{|\{x \in d \mid x=T\}|}$$

$$= \theta^4 (1 - \theta)^1$$

$$p(\theta) = \beta_{5,5}(\theta)$$

Computing expectation of

$$p(\theta \mid d) = \frac{p(d \mid \theta) p(\theta)}{p(d)} = \frac{\theta^4 (1 - \theta)^1 \beta_{5,5}(\theta)}{p(d)}$$

leads to

$$\hat{p}_{\text{MAP}} = 0.6$$
For general priors, we have to solve a 1-dimensional integration problem:

\[ p(\theta | d) = \frac{p(d | \theta) p(\theta)}{p(d)} = \frac{\theta^s (1 - \theta)^t p(\theta)}{p(d)} \]

where \( s := |\{ x \in d \mid x = H \}| \) and \( t := |\{ x \in d \mid x = T \}| \).

nice:
- Problem depends on the data only via summary statistics \( s, t \).

not so nice:
- Complicated priors \( p(\theta) \) may not have analytical solutions.

Definition 1. Gamma function

\[ \Gamma(x) := \int_{0}^{\infty} t^{x-1} e^{-t} dt \]

converging for \( x > 0 \).

Lemma 2 (\( \Gamma \) is generalization of factorial).

(i) \( \Gamma(n) = (n - 1)! \) for \( n \in \mathbb{N} \).

(ii) \( \frac{\Gamma(x+1)}{\Gamma(x)} = x \).
Definition 2. Beta function

\[ \beta(x, y) := \frac{\Gamma(x + y)}{\Gamma(x)\Gamma(y)} \]

defined for \( x, y > 0 \).

Lemma 3 (\( \beta \) is generalization of binomial).

\[ \beta(n-m, m) = \binom{n}{m} \quad \text{for} \quad n, m \in \mathbb{N}, n > m \]

Definition 3. Beta distribution has density

\[ \beta_{a,b}(x) := \frac{1}{\beta(a, b)} x^{a-1}(1 - x)^{b-1} \]

defined on \([0, 1]\).
Lemma 4.

\[ E(\beta_{a,b}(\theta)) = \frac{a}{a+b} \]

Proof.

\[
E(\beta_{a,b}(\theta)) = \int_0^1 \theta \beta_{a,b}(\theta) d\theta
\]

\[
= \int_0^1 \frac{1}{\beta(a,b)} \theta \theta^{a-1}(1-\theta)^{b-1} d\theta
\]

\[
= \frac{\beta(a+1,b)}{\beta(a,b)} \int_0^1 \beta_{a+1,b}(\theta) d\theta
\]

\[
= \frac{\Gamma(a+1)\Gamma(b)\Gamma(a+b)}{\Gamma(a+b+1)\Gamma(b)}
\]

\[
= \frac{a}{a+b}
\]

Lemma 5 (beta is conjugated prior for binomial samples). For a beta prior, the a posterior again is beta:

\[ p(\theta \mid d) = \beta_{s+a,t+b}(\theta) \]

for \( p_{\text{prior}}(\theta) = \beta_{a,b}(\theta) \) and \( s := \{|x \in d \mid x = H\} \) and \( t := \{|x \in d \mid x = T\} \).

Proof.

\[
p(\theta \mid d) = \frac{p(d \mid \theta)p(\theta)}{p(d)}
\]

\[ p(d \mid \theta)p(\theta) = \theta^s (1-\theta)^t \beta_{a,b}(\theta)
\]

\[ = \theta^s (1-\theta)^t \frac{1}{\beta(a,b)} \theta^{a-1}(1-\theta)^{b-1}
\]

\[ = \frac{\beta(s+a,t+b)}{\beta(a,b)} \beta_{s+a,t+b}(\theta)
\]

\[ p(d) = E(\theta^s(1-\theta)^t)
\]

\[ = \int_0^1 \theta^s (1-\theta)^t d\theta
\]

\[ = \int_0^1 \theta^s (1-\theta)^t \frac{1}{\beta(a,b)} \theta^{a-1}(1-\theta)^{b-1} dx
\]

\[ = \frac{\beta(s+a,t+b)}{\beta(a,b)} \int_0^1 \beta_{s+a,t+b}(x) dx
\]

\[ = \frac{\beta(s+a,t+b)}{\beta(a,b)}
\]

\[ \frac{p(\theta \mid d)}{p(d)} = \frac{p(d \mid \theta)p(\theta)}{p(d)} = \beta_{s+a,t+b}(\theta)
\]
If we choose a beta distribution as prior, i.e.,
\[ p(\theta) = \beta_{a,b}(\theta) \]
then we can compute the a posterior analytically:
\[ p(\theta | d) = \beta_{s+a,t+b}(\theta) \]
and by taking expectations, compute parameter values also analytically:
\[ \hat{\theta}_{\text{MAP}} = E(p(\theta | d)) = E(\beta_{s+a,t+b}(\theta)) = \frac{s + a}{s + a + t + b} \]
A closer look at
\[ \hat{\theta}_{\text{MAP}} = \frac{s + a}{s + a + t + b} \]
With
\[ \theta_{\text{prior}} := \frac{a}{a + b}, \quad n_{\text{prior}} := a + b \]
and
\[ \hat{\theta}_{\text{actual}} = \frac{s}{s + t}, \quad n_{\text{actual}} = s + t \]
we have exactly
\[ \hat{\theta}_{\text{MAP}} = \frac{n_{\text{prior}}}{n_{\text{prior}} + n_{\text{actual}}} \theta_{\text{prior}} + \frac{n_{\text{actual}}}{n_{\text{prior}} + n_{\text{actual}}} \hat{\theta}_{\text{actual}} \]
SUFFICIENT STATISTICS
DATA

\[ \begin{array}{c|c|c|c|c}
\text{H} & \text{T} & \text{H} & \text{H} \\
\text{H} & \text{T} & 4 & 1 \\
\end{array} \]

\[ \begin{array}{c|c|c}
\theta & 1-\theta \\
\theta & \beta(5,5) \\
\end{array} \]
Advanced AI Techniques / 2. Bayesian Parameter Estimates / One Variable

DATA

SUFFICIENT STATISTICS

UPDATE

MARGINALIZE TO EMBEDDED BN

Wolfram Burgard, Luc de Raedt, Bernhard Nebel, Lars Schmidt-Thieme, Institute for Computer Science, University of Freiburg, Germany,
Course on Advanced AI Techniques, winter term 2004
1. Maximum Likelihood Parameter Estimates

2. Bayesian Parameter Estimates / One Variable

3. Bayesian Parameter Estimates / Several Variables

More than one variable
Parameter priors are independent (as roots in a BN):

- priors for parameters of different variables (e.g., $\theta$ and $\{\eta_1, \eta_2\}$; **global parameter independence**)

- as well as priors for different parameters of the same variable (e.g., $\eta_1$ and $\eta_2$; **local parameter independence**).

Lemma 6 (global and local parameter posterior independence).

$$p(\theta_{1,1}, \theta_{1,2}, \ldots, \theta_{n,q_n} \mid d) = \prod_{i=1}^{n} \prod_{j=1}^{q_i} p(\theta_{i,j} \mid d)$$

Proof see Theorem 6.12, p. 337 of Neapolitan.

This means:

- we can compute each parameter on its own.

- same technique as for one parameter seen before.
DATA

\[
\begin{array}{ccc}
\text{X} & \text{Y} \\
H & T & H \\
H & T & H \\
T & T & H \\
T & T & H \\
T & T & T \\
T & T & T \\
T & T & T \\
T & T & T \\
\end{array}
\]
Advanced AI Techniques / 3. Bayesian Parameter Estimates / Several Variables

DATA

UPDATE

MARGINALIZE TO EMBEDDED BN

Wolfram Burgard, Luc de Raedt, Bernhard Nebel, Lars Schmidt-Thieme, Institute for Computer Science, University of Freiburg, Germany, Course on Advanced AI Techniques, winter term 2004

34/40
\[ \hat{p}(Y = H) = 0.4 \cdot 0.4 + 0.6 \cdot \frac{4}{7} = 0.503 \]

\[ \hat{p}(Y = H) = 0.4 \cdot \frac{3}{8} + 0.6 \cdot \frac{7}{12} = 0.5 \]
Definition 4. Let $\beta_{i,j,b_{i,j}}$ the priors in an augmented BN $(i = 1, \ldots, n; j = 1, \ldots, q_i)$.

If there is a number $N$ with

$$a_{i,j} + b_{i,j} = p(pa_{i,j}) \cdot N$$

for all $i$ and $j$, the BN is said to have **equivalent sample size** $N$.

If all variables are binary, use equivalent sample size 2 to express an uninformative prior.

### Multinomial Variables

So far we have looked at

- one binary variable (i.e., having two different values)

and

- several binary variables

Now we look at

- one or several multinomial variables (i.e., having $n$ different values)
Definition 5. **Dirichlet distribution** has density

$$\text{Dir}_{a_1,a_2,\ldots,a_n}(x_1, x_2, \ldots, x_{n-1}) :=$$

$$\frac{\Gamma(a_1 + a_2 + \ldots + a_n)}{\prod_{i=1}^{n} \Gamma(a_i)} x_1^{a_1-1} x_2^{a_2-1} \cdots x_{n-1}^{a_{n-1}-1} (1 - x_1 - x_2 - \ldots - x_{n-1})^{a_n-1}$$

defined on \( \{ x \in [0, 1]^{n-1} \mid x_1 + x_2 + \cdots + x_{n-1} \leq 1 \} \).

For the special case of a binary variable \((n = 2)\):

$$\text{Dir}_{a_1,a_2}(x) = \beta_{a_1,a_2}(x)$$

**Lemma 7.** For \(i = 1, \ldots, n\):

$$E_{\text{Dir}_{a_1,a_2,\ldots,a_n}}(X_i) = \frac{a_i}{a_1 + a_2 + \cdots + x_n}$$

(where \(X_n := 1 - X_1 - X_2 - \ldots - X_{n-1}\)).
Lemma 8 (Dirichlet is conjugated prior for multinomial samples). For a Dirichlet prior, the a posterior again is Dirichlet:

\[ p(\theta \mid d) = \text{Dir}_{a_1+s_1, a_2+s_2, \ldots, a_n+s_n}(\theta) \]

for \( p_{\text{prior}}(\theta) = \text{Dir}_{a_1, a_2, \ldots, a_n}(\theta) \) and \( s_i := \#\{x \in d \mid x = i\} \) (\( i = 1, \ldots, n \)).

This means:

- We can compute each parameter estimate by counting, as in the binary case.

- Due to global and local posterior parameter independence, we can estimate each parameter on its own.

\[ \leadsto \text{same procedure as for binary variables seen before.} \]