The EM algorithm for Bayesian networks

Aarhus University, Fall 2003, Lecture 8

Steffen L. Lauritzen, Aalborg University

Entropy

The *entropy* of a discrete probability distribution P is

$$\mathrm{Ent}(P) = -\sum_x p(x) \log p(x).$$

Entropy is a measure of *spread* of the distribution and it is always positive.

The entropy is never larger than the entropy of the uniform distribution:

Let $P_u(x) = 1/|\mathcal{X}|$, then it holds that

$$0 \leq \mathrm{Ent}(P) \leq \mathrm{Ent}(P_u) = \log |\mathcal{X}|.$$

Proof on next overhead.

Uniform distribution has maximal entropy

The information inequality

$$\sum_x p(x) \log p(x) \geq \sum_x p(x) \log q(x)$$

yields

$$\begin{aligned}
\mathrm{Ent}(P) &= -\sum_x p(x) \log p(x) \\
&\leq -\sum_x p(x) \log \frac{1}{|\mathcal{X}|} \\
&= -\sum_x \frac{1}{|\mathcal{X}|} \log \frac{1}{|\mathcal{X}|} = \mathrm{Ent}(P_u).
\end{aligned}$$

Kullback-Leibler divergence

The *KL divergence* between P and Q is

$$KL(P:Q) = \sum_x p(x) \log \frac{p(x)}{q(x)}.$$

Also known as *relative entropy* of Q with respect to P.

Information inequality says that

$KL(P:Q) \geq 0$ and $KL(P:Q) = 0$ if and only if $P = Q$,

so KL divergence defines an (asymmetric) distance measure between probability distributions.

Incomplete observations

Bayesian network with conditional probability distributions only partially known:

$$p(x) = \prod_{v \in V} p(x_v \mid x_{\mathrm{pa}(v)}, \theta)$$

where $\theta \in \Theta \subseteq \mathcal{R}^k$ is unknown parameter.

Instead of *complete data* $(x^1, \ldots, x^n)$, only *incomplete data* $(x^1_{A_1}, \ldots, x^n_{A_n})$ available, where $A_i \subseteq V$.

Example: paternity cases. Unknown parameters: gene frequencies, probability of paternity, possibly mutation rates.

EM algorithm

Complete data x, incomplete data (observed) $y = g(x)$.
Complete data log-likelihood:

$$l(\theta) = \log L(x \mid \theta) = \log p(x \mid \theta).$$

The *marginal log-likelihood* is

$$l_y(\theta) = \log L(\theta \mid y) = \log p(y \mid \theta).$$

Wish to maximize l_y in θ but l_y is unpleasant:

$$l_y(\theta) = \log \sum_{x:g(x)=y} p(x \mid \theta).$$

However, we assume that we know how to maximize l.
How can this be exploited?

E-step and M-step

We let θ^* be arbitrary but fixed.

The **E-step** calculates *expected complete data log-likelihood* $q(\theta \mid \theta^*)$:

$$q(\theta \mid \theta^*) = \mathbf{E}_{\theta^*}\{l(\theta) \mid y\} = \sum_{x:g(x)=y} p(x \mid y, \theta^*) \log p(x \mid \theta).$$

The **M-step** maximizes $q(\cdot \mid \theta^*)$ for fixed θ^*:

The algorithm alternates between an E-step and an M-step.

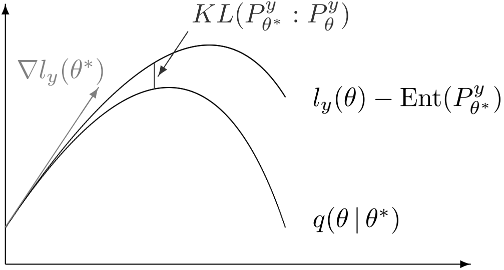
After an E-step and subsequent M-step, the likelihood function has never decreased, as we shall now show.

EM algorithm

Since $p(x \mid y, \theta) = \chi_{g(x)}(y) p(x \mid \theta) / p(y \mid \theta)$ we have

$$\begin{aligned} q(\theta \mid \theta^*) &= \sum_x p(x \mid y, \theta^*) \log\{p(y \mid \theta) p(x \mid y, \theta)\} \\ &= \log p(y \mid \theta) + \sum_x p(x \mid y, \theta^*) \log p(x \mid y, \theta) \\ &= l_y(\theta) - \sum_x p(x \mid y, \theta^*) \log p(x \mid y, \theta^*) \\ &\quad - \sum_x p(x \mid y, \theta^*) \log \frac{p(x \mid y, \theta^*)}{p(x \mid y, \theta)} \\ &= l_y(\theta) - \mathrm{Ent}\, P^y_{\theta^*} - KL(P^y_{\theta^*} : P^y_\theta). \end{aligned}$$

Expected and complete data likelihood



$$l_y(\theta) - \mathrm{Ent}(P^y_{\theta^*}) = q(\theta \,|\, \theta^*) + KL(P^y_{\theta^*} : P^y_\theta)$$

$$\nabla l_y(\theta^*) = \nabla q(\theta^* \,|\, \theta^*)$$

Likelihood monotonicity of EM algorithm

Let $\theta_0 = \theta^*$ and $\theta_{n+1} = \arg\max_\theta q(\theta \,|\, \theta_n)$.

Then

$$
\begin{aligned}
l_y(\theta_{n+1}) &= q(\theta_{n+1} \,|\, \theta_n) + \mathrm{Ent}(P^y_{\theta_n}) + KL(P^y_{\theta_{n+1}} : P^y_{\theta_n}) \\
&\geq q(\theta_n \,|\, \theta_n) + \mathrm{Ent}(P^y_{\theta_n}) = l_y(\theta_n).
\end{aligned}
$$

So likelihood never decreases. Note, this also holds if just $q(\theta_{n+1} \,|\, \theta_n) \geq q(\theta_n \,|\, \theta_n)$.

E-step for Bayesian networks

The complete data likelihood is

$$\log p(x \mid \theta) = \sum_{i=1}^{n} \log p(x^i \mid \theta) = \sum_{x} n(x) \log p(x \mid \theta).$$

where $n(x) = \#\{i : x^i = x\}$. Using factorization we get

$$\begin{aligned} \log p(x \mid \theta) &= \sum_{x} \sum_{v} n(x) \log p(x_v \mid x_{\mathrm{pa}(v)}, \theta) \\ &= \sum_{v} \sum_{x_{v \cup \mathrm{pa}(v)}} n(x_{v \cup \mathrm{pa}(v)}) \log p(x_v \mid x_{\mathrm{pa}(v)}, \theta), \end{aligned}$$

with $n(x_{v \cup \mathrm{pa}(v)}) = \#\{i : x^i_{v \cup \mathrm{pa}(v)} = x_{v \cup \mathrm{pa}(v)}\}$. So E-step equivalent to computing

$$n^*(x_{v \cup \mathrm{pa}(v)}) = \mathbf{E}\{N(x_{v \cup \mathrm{pa}(v)}) \mid y, \theta^*\}.$$

Computing expected counts

We now get

$$\begin{aligned} n^*(x_{v\cup\mathrm{pa}(v)}) &= \mathbf{E}\{N(x_{v\cup\mathrm{pa}(v)}) \,|\, y, \theta^*\} \\ &= \sum_i \mathbf{E}\{\chi_{x_{v\cup\mathrm{pa}(v)}}(x^i_{v\cup\mathrm{pa}(v)}) \,|\, y, \theta^*\} \\ &= \sum_i \mathbf{E}\{\chi_{x_{v\cup\mathrm{pa}(v)}}(x^i_{v\cup\mathrm{pa}(v)}) \,|\, x^i_{A_i}, \theta^*\} \\ &= \sum_i p(x_{v\cup\mathrm{pa}(v)} \,|\, x^i_{A_i}, \theta^*). \end{aligned}$$

Each of the latter terms can be calculated by probability propagation as can the marginal likelihood function:

$$\log p(y \,|\, \theta) = \sum_i \log p(x^i_{A_i} \,|\, \theta).$$

M-step for Bayesian networks

Note the similarity between the complete data likelihood and q:

$$\log p(x \mid \theta) = \sum_v \sum_{x_{v \cup \mathrm{pa}(v)}} n(x_{v \cup \mathrm{pa}(v)}) \log p(x_v \mid x_{\mathrm{pa}(v)}, \theta)$$

whereasx

$$q(\theta \mid \theta^*) = \sum_v \sum_{x_{v \cup \mathrm{pa}(v)}} n^*(x_{v \cup \mathrm{pa}(v)}) \log p(x_v \mid x_{\mathrm{pa}(v)}, \theta).$$

So any algorithm which maximizes the complete data likelihood can be used to maximize q in the M-step.