Statistical Approaches to Learning and Discovery

Variational Approximations

Zoubin Ghahramani & Teddy Seidenfeld
zoubin@cs.cmu.edu & teddy@stat.cmu.edu

CALD / CS / Statistics / Philosophy
Carnegie Mellon University
Spring 2002
Review: The EM algorithm

Given a set of observed (visible) variables $V$, a set of unobserved (hidden / latent / missing) variables $H$, and model parameters $\theta$, optimize the log likelihood:

$$\mathcal{L}(\theta) = \log p(V|\theta) = \log \int p(H, V|\theta) dH,$$

Using Jensen’s inequality, for any distribution of hidden variables $q(H)$ we have:

$$\mathcal{L}(\theta) = \log \int q(H) \frac{p(H, V|\theta)}{q(H)} dH \geq \int q(H) \log \frac{p(H, V|\theta)}{q(H)} dH = \mathcal{F}(q, \theta),$$

defining the $\mathcal{F}(q, \theta)$ functional, which is a lower bound on the log likelihood.

In the EM algorithm, we alternately optimize $\mathcal{F}(q, \theta)$ wrt $q$ and $\theta$, and we can prove that this will never decrease $\mathcal{L}$. 
The E and M steps of EM

The lower bound on the log likelihood:

$$\mathcal{F}(q, \theta) = \int q(H) \log \frac{p(H, V|\theta)}{q(H)} dH = \int q(H) \log p(H, V|\theta) dH + \mathcal{H}(q),$$

where $\mathcal{H}(q) = -\int q(H) \log q(H) dH$ is the entropy of $q$. We iteratively alternate:

**E step:** maximize $\mathcal{F}(q, \theta)$ wrt the distribution over hidden variables given the parameters:

$$q^{(k)}(H) := \arg\max_{q(H)} \mathcal{F}(q(H), \theta^{(k-1)}).$$

**M step:** maximize $\mathcal{F}(q, \theta)$ wrt the parameters given the hidden distribution:

$$\theta^{(k)} := \arg\max_{\theta} \mathcal{F}(q^{(k)}(H), \theta) = \arg\max_{\theta} \int q^{(k)}(H) \log p(H, V|\theta) dH,$$

which is equivalent to optimizing the expected complete-data likelihood $p(H, V|\theta)$, since the entropy of $q(H)$ does not depend on $\theta$. 
EM as Coordinate Ascent in $\mathcal{F}$
Variational Approximations to the EM algorithm

Often \( p(H|V, \theta) \) is computationally \textit{intractable}, so an exact E step is out of the question.

**Assume some simpler form for** \( q(H) \), e.g. \( q \in Q \), the set of fully-factorized distributions over the hidden variables: \( q(H) = \prod_i q(H_i) \)

**E step** (approximate): maximize \( \mathcal{F}(q, \theta) \) wrt the distribution over hidden variables given the parameters:

\[
q^{(k)}(H) := \arg\max_{q(H) \in Q} \mathcal{F}(q(H), \theta^{(k-1)}).
\]

**M step** : maximize \( \mathcal{F}(q, \theta) \) wrt the parameters given the hidden distribution:

\[
\theta^{(k)} := \arg\max_{\theta} \mathcal{F}(q^{(k)}(H), \theta) = \arg\max_{\theta} \int q^{(k)}(H) \log p(H, V | \theta) dH,
\]

This maximizes a lower bound on the log likelihood.

Using the fully factorized form of \( q \) is sometimes called a **mean-field approximation**.
Example: A Multiple Cause Model

Model with binary latent variables $s_i$, real-valued observed vector $y$ and parameters $	heta = \{\{\mu_i, \pi_i\}_{i=1}^K, \sigma^2\}$

$$p(s_1, \ldots, s_K | \pi) = \prod_{i=1}^K p(s_i) = \prod_{i=1}^K \pi_i^{s_i} (1 - \pi_i)^{1-s_i}$$

$$p(y | s_1, \ldots, s_K | \mu, \sigma^2) = \mathcal{N}(\sum_i s_i \mu_i, \sigma^2 I)$$

EM optimizes lower bound on likelihood:

$$\mathcal{F}(q, \theta) = \langle \log p(s, y | \theta) \rangle_{q(s)} - \langle \log q(s) \rangle_{q(s)}$$

where $\langle \rangle_q$ is expectation under $q$.

**Optimum E step:** $q(s) = p(s | y, \theta)$ is exponential in $K$. 
$\mathcal{F}(q, \theta) = \langle \log p(s, y|\theta) \rangle_{q(s)} - \langle \log q(s) \rangle_{q(s)}$

$\log p(s, y|\theta) + c$

$= \sum_{i=1}^{K} s_i \log \pi_i + (1 - s_i) \log (1 - \pi_i) - D \log \sigma - \frac{1}{2\sigma^2}(y - \sum_i s_i \mu_i)^	op (y - \sum_i s_i \mu_i)$

$= \sum_{i=1}^{K} s_i \log \pi_i + (1 - s_i) \log (1 - \pi_i) - D \log \sigma$

$- \frac{1}{2\sigma^2}(y^\top y - 2 \sum_i s_i \mu_i^\top y + \sum_i \sum_j s_i s_j \mu_i^\top \mu_j)$

we therefore need $\langle s_i \rangle$ and $\langle s_i s_j \rangle$ to compute $\mathcal{F}$.

These are the expected sufficient statistics of the hidden variables.
Example: A Multiple Cause Model (cont)

Variational approximation:

\[ q(s) = \prod_i q_i(s_i) = \prod_{i=1}^K \lambda_i^{s_i} (1 - \lambda_i)^{(1-s_i)} \]

Under this approximation we know \( \langle s_i \rangle = \lambda_i \) and \( \langle s_i s_j \rangle = \lambda_i \lambda_j + \delta_{ij}(\lambda_i - \lambda_i^2) \).

\[ \mathcal{F}(\lambda, \theta) = \sum_i \lambda_i \log \frac{\pi_i}{\lambda_i} + (1 - \lambda_i) \log \frac{(1 - \pi_i)}{(1 - \lambda_i)} \\
- D \log \sigma - \frac{1}{2\sigma^2} (y - \sum_i \lambda_i \mu_i)^\top (y - \sum_i \lambda_i \mu_i) + C(\lambda, \mu) \]

where \( C(\lambda, \mu) = -\frac{1}{2\sigma^2} \sum_i (\lambda_i - \lambda_i^2) \mu_i^\top \mu_i \)
Fixed point equations for multiple cause model

Taking derivatives w.r.t. $\lambda_i$:

$$
\frac{\partial F}{\partial \lambda_i} = \log \frac{\pi_i}{1 - \pi_i} - \log \frac{\lambda_i}{1 - \lambda_i} + \frac{1}{\sigma^2} (y - \sum_{j \neq i} \lambda_j \mu_j) \top \mu_i - \frac{1}{2\sigma^2} \mu_i \top \mu_i
$$

Setting to zero we get fixed point equations:

$$
\lambda_i = f \left( \log \frac{\pi_i}{1 - \pi_i} + \frac{1}{\sigma^2} (y - \sum_{j \neq i} \lambda_j \mu_j) \top \mu_i - \frac{1}{2\sigma^2} \mu_i \top \mu_i \right)
$$

where $f(x) = 1/(1 + \exp(-x))$ is the logistic (sigmoid) function.

Learning algorithm:

E step: run fixed point equations until convergence of $\lambda$ for each data point.
M step: re-estimate $\theta$ given $\lambda$s.
KL divergence

Note that

**E step** maximize $\mathcal{F}(q, \theta)$ wrt the distribution over hidden variables, given the parameters:

$$q^{(k)}(H) := \argmax_{q(H) \in Q} \mathcal{F}(q(H), \theta^{(k-1)}).$$

is equivalent to:

**E step** minimize $\mathcal{KL}(q\|p(H|V, \theta))$ wrt the distribution over hidden variables, given the parameters:

$$q^{(k)}(H) := \argmin_{q(H) \in Q} \int q(H) \log \frac{q(H)}{p(H|V, \theta^{(k-1)})} dH$$

So, at each E step, the variational approximation is trying to find the best approximation to $p$ in $Q$.
This is related to ideas in information geometry.
Structured Variational Approximations

$q(H)$ need not be completely factorized.

For example, suppose you can partition $H$ into sets $H_1$ and $H_2$ such that computing the expected sufficient statistics under $q(H_1)$ and $q(H_2)$ is tractable. Then $q(H) = q(H_1)q(H_2)$ is tractable.

If you have a graphical model, you may want to factorize $q(H)$ into a product of trees, which are tractable distributions.

More about this later (after we study graphical models).
**Variational Approximations to Bayesian Learning**

\[
\log p(V) = \log \int \int p(V, H|\theta)p(\theta) \, dH \, d\theta \\
\geq \int \int q(H, \theta) \log \frac{p(V, H, \theta)}{q(H, \theta)} \, dH \, d\theta
\]

Constrain \( q \in Q \) s.t. \( q(H, \theta) = q(H)q(\theta) \).

This results in the **variational Bayesian EM algorithm**.

More about this later (when we study model selection).
How tight is the lower bound?

It is hard to compute a nontrivial general upper bound.

To determine how tight the bound is, one can approximate the true likelihood by a variety of other methods.

One approach is to use the variational approximation as a proposal distribution for \textbf{importance sampling} (but we know how hard importance sampling can be in high dimensions).
Readings

