Learning with Hidden Variables[1]

Dmitry Adamskiy, David Barber

University College London

[1] These slides accompany the book *Bayesian Reasoning and Machine Learning*. The book and demos can be downloaded from `www.cs.ucl.ac.uk/staff/D.Barber/brml`. Feedback and corrections are also available on the site. Feel free to adapt these slides for your own purposes, but please include a link the above website.

Hidden Variables and Missing Data

Missing Data

In practice data entries are often missing resulting in incomplete information to specify a likelihood.

Observational Variables

Observational variables may be split into visible (those for which we actually know the state) and missing (those whose states would nominally be known but are missing for a particular datapoint).

Latent Variables

Another scenario in which not all variables in the model are observed are the so-called hidden or latent variable. In this case there are variables which are essential for the model description but never observed. For example, the underlying physics of a model may contain latent processes which are essential to describe the model, but cannot be directly measured.

Why hidden/missing variables can complicate proceedings

In learning the parameters of models we previously assumed we have complete information to define all variables of the joint model of the data $p(v|\theta)$.

Complete data

Consider the Asbestos-Smoking-Cancer network. In the case of complete data, the likelihood is

$$p(v^n|\theta) = p(a^n, s^n, c^n|\theta) = p(c^n|a^n, s^n, \theta_c)p(a^n|\theta_a)p(s^n|\theta_s)$$

which is factorised in terms of the table entry parameters. We exploited this property to show that table entries θ can be learned by considering only local information, both in the Maximum Likelihood and Bayesian frameworks.

Missing data

Now consider the case that for some of the patients, only partial information is available. For example, for patient n with record $v^n = \{c = 1, s = 1\}$ it is known that the patient has cancer and is a smoker, but whether or not they had exposure to asbestos is unknown. Since we can only use the 'visible' available information is it would seem reasonable to assess parameters using the marginal likelihood

$$p(v^n|\theta) = \sum_a p(a, s^n, c^n|\theta) = \sum_a p(c^n|a, s^n, \theta_c)p(a|\theta_a)p(s^n|\theta_s)$$

The likelihood cannot be written as a product of functions, one for each separate parameter. In this case the maximisation of the likelihood is more complex since the parameters of different tables are coupled.

Bayesian learning

A similar complication holds for Bayesian learning. Under a prior factorised over each CPT θ, the posterior is also factorised. However, in the case of unknown asbestos exposure, a term $p(v^n|\theta)$ as above is introduced, which cannot be written as a product of a functions of $f_s(\theta_s)f_a(\theta_a)f_c(\theta_c)$. The missing variable therefore introduces dependencies in the posterior parameter distribution, making the posterior more complex.

Maximum Likelihood

- For hidden variables h, and visible variables v we still have a well defined likelihood

$$p(v|\theta) = \sum_h p(v, h|\theta)$$

- Our task is to find the parameters θ that optimise $p(v|\theta)$.
- This task is more numerically complex than in the case when all the variables are visible.
- Nevertheless, we can perform numerical optimisation using any routine we wish to find θ.
- The Expectation-Maximisation algorithm is an alternative optimisation algorithm that can be very useful in producing simple and elegant updates for θ that converge to a local optimum.
- Just to hammer this home: We don't 'need' the EM algorithm, but it can be very handy.

Variational EM

The key feature of the EM algorithm is to form an alternative objective function for which the parameter coupling effect discussed is removed, meaning that individual parameter updates can be achieved, akin to the case of fully observed data. The way this works is to replace the marginal likelihood with a lower bound – it is this lower bound that has the decoupled form.

Consider the Kullback-Leibler divergence between a 'variational' distribution $q(h|v)$ and the parametric model $p(h|v,\theta)$:

$$\mathrm{KL}(q(h|v)||p(h|v,\theta)) \equiv \left\langle \log q(h|v) - \log p(h|v,\theta) \right\rangle_{q(h|v)} \geq 0$$

Using Bayes' rule, $p(h|v,\theta) = p(h,v|\theta)/p(v|\theta)$ and the fact that $p(v|\theta)$ does not depend on h,

$$\log p(v|\theta) \geq \underbrace{-\left\langle \log q(h|v) \right\rangle_{q(h|v)}}_{\text{Entropy}} + \underbrace{\left\langle \log p(h,v|\theta) \right\rangle_{q(h|v)}}_{\text{Energy}}$$

The bound is potentially useful since the energy is similar in form to the fully observed case, except that terms with missing data have their log likelihood weighted by a prefactor.

Lower bound

We could rewrite the likelihood indroducing auxillary variational distribution $q(h)$:

$$\log p(v|\theta) = \log \sum_h p(v, h|\theta) = \log \sum_h \frac{q(h)}{q(h)} p(v, h|\theta) = \log \left\langle \frac{p(v, h|\theta)}{q(h)} \right\rangle_{q(h)}$$

Now we can use Jensen's inequality

$$\log \left\langle \frac{p(v, h|\theta)}{q(h)} \right\rangle_{q(h)} \geq \left\langle \log \frac{p(v, h|\theta)}{q(h)} \right\rangle_{q(h)} = \langle \log p(v, h|\theta) \rangle_{q(h)} + H(q)$$

- The right-hand side is called free energy and is a function $F(q, \theta)$ of the parameters of the model and the variational distribution.
- It is a lower bound on the likelihood.
- We will take alternating steps optimizing it in q and θ.

EM idea

- Free energy is a function of q and θ:

$$F(q,\theta) = \langle \log p(v,h|\theta) \rangle_{q(h)} + H(q)$$

- Start at some initial guess θ_0, $q_0(h)$
- Repeat the following steps until convergence:
- E-step: $q_{new} = \arg\max_q F(q, \theta_{old})$
- M-step: $\theta_{new} = \arg\max_\theta F(q_{new}, \theta)$

There are several extensions, but today we'll look at the basic EM-algorithm.

E-step

We can rewrite free energy like this:

$$\begin{aligned}\left\langle \log \frac{p(v,h|\theta)}{q(h)} \right\rangle_{q(h)} &= \left\langle \log \frac{p(h|v,\theta)P(v|\theta)}{q(h)} \right\rangle_{q(h)} \\ &= \log p(v|\theta) - \mathrm{KL}(q(h)||p(h|v,\theta))\end{aligned}$$

This means that the optimal setting is $q(h) = p(h|v,\theta)$.

M-step

In M-step we optimise for the new parameters of the model.

$$\theta_{new} = \arg\max \left\langle \log p(v, h|\theta) \right\rangle_{q(h)} + H(q) = \arg\max_{\theta} \left\langle \log p(v, h|\theta) \right\rangle_{q(h)},$$

where $q(h) = p(h|v, \theta_{old})$. This means that we are only interested in computing expectations under posterior (rather than the full posterior itself).

A one-parameter one-state example

The model is on a single visible variable v and single two-state hidden variable $h \in \{1, 2\}$. We define a model $p(v, h) = p(v|h)p(h)$ with

$$p(v|h, \theta) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(v-\theta h)^2}$$

and $p(h = 1) = p(h = 2) = 0.5$. For an observation $v = 2.75$ and $\sigma^2 = 0.5$ our interest is to find the parameter θ that optimises the likelihood

$$p(v = 2.75|\theta) = \frac{1}{2\sqrt{\pi}} \sum_{h=1,2} e^{-(2.75-\theta h)^2}$$

The lower bound, as a function of θ and q (we need only say $q(h = 2)$ since $q(h = 1) = 1 - q(h = 2)$) is

$$\log p(v = 2.75|\theta) \geq LB(q(h = 2), \theta)$$

$$LB(q(h = 2), \theta) \equiv -q(h = 1) \log q(h = 1) - q(h = 2) \log q(h = 2) \\ - \sum_{h=1,2} q(h) (2.75 - \theta h)^2 + \log 2$$

A one-parameter one-state example

M-step

The M-step is easy to work out analytically in this case with

$$\theta^{new} = v \left\langle h \right\rangle_{q(h)} / \left\langle h^2 \right\rangle_{q(h)}$$

E-step

Similarly, the E-step sets

$$q^{new}(h) = p(h|v, \theta)$$

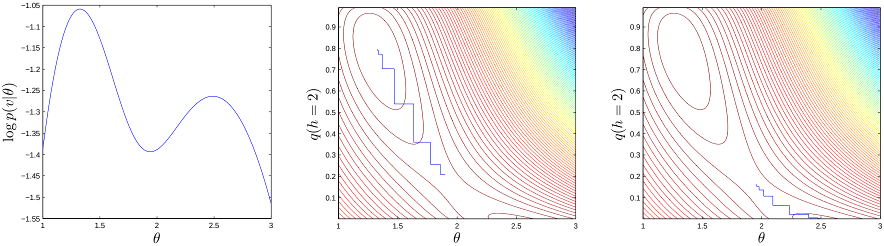
so that

$$q^{new}(h=2) = \frac{p(v=2.75|h=2,\theta)p(h=2)}{p(v=2.75)} = \frac{e^{-(2.75-2\theta)^2}}{e^{-(2.75-2\theta)^2} + e^{-(2.75-\theta)^2}}$$

where we used

$$p(v=2.75) = p(v=2.75|h=1,\theta)p(h=1) + p(v=2.75|h=2,\theta)p(h=2)$$

A one-parameter one-state example



(a)

(b)

(c)

Figure: **(a)**: The log likelihood. **(b)**: Contours of the lower bound $LB(q(h=2), \theta)$. For an initial choice $q(h=2) = 0.5$ and $\theta = 1.9$, successive updates of the E (vertical) and M (horizontal) steps are plotted. **(c)**: Starting at $\theta = 1.95$, the EM algorithm converges to a local optimum.

The EM algorithm increases the likelihood

We use θ' for the new parameters, and θ for the previous parameters in two consecutive iterations. Using $q(h|v) = p(h|v,\theta)$ we see that as a function of the parameters, the lower bound depends on θ and θ':

$$LB(\theta'|\theta) \equiv -\langle \log p(h|v,\theta)\rangle_{p(h|v,\theta)} + \langle \log p(h,v|\theta')\rangle_{p(h|v,\theta)}$$

and

$$\log p(v|\theta') = LB(\theta'|\theta) + KL(p(h|v,\theta)|p(h|v,\theta'))$$

We may write

$$\log p(v|\theta) = LB(\theta|\theta) + \underbrace{KL(p(h|v,\theta)|p(h|v,\theta))}_{0}$$

Hence

$$\log p(v|\theta') - \log p(v|\theta) = \underbrace{LB(\theta'|\theta) - LB(\theta|\theta)}_{\geq 0} + \underbrace{KL(p(h|v,\theta)|p(h|v,\theta'))}_{\geq 0}$$

The first assertion is true since, by definition of the M-step, we search for a θ' which has a higher value for the bound than our starting value θ.

Belief Network example

s	c
1	1
0	0
1	1
1	0
1	1
0	0
0	1

A database containing information about being a Smoker (1 signifies the individual is a smoker), and lung Cancer (1 signifies the individual has lung Cancer). Each row contains the information for an individual, so that there are 7 individuals in the database.

$$p(a, c, s) = p(c|a, s)p(a)p(s)$$

for which the states of a are never observed.

Task

Our goal is to learn the CPTs $p(c|a, s)$ and $p(a)$ and $p(s)$.

Step 0: initialisation

We first assume initial parameters θ_a^0, θ_s^0, θ_c^0.

First E-step, $t = 1$

$$q_{t=1}^{n=1}(a) = p(a|c = 1, s = 1, \theta^0), \qquad q_{t=1}^{n=2}(a) = p(a|c = 0, s = 0, \theta^0)$$

and so on for the 7 training examples, $n = 2, \ldots, 7$. For notational convenience, we write $q_t^n(a)$ in place of $q_t^n(a|v^n)$.

First M-step $t = 1$

The energy term for any iteration t is:

$$\begin{aligned}
E(\theta) &= \sum_{n=1}^{7} \left\langle \log p(c^n|a^n, s^n) + \log p(a^n) + \log p(s^n) \right\rangle_{q_t^n(a)} \\
&= \sum_{n=1}^{7} \left\{ \left\langle \log p(c^n|a^n, s^n) \right\rangle_{q_t^n(a)} + \left\langle \log p(a^n) \right\rangle_{q_t^n(a)} + \log p(s^n) \right\}
\end{aligned}$$

The final term is the log likelihood of the variable s, and $p(s)$ appears explicitly only in this term. Hence, the usual maximum likelihood rule applies, and $p(s = 1)$ is simply given by the relative number of times that $s = 1$ occurs in the database, giving $p(s = 1) = 4/7$, $p(s = 0) = 3/7$.

First M-step $t = 1$

$$E(\theta) = \sum_{n=1}^{7} \left\{ \langle \log p(c^n|a^n, s^n) \rangle_{q_t^n(a)} + \langle \log p(a^n) \rangle_{q_t^n(a)} + \log p(s^n) \right\}$$

The parameter $p(a = 1)$ occurs in the terms

$$\sum_n \{ q_t^n(a = 0) \log p(a = 0) + q_t^n(a = 1) \log p(a = 1) \}$$

which, using the normalisation constraint is

$$\log p(a = 0) \sum_n q_t^n(a = 0) + \log(1 - p(a = 0)) \sum_n q_t^n(a = 1)$$

Differentiating with respect to $p(a = 0)$ and solving for the zero derivative we get

$$p(a = 0) = \frac{\sum_n q_t^n(a = 0)}{\sum_n q_t^n(a = 0) + \sum_n q_t^n(a = 1)} = \frac{1}{N} \sum_n q_t^n(a = 0)$$

Whereas in the standard Maximum Likelihood estimate, we would have the real counts of the data in the above formula, here they have been replaced with our guessed values $q_t^n(a = 0)$ and $q_t^n(a = 1)$.

First M-step $t = 1$

A similar story holds for $p(c = 1|a = 0, s = 1)$. Optimising the bound gives:

$$
\begin{aligned}
&p(c = 1|a = 0, s = 1)\\
&= \frac{\sum_n \mathbb{I}\left[c^n = 1\right] \mathbb{I}\left[s^n = 1\right] q_t^n(a = 0)}{\sum_n \mathbb{I}\left[c^n = 1\right] \mathbb{I}\left[s^n = 1\right] q_t^n(a = 0) + \sum_n \mathbb{I}\left[c^n = 0\right] \mathbb{I}\left[s^n = 1\right] q_t^n(a = 0)}
\end{aligned}
$$

For comparison, the setting in the complete data case is

$$
\begin{aligned}
&p(c = 1|a = 0, s = 1)\\
&= \frac{\sum_n \mathbb{I}\left[c^n = 1\right] \mathbb{I}\left[s^n = 1\right] \mathbb{I}\left[a^n = 0\right]}{\sum_n \mathbb{I}\left[c^n = 1\right] \mathbb{I}\left[s^n = 1\right] \mathbb{I}\left[a^n = 0\right] + \sum_n \mathbb{I}\left[c^n = 0\right] \mathbb{I}\left[s^n = 1\right] \mathbb{I}\left[a^n = 0\right]}
\end{aligned}
$$

There is an intuitive relationship between these updates: in the missing data case we replace the indicators by the assumed distributions q.

E-step t

$$q_t^{n=1}(a) = p(a|c=1, s=1, \theta^{t-1}), \qquad q_t^{n=2}(a) = p(a|c=0, s=0, \theta^{t-1})$$

and so on for the 7 training examples, $n = 2, \ldots, 7$.

Iteration

Iterating the E and M steps, the parameters will converge to a local likelihood optimum.

Example: EM for HMMs

- EM for Hidden Markov Models has a special name: Baum-Welch algorithm.
- Here we sketch the derivation of the updates.

$$F(\theta, \theta_s) = \sum_h \log P(v, h|\theta) P(h|v, \theta_s)$$

Suppose that we have D observations, the d-th of them being a string of v_t^d,$t = 1, \ldots, T$. As we're maximizing in θ we could maximize instead

$$F'(\theta, \theta_s) = \sum_h \log P(v, h|\theta) P(v, h|\theta_s)$$

And the joint nicely factorises:

$$\log P(h, v|\theta) = \sum_d \left[\log \pi_{h_1^d} + \sum_{t=2}^{T} \log P_{h_{t-1}^d h_t^d} + \sum_{t=1}^{T} \log B_{h_t^d x_t^d} \right]$$

To get the updates we could use the idea we derived for Belief Networks (or using Lagrange multipliers technique).

Example: EM for HMMs

The free energy becomes then

$$F'(\theta, \theta_s) = \sum_h \sum_d \log \pi_{h_1^d} P(h, v|\theta_s) + \sum_h \sum_d \sum_{t=2}^{T} \log P_{h_{t-1}^d h_t^d} P(h, v|\theta_s) +$$
$$+ \sum_h \sum_d \sum_{t=1}^{T} \log B_{h_t^d x_t^d} P(h, v|\theta_s)$$

We need to optimise it with respect to π, P and B. Using Lagrange multiplier technique we could get rather intuitive updates.

$$\pi_i \propto \sum_d P(h_1^d = i | v^d, \theta^s)$$
$$P_{ij} = \frac{\sum_d \sum_{t=2}^T P(h_{t-1}^d = i, h_t^d = j | v^d, \theta^s)}{\sum_d \sum_{t=2}^T P(h_{t-1}^d = i | v^d, \theta^s)}$$
$$B_{ij} = \frac{\sum_d \sum_{t=1}^T P(h_t^d = i | v^d, \theta^s) I\left[v_t^d = j\right]}{\sum_d \sum_{t=1}^T P(h_t^d = i | v^d, \theta^s)}$$

Bernoulli Mixtures (example from tutorial)

- Suppose that we have a questionnaire and we suspect that people from different groups would have different answer patterns. The trong assumption we make is that the individual answers are independent given the group. This is modeled by

$$p(\boldsymbol{v}) = \sum_{h=1}^{K} p(h) \prod_{i=1}^{D} p(v_i|h).$$

- We may or may not know the number of groups (clusters), but we don't know conditional probabilities and we never observe the group variable.
- This is very similar to Naive Bayes, only the class label is not observed!

EM training

Recall the EM procedure. Write down the energy:

$$\sum_n \langle \log p(\boldsymbol{v}^n, h) \rangle_{p^{old}(h|v)} = \sum_n \sum_i \langle \log p(v_i^n | h) \rangle_{p^{old}(h|v)} + \sum_n \langle \log p(h) \rangle_{p^{old}(h|v)}$$

For E-step, we could write the update immediately as we set

$$q(h^n | \boldsymbol{v}^n) = p(h^n | \boldsymbol{v}^n, \theta) \propto p(h^n, \boldsymbol{v}^n)$$

For M-step we could write the maximum likelihood update, substituting the hidden variable indicators for the conditional.

EM training for Bernoulli mixtures

So, the update becomes:

M-step

$$p^{new}(v_i=1|h=j)=\frac{\sum_n \mathbb{I}[v_i^n=1]p^{old}(h=j|\boldsymbol{v}^n)}{\sum_n \mathbb{I}[v_i^n=1]p^{old}(h=j|\boldsymbol{v}^n)+\sum_n \mathbb{I}[v_i^n=0]p^{old}(h=j|\boldsymbol{v}^n)}$$

$$p^{new}(h=j) \propto \sum_n p_{old}(h=j|\boldsymbol{v}^n)$$

E-step

$$p^{new}(h=j|\boldsymbol{v}^n) \propto p^{old}(h=j)\prod_{i=1}^{D} p^{old}(v_i^n|h=j)$$