Learning with Hidden Variables

Prof. Dr. H. H. Takada

Quantitative Research – Itaú Asset Management
Institute of Mathematics and Statistics – University of São Paulo

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).

Hidden Variables

Another scenario in which not all variables in the model are observed are the so-called hidden, state 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. The observed variables are also called space variables.

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 (a), smoking (s) and cancer (c) 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 factorized 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.

Maximum Likelihood and Missing Data

Now consider the case that for some of the patients, only partial information is available. For example, patient n with record $v^n = \{c = 1, s = 1\}$ has cancer and is a smoker, but whether or not she had exposure to asbestos is unknown. Since we can only use the 'visible' available information, it is 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 maximization of the likelihood is more complex since the parameters of different tables are coupled.

Bayesian Learning and Missing Data

A similar complication holds for Bayesian learning. Under a prior factorized over each CPT θ, the posterior is also factorized. 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 $f_s(\theta_s)f_a(\theta_a)f_c(\theta_c)$. The missing variable therefore introduces dependencies in the posterior parameter distribution, making it more complex.

Favourite Colour (wrong way)

EZsurvey.org stop men on the street and ask them their favourite colour (blue,green or pink). All men whose favourite colour is pink decline to respond to the question – for any other colour, all men respond to the question. EZsurvey.org attempts to find the histogram with probabilities θ_b,θ_g,θ_p with $\theta_b + \theta_g + \theta_p = 1$. Each respondent produces a visible response x_c with $\mathrm{dom}(x_c) = \{\text{blue}, \text{green}, \text{pink}\}$. Three men are asked their favourite colour, giving data

$$\{x_c^1, x_c^2, x_c^3\} = \{\text{blue}, \text{missing}, \text{green}\}$$

Assuming i.i.d. data, the loglikelihood of the visible data alone (without the missing data) is

$$L(\theta_b, \theta_g, \theta_p) = \log \theta_b + \log \theta_g + \lambda (1 - \theta_b - \theta_g - \theta_p)$$

where the Lagrange term ensures normalization. Maximizing the expression we have

$$\theta_b = \frac{1}{2}, \theta_g = \frac{1}{2}, \theta_p = 0.$$

Favourite Colour (right way)

The correct mechanism that generates all the data (including the missing data) is

$$p(x_c^1 = \text{blue}|\theta)p(x_c^2 = \text{missing}|\theta)p(x_c^3 = \text{green}|\theta) = \theta_b\theta_p\theta_g = \theta_b\left(1-\theta_b-\theta_g\right)\theta_g$$

where we used $p(x_c^2 = \text{missing}|\theta) = \theta_p$ assuming the probability that a datapoint is missing is the same as the probability that the favourite colour is pink. Maximizing the likelihood, we arrive at

$$\theta_b = \frac{1}{3}, \theta_g = \frac{1}{3}, \theta_p = \frac{1}{3}$$

as we would expect. The favourite color example is a not missing at random example.

Missing at random

On the other hand if there is another visible variable, t, denoting the time of day, and the probability that men respond to the question depends only on the time t alone (for example the missing probability is high during rush hour), then we may indeed treat the missing data as missing at random (MAR). If the data is MAR it is OK to use the likelihood of the observed data to learn parameters. We assume the data is MAR for the remainder here.

Maximum Likelihood

- For hidden variables h and a visible variable 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 optimize $p(v|\theta)$.
- This task is more numerically complex than in the case when all the variables are visible.
- Nevertheless, we can perform numerical optimization using any routine we wish to find θ.
- The Expectation-Maximization (EM) algorithm is an alternative optimization algorithm that can be very useful in producing simple and elegant updates for θ that converge to a local optimum.

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.

Single observation

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

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

The term 'variational' refers to the fact that this distribution will be a parameter of an optimization problem. 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,

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

Variational EM

For i.i.d. data $\mathcal{V} = \{v^1, \ldots, v^N\}$

$$\log p(\mathcal{V}|\theta) \geq -\sum_{n=1}^{N} \langle \log q(h^n|v^n) \rangle_{q(h^n|v^n)} + \sum_{n=1}^{N} \langle \log p(h^n, v^n|\theta) \rangle_{q(h^n|v^n)}$$

This suggests an iterative procedure to optimize θ:

E-step For fixed θ, find the distributions $q(h^n|v^n)$ that maximize the bound.

M-step For fixed $\{q(h^n|v^n), n = 1, \ldots, N\}$, find the parameters θ that maximize the bound.

Classical EM

In the variational E-step above, the fully optimal setting is

$$q(h^n|v^n) = p(h^n|v^n, \theta).$$

Classical EM

Input: a distribution $p(x|\theta)$ and dataset $\mathcal{V} = \{v^1, ..., v^N\}$.

$t = 0$ ▷ Iteration counter

Choose an initial setting for the parameters θ^0. ▷ Initialization

while θ not converged (or likelihood not converged) **do**

  $t \leftarrow t + 1$

  for $n = 1$ to N **do** ▷ Run over all datapoints

    $q_t^n(h^n|v^n) = p(h^n|v^n, \theta^{t-1})$ ▷ E step

  end for

  $\theta^t = \arg\max_\theta \sum_{n=1}^N \left\langle \log p(h^n, v^n|\theta) \right\rangle_{q_t^n(h^n|v^n)}$ ▷ M step

end while

return θ^t ▷ The max likelihood parameter estimate

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 optimizes 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

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

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

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

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

so that

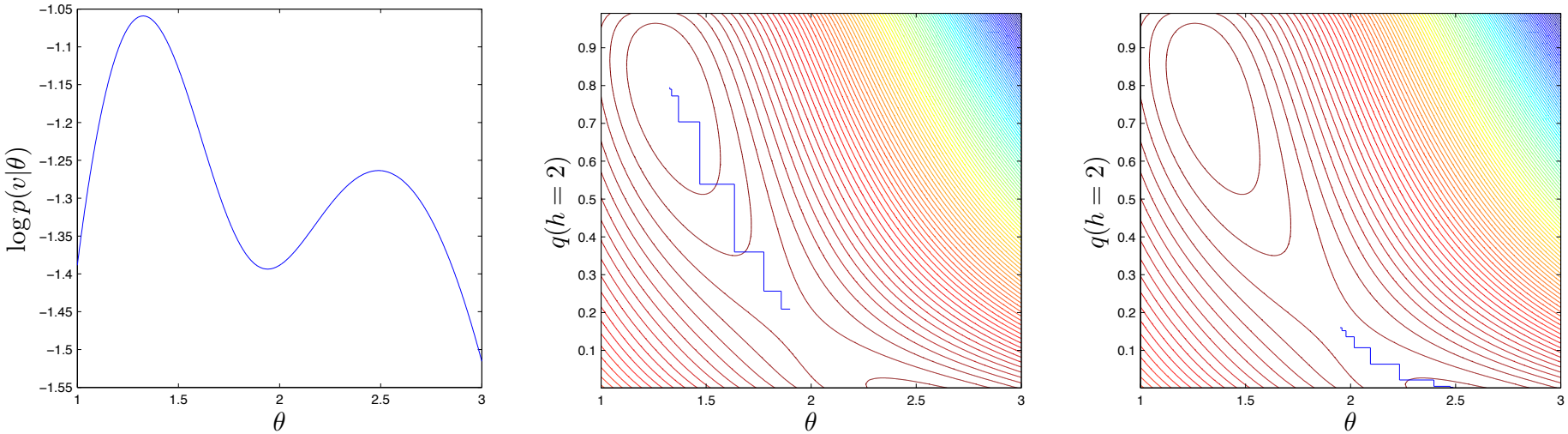
$$q^{new}(h = 2|v = 2.75) = \frac{p(v = 2.75|h = 2, \theta)p(h = 2)}{p(v = 2.75)}$$

$$= \frac{e^{-(2.75-2\theta)^2}}{e^{-(2.75-2\theta)^2} + e^{-(2.75-\theta)^2}}$$

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 loglikelihood. **(b)**: Contours of the lower bound LB$(q(h=2),\theta)$. For an initial choice $\theta = 1.9$, 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':

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

and

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

We may write

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

Hence

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

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:

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

The 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 normalization 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 $p(c = 1|a = 0, s = 1)$. Optimizing the bound gives:

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

For comparison, the setting in the complete data case is

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

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.