Learning with Hidden Variables

(Chapters 10,11)[1]

David Barber

(adapted by Miguel Palomino)

[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.

Maximum Likelihood Training of Belief Networks

Consider the following model of the relationship between exposure to asbestos (a), being a smoker (s) and the incidence of lung cancer (c)

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

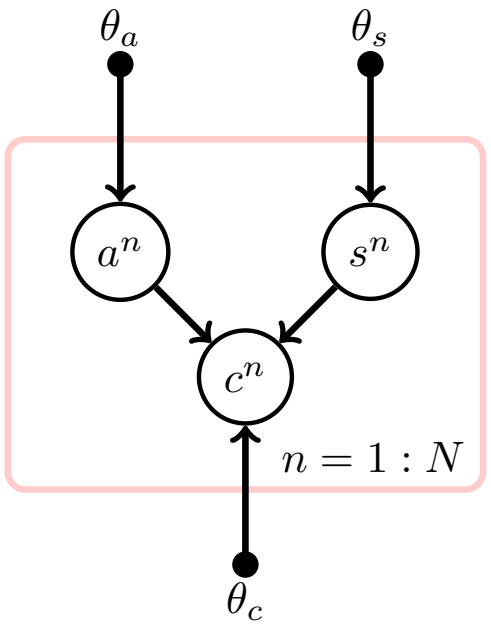
Each variable is binary, $\mathrm{dom}(a) = \{0, 1\}$, $\mathrm{dom}(s) = \{0, 1\}$, $\mathrm{dom}(c) = \{0, 1\}$. Furthermore, we assume that we have a list of patient records, where each row represents a patient's data.

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

A database containing information about the Asbestos exposure (1 signifies exposure), 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.

Learning the table

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



To learn the table entries $p(c|a, s)$ we can do so by counting the number of times c is in state 1 for each of the 4 parental states of a and s:

$$p(c = 1|a = 0, s = 0) = 0, \quad p(c = 1|a = 0, s = 1) = 0.5$$
$$p(c = 1|a = 1, s = 0) = 0.5 \quad p(c = 1|a = 1, s = 1) = 1$$

Similarly, based on counting, $p(a = 1) = 4/7$, and $p(s = 1) = 4/7$. These three CPTs then complete the full distribution specification.

Máxima verosimilitud

¿Qué justifica la elección anterior? Una RB tiene la forma:

$$p(x) = \prod_{i=1}^{K} p(x_i|\mathrm{pa}\,(x_i))$$

Las relaciones de independencia nos proporcionan el grafo subyacente pero ¿cómo obtener $p(x_i|\mathrm{pa}\,(x_i))$?

Dados datos $\mathcal{X} = \{x^1, \ldots, x^N\}$, se trata de maximizar su verosimilitud

$$p(\mathcal{X}) = \prod_{n=1}^{N} p(x^n) = \prod_{n=1}^{N} \prod_{i=1}^{K} p(x_i^n|\mathrm{pa}\,(x_i^n))$$

$$= \prod_{i=1}^{K} \left[\prod_{n=1}^{N} p(x_i^n|\mathrm{pa}\,(x_i^n)) \right]$$

Conteo basado en máxima verosimilitud

Podemos maximizar la verosimilitud de cada variable x_i por separado:

$$L(x, i) = \prod_{n=1}^{N} p(x_i^n = \mathsf{s}^n | \mathrm{pa}\,(x_i^n) = \mathsf{t}^n) = \prod_{n=1}^{N} \theta_{\mathsf{s}^n|\mathsf{t}^n}$$

Derivando con respecto a los $\boldsymbol{\theta}$ y bajo la restricción

$$\sum_{s} \theta_{s|\mathsf{t}} = 1 \qquad \text{para todo t}$$

se llega a que los parámetros óptimos son

$$p(x_i = \mathsf{s} | \mathrm{pa}\,(x_i) = \mathsf{t}) = \frac{\sum_{n=1}^{N} \mathbb{I}\left[x_i^n = \mathsf{s}\right] \prod_{x_j \in \mathrm{pa}(x_i)} \mathbb{I}\left[x_j^n = \mathsf{t}^j\right]}{\sum_{\mathsf{s}} \sum_{n=1}^{N} \mathbb{I}\left[x_i^n = \mathsf{s}\right] \prod_{x_j \in \mathrm{pa}(x_i)} \mathbb{I}\left[x_j^n = \mathsf{t}^j\right]}$$

El valor de $p(x_i|\mathrm{pa}\,(x_i))$ se establece contando el número de veces que el estado $\{x_i = \mathsf{s}, \mathrm{pa}\,(x_i) = \mathsf{t}\}$ ocurre entre los datos (donde t es un vector de estados antecesores). La tabla se rellena por el número relativo de apariciones en el estado s comparado con los otros estados s′, para un estado antecesor fijo t.

Naive Bayes Classifier

A joint model of observations $\mathbf{x}$ and the corresponding class label c using a Belief network of the form

$$p(\mathbf{x}, c) = p(c) \prod_{i=1}^{D} p(x_i|c)$$

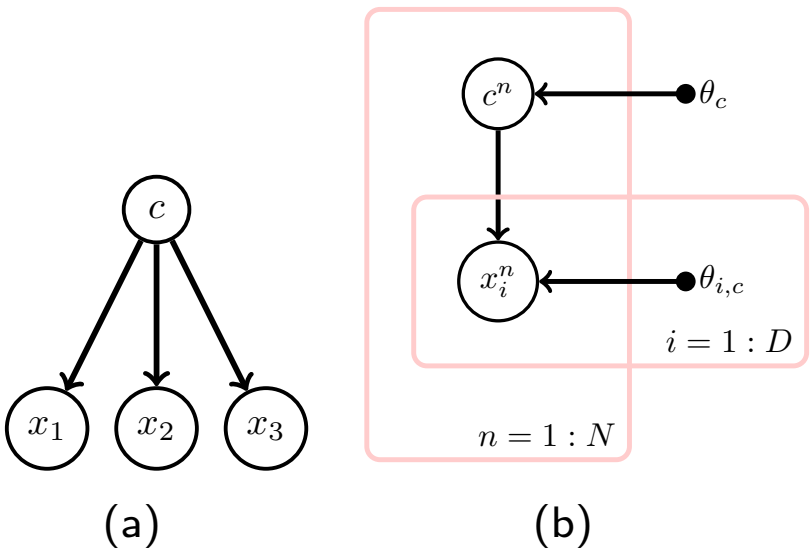


Figure: Naive Bayes classifier. **(a)**: The central assumption is that given the class c, the attributes x_i are independent. **(b)**: Assuming the data is i.i.d., Maximum Likelihood learns the optimal parameters of the distribution $p(c)$ and the class-dependent attribute distributions $p(x_i|c)$.

Coupled with a suitable choice for each conditional distribution $p(x_i|c)$, we can then use Bayes' rule to form a classifier for a novel attribute vector $\mathbf{x}^*$:

$$p(c|\mathbf{x}^*) = \frac{p(\mathbf{x}^*|c)p(c)}{p(\mathbf{x}^*)} = \frac{p(\mathbf{x}^*|c)p(c)}{\sum_c p(\mathbf{x}^*|c)p(c)}$$

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 in the Maximum Likelihood framework (as well as in the Bayesian framework, which we did not cover).

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, 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.

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 $\theta_b, \theta_g, \theta_p$ with $\theta_b + \theta_g + \theta_p = 1$. Each respondent produces a visible response x_c with $\mathrm{dom}(x_c) = \{\text{blue}, \text{green}, \text{pink}\}$, otherwise $m_c = 1$ if there is no response. 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 likelihood of the visible data alone 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 normalisation. Maximising 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 the data (including the missing data is)

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

where we used $p(m_c^2 = 1|\theta) = \theta_p$ since the probability that a datapoint is missing is the same as the probability that the favourite colour is pink. Maximising the likelihood, we arrive at

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

as we would expect.

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.

If the data is MAR it is OK to use the likelihood of the observed data to learn parameters (see textbook). We assume the data is MAR for the remainder here.

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.

Divergencia de Kullback-Leibler

- $\mathrm{KL}(q(x)|p(x))$ mide la "diferencia" entre distribuciones p y q:

$$\mathrm{KL}(q(x)|p(x)) = \left\langle \log \frac{q(x)}{p(x)} \right\rangle_{q(x)} = E_q \left[\log \frac{q(x)}{p(x)} \right]$$

- $\mathrm{KL}(q(x)|p(x)) \geq 0$ y $\mathrm{KL}(q(x)|p(x)) = 0 \iff p = q.$
- $\mathrm{KL}(q(x)|p(x)) \neq \mathrm{KL}(p(x)|q(x))$.

Ejemplo

Para la distribución p correspondiente a una moneda trucada: $p(\text{cara}) = 0.9$, $p(\text{cruz}) = 0.1$:

$$\begin{aligned}
q_1(\text{cara}) = 0.89, q_1(\text{cruz}) = 0.11, &\quad \text{entonces } \mathrm{KL}(p|q_1) = 0.0005. \\
q_2(\text{cara}) = 0.50, q_2(\text{cruz}) = 0.50, &\quad \text{entonces } \mathrm{KL}(p|q_2) = 0.37. \\
q_3(\text{cara}) = 0.01, q_3(\text{cruz}) = 0.99, &\quad \text{entonces } \mathrm{KL}(p|q_3) = 3.82.
\end{aligned}$$

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 $p(h|v,\theta)$:

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

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,

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

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.

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 optimise θ:

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

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

Classical EM

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

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

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':

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

and

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

We may write

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

Hence

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

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 for which the states of a are never observed.

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

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, $t = 1$

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

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

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 $\theta_a = p(a = 1)$ 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:

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

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.