T-61.5140 Machine Learning: Advanced Probabilistic Methods
Pekka Marttinen, Sami Remes (Spring 2015)
Exercise solutions, round 6, Tuesday, 3rd March, 2015
https://noppa.aalto.fi/noppa/kurssi/t-61.5140

Problem 1. *"Deriving VB for a simple model, part 1."*

Consider the variational Bayesian approximation for the example model from the lecture (see the attached `simple_vb_example.pdf`). Derive the VB update for the factor $q(\pi)$ in the example and implement it in `simple_vb_ex_template.m`.

Solution.

$$
\begin{aligned}
\log q(\pi) &= E_{z,\theta}[\log p(x,z,\pi,\theta)] + C \\
&= \log p(\pi) + E_z[\log p(z|\pi)] + C \\
&= (\alpha_0 - 1)\log \pi + (\alpha_0 - 1)\log(1-\pi) + E_z[\sum_{i=1}^{N}(z_{i2}\log \pi + z_{i1}\log(1-\pi)] + C \\
&= (\alpha_0 + \sum_{i=1}^{N} z_{i2} - 1)\log \pi + (\alpha_0 + \sum_{i=1}^{N} z_{i1} - 1)\log(1-\pi) + C \\
&= (\alpha_0 + N_2 - 1)\log \pi + (\alpha_0 + N_1 - 1)\log(1-\pi) + C \\
\Longrightarrow q(\pi) &= Beta(\pi|\alpha_0 + N_2, \alpha_0 + N_1),
\end{aligned}
$$

where we have defined $N_k = \sum_{i=1}^{N} E[z_{ik}]$, and C is an arbitrary constant (not depending on π) on each row.

Problem 2. *"Deriving VB for a simple model, part 2."*

As in Problem 1, consider the variational Bayesian approximation for the example model from the lecture (see the attached `simple_vb_example.pdf`). Now, derive the VB update for the factor $q(\theta)$ in the example and implement it in `simple_vb_ex_template.m`.

Solution.

$$
\begin{aligned}
\log q(\theta) &= E_{z,\pi}[\log p(x,z,\pi,\theta)] + C \\
&= \log p(\theta) + E_z[\log p(x|z,\theta)] + C \\
&= \log \mathcal{N}(\theta|0,\beta_0^{-1}) + E_z[\sum_{i=1}^{N} z_{i2} \log \mathcal{N}(x_i|\theta,1)] + C \\
&= -\frac{\beta_0}{2}\theta^2 - \frac{1}{2}\sum_{i=1}^{N} r_{i2}(x_i-\theta)^2 + C \\
&= -\frac{\beta_0}{2}\theta^2 - \frac{1}{2}\sum_{i=1}^{N} r_{i2}(x_i^2 - 2x_i\theta + \theta^2) + C \\
&= -\frac{1}{2}\theta^2(\beta_0 + N_2) + \theta \underbrace{\sum_{i=1}^{N} r_{i2}x_i}_{=N_2\bar{x}_2} + C \\
&= -\frac{1}{2}\theta^2(\beta_0 + N_2) + \theta N_2\bar{x}_2 + C \\
&= -\frac{\beta_0+N_2}{2}[\theta^2 - 2\frac{N_2\bar{x}_2}{\beta_0+N_2}\theta] + C \\
&= -\frac{\beta_0+N_2}{2}[\theta - \frac{N_2\bar{x}_2}{\beta_0+N_2}]^2 + C \\
\Longrightarrow q(\theta) &= \mathcal{N}(\theta|\frac{N_2\bar{x}_2}{\beta_0+N_2}, (\beta_0+N_2)^{-1}),
\end{aligned}
$$

where $\bar{x}_2 = \frac{1}{N_2}\sum_{i=1}^{N} r_{i2}x_i$ and C is an arbitrary constant on each row.

Problem 3. *"Variational approximation for a simple distribution."*

Consider a model with two binary random variables x_1 and x_2, defined by the distributions:

$p(x_1)$	
x_1=0	0.4
x_1=1	0.6

$p(x_2 \mid x_1)$	x_1=0	x_1=1
x_2=0	0.5	0.9
x_2=1	0.5	0.1

Find a fully factorized distribution $q(x_1,x_2) = q_1(x_1)q_2(x_2)$ that best approximates the joint $p(x_1,x_2)$, in the sense of minimizing KL $(p \parallel q)$.

> **Note:** For "normal" variational inference, we would rather minimize KL $(q \parallel p)$; recall that, in general, KL $(p \parallel q) \neq$ KL $(q \parallel p)$. See (Barber, 2012), Figure 28.1 as well as Chapter 28.3.4 and 28.3.5, for the dramatically different solutions that can result by minimizing the different quantities, as well as commentary on their relative usefulness for approximate inference. Here, we'll minimize KL $(p \parallel q)$, as that is algebraically simpler.

Solution. Let us parametrize q such that q_1 and q_2 have the Bernoulli parameters $a = q_1(x_1 = 1)$ and $b = q_2(x_2 = 1)$. Now we have the following joint distributions $p(x_1,x_2)$ and $q(x_1,x_2)$:

$p(x_1, x_2)$	x_1=0	x_1=1
x_2=0	0.20	0.54
x_2=1	0.20	0.06

$q(x_1, x_2)$	x_1=0	x_1=1
x_2=0	$(1-a)(1-b)$	$a(1-b)$
x_2=1	$(1-a)b$	ab

Then we write $\mathrm{KL}\,(p \parallel q)$, denoting by C an arbitrary constant; and find the zeros of the partial derivatives w.r.t. a and b:

$$\begin{aligned}
\mathrm{KL}\,(p \parallel q) &= \sum_{\mathbf{x}} p(\mathbf{x}) \ln \frac{p(\mathbf{x})}{q(\mathbf{x})} = -\sum_{\mathbf{x}} p(\mathbf{x}) \ln \frac{q(\mathbf{x})}{p(\mathbf{x})} \\
&= -0.2 \ln \frac{(1-a)(1-b)}{0.2} - 0.54 \ln \frac{a(1-b)}{0.6} - 0.2 \ln \frac{(1-a)b}{0.2} - 0.06 \ln \frac{ab}{0.06} \\
&= C - 0.4 \ln(1-a) - 0.6 \ln(a) - 0.74 \ln(1-b) - 0.26 \ln(b) \\
\frac{\partial}{\partial a}\mathrm{KL}\,(p \parallel q) &= \frac{0.4}{1-a} - \frac{0.6}{a} = 0 \qquad \Rightarrow a = 0.6 \\
\frac{\partial}{\partial b}\mathrm{KL}\,(p \parallel q) &= \frac{0.74}{1-b} - \frac{0.26}{b} = 0 \qquad \Rightarrow b = 0.26
\end{aligned}$$

In this case, the result for minimizing $\mathrm{KL}\,(q \parallel p)$, solved below, is not dramatically different.

$\min \mathrm{KL}\,(p \parallel q)$

$q(x_1, x_2)$	x_1=0	x_1=1
x_2=0	0.296	0.444
x_2=1	0.104	0.156

$\min \mathrm{KL}\,(q \parallel p)$

$q(x_1, x_2)$	x_1=0	x_1=1
x_2=0	0.29192	0.50984
x_2=1	0.07218	0.12606

In theory, $\mathrm{KL}\,(q \parallel p)$ can be minimized in a similar way; however, the analytical solution seems more challenging. As we are assuming a fully factorized approximation $q(x_1, x_2) = q_1(x_1)\, q_2(x_2)$, let us therefore look into how the general mean field equations (see Barber, 2012, ch. 28.4.2) can be applied to this problem.

For a fully factorized $q(\mathbf{x}) = \prod_i q(x_i)$ approximating $p(\mathbf{x})$, the optimal mean-field update for a single $q(x_i)$ is to set its parameters so that

$$q(x_i) \propto \exp\left(\langle \log p(\mathbf{x}) \rangle_{\prod_{j \neq i} q(x_j)} \right). \tag{3.1}$$

Given an arbitrary constant K (omitting the details of b for brevity), we have

$$\begin{cases} q_1(0) = (1-a) = K \exp\left[(1-b)\log 0.20 + b \log 0.20\right] = 0.20K, \\ q_1(1) = a = K \exp\left[(1-b)\log 0.54 + b \log 0.06\right] = 0.54K \left(\dfrac{0.06}{0.54}\right)^b, \end{cases} \tag{3.2}$$

$$\frac{q_1(1)}{q_1(0)} = \frac{a}{1-a} = \frac{0.54K\left(\frac{0.06}{0.54}\right)^b}{0.20K} = \frac{0.54}{0.20}\left(\frac{0.06}{0.54}\right)^b, \tag{3.3}$$

$$a = \frac{0.54}{0.20}\left(\frac{0.06}{0.54}\right)^b \Bigg/ \left[1 + \frac{0.54}{0.20}\left(\frac{0.06}{0.54}\right)^b\right], \tag{3.4}$$

$$b = \left(\frac{0.06}{0.54}\right)^a \Bigg/ \left[1 + \left(\frac{0.06}{0.54}\right)^a\right]. \tag{3.5}$$

Problem 4. *"Variational inference in a Markov network."*

Consider a model with four random variables w, x, y, z and six states forming a Markov network, find a fully factorized distribution in the sense of minimizing $\mathrm{KL}(p||q)$ and report the KL-divergence between the original distribution (p) and the approximate distribution (q). Complete the template `ex6_4_template.m` with your own code. In the code, the parameter α control the complexity of the original distribution. Demonstrate the effect of α on the factorized distribution.

Hint: `demoMFBPGibbs.m`[1] shows an example on how to do this. Go through the demo, and copy/paste/modify relevant parts of the code into your solution (so, you don't need to write your own code for $\mathrm{KL}(p||q)$, unless you want...). For a useful detailed description, see Barber (2012), in Chapter 28 Example 28.4, Figures 28.10 and 28.11.

Solution.

[1] Freely downloadable from `www.cs.ucl.ac.uk/staff/D.Barber/brml`