T.61.5140 Machine Learning: Advanced Probablistic Methods
Hollmén, Raiko (Spring 2008)
Problem session, 18th of April, 2008
http://www.cis.hut.fi/Opinnot/T-61.5140/

On notation:

Kullback-Leibler divergence can be written in two ways:

$$\mathrm{KL}\left(q(x) \parallel p(x)\right) = E_{q(x)}\left\{\ln \frac{q(x)}{p(x)}\right\} \tag{1}$$

$$= -E_{q(x)}\left\{\ln \frac{p(x)}{q(x)}\right\} \tag{2}$$

Tapani has used the former and Bishop's book uses the latter. The lower bound $\mathcal{L}(q)$ on page 463 is the same as $-C_{\mathrm{vb}}$. Maximization of $\mathcal{L}(q)$ is thus the same thing as minimization of C_{vb}. Also, Bishop absorbs the parameters $\boldsymbol{\theta}$ into $\mathbf{Z}$ when describing variational inference. This makes some of the formulas nicer, but the VB-EM algorithm cannot be described easily.

Let us use the shorthand $\langle \cdot \rangle = E_q\{\cdot\}$ to denote the expectation over the posterior approximation q.

1. Consider the probabilistic principal component analysis (PPCA) model

$$\mathbf{x}(t) = \mathbf{A}s(t) + \mathbf{n}(t) \tag{3}$$

with 3-dimensional data $\mathbf{x}(t)$, 1-dimensional source $s(t)$, and Gaussian noise $p(n_i(t)) = N(n_i(t) \mid 0, \sigma_i^2)$. We would like to estimate the noise level σ_i^2 for each data dimension $i = 1, 2, 3$ using the maximum likelihood estimator. What happens when the weight matrix $\mathbf{A}$ goes to $\mathbf{A} = \begin{pmatrix} 1 & 0 & 0 \end{pmatrix}^T$, and the source $s(t)$ copies the first dimension of the data: $s(t) = x_1(t)$?

Solution:

First, let us compute the reconstruction error, or the noise term

$$\mathbf{n}(t) = \mathbf{x}(t) - \mathbf{A}s(t) = \begin{pmatrix} 0 \\ x_2(t) \\ x_3(t) \end{pmatrix}, \tag{4}$$

which hints that there might be a problem with the variance parameters σ_1^2. The likelihood of the data is

$$p(\mathbf{x}(t) \mid \mathbf{A}, s(t)) = \prod_{t=1}^{T} \prod_{i=1}^{3} \frac{1}{\sqrt{2\pi\sigma_i^2}} \exp\left[-\frac{(x_i(t) - \mathbf{A}_i s(t))^2}{2\sigma_i^2}\right] \tag{5}$$

$$= \prod_{t=1}^{T} \frac{1}{\sqrt{2\pi\sigma_1^2}} \exp\left[-\frac{0}{2\sigma_1^2}\right] \prod_{i=2}^{3} \frac{1}{\sqrt{2\pi\sigma_i^2}} \exp\left[-\frac{(x_i(t) - \mathbf{A}_i s(t))^2}{2\sigma_i^2}\right] \tag{6}$$

$$= \text{const} \frac{1}{\sigma_1} \tag{7}$$

so we notice that when σ_1 goes to zero, the likelihood goes to infinity.

Maximum likelihood criterion considers this degenerate solution infinitely good, while it is in fact useless. This is an extreme case of overfitting, which is related to the fact that the posterior distribution has a high but narrow peak at the mentioned solution. Using any methodology that is sensitive to probability mass rather than density avoids this problem.

2. Variational Bayesian cost function C_{vb} is often a sum of simple terms, one for each variable. Consider the variable s whose model is $p(s \mid m, v) = N(s \mid m, \exp(-v))$. Assume that s, m, v are independent aposteriori, that is, $q(s, m, v) = q(s)q(m)q(v)$ and that $q(s) = N(s \mid \overline{s}, \widetilde{s})$. Show that the term of the cost for s, $C_{\text{vb}}(s) = E_q\left\{\ln \frac{q(s)}{p(s|m,v)}\right\}$, is

$$C_{\text{vb}}(s) = \frac{1}{2}\left(E_q\{\exp v\}\left[(\overline{s} - E_q\{m\})^2 + \underset{q}{\text{Var}}\{m\} + \widetilde{s}\right] - E_q\{v\} + \ln(2\pi)\right)$$
$$- \frac{1}{2}\ln(2\pi e\widetilde{s}). \tag{8}$$

Solution:

Let us first show that $\langle mv \rangle = \langle m \rangle \langle v \rangle$ when $q(m, v) = q(m)q(v)$.

$$\langle mv \rangle = \int\int q(m, v) mv \; dm \; dv \tag{9}$$

$$= \int q(m) m \; dm \int q(v) v \; dv \tag{10}$$

$$= \langle m \rangle \langle v \rangle \tag{11}$$

Let us address the cost function $C_{\mathrm{vb}}(s) = \langle \ln q(s) \rangle - \langle \ln p(s \mid m, v) \rangle$ in two parts.

$$
\begin{aligned}
-\langle \ln p\,(s \mid m, v) \rangle &= -\langle \ln \mathrm{N}\,(s \mid m, \exp(-v)) \rangle && (12)\\
&= -\left\langle \ln \left[(2\pi \exp(-v))^{-1/2} \exp \frac{-(s-m)^2}{2\exp(-v)} \right] \right\rangle && (13)\\
&= -\left\langle \ln \left[2\pi \exp(-v) \right]^{-1/2} \right\rangle - \left\langle \ln \exp \left[-\frac{1}{2}(s-m)^2 \exp v \right] \right\rangle && (14)\\
&= \frac{1}{2} \ln 2\pi - \frac{1}{2} \langle v \rangle + \frac{1}{2} \left\langle (s-m)^2 \right\rangle \langle \exp v \rangle . && (15)
\end{aligned}
$$

We still need to compute the expectation of $(s-m)^2$

$$
\begin{aligned}
\left\langle (s-m)^2 \right\rangle &= \left\langle s^2 - 2sm + m^2 \right\rangle && (16)\\
&= \left\langle s^2 \right\rangle - 2 \langle sm \rangle + \left\langle m^2 \right\rangle && (17)\\
&= \overline{s}^2 + \widetilde{s} - 2\overline{s} \langle m \rangle + \langle m \rangle^2 + \mathrm{Var}\,\{m\} && (18)\\
&= (\overline{s} - \langle m \rangle)^2 + \widetilde{s} + \mathrm{Var}\,\{m\} . && (19)
\end{aligned}
$$

Substituting $\left\langle (s-m)^2 \right\rangle$ back in, we get:

$$
-\langle \ln p\,(s \mid m, v) \rangle = \frac{1}{2} \left\{ \langle \exp v \rangle \left[(\overline{s} - \langle m \rangle)^2 + \mathrm{Var}\,\{m\} + \widetilde{s} \right] - \langle v \rangle + \ln 2\pi \right\} . \tag{20}
$$

The second part of the cost function is

$$
\begin{aligned}
\langle \ln q(s) \rangle &= \langle \ln \mathrm{N}\,(s \mid \overline{s}, \widetilde{s}) \rangle && (21)\\
&= \left\langle \ln \left[(2\pi\widetilde{s})^{-1/2} \exp \frac{-(s-\overline{s})^2}{2\widetilde{s}} \right] \right\rangle && (22)\\
&= -\frac{1}{2} \ln 2\pi\widetilde{s} + \frac{-\left\langle (s-\overline{s})^2 \right\rangle}{2\widetilde{s}} && (23)\\
&= -\frac{1}{2} \ln 2\pi\widetilde{s} + \frac{-\widetilde{s}}{2\widetilde{s}} && (24)\\
&= -\frac{1}{2} \ln 2\pi e\widetilde{s}. && (25)
\end{aligned}
$$

3. Consider the following model:

$$p(x(t) \mid m, v) = N(x(t) \mid m, \exp(-v)) \tag{26}$$
$$p(m) = N(m \mid 0, \exp(5)) \tag{27}$$
$$p(v) = N(v \mid 0, \exp(5)), \tag{28}$$

where $x(t), t = 1, \ldots, T$ are the observed data and m and v are latent variables. Use the posterior approximation $q(m, v) = q(m)q(v) = N(m \mid \overline{m}, \widetilde{m})N(v \mid \overline{v}, \widetilde{v})$. Assuming that $q(v)$ is fixed, find $q(m)$ that minimizes $C_{\text{vb}} = E_q \left\{ \ln \frac{q(m,v)}{p(\{x(t)\}_{t=1}^T, m, v)} \right\}$.

Solution:

Let us first write the whole cost function.

$$\begin{aligned} C_{\text{vb}} &= \left\langle \ln \frac{q(m, v)}{p(\{x(t)\}_{t=1}^T, m, v)} \right\rangle && (29) \\ &= \langle \ln q(m) \rangle + \langle \ln q(v) \rangle + \sum_{t=1}^{T} \\ &\quad - \langle \ln p(x(t) \mid m, v) \rangle - \langle \ln p(m) \rangle - \langle \ln p(v) \rangle && (30) \\ &= -\frac{1}{2} \ln(2\pi e \widetilde{m}) - \frac{1}{2} \ln(2\pi e \widetilde{v}) \\ &\quad + \sum_{t=1}^{T} \frac{1}{2} \left\{ \langle \exp v \rangle \left[(x(t) - \overline{m})^2 + \widetilde{m} \right] - \langle v \rangle + \ln 2\pi \right\} \\ &\quad + \frac{1}{2} \left[\langle \exp 5 \rangle \, (\overline{m}^2 + \widetilde{m}) - \langle -5 \rangle + \ln 2\pi \right] \\ &\quad + \frac{1}{2} \left[\langle \exp 5 \rangle \, (\overline{v}^2 + \widetilde{v}) - \langle -5 \rangle + \ln 2\pi \right] && (31) \end{aligned}$$

Now if we only leave the terms that depend on $q(m)$ we have

$$\begin{aligned} C_{\text{vb}} &= \text{const} - \frac{1}{2} \ln(2\pi e \widetilde{m}) + \sum_{t=1}^{T} \frac{1}{2} \langle \exp v \rangle \left[(x(t) - \overline{m})^2 + \widetilde{m} \right] \\ &\quad + \frac{1}{2} \langle \exp 5 \rangle \, (\overline{m}^2 + \widetilde{m}). && (32) \end{aligned}$$

The minimum of this can be found at the zero of the gradient with respect to mean $\overline{m}$ and the variance $\widetilde{m}$:

$$\frac{\partial C_{\mathrm{vb}}}{\partial \overline{m}} = \sum_{t=1}^{T} \langle \exp v \rangle \, (\overline{m} - x(t)) + \exp(-5)\overline{m} = 0 \tag{33}$$

$$\overline{m} = \frac{\sum_{t=1}^{T} x(t)}{\frac{\exp(-5)}{\exp v} + T} \tag{34}$$

$$\frac{\partial C_{\mathrm{vb}}}{\partial \widetilde{m}} = -\frac{1}{2\overline{m}} + \sum_{t=1}^{T} \frac{1}{2} \langle \exp v \rangle + \frac{1}{2} \exp(-5) = 0 \tag{35}$$

$$\widetilde{m} = \frac{1}{\exp(-5) + T \exp v}. \tag{36}$$

We can note that the mean $\overline{m}$ is close to the average of the data, only the prior makes it go slightly towards zero. Also we note that the posterior variance $\widetilde{m}$ is the same as prior variance in case that there are no observations, or $T = 0$, and with each observation, the uncertainty (or variance) decreases. It is interesting that the actual data values $x(t)$ do not affect $\widetilde{m}$.

The fourth problem will be solved next week.