T-61.5040 Oppivat mallit ja menetelmät
T-61.5040 Learning Models and Methods
Pajunen, Viitaniemi

Solutions to exercise 9, 23.3.2007

Problem 1.

i) In variational learning, we maximize

$$C = \int \log\left(\frac{p(y|\theta)p(\theta)}{q(\theta)}\right) q(\theta)d\theta$$

with respect to the distribution $q(\theta)$. C can be written as

$$\begin{aligned}\int \log\left(\frac{p(y|\theta)p(\theta)}{q(\theta)}\right) q(\theta)d\theta &= \int \log\left(\frac{p(\theta|y)p(y)}{q(\theta)}\right) q(\theta)d\theta \\ &= \int \log\left(\frac{p(\theta|y)}{q(\theta)}\right) q(\theta)d\theta + \int \log p(y)\, q(\theta)d\theta \\ &= -D(q(\theta)||p(\theta|y)) + \log p(y)\end{aligned}$$

Instead of directly minimizing the Kullback-Leibler divergence $D(q(\theta)||p(\theta|y))$, in variational learning one often maximizes C because $D(q(\theta)||p(\theta|y))$ cannot be computed when $p(y)$ is unknown. But we see that maximizing C with respect to $q(\theta)$ results in minimizing $D(q(\theta)||p(\theta|y))$, because the term $\log p(y)$ does not depend on $q(\theta)$.

ii) The data evidence $p(y)$ is maximized when $\log p(y)$ is maximized. Write the log of evidence as

$$\begin{aligned}\log p(y) &= \log \int p(y,\theta)d\theta \\ &= \log \int \frac{p(y|\theta)p(\theta)}{q(\theta)} q(\theta)d\theta \\ &\geq \int \log\left(\frac{p(y|\theta)p(\theta)}{q(\theta)}\right) q(\theta)d\theta \\ &= C\end{aligned}$$

The inequality is due to Jensen's inequality. So maximizing C maximizes a lower bound for $\log p(y)$.

Problem 2.

The true posterior is $p(\theta|y) = a_1 p_1(\theta|y) + a_2 p_2(\theta|y)$ where p_i is a Normal distribution $N(\mu_i, \sigma_i^2)$, $i = 1, 2$.

In this problem we're going to minimise the KL-divergence between q and p. We argue that since the mixture components are well-separated, it is enough to consider the cost function over each of the mixture components separately as a good posterior approximation q to one of the mixture components is almost zero at the other component,

So assume $q(\theta) = N(\theta|\mu_0, \sigma_0^2)$ and consider one of $p_i(\theta|y) = N(\theta|\mu_i, \sigma_i^2)$ with coefficient a_i. In this case we use the KL-divergence between $q(\theta)$ and the true posteriori component $a_i p_i(\theta|y)$ directly as p_i is known. We minimise the cost function

$$C = D(q(\theta)|a_i p_i(\theta|y)) = \int \log\left[\frac{q(\theta)}{a_i p_i(\theta|y)}\right] q(\theta)d\theta = D(q(\theta)|p_i(\theta|y)) - \log a_i$$

w.r.t. $q(\theta)$. Now the KL-divergence term attains its minimum value 0 iff $q = p_i$. As the term $-\log a_i$ is constant in terms of q, the whole cost function is minimised by $q = p_i = N(\theta|\mu_i, \sigma_i^2)$, i.e. $\mu_0 = \mu_i$ and $\sigma_0^2 = \sigma_i^2$.

So far we have seen that fitting a Normal distribution $N(\mu_0, \sigma_0^2)$ to another Normal distribution $N(\mu, \sigma^2)$ gives the correct parameters.

What about the mixture components? Comparing the fits to each mixture component, it is clear that the one minimizing $-\log a_i$ wins, and this is the one with larger a_i, regardless of the mean and variance. Since a_i measures directly the posterior mass contained in the component distribution, it seems that well-separated Normal-like modes are handled correctly by variational learning: regardless of variance and mean, the mode with the largest posterior mass is found.

Problem 3.

i) Laplace approximation fits a Normal distribution to the posterior distribution. The approximating distribution is centered at the posterior mode.

First we need to find the posterior mode. This is the value λ_0 that maximizes

$$p(\lambda|k) \propto p(k|\lambda)p(\lambda).$$

Since $p(\lambda|k) \propto e^{-\lambda}\lambda^{k-1}$, its derivative is

$$p'(\lambda|k) \propto -e^{-\lambda}\lambda^{k-1} + e^{-\lambda}\lambda^{k-2}(k-1).$$

Setting it to zero gives the mode $\lambda_0 = k - 1$. This becomes the mean of the approximating distribution.

Next we need the variance of the approximating distribution. The inverse of the variance σ^2 is calculated as $\sigma^{-2} = [-\log p(\lambda|k)]''|_{\lambda=\lambda_0}$. First we need to calculate $(\log p)''$:

$$\begin{aligned}\log p(\lambda|k) &= -\lambda + (k-1)\log\lambda \\ (\log p)' &= -1 + (k-1)\lambda^{-1} \\ (\log p)'' &= -(k-1)\lambda^{-2}.\end{aligned}$$

Substituting the posterior mode $\lambda_0 = k - 1$ gives

$$-(k-1)(k-1)^{-2} = -(k-1)^{-1}.$$

This gives us $\sigma^2 = k - 1$. This gives us the Laplace approximation $N(\lambda|k-1, k-1)$.

ii) Write $l = \log\lambda$. Now we have

$$p(l|k) \propto p(k|l) \propto e^{-e^l}e^{kl}$$

because the prior $p(l)$ is constant. The posterior mode is obtained by setting the derivative equal to zero:

$$\begin{aligned} -e^l e^{-e^l} e^{lk} + e^{-e^l} k e^{lk} &= 0 \\ \implies e^l e^{-e^l} e^{lk} &= e^{-e^l} k e^{lk} \\ \implies e^l &= k \\ \implies l &= \log k. \end{aligned}$$

This gives us the mode $l_0 = \log k$.

The logarithm of the posterior is

$$\log p \propto -e^l + lk,$$

the first derivative is

$$(\log p)' \propto -e^l + k,$$

and the second derivative is

$$(\log p)'' \propto -e^l.$$

Substituting the posterior mode $l_0 = \log k$, we obtain

$$(\log p)''|_{l=\log k} = -k.$$

Now the variance is $\sigma^2 = k^{-1}$, and the Laplace approximation is $N(l|\log k, k^{-1})$.

Comments: we can compare the two approximations by comparing their means. The first gives $\lambda = k-1$ and the second $l = \log \lambda = \log k \implies \lambda = k$. This example demonstrates that the parameterization matters when computing Laplace (and most other) approximations.

Problem 4.

i) Minimize KL divergence

$$D(p\|q) = \int p \log(p/q) d\theta$$

Since $\int p \log p \, d\theta$ is constant with respect to θ_0, one has to maximize

$$\int p \log q \, d\theta \quad (*)$$

Since

$$q(\theta) = N(\theta|\theta_0, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp(-\frac{1}{2\sigma^2}(\theta - \theta_0)^2),$$

its logarithm is

$$-\frac{1}{2\sigma^2}(\theta - \theta_0)^2$$

where constant terms are ignored. Then the maximization of $(*)$ is equivalent to minimization of

$$\int p(\theta - \theta_0)^2 d\theta = \mathrm{E}((\theta - \theta_0)^2)$$

So the solution is to choose θ_0 as the minimum mean-square estimate of θ.

ii) Minimize

$$D(q\|p) = \int q \log(q/p) d\theta$$

Since $\int q \log q \, d\theta = E_q(\log q) \propto E_q((\theta - \theta_0)^2)/\sigma^2 = 1$ is constant with respect to θ_0, it remains to maximize

$$\int q \log p \, d\theta. \quad (**)$$

Taylor-expand $\log p(\theta|y)$ at θ_0, which gives

$$\log p(\theta|y) = \log p(\theta_0|y) + (\theta - \theta_0)(\log p)'_{\theta=\theta_0} + \frac{1}{2}(\theta - \theta_0)^2 (\log p)''_{\theta=\theta_0} + \text{ higher terms}$$

The integral $(**)$ is an expectation $\mathrm{E}(\log p)$ over the Normal distribution $N(\theta|\theta_0, \sigma^2)$. Approximate by dropping the higher terms, which gives

$$\int q \log p \, d\theta = \log p(\theta_0|y) + \frac{1}{2}\sigma^2 (\log p)''$$

The value θ_0 is chosen to maximize this expression. Note that when p itself is Normal, then $(\log p)''$ is constant w.r.t. θ_0 and the solution is to maximize $p(\theta_0|y)$. Other distributions can have non-constant $(\log p)''$ so the posterior mode is not always optimal θ_0.