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

Solutions to exercise 6, 23.2.2007

Problem 1.

The data vectors $y = y_1, \ldots, y_n$ are independent and identically distributed. The distribution is $p(y_i|\theta, V) = N(y_i|\theta, V)$, where V is known. The unknown mean θ is a vector and it has a prior distribution $p(\theta) = N(\theta|\theta_0, V_0)$. The posterior distribution $p(\theta|y, V)$ is

$$\begin{aligned} p(\theta|y,V) &\propto p(\theta)p(y|\theta,V) = N(\theta|\theta_0,V_0)\prod_{i=1}^{n} N(y_i|\theta,V) \\ &\propto \exp(-\frac{1}{2}(\theta-\theta_0)^T V_0^{-1}(\theta-\theta_0))\exp(-\frac{1}{2}\sum_{i=1}^{n}(y_i-\theta)^T V^{-1}(y_i-\theta)). \end{aligned}$$

The posterior $p(\theta|y, V)$ is seen to be a normal distribution, as we expect the product of normal distributions to be. But how do we extract the parameters of the distribution? Let's recall the formula for the pdf of a multivariate normal distribution

$$\begin{aligned} N(\theta|\mu,S) &= |2\pi S|^{-1/2}\exp(-\frac{1}{2}(\theta-\mu)^T S^{-1}(\theta-\mu)) \\ \Rightarrow \frac{\ln N(\theta|\mu,S)}{-1/2} &= (\theta-\mu)^T S^{-1}(\theta-\mu) + K_1 = \theta^T S^{-1}\theta - 2\theta^T S^{-1}\mu + K_2. \end{aligned}$$

We may thus extract the parameters by writing the exponent as polynomial of the variable θ. S^{-1} is directly the coefficient of the second order term and μ is obtained from the first order term coefficient by multiplying with $-S/2$.

Let's proceed:

$$\begin{aligned} \frac{\ln p(\theta|y,V)}{-1/2} &= (\theta-\theta_0)^T V_0^{-1}(\theta-\theta_0) + \sum_{i=1}^{n}(y_i-\theta)^T V^{-1}(y_i-\theta) + K_1 \\ &= \theta^T V_0^{-1}\theta - 2\theta^T V_0^{-1}\theta_0 + \sum_{i=1}^{n}(\theta^T V^{-1}\theta - 2\theta^T V^{-1}y_i) + K_2 \\ &= \theta^T \underbrace{(V_0^{-1} + \sum_{i=1}^{n} V^{-1})}_{S^{-1}}\theta - 2\theta^T S^{-1}\underbrace{S(V_0^{-1}\theta_0 + V^{-1}\sum_{i=1}^{n} y_i)}_{\mu} + K_2. \end{aligned}$$

From here we get the parameters of the posterior $p(\theta|y, V) = N(\theta|\theta_p, V_p)$ as

$$V_p = (nV^{-1} + V_0^{-1})^{-1}$$

and

$$\theta_p = V_p(V^{-1}\sum_i y_i + V_0^{-1}\theta_0).$$

Now we have solved the problem. However, some insight to the problem can be gained by considering an alternative solution method. The solution is based on the fact that both prior and posterior distributions are same type of distributions, just the parameter values differ.

Thus we may obtain the solution sequentially by finding out the posterior parameters after one observation and then using these parameters as the prior parameters for the second observation. Omitting the similar algebraic manipulation, the posterior precision after one observation is

$$V_p^{-1} = V^{-1} + V_0^{-1}, \tag{1}$$

sum of prior and data precisions. After the second observations, the precision is again the sum of prior precision (which is now V_p^{-1}) and data precision. When this is repeated n times, the posterior precision becomes $V_0^{-1} + nV^{-1}$.

The mean of the posterior $p(\theta|y, V)$ after one observation y_1 is

$$\theta_p = \left(V_0^{-1} + V^{-1}\right)^{-1}\left(V_0^{-1}\theta_0 + V^{-1}y_1\right) \tag{2}$$

where θ_0 is the prior mean of θ. For the next observation y_2 we insert θ_p and V_p^{-1} in place of θ_0 and V_0^{-1} and obtain

$$\theta_p = \left(V_0^{-1} + 2V^{-1}\right)^{-1}\left(V_0^{-1}\theta_0 + V^{-1}(y_1 + y_2)\right).$$

Repeating the above steps, we get

$$\theta_p = \left(V_0^{-1} + nV^{-1}\right)^{-1}\left(V_0^{-1}\theta_0 + V^{-1}\sum_{i=1}^{n} y_i\right).$$

Some useful formulas for this problem are found in
http://www.cs.toronto.edu/˜roweis/notes/gaussid.pdf

Problem 2.

We calculate the posterior mean and variance using the Formulas (1) and (2) in Problem 1. We have only a single scalar observation y, and thus we may replace the covariance matrix V with the simple variance σ^2. Then

$$\sigma_p^{-2} = \sigma^{-2} + \sigma_0^{-2}.$$

Thus the posterior precision is the sum of prior and data precisions. Also,

$$\theta_p = \frac{\sigma_0^{-2}\theta_0 + \sigma^{-2}y}{\sigma^{-2} + \sigma_0^{-2}}.$$

This gives the posterior mean as a weighted average of the prior mean θ_0 and the observation y.

We now proceed to write the posterior mean as $\theta_p = \theta_0 + (y - \theta_0)C$.

$$\begin{aligned}\theta_p &= \frac{\sigma^{-2}y + \sigma_0^{-2}\theta_0 + \sigma^{-2}\theta_0 - \sigma^{-2}\theta_0}{\sigma^{-2} + \sigma_0^{-2}} \\ &= \frac{\sigma^{-2}(y - \theta_0) + (\sigma^{-2} + \sigma_0^{-2})\theta_0}{\sigma^{-2} + \sigma_0^{-2}} \\ &= \frac{\sigma^{-2}(y - \theta_0)}{\sigma^{-2} + \sigma_0^{-2}} + \theta_0 \\ &= \frac{\sigma_0^2(y - \theta_0)}{\sigma^2 + \sigma_0^2} + \theta_0 \\ &= \theta_0 + (y - \theta_0)\frac{\sigma_0^2}{\sigma^2 + \sigma_0^2}.\end{aligned}$$

Thus the "step size" $C = \frac{\sigma_0^2}{\sigma^2+\sigma_0^2}$.

If the prior variance σ_0^2 is much smaller than the data variance σ^2, the step size C is close to zero and the posterior mean θ_p is close to the prior mean θ_0.
On the other hand, if $\sigma_0^2 >> \sigma^2$, the step size C is close to 1 and the posterior mean is close to the observation y.

Problem 3.

i) y is Normal so

$$p(y|\sigma^2, \mu) = \frac{1}{\sqrt{2\pi}\sigma}\exp(-\frac{1}{2\sigma^2}(y-\mu)^2) \propto \sigma^{-1}\exp(-\frac{1}{2\sigma^2}(y-\mu)^2)$$

Then the posterior is proportional to

$$p(\sigma^2|y,\mu) \propto p(y|\sigma^2,\mu)p(\sigma^2|\mu) \propto \sigma^{-1}\exp(-\frac{1}{2\sigma^2}(y-\mu)^2)\sigma^{-2(a+1)}e^{-b/\sigma^2} = \sigma^{-2a-3}e^{-(\frac{1}{2}(y-\mu)^2+b)/\sigma^2}$$

ii) By writing the posterior as

$$p(\sigma^2|\mu, y) \propto \sigma^{-2(a+1/2+1)}e^{-(\frac{1}{2}(y-\mu)^2+b)/\sigma^2}$$

we can see that now $a \rightarrow a + 1/2$ and $b \rightarrow b + \frac{1}{2}(y-\mu)^2$. This is an *inverse gamma distribution*, and the parameter a is its *shape* and b its *scale*. An inverse gamma distribution with parameters a and b has mean $b/(a-1)$ and variance $b^2/(a-1)^2(a-2)$.

Increasing a decreases the mean and variance, and increasing b increases the mean and variance. For example, if we keep observing $y = \mu$, then b does not increase but a does. This makes the posterior mean and variance converge to zero, as they should.

The result of this problem shows that inverse gamma distribution is a conjugate distribution for the Normal model with unknown variance and known mean.

Problem 4.

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

Then $\log p(y|\theta) = C - \frac{1}{2\sigma^2}(y-\theta)^2$ and $\frac{\partial}{\partial\theta}\log p(y|\theta) = (\theta - y)/\sigma^2$. Fisher information is the expectation

$$\mathrm{E}[(\theta - y)^2/\sigma^4] = \sigma^{-2}$$

since $\mathrm{E}[(\theta-y)^2] = \mathrm{E}[(y-\theta)^2]$ is the variance of y given θ. Then the prior $p(\theta) \propto \sqrt{I(\theta)} = \sigma^{-1}$ which implies that the prior is constant. This cannot be normalized, so the value of σ is irrelevant.

ii) $N(y|\mu, \theta^2) = \frac{1}{\sqrt{2\pi}\theta}\exp(-\frac{1}{2\theta^2}(y-\mu)^2)$

The logarithm is

$$\log p(y|\theta) = -\log\sqrt{2\pi}\theta - \frac{1}{2\theta^2}(y-\mu)^2$$

and its derivative is

$$\frac{\partial}{\partial\theta}\log p(y|\theta) = -\theta^{-1} + \theta^{-3}(y-\mu)^2$$

Expectation of the square is

$$\mathrm{E}[\theta^{-2} + \theta^{-6}(y-\mu)^4 - 2\theta^{-4}(y-\mu)^2] = \theta^{-2} + 3\theta^{-2} - 2\theta^{-4}\theta^2 = 2\theta^{-2}$$

The expectation $\mathrm{E}[(y-\mu)^4]$ is the 4th central moment and equals $3\theta^4$ for a Normal distribution. This gives the Jeffreys' prior

$$p(\theta) \propto \sqrt{2\theta^{-2}} = \sqrt{2}\theta^{-1}$$

which again cannot be normalized.

Additional information: How does one compute the **central moments**?

can be obtained from the *cumulant generating function*

$$c_X(t) = \log \mathrm{E}(e^{tX})$$

by evaluating the derivatives of various orders at $t = 0$. Using power series expansions for e^x and $\log(1+x)$ we get

$$c_X(t) = (\mathrm{E}(tX) + \mathrm{E}((tX)^2/2) + \ldots) - 1/2()^2 + 1/3()^3 - \ldots$$

where $()^k$ denotes $(\mathrm{E}(tX)+\mathrm{E}((tX)^2/2)+\ldots)^k$. Collect the terms with the same multiplier t^k to obtain

$$c_X(t) = t\mathrm{E}(X) + t^2(1/2\mathrm{E}(X^2) - 1/2(\mathrm{E}(X))^2) + t^3 \ldots.$$

Notice that

$$\begin{aligned}c'_X(0) &= \mathrm{E}(X-\mu) = 0 \\ c''_X(0) &= \mathrm{E}((X-\mu)^2) \\ c'''_X(0) &= \mathrm{E}((X-\mu)^3)\end{aligned}$$

For higher derivatives, the central moments are not directly obtained. But it holds that

$$c_X^{(4)}(0) = \mathrm{E}((X-\mu)^4) - 3 * \mathrm{E}((X-\mu)^2)$$