Machine Learning Worksheet Solution 12

Julius Jankowski

1 KL divergence

Problem 1:

The KL-Divergence can be rewritten:

$$KL(p||q) = E_p[log(p(\mathbf{z}))] - E_p[log(q(\mathbf{z}))] \tag{1.1}$$

Then we can take a look at each term individually, so the first term is computed:

$$\begin{aligned}
E_p[log(p(\mathbf{z}))] &= -log((2\pi)^{\frac{d}{2}} det(\mathbf{\Sigma}_1)) - \frac{1}{2}(E_p[\mathbf{z}^T \mathbf{\Sigma}_1^{-1} \mathbf{z}] + E_p[\mu_1^T \mathbf{\Sigma}_1^{-1} \mu_1]) + E_p[\mu_1^T \mathbf{\Sigma}_1^{-1} \mathbf{z}] && (1.2)\\
&= -log((2\pi)^{\frac{d}{2}} det(\mathbf{\Sigma}_1)) - \frac{1}{2}(d + \mu_1^T \mathbf{\Sigma}_1^{-1} \mu_1 + \mu_1^T \mathbf{\Sigma}_1^{-1} \mu_1) + \mu_1^T \mathbf{\Sigma}_1^{-1} \mu_1 && (1.3)\\
&= -log((2\pi)^{\frac{d}{2}} det(\mathbf{\Sigma}_1)) - \frac{d}{2} && (1.4)
\end{aligned}$$

with $\mathbf{z} \in R^d$. The solution for the second term is computed almost similarly:

$$\begin{aligned}
E_p[log(q(\mathbf{z}))] &= -log((2\pi)^{\frac{d}{2}} det(\mathbf{\Sigma}_2)) - \frac{1}{2}(E_p[\mathbf{z}^T \mathbf{\Sigma}_2^{-1} \mathbf{z}] + E_p[\mu_2^T \mathbf{\Sigma}_2^{-1} \mu_2]) + E_p[\mu_2^T \mathbf{\Sigma}_2^{-1} \mathbf{z}] && (1.5)\\
&= -log((2\pi)^{\frac{d}{2}} det(\mathbf{\Sigma}_2)) - \frac{1}{2}(trace(\mathbf{\Sigma}_1 \mathbf{\Sigma}_2^{-1}) + \mu_1^T \mathbf{\Sigma}_2^{-1} \mu_1 + \mu_2^T \mathbf{\Sigma}_2^{-1} \mu_2) + \mu_2^T \mathbf{\Sigma}_2^{-1} \mu_1 && (1.6)\\
&= -log((2\pi)^{\frac{d}{2}} det(\mathbf{\Sigma}_2)) - \frac{1}{2} trace(\mathbf{\Sigma}_1 \mathbf{\Sigma}_2^{-1}) - \frac{1}{2}(\mu_1 - \mu_2)^T \mathbf{\Sigma}_2^{-1} (\mu_1 - \mu_2) && (1.7)
\end{aligned}$$

The result is yield by putting the two terms together:

$$KL(p||q) = log(\frac{det(\mathbf{\Sigma}_2)}{det(\mathbf{\Sigma}_1)}) - \frac{1}{2}(d - trace(\mathbf{\Sigma}_1 \mathbf{\Sigma}_2^{-1}) - (\mu_1 - \mu_2)^T \mathbf{\Sigma}_2^{-1} (\mu_1 - \mu_2)) \tag{1.8}$$

Problem 2:

The objective of the optimization problem is to find a parameter μ such that the isotropic Gaussian $q(\mathbf{x}|\mu)$ is close to the arbitrary distribution $p(\mathbf{x})$ with respect to the KL-Divergence. Hence we have to minimize the KL-Divergence with respect to the parameter μ:

$$\mu^* = \underset{\mu}{\text{argmin}} KL(p||q) = \underset{\mu}{\text{argmin}}(E_p[log(p(\mathbf{x}))] - E_p[log(q(\mathbf{x}|\mu))]) = \underset{\mu}{\text{argmin}}(-E_p[log(q(\mathbf{x}|\mu))]) \tag{1.9}$$

Before taking the derivative w.r.t. μ we can simplify the objective first:

$$E_p[log(q(\mathbf{x}|\mu))] = c + \mu^T E_p[\mathbf{x}] - \frac{1}{2}\mu^T\mu \tag{1.10}$$

Now after deriving w.r.t. μ we can easily see that the optimal solution for the mean of the approximating distribution is equal to the mean of the original arbitrary distribution:

$$\frac{\delta E_p[log(q(\mathbf{x}|\mu))]}{\delta\mu} = E_p[\mathbf{x}] - \mu = 0 \tag{1.11}$$

$$\implies \mu^* = E_p[\mathbf{x}] \tag{1.12}$$

2 Mean-field variational inference

Problem 3:

The posterior distribution can be computed according to Bayes theorem:

$$p(\mathbf{z}|x) = \frac{p(x|\mathbf{z})p(\mathbf{z})}{p(x)} \tag{2.1}$$

where $p(x)$ is unknown and can be computed by marginalizing over $\mathbf{z}$:

$$p(\mathbf{z}|x) = \frac{p(x|\mathbf{z})p(\mathbf{z})}{\int p(x|\mathbf{z})p(\mathbf{z})d\mathbf{z}} \tag{2.2}$$

By inserting the given distributions, we obtain a constant factor c and an exponential function:

$$p(\mathbf{z}|x) = c \cdot exp[-\frac{1}{2}(z_1^2(1+\theta_1^2) + z_2^2(1+\theta_2^2) - 2x(\theta_1 z_1 + \theta_2 z_2) + 2\theta_1\theta_2 z_1 z_2)] \tag{2.3}$$

Since there is a product term of z_1 and z_2 in the exponential function, we can not express the posterior distribution exactly by factorizing.

Problem 4:

The task is to find the optimal factorized distribution such that the KL-Divergence between $q(\mathbf{z})$ and $p(\mathbf{z}|x)$ is minimized. Starting from the ELBO, we can split up into four individual parts:

$$L(q) = E_q[log(p(\mathbf{z}, x)) - log(q(\mathbf{z}))] \tag{2.4}$$
$$= E_q[log(p(\mathbf{z}))] + E_q[log(p(x|\mathbf{z}))] - E_q[log(q_1(z_1))] - E_q[log(q_2(z_2))] \tag{2.5}$$

which are now resolved individually by using the given distributions. First one:

$$E_q[log(p(\mathbf{z}))] = -E_q[log(2\pi)] - \frac{1}{2}E_q[z_1^2 + z_2^2] = -log(2\pi) - \frac{1}{2}(m_1^2 + s_1^2 + m_2^2 + s_2^2) \tag{2.6}$$

Second one:

$$E_q[log(p(x|\mathbf{z}))] = -\frac{1}{2}(log(2\pi)+E_q[(x-\theta^T\mathbf{z})^2]) = -\frac{1}{2}(log(2\pi)+(x-\theta^T\mathbf{m})^2+\theta_1^2 s_1^2+\theta_2^2 s_2^2) \quad (2.7)$$

Third one:

$$E_q[log(q_1(z_1))] = -\frac{1}{2}(log(2\pi) - log(s_1) - \frac{1}{2s_1^2} \cdot E_q[z_1^2 - 2m_1 z_1 + m_1^2] = -\frac{1}{2} - \frac{1}{2}log(2\pi) - log(s_1) \quad (2.8)$$

The fourth term is computed analog to the third one:

$$E_q[log(q_2(z_2))] = -\frac{1}{2} - \frac{1}{2}log(2\pi) - log(s_2) \quad (2.9)$$

The resulting ELBO is:

$$L(q) = -\frac{1}{2}[(m_1^2 + s_1^2)(1+\theta_1^2) + (m_2^2 + s_2^2)(1+\theta_2^2) + log(2\pi) + x^2] \quad (2.10)$$

$$-\theta_1\theta_2 m_1 m_2 + x\theta_1 m_1 + x\theta_2 m_2 + log(s_1) + log(s_2) \quad (2.11)$$

Now we can compute the derivatives w.r.t. the parameters m and s and set them to zero in order to obtain the optimal solution:

$$\frac{\delta L(q)}{\delta m_1} = -m_1(1+\theta_1^2) - \theta_1\theta_2 m_2 + x\theta_1 = 0 \quad (2.12)$$

$$\implies m_1^* = \theta_1 \frac{x - \theta_2 m_2}{1+\theta_1^2} \quad (2.13)$$

$$\frac{\delta L(q)}{\delta s_1} = -s_1(1+\theta_1^2) + \frac{1}{s_1} = 0 \quad (2.14)$$

$$\implies s_1^{2^*} = \frac{1}{1+\theta_1^2} \quad (2.15)$$

while the two other parameters m_2 and s_2 are computed similarly by switching the indices.