Machine Learning Homework Sheet 12

Variational Inference

1 KL divergence

Problem 1: Compute the KL divergence between two Gaussian distributions $\mathcal{N}(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1)$ and $\mathcal{N}(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2)$ with diagonal covariance matrices.

Hint: If you use the facts you know about normal distribution, you can save yourself a lot of work before taking the straightforward path.

Let $p(\boldsymbol{x})$ and $q(\boldsymbol{x})$ denote the respective densities. Each distribution is parametrized by

$$\boldsymbol{\mu}_i = (\mu_{i,1}, \dots, \mu_{i,D}), \boldsymbol{\Sigma}_i = \operatorname{diag}\left(\sigma_{i,1}^2, \dots, \sigma_{i,D}^2\right).$$

We know that for Gaussians with a diagonal covariance, the PDF simply decomposes into a product of D independent Gaussians

$$p(\boldsymbol{x}) = \prod_j p_j(x_j) = \prod_j \mathcal{N}\left(x_j \mid \mu_{1,j}, \sigma_{1,j}^2\right),$$

and similarly for $q(\boldsymbol{x})$. Now

$$\begin{aligned}\mathbb{KL}(p \,||\, q) &= \int p(\boldsymbol{x}) \ln \frac{p(\boldsymbol{x})}{q(\boldsymbol{x})} \, \mathrm{d}\boldsymbol{x} \\ &= \mathbb{E}_p[\log p(\boldsymbol{x})] - \mathbb{E}_p[\log q(\boldsymbol{x})]\end{aligned}$$

Since p and q factorize, the logarithm of the fraction turns into a sum of log fractions. Linearity of expectation then gives us that the KL decomposes into a sum of KL divergences of the components:

$$\mathbb{KL}(p \,||\, q) = \sum_j \mathbb{KL}(p_j \,||\, q_j)$$

We have reduced the problem to the one-dimensional case, which is less bothersome.

$$\mathbb{KL}(p_j \| q_j) = \underbrace{-\int p_j(x) \log q_j(x) dx}_{(i)} + \underbrace{\int p_j(x) \log p_j(x) dx}_{(ii)}$$

We notice, that (ii) is just the negative entropy of a univariate Gaussian $p_j(x)$

$$\begin{aligned}\int p_j(x) \log p_j(x) dx &= -\mathbb{H}[p_j] \\ &= -\frac{1}{2}\log(2\pi\sigma_{1,j}^2) - \frac{1}{2}\end{aligned}$$

As for the first term (i), we get

$$-\int p_j(x)\log q_j(x)dx = \mathbb{E}_{p_j}\left[-\log q_j(x)\right]$$
$$= \mathbb{E}_{p_j}\left[\frac{1}{2}\log(2\pi\sigma_{2,j}^2) + \frac{(x-\mu_{2,j})^2}{2\sigma_{2,j}^2}\right]$$

By linearity of expectation

$$= \frac{1}{2}\log(2\pi\sigma_{2,j}^2) + \frac{\mathbb{E}_{p_j}\left[x^2\right] - 2\mathbb{E}_{p_j}\left[x\right]\mu_{2,j} + \mu_{2,j}^2}{2\sigma_{2,j}^2}$$
$$= \frac{1}{2}\log(2\pi\sigma_{2,j}^2) + \frac{\mu_{1,j}^2 + \sigma_{1,j}^2 - 2\mu_{1,j}\mu_{2,j} + \mu_{2,j}^2}{2\sigma_{2,j}^2}$$
$$= \frac{1}{2}\log(2\pi\sigma_{2,j}^2) + \frac{\sigma_{1,j}^2 + (\mu_{1,j}-\mu_{2,j})^2}{2\sigma_{2,j}^2}$$

Putting (i) and (ii) together, we obtain

$$\mathbb{KL}(p_j\|q_j) = \frac{1}{2}\log(2\pi\sigma_{2,j}^2) + \frac{\sigma_{1,j}^2 + (\mu_{1,j}-\mu_{2,j})^2}{2\sigma_{2,j}^2} - \frac{1}{2}\log(2\pi\sigma_{1,j}^2) - \frac{1}{2}$$
$$= \log\frac{\sigma_{2,j}}{\sigma_{1,j}} + \frac{\sigma_{1,j}^2 + (\mu_{1,j}-\mu_{2,j})^2}{2\sigma_{2,j}^2} - \frac{1}{2}$$

Finally, we can conclude that

$$\mathbb{KL}(p \mid\mid q) = \sum_j \mathbb{KL}(p_j \mid\mid q_j) = -\frac{D}{2} + \sum_j\left(\log\frac{\sigma_{2,j}}{\sigma_{1,j}} + \frac{\sigma_{1,j}^2 + (\mu_{1,j}-\mu_{2,j})^2}{2\sigma_{2,j}^2}\right).$$

Problem 2: Consider that $p(\boldsymbol{x})$ is some arbitrary fixed distribution that we wish to approximate using an isotropic Gaussian distribution $q(\boldsymbol{x}) = \mathcal{N}(\boldsymbol{x} \mid \boldsymbol{\mu}, \boldsymbol{I})$ (covariance matrix is identity matrix).

By writing down the KL divergence $\mathbb{KL}(p\|q)$ and then differentiating w.r.t. $\boldsymbol{\mu}$, show that the optimal setting of the parameter is

$$\boldsymbol{\mu}^* = \arg\min_{\boldsymbol{\mu}} \mathbb{KL}(p\|q) = \mathbb{E}_p[\boldsymbol{x}]$$

Write down the KL divergence

$$\mathbb{KL}(p\|q) = -\int p(\boldsymbol{x})\log q(\boldsymbol{x})d\boldsymbol{x} + \int p(\boldsymbol{x})\log p(\boldsymbol{x})d\boldsymbol{x}.$$

The second term doesn't depend on $q(\boldsymbol{x})$, so we can absorb it into const.

Plugging in the (Gaussian) density of $q(\boldsymbol{x})$

$$= -\int p(\boldsymbol{x}) \left(-\frac{D}{2} \log 2\pi - \frac{1}{2} \log |\boldsymbol{I}| - \frac{1}{2} (\boldsymbol{x} - \boldsymbol{\mu})^T \boldsymbol{I}^{-1} (\boldsymbol{x} - \boldsymbol{\mu}) \right) d\boldsymbol{x} + \text{const.}$$

Absorbing the constant terms

$$= -\int p(\boldsymbol{x}) \left(-\frac{1}{2} (\boldsymbol{x} - \boldsymbol{\mu})^T (\boldsymbol{x} - \boldsymbol{\mu}) \right) d\boldsymbol{x} + \text{const.}$$

Notice, that this is just an expectation w.r.t. $p(\boldsymbol{x})$. By linearity of expectation

$$\begin{aligned} &= \frac{1}{2} (\mathbb{E}_p [\boldsymbol{x}] - \boldsymbol{\mu})^T (\mathbb{E}_p [\boldsymbol{x}] - \boldsymbol{\mu}) + \text{const.} \\ &= \frac{1}{2} \boldsymbol{\mu}^T \boldsymbol{\mu} - \mathbb{E}_p [\boldsymbol{x}]^T \boldsymbol{\mu} + \text{const.} \end{aligned}$$

Compute the gradient w.r.t. $\boldsymbol{\mu}$

$$\begin{aligned} \nabla_{\boldsymbol{\mu}} \mathbb{KL}(p \| q) &= \nabla_{\boldsymbol{\mu}} \left(\frac{1}{2} \boldsymbol{\mu}^T \boldsymbol{\mu} - \mathbb{E}_p [\boldsymbol{x}]^T \boldsymbol{\mu} + \text{const.} \right) \\ &= \boldsymbol{\mu} - \mathbb{E}_p [\boldsymbol{x}] \end{aligned}$$

Setting the gradient to zero, we obtain the solution

$$\boldsymbol{\mu}^* = \arg\min_{\boldsymbol{\mu}} \mathbb{KL}(p \| q) = \mathbb{E}_p [\boldsymbol{x}]$$

2 Mean-field variational inference

Consider a very simple probabilistic model with a 2-D latent variable $\boldsymbol{z} \in \mathbb{R}^2$ and an observed variable $x \in \mathbb{R}$.

The prior over the latent variable is

$$p(\boldsymbol{z}) = \mathcal{N}(\boldsymbol{z} \mid \boldsymbol{0}, \boldsymbol{I}) = \mathcal{N}(z_1 \mid 0, 1) \cdot \mathcal{N}(z_2 \mid 0, 1),$$

and the likelihood is

$$p(x \mid \boldsymbol{z}) = \mathcal{N}(x \mid \boldsymbol{\theta}^T \boldsymbol{z}, 1),$$

where $\boldsymbol{\theta} \in \mathbb{R}^2$ is a known and fixed parameter.

Both Problem 3 and Problem 4 are about this model.

Problem 3: Write down the true posterior distribution $p(\boldsymbol{z} \mid x)$ up to the normalizing constant.

Can the posterior be factorized over z_1 and z_2? (i.e. can it be expressed as $p(z_1 \mid x) p(z_2 \mid x)$?)

The posterior distribution is

$$
\begin{aligned}
p(\boldsymbol{z} \mid x) &\propto p(\boldsymbol{z}, x) \\
&= p(z_1)p(z_2)p(x \mid \boldsymbol{z}) \\
&\propto \exp\left(-\frac{1}{2}(z_1^2 + z_2^2 + (x - \theta_1 z_1 - \theta_2 z_2)^2)\right) \\
&= \exp\left(-\frac{1}{2}(z_1^2 + z_2^2 + x^2 + \theta_1^2 z_1^2 + \theta_2^2 z_2^2 - 2x\theta_1 z_1 - 2x\theta_2 z_2 + 2\theta_1 z_1 \theta_2 z_2)\right)
\end{aligned}
$$

Because of the presence of term $2\theta_1 z_1 \theta_2 z_2$ we are not able to write the posterior as the product

$$p(\boldsymbol{z} \mid x) = p(z_1 \mid x)p(z_2 \mid x).$$

Problem 4: We approximate the true posterior using a mean-field variational distribution

$$q(\boldsymbol{z}) = q_1(z_1)q_2(z_2) = \mathcal{N}(z_1 \mid m_1, s_1^2) \cdot \mathcal{N}(z_2 \mid m_2, s_2^2)$$

Your task is to derive the optimal updates for q_1 and q_2.

Is $q(\boldsymbol{z})$ able to match the true posterior $p(\boldsymbol{z} \mid x)$?

Applying the formula for the optimal mean-field update for $q(z_1)$, we obtain

$$
\begin{aligned}
q_1^*(z_1) &\propto \exp\left(\mathbb{E}_{q_2(z_2)}\left[\log p(\boldsymbol{z}, x)\right]\right) \\
&= \exp\left(-\frac{1}{2}\mathbb{E}_{q_2}\left[z_1^2 + z_2^2 + x^2 + \theta_1^2 z_1^2 + \theta_2^2 z_2^2 - 2x\theta_1 z_1 - 2x\theta_2 z_2 + 2\theta_1 z_1 \theta_2 z_2\right]\right) \\
&= \exp\left(-\frac{1}{2}(z_1^2 + \mathbb{E}_{q_2}\left[z_2^2\right] + x^2 + \theta_1^2 z_1^2 + \theta_2^2 \mathbb{E}_{q_2}\left[z_2^2\right]\right. \\
&\qquad \left. - 2x\theta_1 z_1 - 2x\theta_2 \mathbb{E}_{q_2}\left[z_2\right] + 2\theta_1 z_1 \theta_2 \mathbb{E}_{q_2}\left[z_2\right])\right).
\end{aligned}
$$

Grouping together the terms dependent on z_1, and absorbing the rest into const

$$\propto \exp\left(-\frac{1}{2}((1 + \theta_1^2)z_1^2 - 2\theta_1 z_1(x - \theta_2 \mathbb{E}_{q_2}\left[z_2\right]))\right). \qquad (\star)$$

Plugging in $\mathbb{E}_{q_2}\left[z_2\right] = \mu_2$

$$\propto \exp\left(-\frac{1}{2}((1 + \theta_1^2)z_1^2 - 2\theta_1 z_1(x - \theta_2 \mu_2))\right). \qquad (\star)$$

We recognize that this is a squared exponential function of z_1, hence $q_1(z_1)$ must be a Gaussian distribution, which matches our initial assumption.

We can find its parameters μ_1 and σ_1^2 by completing the square. A univariate Gaussian density can be written as

$$\begin{aligned}\mathcal{N}(z_1 \mid \mu_1, \sigma_1^2) &= \exp\left(-\frac{1}{2}\frac{(z_1-\mu_1)^2}{\sigma_1^2}\right) \\ &= \exp\left(-\frac{1}{2}\left(\frac{1}{\sigma_1^2}z_1^2 - \frac{2\mu_1}{\sigma_1^2}z_1 + \frac{\mu_1^2}{\sigma_1^2}\right)\right). \qquad (\star\star)\end{aligned}$$

Comparing $(\star)$ and $(\star\star)$, we observe that

$$\begin{aligned}\frac{1}{\sigma_1^2}z_1^2 &\overset{!}{=} (1+\theta_1^2)z_1^2 \\ \implies \sigma_1^2 &= \frac{1}{1+\theta_1^2}.\end{aligned}$$

Furthermore,

$$\begin{aligned}\frac{-2\mu_1}{\sigma_1^2}z_1 &= -2(1+\theta_1^2)\mu_1 z_1 \overset{!}{=} -2\theta_1(x-\theta_2\mu_2)z_1 \\ \implies \mu_1 &= \frac{\theta_1(x-\theta_2\mu_2)}{1+\theta_1^2}.\end{aligned}$$

Using the same line of reasoning, we find that $q_2(z_2)$ is indeed as well a Gaussian, with the optimal update given as

$$\begin{aligned}\mu_2 &= \frac{\theta_2(x-\theta_1\mu_1)}{1+\theta_2^2} \\ \sigma_2^2 &= \frac{1}{1+\theta_2^2}.\end{aligned}$$

As we noticed when solving Problem 3, the true posterior $p(\boldsymbol{z} \mid x)$ cannot be factorized as

$$p(\boldsymbol{z} \mid x) = p(z_1 \mid x)p(z_2 \mid x).$$

Therefore, obviously, a factorized variational distribution

$$q(\boldsymbol{z}) = q(z_1)q(z_2)$$

is not able to match it.