TU München
Fakultät für Informatik
PD Dr. Rudolph Triebel
John Chiotellis, Maximilian Denninger

Machine Learning for Computer Vision
Winter term 2018

January 25, 2019
Topic: Variational Inference

Exercise 1: Kullback-Leibler divergence

a) *What does the KL divergence describe? What are its key properties?*

The Kullback-Leibler divergence is a measure of (dis-)similarity between probability distributions. It is the extra amount of information needed when a distribution q is used to approximate a distribution p. It is not symmetric ($KL(p||q) \neq KL(q||p)$) and non-negative ($KL(p||q) \geq 0$). It is minimized (zero) when the two distributions are identical. By the definition we have:

$$\begin{aligned} KL(p||q) &= \int p(x) \log \frac{p(x)}{q(x)} dx \\ &= \int p(x) \log p(x) dx - \int p(x) \log q(x) dx \\ &= -H(p) + H(p,q) \\ &= \textit{negative entropy of } p + \textit{cross entropy between } p \textit{ and } q \end{aligned}$$

b) *Compute the KL-divergence of two univariate normal distributions.*
What if they have the same mean? What if they have the same variance?

Let us define $p_1(x) = \mathcal{N}(x|\mu_1, \sigma_1)$ and $p_2(x) = \mathcal{N}(x|\mu_2, \sigma_2)$. We then have

$$KL(p_1||p_2) = \int p_1(x) \log\{\frac{p_1(x)}{p_2(x)}\} dx$$

First let us simplify the fraction

$$\begin{aligned} \frac{p_1(x)}{p_2(x)} &= \frac{\frac{1}{\sqrt{2\pi\sigma_1^2}} \exp(-\frac{(x-\mu_1)^2}{2\sigma_1^2})}{\frac{1}{\sqrt{2\pi\sigma_2^2}} \exp(-\frac{(x-\mu_2)^2}{2\sigma_2^2})} = \frac{\sigma_2}{\sigma_1} \frac{\exp(-\frac{(x-\mu_1)^2}{2\sigma_1^2})}{\exp(-\frac{(x-\mu_2)^2}{2\sigma_2^2})} \\ &= \frac{\sigma_2}{\sigma_1} \exp(-\frac{(x-\mu_1)^2}{2\sigma_1^2} + \frac{(x-\mu_2)^2}{2\sigma_2^2}) \end{aligned}$$

Taking the logarithm of this gives us

$$\log(\frac{p_1(x)}{p_2(x)}) = \log(\frac{\sigma_2}{\sigma_1}) + \left(\frac{(x-\mu_2)^2}{2\sigma_2^2} - \frac{(x-\mu_1)^2}{2\sigma_1^2}\right)$$

Now plugging this in the KL-divergence definition we get

$$\begin{aligned}
KL(p_1||p_2) &= \int p_1(x)\log(\frac{\sigma_2}{\sigma_1})dx + \int p_1(x)\left(\frac{(x-\mu_2)^2}{2\sigma_2^2} - \frac{(x-\mu_1)^2}{2\sigma_1^2}\right)dx \\
&= \log(\frac{\sigma_2}{\sigma_1})\int p_1(x)dx + \int p_1(x)\frac{(x-\mu_2)^2}{2\sigma_2^2}dx - \int p_1(x)\frac{(x-\mu_1)^2}{2\sigma_1^2}dx \\
&= \log(\frac{\sigma_2}{\sigma_1}) + \frac{1}{2\sigma_2^2}\int p_1(x)(x-\mu_2)^2dx - \frac{1}{2\sigma_1^2}\int p_1(x)(x-\mu_1)^2dx \\
&= \log(\frac{\sigma_2}{\sigma_1}) + \frac{1}{2\sigma_2^2}\int p_1(x)(x-\mu_1+\mu_1-\mu_2)^2dx - \frac{\sigma_1^2}{2\sigma_1^2} \\
&= \log(\frac{\sigma_2}{\sigma_1}) + \frac{1}{2\sigma_2^2}\left(\int p_1(x)(x-\mu_1)^2dx + 2\int p_1(x)(x-\mu_1)(\mu_1-\mu_2)dx + \int p_1(x)(\mu_1-\mu_2)^2dx\right) - \frac{1}{2} \\
&= \log(\frac{\sigma_2}{\sigma_1}) + \frac{1}{2\sigma_2^2}\left(\sigma_1^2 + 2(\mu_1-\mu_2)\int p_1(x)(x-\mu_1)dx + (\mu_1-\mu_2)^2\int p_1(x)dx\right) - \frac{1}{2} \\
&= \log(\frac{\sigma_2}{\sigma_1}) + \frac{1}{2\sigma_2^2}\left(\sigma_1^2 + (\mu_1-\mu_2)^2\right) - \frac{1}{2}
\end{aligned}$$

If two distributions only differ in their mean values ($\sigma_1 = \sigma_2$) then the KL-divergence is proportional to the square of their means difference,

$$KL(p_1||p_2) = \frac{(\mu_1-\mu_2)^2}{2\sigma_2^2}.$$

If they have equal mean but different variances ($\mu_1 = \mu_2$) then the KL-divergence is a function of the ratio of their variances:

$$KL(p_1||p_2) = \log(\frac{\sigma_2}{\sigma_1}) + \frac{\sigma_1^2}{2\sigma_2^2} - \frac{1}{2} = \frac{\sigma_1^2}{2\sigma_2^2} - \log(\frac{\sigma_1}{\sigma_2}) - \frac{1}{2}$$

c) *Consider a factorized variational distribution $q(Z)$. By using the technique of Lagrange multipliers, verify that minimization of $KL(p||q)$ with respect to one of the factors $q_i(Z_i)$ keeping all other factors fixed, leads to the solution:*

$$q_j^*(Z_j) = \int p(Z)\prod_{i\neq j} dZ_i = p(Z_j)$$

$$
\begin{aligned}
KL(p||q) &= \int p(Z) \ln \frac{p(Z)}{q(Z)} dZ \\
&= \int p(Z) \ln p(Z) dZ - \int p(Z) \ln q(Z) dZ \\
&= \int p(Z) \ln p(Z) dZ - \int p(Z) \ln \prod_i q_i(Z_i) dZ \\
&= - \int p(Z) \sum_{i=1}^{M} \ln q_i(Z_i) dZ + const. \\
&= - \int (p(Z) \ln q_j(Z_j) + p(Z) \sum_{i \neq j} \ln q_i(Z_i)) dZ + const. \\
&= - \int p(Z) \ln q_j(Z_j) dZ + const. \\
&= - \int \ln q_j(Z_j) \left(\int p(Z) \prod_{i \neq j} dZ_i \right) dZ_j + const.
\end{aligned}
$$

Note that by *const.* we imply w.r.t. q_j. We want to minimize this and at the same time enforce the constraint

$$
\int q_j(Z_j) dZ_j = 1.
$$

Therefore we add a Lagrange multiplier and our objective function becomes

$$
\mathcal{L}(q_j(Z_j)) = - \int \ln q_j(Z_j) \left(\int p(Z) \prod_{i \neq j} dZ_i \right) dZ_j + \lambda \left(\int q_j(Z_j) dZ_j - 1 \right)
$$

Taking the derivative w.r.t. $q_j(Z_j)$ and setting it equal to zero we get

$$
\frac{\partial \mathcal{L}(q_j(Z_j))}{\partial q_j(Z_j)} = - \frac{\int p(Z) \prod_{i \neq j} dZ_i}{q_j(Z_j)} + \lambda \overset{!}{=} 0
$$

We solve for λ

$$
\begin{aligned}
\lambda q_j(Z_j) &= \int p(Z) \prod_{i \neq j} dZ_i \\
\lambda \int q_j(Z_j) dZ_j &= \int \left(\int p(Z) \prod_{i \neq j} dZ_i \right) dZ_j \\
\lambda &= 1
\end{aligned}
$$

And thus

$$
q_j^*(Z_j) = \int p(Z) \prod_{i \neq j} dZ_i = p(Z_j)
$$