[HW1] KL Divergence between Two Gaussian Distribution

0616014 楊政道

May 8, 2020

Question

Given the prior $p(z) \sim N(0, I)$ and the posterior approximation $q(z|x;\theta) \sim N(\mu_\theta(x), \Sigma_\theta(x))$, prove that $KL(q(z|x;\theta)\|p(z))$ is tractable; that is, it can be the functions of $\mu_\theta(x)$ and $\Sigma_\theta(x)$, expressed as a closed-form expression. Both dimensions of multivariate Gaussian are n where mean $\mu_\theta(x)$ and covariance matrix $\Sigma_\theta(x) = diag(\sigma_1^2, \ldots, \sigma_n^2)$ are functions of x and the parameters θ of a neural network.

Solution

Given two gaussian distribution $p(z) \sim N(0, I)$ and $q(z|x;\theta) \sim N(\mu_\theta(x), \Sigma_\theta(x))$. Evaluate $KL(q(z|x;\theta)\|p(z))$.

$$KL(q(z|x;\theta)\|p(z)) = -\int q(z|x;\theta)ln[p(z)]dz + \int q(z|x;\theta)ln[q(z|x;\theta)]dz$$

Evaluate $-\int q(z|x;\theta)ln[p(z)]dz$

$$\begin{aligned}
-\int q(z|x;\theta)ln[p(z)]dz &= -\int q(z|x;\theta)ln\left[\frac{1}{\sqrt{(2\pi)^n|I|}}e^{-\frac{1}{2}(z-0)^T I^{-1}(z-0)}\right]dz \\
&= -\int q(z|x;\theta)\left[ln(2\pi)^{-\frac{n}{2}} + ln(e^{-\frac{1}{2}z^T z})\right]dz \\
&= -\int q(z|x;\theta)(-\frac{n}{2}ln(2\pi) - \frac{1}{2}z^T z)dz \\
&= \frac{n}{2}ln(2\pi)\int q(z|x;\theta)dz + \frac{1}{2}\int q(z|x;\theta)z^T z)dz \\
&= \frac{n}{2}ln(2\pi) + \frac{1}{2}E_{z\sim q(z|x;\theta)}[z^T z] \\
&= \frac{n}{2}ln(2\pi) + \frac{1}{2}(|\Sigma_\theta(x)| + \left[E_{z\sim q(z|x;\theta)}[z]\right]^T\left[E_{z\sim q(z|x;\theta)}[z]\right]) \\
&= \frac{n}{2}ln(2\pi) + \frac{1}{2}|\Sigma_\theta(x)| + \frac{1}{2}\mu_\theta(x)^T\mu_\theta(x)
\end{aligned} \tag{1}$$

Evaluate $\int q(z|x;\theta)ln[q(z|x;\theta)]dz$

$$\begin{aligned}
\int q(z|x;\theta)ln[q(z|x;\theta)]dz &= \int q(z|x;\theta)ln\left[\frac{1}{\sqrt{(2\pi)^n|\Sigma_\theta(x)|}}e^{-\frac{1}{2}(z-\mu_\theta(x))^T\Sigma_\theta^{-1}(x)(z-\mu_\theta(x))}\right]dz \\
&= \int q(z|x;\theta)\left(ln(2\pi)^{-\frac{n}{2}} + ln[|\Sigma_\theta(x)|]^{-\frac{1}{2}} + ln\left[e^{-\frac{1}{2}(z-\mu_\theta(x))^T\Sigma_\theta^{-1}(x)(z-\mu_\theta(x))}\right]\right)dz \\
&= \int q(z|x;\theta)\left[-\frac{n}{2}ln(2\pi) - \frac{1}{2}ln[|\Sigma_\theta(x)|] - \frac{1}{2}(z-\mu_\theta(x))^T\Sigma_\theta^{-1}(x)(z-\mu_\theta(x))\right]dz \\
&= -\frac{n}{2}ln(2\pi) - \frac{1}{2}ln[|\Sigma_\theta(x)|] - \frac{1}{2}|\Sigma_\theta^{-1}(x)|\int q(z|x;\theta)\left[(z-\mu_\theta(x))^T(z-\mu_\theta(x))\right]dz \\
&= -\frac{n}{2}ln(2\pi) - \frac{1}{2}ln[|\Sigma_\theta(x)|] - \frac{1}{2}|\Sigma_\theta^{-1}(x)||\Sigma_\theta(x)| \\
&= -\frac{n}{2}ln(2\pi) - \frac{1}{2}ln[|\Sigma_\theta(x)|] - \frac{1}{2}
\end{aligned} \tag{2}$$

Combine (1) and (2) to get the closed-form expression of $KL(q(z|x;\theta)\|p(z))$

$$\begin{aligned}
KL(q(z|x;\theta)\|p(z)) &= -\int q(z|x;\theta)ln[p(z)]dz + \int q(z|x;\theta)ln[q(z|x;\theta)]dz \\
&= \frac{n}{2}ln(2\pi) + \frac{1}{2}|\Sigma_\theta(x)| + \frac{1}{2}\mu_\theta(x)^T\mu_\theta(x) - \frac{n}{2}ln(2\pi) - \frac{1}{2}ln[|\Sigma_\theta(x)|] - \frac{1}{2} \\
&= \frac{1}{2}|\Sigma_\theta(x)| + \frac{1}{2}\mu_\theta(x)^T\mu_\theta(x) - \frac{1}{2}ln[|\Sigma_\theta(x)|] - \frac{1}{2} \\
&= \frac{1}{2}\prod_{i=1}^{n}\sigma_i^2 + \frac{1}{2}\mu_\theta(x)^T\mu_\theta(x) - \frac{1}{2}ln[\prod_{i=1}^{n}\sigma_i^2] - \frac{1}{2} \\
&= \frac{1}{2}\prod_{i=1}^{n}\sigma_i^2 + \frac{1}{2}\mu_\theta(x)^T\mu_\theta(x) - \sum_{i=1}^{n}ln(\sigma_i) - \frac{1}{2}
\end{aligned} \tag{3}$$