Introduction to Deep Learning M2177.0043
Seoul National University

Homework #(**4**)
Seungyong Moon

INSTRUCTIONS

- Anything that is received after the deadline will be considered to be late and we do not receive late homeworks. We do however ignore your lowest homework grade.
- Answers to every theory questions need to be submitted electronically on ETL. Only PDF generated from LaTex is accepted.
- Make sure you prepare the answers to each question separately. This helps us dispatch the problems to different graders.
- Collaboration on solving the homework is allowed. Discussions are encouraged but you should think about the problems on your own.
- If you do collaborate with someone or use a book or website, you are expected to write up your solution independently. That is, close the book and all of your notes before starting to write up your solution.

1 VAE

Implement functionalities in VAE.ipynb file.

sol) See hw4_results.zip

2 KL divergence between two multivariate Gaussians

Derive the closed form expression for the KL divergence between two multivariate Gaussian distributions $D_{KL}(p,q)$ where each distribution is parameterized by (μ_1, Σ_1) and (μ_2, Σ_2) respectively.

sol) Suppose that $p \sim N(\mu_1, \Sigma_1)$, $q \sim N(\mu_2, \Sigma_2)$. we already know that the pdf of normal distribution $N(\mu, \Sigma)$ is

$$f(\mathbf{x}) = \frac{exp(-\frac{1}{2}(\mathbf{x}-\mu)^T\Sigma^{-1}(\mathbf{x}-\mu))}{\sqrt{(2\pi)^k|\Sigma|}} \tag{1}$$

where k is the dimension of $\mathbf{x}$.
Let f_p and f_q be the pdf of the p and q. By the definition of KL divergence, we can get the followings.

$$\begin{aligned}
D_{KL}(p,q) &= \int_{\mathbb{R}^k} f_p(x)\log\frac{f_p(x)}{f_q(x)} \\
&= \int_{\mathbb{R}^k} f_p(x)\log f_p(x) - \int_{\mathbb{R}^k} f_p(x)\log f_q(x) \\
&= E_p[\log f_p] - E_p[\log f_q] \\
&= E_p\Big[-\frac{1}{2}(p-\mu_1)^T\Sigma_1^{-1}(p-\mu_1) - \frac{k}{2}\log(2\pi) - \frac{1}{2}\log(|\Sigma_1|)\Big] \\
&\quad - E_p\Big[-\frac{1}{2}(p-\mu_2)^T\Sigma_2^{-1}(p-\mu_2) - \frac{k}{2}\log(2\pi) - \frac{1}{2}\log(|\Sigma_2|)\Big] \\
&= -\frac{1}{2}E_p[(p-\mu_1)^T\Sigma_1^{-1}(p-\mu_1)] + \frac{1}{2}E_p[(p-\mu_2)^T\Sigma_2^{-1}(p-\mu_2)] + \frac{1}{2}log\frac{|\Sigma_2|}{|\Sigma_1|} \\
&= -\frac{1}{2}E_p[tr(\Sigma_1^{-1}(p-\mu_1)(p-\mu_1)^T)] + \frac{1}{2}E_p[tr(\Sigma_2^{-1}(p-\mu_2)(p-\mu_2)^T)] + \frac{1}{2}log\frac{|\Sigma_2|}{|\Sigma_1|}
\end{aligned} \tag{2}$$

Introduction to Deep Learning M2177.0043
Seoul National University

Homework #(**4**)
Seungyong Moon

Here, some properties of trace are used, $tr(k) = k$ and $tr(AB) = tr(BA)$, where k is scalar value and A and B are Matrix.

Since trace and matrix multiplication is linear mapping, we can exchange trace and matrix multiplication with expectation.

$$
\begin{aligned}
D_{KL}(p,q) &= \int_{\mathbb{R}^k} f_p(x) \log \frac{f_p(x)}{f_q(x)} \\
&= -\frac{1}{2}E_p[tr(\Sigma_1^{-1}(p-\mu_1)(p-\mu_1)^T)] + \frac{1}{2}E_p[tr(\Sigma_2^{-1}(p-\mu_2)(p-\mu_2)^T)] + \frac{1}{2}log\frac{|\Sigma_2|}{|\Sigma_1|} \\
&= -\frac{1}{2}tr(\Sigma_1^{-1}E_p[(p-\mu_1)(p-\mu_1)^T]) + \frac{1}{2}tr(\Sigma_2^{-1}E_p[(p-\mu_2)(p-\mu_2)^T]) + \frac{1}{2}log\frac{|\Sigma_2|}{|\Sigma_1|} \\
&= -\frac{1}{2}tr(\Sigma_1^{-1}\Sigma_1) + \frac{1}{2}tr(\Sigma_2^{-1}E_p[pp^T - p\mu_2^T - \mu_2 p^T + \mu_2\mu_2^T]) + \frac{1}{2}log\frac{|\Sigma_2|}{|\Sigma_1|} \\
&= -\frac{1}{2}tr(I) + \frac{1}{2}tr(\Sigma_2^{-1}(\Sigma_1 + \mu_1\mu_1^T - \mu_1\mu_2^T - \mu_2\mu_1^T + \mu_2\mu_2^T)) + \frac{1}{2}log\frac{|\Sigma_2|}{|\Sigma_1|}
\end{aligned}
\tag{3}
$$

Here, I use the fact $\Sigma_1 = E_p[pp^T] - \mu_1\mu_1^T$.

Since $tr(I) = k$ and $tr(A) = tr(A^T)$, we can get the followings.

$$
\begin{aligned}
D_{KL}(p,q) &= \int_{\mathbb{R}^k} f_p(x) \log \frac{f_p(x)}{f_q(x)} \\
&= -\frac{1}{2}tr(I) + \frac{1}{2}tr(\Sigma_2^{-1}(\Sigma_1 + \mu_1\mu_1^T - \mu_1\mu_2^T - \mu_2\mu_1^T + \mu_2\mu_2^T)) + \frac{1}{2}log\frac{|\Sigma_2|}{|\Sigma_1|} \\
&= \frac{1}{2}\Big(tr(\Sigma_2^{-1}\Sigma_1 + \Sigma_2^{-1}\mu_1\mu_1^T - \Sigma_2^{-1}\mu_1\mu_2^T - \Sigma_2^{-1}\mu_2\mu_1^T + \Sigma_2^{-1}\mu_2\mu_2^T) + log\frac{|\Sigma_2|}{|\Sigma_1|} - k\Big) \\
&= \frac{1}{2}\Big(tr(\Sigma_2^{-1}\Sigma_1) + tr(\Sigma_2^{-1}\mu_1\mu_1^T) - tr(\Sigma_2^{-1}\mu_1\mu_2^T) - tr(\Sigma_2^{-1}\mu_2\mu_1^T) + tr(\Sigma_2^{-1}\mu_2\mu_2^T) + log\frac{|\Sigma_2|}{|\Sigma_1|} - k\Big) \\
&= \frac{1}{2}\Big(tr(\Sigma_2^{-1}\Sigma_1) + tr(\mu_1^T\Sigma_2^{-1}\mu_1) - tr(\mu_2^T\Sigma_2^{-1}\mu_1) - tr(\mu_1^T\Sigma_2^{-1}\mu_2) + tr(\mu_2^T\Sigma_2^{-1}\mu_2) + log\frac{|\Sigma_2|}{|\Sigma_1|} - k\Big) \\
&= \frac{1}{2}\Big(tr(\Sigma_2^{-1}\Sigma_1) + tr(\mu_1^T\Sigma_2^{-1}\mu_1 - 2\mu_2^T\Sigma_2^{-1}\mu_1 + \mu_2^T\Sigma_2^{-1}\mu_2) + log\frac{|\Sigma_2|}{|\Sigma_1|} - k\Big) \\
&= \frac{1}{2}\Big(tr(\Sigma_2^{-1}\Sigma_1) + tr((\mu_2-\mu_1)^T\Sigma_2^{-1}(\mu_2-\mu_1)) + log\frac{|\Sigma_2|}{|\Sigma_1|} - k\Big) \\
&= \frac{1}{2}\Big(tr(\Sigma_2^{-1}\Sigma_1) + (\mu_2-\mu_1)^T\Sigma_2^{-1}(\mu_2-\mu_1) + log\frac{|\Sigma_2|}{|\Sigma_1|} - k\Big)
\end{aligned}
\tag{4}
$$