KL divergence between two multivariate Gaussians

jojonki

May 2018

1　導出するよ

VAEでは多次元正規分布$\mathcal{N}(\mathbf{0}, \boldsymbol{I})$ と$\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\sigma^2})$のKLダイバージェンスを求めたが，$\mathcal{N}(\boldsymbol{\mu}_1, \boldsymbol{\sigma_1^2})$と$\mathcal{N}(\boldsymbol{\mu}_2, \boldsymbol{\sigma_2^2})$のKLダイバージェンスを求めたいと思う．それぞれ$q(z)$，$p(z)$として考えてみる．

$$
\begin{aligned}
D_{KL}(q(z)||p(z)) &= \int q(z) \log \frac{q(z)}{p(z)} dz \\
&= \int q(z) \Big\{ \log q(z) - \log p(z) \Big\} dz \\
&= \int q(z) \log q(z) dz - \int q(z) \log p(z) dz
\end{aligned} \tag{1}
$$

第1項と第2項でそれぞれ計算する．第1項は以前にVAEの計算のときに求めた方法と同じ．第2項が異なる分布間（4変数）なので少しだけトリッキー．

第1項

$$
\begin{aligned}
\int q(z)\log q(z)dz &= \int \mathcal{N}(z;\mu_1,\sigma_1^2)\log \mathcal{N}(z;\mu_1,\sigma_1^2)dz \\
&= \sum_j^J E_{q(z_j)}\Big[\log \frac{1}{\sqrt{2\pi\sigma_{1,j}^2}}\exp\big(-\frac{(z_j-\mu_{1,j})^2}{2\sigma_{1,j}^2}\big)\Big] \\
&= \sum_j^J E_{q(z_j)}\Big[-\frac{1}{2}\log(2\pi\sigma_{1,j}^2)-\frac{(z_j-\mu_{1,j})^2}{2\sigma_{1,j}^2}\Big] \\
&= \sum_j^J \Big\{-\frac{1}{2}\log(2\pi\sigma_{1,j}^2)-E_{q(z_j)}\Big[\frac{(z_j-\mu_{1,j})^2}{2\sigma_{1,j}^2}\Big]\Big\} \\
&= -\frac{J}{2}\log(2\pi)-\frac{1}{2}\sum_{j=1}^J\log\sigma_{1,j}^2-\frac{1}{2}\sum_{j=1}^J E_{q(z_j)}\Big[\frac{(z_j-\mu_{1,j})^2}{\sigma_{1,j}^2}\Big] \qquad (2)\\
&= -\frac{J}{2}\log(2\pi)-\frac{1}{2}\sum_{j=1}^J\log\sigma_{1,j}^2-\frac{1}{2}\sum_{j=1}^J \frac{1}{\sigma_{1,j}^2}E_{q(z_j)}\Big[(z_j-\mu_{1,j})^2\Big] \\
&\quad \text{第3項の右側は分散の定義そのもの.} \\
&= -\frac{J}{2}\log(2\pi)-\frac{1}{2}\sum_{j=1}^J\log\sigma_{1,j}^2-\frac{1}{2}\sum_{j=1}^J \frac{1}{\sigma_{1,j}^2}\sigma_{1,j}^2 \\
&= -\frac{J}{2}\log(2\pi)-\frac{1}{2}\sum_{j=1}^J(\log\sigma_{1,j}^2+1)
\end{aligned}
$$

第2項

$$
\begin{aligned}
\int q(z)\log p(z)dz &= \int \mathcal{N}(z;\mu_1,\sigma_1^2)\log\mathcal{N}(z;\mu_2,\sigma_2^2)dz \\
&= \sum_j^J E_{q(z_j)}\Big[\log\frac{1}{\sqrt{2\pi\sigma_{2,j}^2}}\exp\big(-\frac{(z_j-\mu_{2,j})^2}{2\sigma_{2,j}^2})\Big] \\
&= \sum_j^J E_{q(z_j)}\Big[-\frac{1}{2}\log(2\pi\sigma_{2,j}^2)-\frac{(z_j-\mu_{2,j})^2}{2\sigma_{2,j}^2}\Big] \\
&= \sum_j^J \Big\{-\frac{1}{2}\log(2\pi\sigma_{2,j}^2)-E_{q(z_j)}\Big[\frac{(z_j-\mu_{2,j})^2}{2\sigma_{2,j}^2}\Big]\Big\} \\
&= -\frac{J}{2}\log(2\pi)-\frac{1}{2}\sum_{j=1}^J\log\sigma_{2,j}^2-\frac{1}{2}\sum_{j=1}^J E_{q(z_j)}\Big[\frac{(z_j-\mu_{2,j})^2}{\sigma_{2,j}^2}\Big] \\
&= -\frac{J}{2}\log(2\pi)-\frac{1}{2}\sum_{j=1}^J\log\sigma_{2j}^2-\frac{1}{2}\sum_{j=1}^J \frac{1}{\sigma_{2,j}^2}E_{q(z_j)}\Big[(z_j-\mu_{2,j})^2\Big]
\end{aligned}
$$

ここから先程と違う．第3項を展開していく．

$$
\begin{aligned}
&= -\frac{J}{2}\log(2\pi)-\frac{1}{2}\sum_{j=1}^J\log\sigma_{2j}^2-\frac{1}{2}\sum_{j=1}^J \frac{1}{\sigma_{2,j}^2}E_{q(z_j)}\Big[z_j^2-2z_j\mu_{2,j}+\mu_{2,j}^2\Big] \qquad (3) \\
&= -\frac{J}{2}\log(2\pi)-\frac{1}{2}\sum_{j=1}^J\log\sigma_{2j}^2-\frac{1}{2}\sum_{j=1}^J \frac{1}{\sigma_{2,j}^2}\Big[E_{q(z_j)}z_j^2-2E_{q(z_j)}z_j\mu_{2,j}+E_{q(z_j)}\mu_{2,j}^2\Big]
\end{aligned}
$$

$q(z)\sim\mathcal{N}(\boldsymbol{\mu}_1,\boldsymbol{\sigma_1^2})$に関する期待値を$\langle X\rangle$の記法で便宜上書く．

$$
= -\frac{J}{2}\log(2\pi)-\frac{1}{2}\sum_{j=1}^J\log\sigma_{2j}^2-\frac{1}{2}\sum_{j=1}^J \frac{1}{\sigma_{2,j}^2}\Big[\langle z_j^2\rangle-2\langle z_j\rangle\mu_{2,j}+\mu_{2,j}^2\Big]
$$

分散の公式から$\langle z_j^2\rangle=\sigma_{1,j}^2+\mu_{1,j}^2$であり，$\langle z_j\rangle=\mu_j$であるから，

$$
\begin{aligned}
&= -\frac{J}{2}\log(2\pi)-\frac{1}{2}\sum_{j=1}^J\log\sigma_{2j}^2-\frac{1}{2}\sum_{j=1}^J \frac{1}{\sigma_{2,j}^2}\Big[\sigma_{1,j}^2+\mu_{1,j}^2-2\mu_{1,j}\mu_{2,j}+\mu_{2,j}^2\Big] \\
&= -\frac{J}{2}\log(2\pi)-\frac{1}{2}\sum_{j=1}^J\log\sigma_{2j}^2-\frac{1}{2}\sum_{j=1}^J \frac{1}{\sigma_{2,j}^2}\Big[\sigma_{1,j}^2+(\mu_{1,j}-\mu_{2,j})^2\Big] \\
&= -\frac{J}{2}\log(2\pi)-\frac{1}{2}\sum_{j=1}^J\log\sigma_{2j}^2-\frac{1}{2}\sum_{j=1}^J \Big[\frac{\sigma_{1,j}^2}{\sigma_{2,j}^2}+\frac{(\mu_{1,j}-\mu_{2,j})^2}{\sigma_{2,j}^2}\Big]
\end{aligned}
$$

まとめ

第1項から第2項を引けば良いので，

$$
\begin{aligned}
D_{KL}(q(z)||p(z)) &= \int q(z)\log\frac{q(z)}{p(z)}dz \\
&= \int q(z)\Big\{\log q(z) - \log p(z)\Big\}dz \\
&= \int q(z)\log q(z)dz - \int q(z)\log p(z)dz \\
&= \Bigg\{-\frac{J}{2}\log(2\pi) - \frac{1}{2}\sum_{j=1}^{J}(\log\sigma_{1,j}^2 + 1)\Bigg\} - \Bigg\{-\frac{J}{2}\log(2\pi) - \frac{1}{2}\sum_{j=1}^{J}\log\sigma_{2j}^2 - \frac{1}{2}\sum_{j=1}^{J}\Big[\frac{\sigma_{1,j}^2}{\sigma_{2,j}^2} + \frac{(\mu_{1,j}-\mu_{2,j})^2}{\sigma_{2,j}^2}\Big] \\
&= -\frac{1}{2}\sum_{j=1}^{J}\Big[\log\frac{\sigma_{1,j}^2}{\sigma_{2,j}^2} - \frac{\sigma_{1,j}^2}{\sigma_{2,j}^2} - \frac{(\mu_{1,j}-\mu_{2,j})^2}{\sigma_{2,j}^2} + 1\Big]
\end{aligned}
\tag{4}
$$

ちなみにVAEのときを復讐すると，$p(z)$はVAEの設定から $\sim \mathcal{N}(\mathbf{0}, \boldsymbol{I})$.であり，$q(z|x)$は$\sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\sigma^2})$となる．これは今回求めたものに，$\mu_1 = 0, \sigma_1 = 1$を代入すれば同じ値になることが確認できる．

$$
\begin{aligned}
D_{KL}(q(z|x)||p(z)) &= \int q(z|x)\log\frac{q(z|x)}{p(z)}dz \\
&= \int q(z|x)\{\log q(z|x) - \log p(z)\}dz \\
&= \int q(z|x)\log q(z|x)dz - \int q(z|x)\log p(z)dz \\
&= \Bigg(-\frac{J}{2}\log(2\pi) - \frac{1}{2}\sum_{j=1}^{J}(\log\sigma_j^2 + 1)\Bigg) - \Bigg(-\frac{J}{2}\log(2\pi) - \frac{1}{2}\sum_{j=1}^{J}(\sigma_j^2 + \mu_j^2)\Bigg) \\
&= -\frac{1}{2}\sum_{j=1}^{J}\Bigg(1 + \log\sigma^2 - \sigma_j^2 - \mu_j^2\Bigg)
\end{aligned}
\tag{5}
$$

2 参考

- `https://stats.stackexchange.com/questions/7440/kl-divergence-between-two-univariate-gaussians`