変分オートエンコーダ (Variational Auto-Encoder)

目次

1 多変量ベルヌーイ分布

n 次元のベクトルの各要素が 0 か 1 のみをとるとしたとき、その確率分布を多変量ベルヌーイ分布という。ここで、各要素は独立に定まるとすると、i 番目の要素が 1 となる確率を p_i として、$\mathbf{x} = (x_1, x_2, \ldots, x_n)$ が現れる確率 $p(\mathbf{x})$ は

$$p(\mathbf{x}) = \prod_{i=1}^{n} p_i^{x_i} (1 - p_i)^{1 - x_i}$$

と書くことができる。よって、対数確率は

$$\log p(\mathbf{x}) = \sum_{i=1}^{n} [x_i \log p_i + (1 - x_i) \log(1 - p_i)]$$

となる。

2　正規分布同士の KL-divergence

確率分布 $p(\mathbf{x})$ と $q(\mathbf{x})$ の間の KL-divergence は

$$D_{\mathrm{KL}}(p(\mathbf{x})||q(\mathbf{x})) = \int p(\mathbf{x}) \log \frac{q(\mathbf{x})}{p(\mathbf{x})} d\mathbf{x}$$

と定義される。KL-divergence は非負な値になることが示され、$p(\mathbf{x})$ と $q(\mathbf{x})$ が等しいときのみゼロとなることから、確率分布同士の近さをはかる汎関数であるといえる。

この $p(\mathbf{x})$ と $q(\mathbf{x})$ が正規分布に従うとして、KL-divergence を解析的に計算する。まず

$$p(\mathbf{x}) = N(\mu_1, \Sigma_1)$$
$$q(\mathbf{x}) = N(\mu_2, \Sigma_2)$$

とする。

よって

$$D_{\mathrm{KL}}(p(\mathbf{x})||q(\mathbf{x})) = \frac{1}{2}\left[\log \frac{\det \Sigma_2}{\det \Sigma_1} - n + \mathrm{Tr}(\Sigma_2^{-1}\Sigma_1) + (\mu_2 - \mu_1)^{\mathrm{T}} \Sigma_2^{-1} (\mu_2 - \mu_1)\right]$$

である。

3　前提

エンコーダのパラメータを ϕ、デコーダのパラメータを θ とする。そして、エンコーダが $\mathbf{x}$ を与えられて、潜在変数 $\mathbf{z}$ を出力する事後確率を $q_\phi(\mathbf{z} \mid \mathbf{x})$ とする。また、デコーダが $\mathbf{x}$ を出力したときに、与えられている潜在変数が $\mathbf{z}$ である事後確率を $p_\theta(\mathbf{z} \mid \mathbf{x})$ とする。

デコーダの確率分布 p_θ をエンコーダの確率分布 q_ϕ で近似することを考える。まず、ある $\mathbf{x}$ に対して p_θ と q_ϕ がどれだけ近い分布かを評価する方法として KL-divergence を用いて

$$D_{\mathrm{KL}}(q_\phi(\mathbf{z} \mid \mathbf{x})||p_\theta(\mathbf{z} \mid \mathbf{x})) = \int q_\phi(\mathbf{z} \mid \mathbf{x}) \log \frac{q_\phi(\mathbf{z} \mid \mathbf{x})}{p_\theta(\mathbf{z} \mid \mathbf{x})} d\mathbf{z}$$

とする。KL-divergence は小さければ小さいほど 2 つの確率分布が近いことを示す汎関数であるため、この値を最小化する問題へ帰着することがわかる。しかし、KL-divergence そのものを最小化するのは難しい。ここで、ベイズの定理を用いると

$$p_\theta = \frac{p_\theta(\mathbf{z}, \mathbf{x})}{p(\mathbf{x})}$$

となり、KL-divergence へ代入することで

$$\begin{aligned}
D_{\mathrm{KL}}(q_\phi(\mathbf{z}\mid\mathbf{x})||p_\theta(\mathbf{z}\mid\mathbf{x})) &= \int q_\phi(\mathbf{z}\mid\mathbf{x})\log\frac{q_\phi(\mathbf{z}\mid\mathbf{x})}{p_\theta(\mathbf{z}\mid\mathbf{x})}d\mathbf{z} \\
&= \int q_\phi(\mathbf{z}\mid\mathbf{x})\log\frac{q_\phi(\mathbf{z}\mid\mathbf{x})p(\mathbf{x})}{p_\theta(\mathbf{z},\mathbf{x})}d\mathbf{z} \\
&= \int q_\phi(\mathbf{z}\mid\mathbf{x})\log\frac{q_\phi(\mathbf{z}\mid\mathbf{x})}{p_\theta(\mathbf{z},\mathbf{x})}d\mathbf{z} + \int q_\phi(\mathbf{z}\mid\mathbf{x})\log p(\mathbf{x})d\mathbf{z} \\
&= -\int q_\phi(\mathbf{z}\mid\mathbf{x})\log\frac{p_\theta(\mathbf{z},\mathbf{x})}{q_\phi(\mathbf{z}\mid\mathbf{x})}d\mathbf{z} + \log p(\mathbf{x})\int q_\phi(\mathbf{z}\mid\mathbf{x})d\mathbf{z} \\
&= -\int q_\phi(\mathbf{z}\mid\mathbf{x})\log\frac{p_\theta(\mathbf{z},\mathbf{x})}{q_\phi(\mathbf{z}\mid\mathbf{x})}d\mathbf{z} + \log p(\mathbf{x})
\end{aligned}$$

となる。ここで $p(\mathbf{x})$ は周辺尤度である。L を

$$L[q_\phi, p_\theta, \mathbf{x}] = \int q_\phi(\mathbf{z}\mid\mathbf{x})\log\frac{p_\theta(\mathbf{z},\mathbf{x})}{q_\phi(\mathbf{z}\mid\mathbf{x})}d\mathbf{z}$$

とすることで

$$\log p(\mathbf{x}) = L[q_\phi, p_\theta, \mathbf{x}] + D_{\mathrm{KL}}(q_\phi(\mathbf{z}\mid\mathbf{x})||p_\theta(\mathbf{z}\mid\mathbf{x}))$$

が成立する。$p(\mathbf{x})$ はパラメータによらないため今考えている最適化問題においては定数である。 したがって、KL-divergece を最小化することと L を最大化することは同値となる。以降、L を最大化する最適化問題を考える。また、$L[q_\phi, p_\theta, \mathbf{x}]$ のことを変分限界、もしくは変分下限 (variational lower bound) という。

一つのデータ $\mathbf{x}$ が与えられたときの目的関数は $L[q_\phi, p_\theta, \mathbf{x}]$ である。N 個のデータ $X = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N\}$ がそれぞれ独立に与えられたとすると、X の事前確率 $p(X)$ は

$$p(X) = \sum_{i=1}^{N} p(\mathbf{x}_i)$$

と表すことができる。これをふまえると

$$\sum_{i=1}^{N}\log p(\mathbf{x}_i) = \sum_{i=1}^{N} L[q_\phi, p_\theta, \mathbf{x}_i] + D_{\mathrm{KL}}(q_\phi(\mathbf{z}\mid\mathbf{x}_i)||p_\theta(\mathbf{z}\mid\mathbf{x}_i))$$

が成立する。これに従って、N 個のデータ X が与えられたときの目的関数は

$$\sum_{\mathbf{x}\in X} L[q_\phi, p_\theta, \mathbf{x}] = \sum_{\mathbf{x}\in X}\int q_\phi(\mathbf{z}\mid\mathbf{x})\log\frac{p_\theta(\mathbf{z},\mathbf{x})}{q_\phi(\mathbf{z}\mid\mathbf{x})}d\mathbf{z}$$

とすることとする。

4　変分限界の変形

定義から変分限界は

$$\begin{aligned}
L[q_\phi, p_\theta, \mathbf{x}] &= \int q_\phi(\mathbf{z} \mid \mathbf{x}) \log \frac{p_\theta(\mathbf{z}, \mathbf{x})}{q_\phi(\mathbf{z} \mid \mathbf{x})} d\mathbf{z} \\
&= \int q_\phi(\mathbf{z} \mid \mathbf{x}) \log \frac{p(\mathbf{z}) p_\theta(\mathbf{x} \mid \mathbf{z})}{q_\phi(\mathbf{z} \mid \mathbf{x})} d\mathbf{z} \\
&= \int q_\phi(\mathbf{z} \mid \mathbf{x}) \log \frac{p(\mathbf{z})}{q_\phi(\mathbf{z} \mid \mathbf{x})} d\mathbf{z} + \int q_\phi(\mathbf{z} \mid \mathbf{x}) \log p_\theta(\mathbf{x} \mid \mathbf{z}) d\mathbf{z} \\
&= \int q_\phi(\mathbf{z} \mid \mathbf{x}) \log \frac{p(\mathbf{z})}{q_\phi(\mathbf{z} \mid \mathbf{x})} d\mathbf{z} + \int q_\phi(\mathbf{z} \mid \mathbf{x}) \log p_\theta(\mathbf{x} \mid \mathbf{z}) d\mathbf{z} \\
&= -\int q_\phi(\mathbf{z} \mid \mathbf{x}) \log \frac{q_\phi(\mathbf{z} \mid \mathbf{x})}{p(\mathbf{z})} d\mathbf{z} + \int q_\phi(\mathbf{z} \mid \mathbf{x}) \log p_\theta(\mathbf{x} \mid \mathbf{z}) d\mathbf{z}
\end{aligned}$$

と変形することができる。上式の第一項目は KL-divergece であり、第二項目は $q_\phi(\mathbf{z} \mid \mathbf{x})$ による期待値とみることができるため

$$L[q_\phi, p_\theta, \mathbf{x}] = -D_{\mathrm{KL}}(q_\phi(\mathbf{z} \mid \mathbf{x}) || p(\mathbf{x})) + \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}[\log p_\theta(\mathbf{x} \mid \mathbf{z})]$$

と書くことができる。ここで、ベイズの定理によって $p_\theta(\mathbf{z}, \mathbf{x}) = p(\mathbf{z} p_\theta(\mathbf{x} \mid \mathbf{z}))$ とした。

5　リパラメタライズ

$\mathbf{z}$ が $\mathbf{x}$ から生成される確率変数であるため、微分によって勾配を求めることが容易でない。したがって決定的な微分可能関数 g_ϕ を用いて $\mathbf{x}$ を

$$\mathbf{z} = g_\phi(\epsilon, \mathbf{x}), \quad \epsilon \sim p(\epsilon)$$

と再定義する。すると

$$\begin{aligned}
\int q_\phi(\mathbf{z} \mid \mathbf{x}) f(\mathbf{z}) d\mathbf{z} &= \int q_\phi(g_\phi(\epsilon, \mathbf{x}) \mid \mathbf{x}) f(g_\phi(\epsilon, \mathbf{x})) d\epsilon \\
&\simeq \frac{1}{L} \sum_{l=1}^{L} f(g_\phi(\epsilon^{(l)}, \mathbf{x}))
\end{aligned}$$

と近似することができる。ここで $\epsilon^{(l)} \sim p(\epsilon)$ であり、L はある程度大きな自然数である。

例えば g_ϕ の例として

$$g_\phi(\epsilon, \mu, \sigma) = \mu + \epsilon \odot \sigma, \quad \epsilon \sim N(\mathbf{0}, I)$$

とすれば $\mathbf{z} = g_\phi(\epsilon, \mu, \sigma)$ は平均ベクトルが μ、各要素が独立で分散が $\sigma \odot \sigma$ である正規分布に従う。ここで $\odot$ はベクトルの要素同士積、I は単位行列とした。

この近似によって

$$\begin{aligned}\mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}[\log p_\theta(\mathbf{x} \mid \mathbf{z})] &= \int q_\phi(\mathbf{z} \mid \mathbf{x}) \log p_\theta(\mathbf{x} \mid \mathbf{z}) d\mathbf{z} \\ &\simeq \frac{1}{L} \sum_{l=1}^{L} \log p_\theta(\mathbf{x} \mid \mathbf{z}^{(l)})\end{aligned}$$

となるため、変分限界は

$$L[q_\phi, p_\theta, \mathbf{x}] \simeq -D_{\mathrm{KL}}(q_\phi(\mathbf{z} \mid \mathbf{x}) || p(\mathbf{x})) + \frac{1}{L} \sum_{l=1}^{L} \log p_\theta(\mathbf{x} \mid \mathbf{z}^{(l)})$$

と近似できる。ここで $\mathbf{z}^{(l)} = g_\phi(\epsilon^{(l)}, \mathbf{x}), \epsilon^{(l)} \sim p(\epsilon^{(l)})$ である。

また、変分限界は

$$\begin{aligned}L[q_\phi, p_\theta, \mathbf{x}] &= \int q_\phi(\mathbf{z} \mid \mathbf{x}) \log \frac{p_\theta(\mathbf{z}, \mathbf{x})}{q_\phi(\mathbf{z} \mid \mathbf{x})} d\mathbf{z} \\ &= \int q_\phi(\mathbf{z} \mid \mathbf{x}) [\log p_\theta(\mathbf{z}, \mathbf{x}) - \log q_\phi(\mathbf{z} \mid \mathbf{x})] d\mathbf{z} \\ &= \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}[\log p_\theta(\mathbf{z}, \mathbf{x}) - \log q_\phi(\mathbf{z} \mid \mathbf{x})]\end{aligned}$$

とも変形できるため

$$L[q_\phi, p_\theta, \mathbf{x}] \simeq \frac{1}{L} \sum_{l=1}^{L} [\log p_\theta(\mathbf{x}, \mathbf{z}^{(l)}) - \log q_\phi(\mathbf{z}^{(l)} \mid \mathbf{x})]$$

という近似も成立する。したがって、KL-divergence が計算できる場合は L_A、難しい場合は L_B

$$L_A[q_\phi, p_\theta, \mathbf{x}] = -D_{\mathrm{KL}}(q_\phi(\mathbf{z} \mid \mathbf{x}) || p(\mathbf{x})) + \frac{1}{L} \sum_{l=1}^{L} \log p_\theta(\mathbf{x} \mid \mathbf{z}^{(l)})$$

$$L_B[q_\phi, p_\theta, \mathbf{x}] = \frac{1}{L} \sum_{l=1}^{L} [\log p_\theta(\mathbf{x}, \mathbf{z}^{(l)}) - \log q_\phi(\mathbf{z}^{(l)} \mid \mathbf{x})]$$

$$\mathbf{z}^{(l)} = g_\phi(\epsilon^{(l)}, \mathbf{x}), \quad \epsilon^{(l)} \sim p(\epsilon^{(l)})$$

をそれぞれ計算し変分限界を近似することができる。そして、近似した変分限界の勾配を用いることでパラメータを更新すればよい。

6 まとめ

データの集合 X が与えられたとしたときの目的関数は

$$\sum_{\mathbf{x} \in X} L[q_\phi, p_\theta, \mathbf{x}]$$

であり、L は L_A もしくは L_B

$$L_A[q_\phi, p_\theta, \mathbf{x}] = -D_{\mathrm{KL}}(q_\phi(\mathbf{z} \mid \mathbf{x}) || p(\mathbf{x})) + \frac{1}{L} \sum_{l=1}^{L} \log p_\theta(\mathbf{x} \mid \mathbf{z}^{(l)})$$

$$L_B[q_\phi, p_\theta, \mathbf{x}] = \frac{1}{L} \sum_{l=1}^{L} [\log p_\theta(\mathbf{x}, \mathbf{z}^{(l)}) - \log q_\phi(\mathbf{z}^{(l)} \mid \mathbf{x})]$$

$$\mathbf{z}^{(l)} = g_\phi(\epsilon^{(l)}, \mathbf{x}), \quad \epsilon^{(l)} \sim p(\epsilon^{(l)})$$

によって

$$\sum_{\mathbf{x} \in X} L[q_\phi, p_\theta, \mathbf{x}] \simeq \sum_{\mathbf{x} \in X} L_A[q_\phi, p_\theta, \mathbf{x}]$$

$$\sum_{\mathbf{x} \in X} L[q_\phi, p_\theta, \mathbf{x}] \simeq \sum_{\mathbf{x} \in X} L_B[q_\phi, p_\theta, \mathbf{x}]$$

と近似できる。

また、ミニバッチとして $X' \subset X$ が与えられたときの目的関数は

$$\sum_{\mathbf{x} \in X} L[q_\phi, p_\theta, \mathbf{x}] \simeq \frac{|X|}{|X'|} \sum_{\mathbf{x} \in X'} L[q_\phi, p_\theta, \mathbf{x}]$$

と近似できる。