生成モデル

EM アルゴリズム, 変分オートエンコーダ

中田和秀

東京工業大学 工学院 経営工学系

機械学習入門

http://www.iee.e.titech.ac.jp/~nakatalab/text/lecture.html

概要

ここでは、データからデータの生成モデルを構築する方法について説明をする。一般的にこの構築は難易度が高いことが知られている。しかし、適切な生成モデルを作ることができれば、様々な予測タスクやクラスタリングやデータ生成など多様な形で応用することができる。

目次：

記号の使い方：

- $A := B$ は、B で A を定義する、B を A に代入することを意味する
- $[n]$ は n までのインデックスの集合を表し $[n] := \{1, 2, \cdots, n\}$

潜在変数を含んだ確率分布と EM アルゴリズム

潜在変数を含んだ確率変数

- $\boldsymbol{z}$：潜在変数
- $\boldsymbol{x}$：確率変数

$$
\begin{aligned}
p(\boldsymbol{x}) &= \sum_{\boldsymbol{z}} p(\boldsymbol{z})p(\boldsymbol{x}|\boldsymbol{z}) && \boldsymbol{z}\text{ が離散分布の場合} \\
&= \int p(\boldsymbol{z})p(\boldsymbol{x}|\boldsymbol{z})d\boldsymbol{z} && \boldsymbol{z}\text{ が連続分布の場合}
\end{aligned}
$$

- 現実世界には、観測できない潜在的な何かに依存して確率が決まる状況は多い
- （潜在的な何かの存在はともかく）複雑な分布 $p(\boldsymbol{x})$ を単純な２つの分布 $p(\boldsymbol{z})$ と $p(\boldsymbol{x}|\boldsymbol{z})$ で表現できる

以下では、連続分布の場合で説明するが、離散分布の場合でも $\int$ が $\sum$ に変わるだけでまったく同じ議論ができる。

基本関係式

$p(\boldsymbol{x})$ はパラメタ $\boldsymbol{\theta}$ を持つ確率分布 $p(\boldsymbol{x};\boldsymbol{\theta})$ とする。

$q(\boldsymbol{z})$：潜在変数に関する任意の確率分布 [1]

$$\int q(\boldsymbol{z})\,d\boldsymbol{z} = 1,\ \ q(\boldsymbol{z}) \geq 0.$$

基本関係式

$$\log p(\boldsymbol{x};\boldsymbol{\theta}) = LB(q(\boldsymbol{z}),\boldsymbol{\theta}) + KL(q(\boldsymbol{z}), p(\boldsymbol{z}|\boldsymbol{x};\boldsymbol{\theta}))$$

ただし、 $LB(q(\boldsymbol{z}),\boldsymbol{\theta}) := \mathbb{E}_{q(\boldsymbol{z})}\left[\log \dfrac{p(\boldsymbol{x},\boldsymbol{z};\boldsymbol{\theta})}{q(\boldsymbol{z})}\right]$

$KL(q(\boldsymbol{z}), p(\boldsymbol{z}|\boldsymbol{x};\boldsymbol{\theta}))$ $\cdots\cdots$ KL ダイバージェンス

※ $KL(q(\boldsymbol{z}), p(\boldsymbol{z}|\boldsymbol{x};\boldsymbol{\theta})) \geq 0$ より、 $\log p(\boldsymbol{x};\boldsymbol{\theta}) \geq LB(q(\boldsymbol{z}),\boldsymbol{\theta})$.

つまり、$LB(q(\boldsymbol{z}),\boldsymbol{\theta})$ は、$\log p(\boldsymbol{x};\boldsymbol{\theta})$ の下界 (Lower Bound)

[1] 真の確率分布ではないことに注意

式変形

$$
\begin{aligned}
&LB(q(\boldsymbol{z}), \boldsymbol{\theta}) + KL(q(\boldsymbol{z}), p(\boldsymbol{z}|\boldsymbol{x}; \boldsymbol{\theta})) \\
&= \mathbb{E}_{q(\boldsymbol{z})}\left[\log \frac{p(\boldsymbol{x}, \boldsymbol{z}; \boldsymbol{\theta})}{q(\boldsymbol{z})}\right] + \int q(\boldsymbol{z}) \log \frac{q(\boldsymbol{z})}{p(\boldsymbol{z}|\boldsymbol{x}; \boldsymbol{\theta})} \, d\boldsymbol{z} \\
&= \int q(\boldsymbol{z}) \log \frac{p(\boldsymbol{x}, \boldsymbol{z}; \boldsymbol{\theta})}{q(\boldsymbol{z})} \, d\boldsymbol{z} + \int q(\boldsymbol{z}) \log \frac{q(\boldsymbol{z})}{p(\boldsymbol{z}|\boldsymbol{x}; \boldsymbol{\theta})} \, d\boldsymbol{z} \\
&= \int q(\boldsymbol{z}) \left(\log \frac{p(\boldsymbol{x}, \boldsymbol{z}; \boldsymbol{\theta})}{q(\boldsymbol{z})} + \log \frac{q(\boldsymbol{z})}{p(\boldsymbol{z}|\boldsymbol{x}; \boldsymbol{\theta})}\right) d\boldsymbol{z} \\
&= \int q(\boldsymbol{z}) \log \frac{p(\boldsymbol{x}, \boldsymbol{z}; \boldsymbol{\theta})}{p(\boldsymbol{z}|\boldsymbol{x}; \boldsymbol{\theta})} \, d\boldsymbol{z} \\
&= \int q(\boldsymbol{z}) \log \frac{p(\boldsymbol{x}; \boldsymbol{\theta}) p(\boldsymbol{z}|\boldsymbol{x}; \boldsymbol{\theta})}{p(\boldsymbol{z}|\boldsymbol{x}; \boldsymbol{\theta})} \, d\boldsymbol{z} \\
&= \int q(\boldsymbol{z}) \log p(\boldsymbol{x}; \boldsymbol{\theta}) \, d\boldsymbol{z} \\
&= \log p(\boldsymbol{x}; \boldsymbol{\theta}) \int q(\boldsymbol{z}) \, d\boldsymbol{z} \\
&= \log p(\boldsymbol{x}; \boldsymbol{\theta})
\end{aligned}
$$

LB の性質

$LB(q(\boldsymbol{z}), \boldsymbol{\theta})$ に関しては次の 2 つの性質を使うので、先に説明しておく。

LB の性質 1

$$LB(q(\boldsymbol{z}), \boldsymbol{\theta}) = \mathbb{E}_{q(\boldsymbol{z})}\left[\log p(\boldsymbol{x}, \boldsymbol{z}; \boldsymbol{\theta})\right] + H(q(\boldsymbol{z}))$$

ただし、$H(q(\boldsymbol{z})) := -\int q(\boldsymbol{z}) \log q(\boldsymbol{z})\ d\boldsymbol{z}$ エントロピー

$\boldsymbol{\theta} = \{\boldsymbol{\theta}_1, \boldsymbol{\theta}_2\}$ で、$p(\boldsymbol{x}, \boldsymbol{z}; \boldsymbol{\theta}) = p(\boldsymbol{z}; \boldsymbol{\theta}_1)\, p(\boldsymbol{x}|\boldsymbol{z}; \boldsymbol{\theta}_2)$ のとき、次の性質が成り立つ。

LB の性質 2

$$LB(q(\boldsymbol{z}), \boldsymbol{\theta}) = \mathbb{E}_{q(\boldsymbol{z})}\left[\log p(\boldsymbol{x}|\boldsymbol{z}; \boldsymbol{\theta}_2)\right] - KL(q(\boldsymbol{z}), p(\boldsymbol{z}; \boldsymbol{\theta}_1))$$

式変形

$$
\begin{aligned}
LB(q(\boldsymbol{z}), \boldsymbol{\theta}) &= \mathbb{E}_{q(\boldsymbol{z})}\left[\log \frac{p(\boldsymbol{x}, \boldsymbol{z}; \boldsymbol{\theta})}{q(\boldsymbol{z})}\right] = \int q(\boldsymbol{z}) \log \frac{p(\boldsymbol{x}, \boldsymbol{z}; \boldsymbol{\theta})}{q(\boldsymbol{z})}\, d\boldsymbol{z} \\
&= \int q(\boldsymbol{z}) \log p(\boldsymbol{x}, \boldsymbol{z}; \boldsymbol{\theta})\, d\boldsymbol{z} - \int q(\boldsymbol{z}) \log q(\boldsymbol{z})\, d\boldsymbol{z} \\
&= \mathbb{E}_{q(\boldsymbol{z})}\left[\log p(\boldsymbol{x}, \boldsymbol{z}; \boldsymbol{\theta})\right] + H(q(\boldsymbol{z}))
\end{aligned}
$$

$$
\begin{aligned}
LB(q(\boldsymbol{z}), \boldsymbol{\theta}) &= \mathbb{E}_{q(\boldsymbol{z})}\left[\log \frac{p(\boldsymbol{x}, \boldsymbol{z}; \boldsymbol{\theta})}{q(\boldsymbol{z})}\right] = \int q(\boldsymbol{z}) \log \frac{p(\boldsymbol{x}, \boldsymbol{z}; \boldsymbol{\theta})}{q(\boldsymbol{z})}\, d\boldsymbol{z} \\
&= \int q(\boldsymbol{z}) \log \frac{p(\boldsymbol{z}; \boldsymbol{\theta}_1)\, p(\boldsymbol{x} | \boldsymbol{z}; \boldsymbol{\theta}_2)}{q(\boldsymbol{z})}\, d\boldsymbol{z} \\
&= \int q(\boldsymbol{z}) \log p(\boldsymbol{x} | \boldsymbol{z}; \boldsymbol{\theta}_2)\, d\boldsymbol{z} - \int q(\boldsymbol{z}) \log \frac{q(\boldsymbol{z})}{p(\boldsymbol{z}; \boldsymbol{\theta}_1)}\, d\boldsymbol{z} \\
&= \mathbb{E}_{q(\boldsymbol{z})}\left[\log p(\boldsymbol{x} | \boldsymbol{z}; \boldsymbol{\theta}_2)\right] - KL(q(\boldsymbol{z}), p(\boldsymbol{z}; \boldsymbol{\theta}_1))
\end{aligned}
$$

最尤推定

- 独立にサンプリングされたデータ　$\mathcal{D} := \{\boldsymbol{x}_d\}_{d \in [D]}$
- 尤度：$p(\mathcal{D}) = \prod_{d \in [D]} p(\boldsymbol{x}_d; \boldsymbol{\theta})$

対数尤度の最大化

$$\max_{\boldsymbol{\theta}} \sum_{d \in [D]} \log p(\boldsymbol{x}_d; \boldsymbol{\theta})$$

$$\begin{aligned}
\sum_{d \in [D]} \log p(\boldsymbol{x}_d; \boldsymbol{\theta}) &= \sum_{d \in [D]} LB_d(q_d(\boldsymbol{z}), \boldsymbol{\theta}) + KL(q_d(\boldsymbol{z}), p(\boldsymbol{z} | \boldsymbol{x}_d; \boldsymbol{\theta})) \\
&\geq \sum_{d \in [D]} LB_d(q_d(\boldsymbol{z}), \boldsymbol{\theta}) \qquad (\because KL(q_d(\boldsymbol{z}), p(\boldsymbol{z} | \boldsymbol{x}_d; \boldsymbol{\theta})) \geq 0)
\end{aligned}$$

EM アルゴリズムの導出

(1) $\sum_{d\in[D]} \log p(\boldsymbol{x}_d;\boldsymbol{\theta})$ の代わりに下界 $\sum_{d\in[D]} LB_d(q(\boldsymbol{z});\boldsymbol{\theta})$ の最大化

$$\rightarrow \qquad \max_{\boldsymbol{\theta}} \sum_{d\in[D]} \log p(\boldsymbol{x}_d;\boldsymbol{\theta}) \implies \max_{\boldsymbol{\theta}} \sum_{d\in[D]} LB_d(q_d(\boldsymbol{z});\boldsymbol{\theta})$$

(2) できるだけタイトな下界となるように $q_d(\boldsymbol{z})$ を決めたい。

$q_d(\boldsymbol{z}) := p(\boldsymbol{z}|\boldsymbol{x}_d;\boldsymbol{\theta})$ のとき、$KL(q_d(\boldsymbol{z}), p(\boldsymbol{z}|\boldsymbol{x}_d;\boldsymbol{\theta})) = 0$ より、

$$\sum_{d\in[D]} \log p(\boldsymbol{x}_d;\boldsymbol{\theta}) = \sum_{d\in[D]} LB_d(q(\boldsymbol{z});\boldsymbol{\theta})$$

一番よい下界となっている。

EM アルゴリズム 1

EM アルゴリズム 1

ステップ 0　初期点 $\boldsymbol{\theta}$ を定める。

ステップ 1　$q_d(\boldsymbol{z}) := p(\boldsymbol{z}|\boldsymbol{x}_d;\boldsymbol{\theta})$　　$d \in [D]$

ステップ 2　$\boldsymbol{\theta} := \operatorname*{argmax}_{\boldsymbol{\theta}} \sum_{d\in[D]} LB_d(q_d(\boldsymbol{z}),\boldsymbol{\theta})$

ステップ 3　終了条件を満たしていなければ、ステップ 1 に戻る。

前の反復点を $\widetilde{\boldsymbol{\theta}}$ とすると、$q_d(\boldsymbol{z}) := p(\boldsymbol{z}|\boldsymbol{x}_d;\widetilde{\boldsymbol{\theta}})$ である。LB の性質 1 より、

$$LB_d(p(\boldsymbol{z}|\boldsymbol{x}_d;\widetilde{\boldsymbol{\theta}}),\boldsymbol{\theta}) = \mathbb{E}_{p(\boldsymbol{z}|\boldsymbol{x}_d;\widetilde{\boldsymbol{\theta}})}[\log p(\boldsymbol{x}_d,\boldsymbol{z};\boldsymbol{\theta})] + H(p(\boldsymbol{z}|\boldsymbol{x}_d;\widetilde{\boldsymbol{\theta}}))$$

よって、

$$\max_{\boldsymbol{\theta}} \sum_{d\in[D]} LB_d(p(\boldsymbol{z}|\boldsymbol{x}_d;\widetilde{\boldsymbol{\theta}}),\boldsymbol{\theta}) \iff \max_{\boldsymbol{\theta}} \sum_{d\in[D]} \mathbb{E}_{p(\boldsymbol{z}|\boldsymbol{x}_d;\widetilde{\boldsymbol{\theta}})}[\log p(\boldsymbol{x}_d,\boldsymbol{z};\boldsymbol{\theta})]$$

※ $H(p(\boldsymbol{z}|\boldsymbol{x}_d;\widetilde{\boldsymbol{\theta}}))$ は最適化に関係ない

EM アルゴリズム 2

EM アルゴリズムは次のように書くことができる
（この形で紹介されることが多い）

EM アルゴリズム 2

ステップ 0　初期点 $\widetilde{\boldsymbol{\theta}}$ を定める。

ステップ 1　$p(\boldsymbol{z}|\boldsymbol{x}_d;\widetilde{\boldsymbol{\theta}})$ の計算　$d \in [D]$

ステップ 2　$\boldsymbol{\theta} := \operatorname*{argmax}_{\boldsymbol{\theta}} Q(\widetilde{\boldsymbol{\theta}}, \boldsymbol{\theta})$

ステップ 3　終了条件を満たしていなければ、$\widetilde{\boldsymbol{\theta}} := \boldsymbol{\theta}$ としてステップ 1 に戻る。

ただし、$Q(\widetilde{\boldsymbol{\theta}}, \boldsymbol{\theta}) := \sum_{d\in[D]} \mathbb{E}_{p(\boldsymbol{z}|\boldsymbol{x}_d;\tilde{\boldsymbol{\theta}})}[\log p(\boldsymbol{x}_d, \boldsymbol{z}; \boldsymbol{\theta})]$

混合モデルの場合

$\boldsymbol{z}$ が（有限）離散分布

- 潜在変数 $\boldsymbol{z}$ の分布： k 番目の分布を取る確率 π_k　　カテゴリ分布
- k 番目の分布：$p_k(\boldsymbol{x};\boldsymbol{\theta}_k)$

- EM アルゴリズム１のステップ１：

$$q_d(\boldsymbol{z}) := p(\boldsymbol{z}|\boldsymbol{x}_d;\boldsymbol{\theta}) = \frac{p(\boldsymbol{z};\boldsymbol{\theta})\,p(\boldsymbol{x}|\boldsymbol{z};\boldsymbol{\theta})}{\sum_{\boldsymbol{z}'} p(\boldsymbol{z}';\boldsymbol{\theta})\,p(\boldsymbol{x}_d|\boldsymbol{z}';\boldsymbol{\theta})}$$

$d \in [D]$ に対し、$\boldsymbol{x}_d$ が k 番目の分布に属する確率は

$$r_{dk} = \frac{\pi_k\, p_k(\boldsymbol{x}_d;\boldsymbol{\theta}_k)}{\sum_{k'} \pi_{k'}\, p_{k'}(\boldsymbol{x}_d;\boldsymbol{\theta}_{k'})} \qquad (k \in [K])$$

- EM アルゴリズム１のステップ２：

$$\sum_{d\in[D]} LB_d(q_d(\boldsymbol{z}),\boldsymbol{\theta}) := \sum_{d\in[D]} \mathbb{E}_{q_d(\boldsymbol{z})}\left[\log \frac{p(\boldsymbol{x}_d,\boldsymbol{z};\boldsymbol{\theta})}{q_d(\boldsymbol{z})}\right]$$

$$= \sum_{d\in[D]} \sum_{k\in[K]} r_{dk} \log \frac{\pi_k\, p_k(\boldsymbol{x}_d;\boldsymbol{\theta}_k)}{r_{dk}}$$

混合モデルに対する EM アルゴリズム

補助関数法と同様に $\displaystyle\sum_{d\in[D]}\sum_{k\in[K]} r_{dk}\log\frac{\pi_k\, p_k(\boldsymbol{x}_d;\boldsymbol{\theta}_k)}{r_{dk}}$ の最適化は分割できる。

混合モデルに対する EM アルゴリズム

ステップ 0 初期点 $\boldsymbol{\pi},\boldsymbol{\theta}_k\ (k\in[K])$ を決める。

ステップ 1 $r_{dk} := \dfrac{\pi_k\, p(\boldsymbol{x}_d;\boldsymbol{\theta}_k)}{\sum_{k'\in[K]}\pi_{k'}\, p(\boldsymbol{x}_d;\boldsymbol{\theta}_{k'})}$ $\quad d\in[D],\ k\in[K]$

ステップ 2

$$\pi_k := \frac{1}{D}\sum_{d\in[D]} r_{dk}\quad (k\in[K])$$

$k\in[K]$ に対して、重み付き最尤推定を行う。

$$\max_{\boldsymbol{\theta}_k}\sum_{d\in[D]} r_{dk}\log p_k(\boldsymbol{x}_d;\boldsymbol{\theta}_k)$$

ステップ 3 終了条件を満たしていなければ、ステップ 1 に戻る

潜在変数が離散でカテゴリ分布のとき、EM アルゴリズムは補助関数法と等価

変分オートエンコーダー

潜在変数が離散分布でなく連続分布の場合を考える。

対数最尤の最大化

$$\max_{\boldsymbol{\theta}} \sum_{d \in [D]} \log p(\boldsymbol{x}_d; \boldsymbol{\theta})$$

$$p(\boldsymbol{x}_d; \boldsymbol{\theta}) = \int p(\boldsymbol{z}; \boldsymbol{\theta})\, p(\boldsymbol{x}_d | \boldsymbol{z}; \boldsymbol{\theta})\, d\boldsymbol{z}$$

EM アルゴリズムを適用する場合は次の計算をする

EM アルゴリズム１（再掲）

ステップ０　適当に $\boldsymbol{\theta}$ を定める。

ステップ１　$q_d(\boldsymbol{z}) := p(\boldsymbol{z} | \boldsymbol{x}_d; \boldsymbol{\theta})$　　$d \in [D]$

ステップ２　$\boldsymbol{\theta} := \underset{\boldsymbol{\theta}}{\operatorname{argmax}} \sum_{d \in [D]} LB_d(q_d(\boldsymbol{z}), \boldsymbol{\theta})$

ステップ３　終了条件を満たさなければ、ステップ１に戻る

枠組み 1

ステップ１の計算が困難

- 分母の積分が計算できない
- $\boldsymbol{z}$ の取れる値が無限にある

$$q_d(\boldsymbol{z}) = p(\boldsymbol{z}|\boldsymbol{x}_d;\boldsymbol{\theta}) = \frac{p(\boldsymbol{x}_d,\boldsymbol{z};\boldsymbol{\theta})}{p(\boldsymbol{x}_d;\boldsymbol{\theta})} = \frac{p(\boldsymbol{z};\boldsymbol{\theta})\,p(\boldsymbol{x}_d|\boldsymbol{z};\boldsymbol{\theta})}{\int p(\boldsymbol{z};\boldsymbol{\theta})\,p(\boldsymbol{x}_d|\boldsymbol{z};\boldsymbol{\theta})\,d\boldsymbol{z}}$$

$\Longrightarrow$ $p(\boldsymbol{z}|\boldsymbol{x}_d;\boldsymbol{\theta}) \simeq q(\boldsymbol{z}|\boldsymbol{x}_d;\boldsymbol{\phi})$ となる分布を $q_d(\boldsymbol{z})$ とする

$\boldsymbol{x}_d$ から $\boldsymbol{z}$ の分布を作り、その関係性をパラメタ ϕ で表す。

- ２つの分布の近さ・遠さは KL ダイバージェンスで考える
- $KL(q(\boldsymbol{z}|\boldsymbol{x}_d;\boldsymbol{\phi}), p(\boldsymbol{z}|\boldsymbol{x}_d;\boldsymbol{\theta}))$ が最小となる分布 $q(\boldsymbol{z}|\boldsymbol{x}_d;\boldsymbol{\phi})$ を探す。

枠組み 2

$$\min_{\boldsymbol{\phi}} \; KL(q(\boldsymbol{z}|\boldsymbol{x}_d;\boldsymbol{\phi}), p(\boldsymbol{z}|\boldsymbol{x}_d;\boldsymbol{\theta}))$$

基本関係式：

$$\log p(\boldsymbol{x}_d;\boldsymbol{\theta}) = LB_d(q(\boldsymbol{z}|\boldsymbol{x}_d;\boldsymbol{\phi}),\boldsymbol{\theta}) + KL(q(\boldsymbol{z}|\boldsymbol{x}_d;\boldsymbol{\phi}), p(\boldsymbol{z}|\boldsymbol{x}_d;\boldsymbol{\theta}))$$

左辺は ϕ に依存しない。
よって、KL ダイバージェンスの最小化は次の最大化と同等。

$$\max_{\boldsymbol{\phi}} \; \sum_{d\in[D]} LB_d(q(\boldsymbol{z}|\boldsymbol{x}_d;\boldsymbol{\phi}),\boldsymbol{\theta})$$

最適化問題

EM アルゴリズム

ステップ 1

$$q_d(\boldsymbol{z}) := p(\boldsymbol{z}|\boldsymbol{x}_d;\boldsymbol{\theta}) \implies \max_{\boldsymbol{\phi}} \sum_{d\in[D]} LB_d(q(\boldsymbol{z}|\boldsymbol{x}_d;\boldsymbol{\phi}),\boldsymbol{\theta})$$

ステップ 2

$$\max_{\boldsymbol{\theta}} \sum_{d\in[D]} LB_d(q(\boldsymbol{z}|\boldsymbol{x}_d;\boldsymbol{\phi}),\boldsymbol{\theta})$$

これらをまとめて、次の最適化問題を解くことになる。

最適化問題

$$\max_{\boldsymbol{\theta},\boldsymbol{\phi}} \sum_{d\in[D]} LB_d(q(\boldsymbol{z}|\boldsymbol{x}_d;\boldsymbol{\phi}),\boldsymbol{\theta})$$

変分オートエンコーダー

オートエンコーダー

$$\boldsymbol{x} \xrightarrow{q(\boldsymbol{z}|\boldsymbol{x};\boldsymbol{\phi})} \boldsymbol{z} \xrightarrow{p(\boldsymbol{x}|\boldsymbol{z};\boldsymbol{\theta})} \boldsymbol{x}$$

エンコーダ　デコーダ

$\boldsymbol{x}$ を復元できるような、エンコーダとデコーダのセットを構築すると捉えることができる。

変分オートエンコーダー（Variational Auto-Encoder; VAE）と呼ぶ

必要な分布

- $p(\boldsymbol{z};\boldsymbol{\theta})$ と $p(\boldsymbol{x}|\boldsymbol{z};\boldsymbol{\theta})$
  $$\boldsymbol{p}(\boldsymbol{x};\boldsymbol{\theta}) = \int p(\boldsymbol{x},\boldsymbol{z};\boldsymbol{\theta})\,d\boldsymbol{z} = \int p(\boldsymbol{z};\boldsymbol{\theta})\,p(\boldsymbol{x}|\boldsymbol{z};\boldsymbol{\theta})\,d\boldsymbol{z}$$
- $p(\boldsymbol{z}|\boldsymbol{x};\boldsymbol{\phi})$

デコーダーの設計

$p(\boldsymbol{x}, \boldsymbol{z}; \boldsymbol{\theta}) = p(\boldsymbol{z}; \boldsymbol{\theta})\, p(\boldsymbol{x}|\boldsymbol{z}; \boldsymbol{\theta})$ を次のように定める。

(1)–A $p(\boldsymbol{z}; \boldsymbol{\theta})$

$\boldsymbol{z} \sim N(\mathbf{0}, \boldsymbol{I})$ とする。つまりパラメタ $\boldsymbol{\theta}$ に依存させない。

(1)–B $p(\boldsymbol{x}|\boldsymbol{z}; \boldsymbol{\theta})$

- 深層学習によって、$\boldsymbol{z}$ から正規分布の平均ベクトルを生成

$$
\begin{aligned}
&F_\mu : \mathbb{R}^k \to \mathbb{R}^n, \\
&\boldsymbol{\mu}_D := F_\mu(\boldsymbol{z}; \boldsymbol{\theta})
\end{aligned}
$$

  ※ F_μ は深層学習の回帰関数で $\boldsymbol{\theta}$ はパラメタ

- $\boldsymbol{x}$ は正規分布 $N(\boldsymbol{\mu}_D, \boldsymbol{I})$ に従う。

エンコーダーの設計

$p(\boldsymbol{z}|\boldsymbol{x};\boldsymbol{\theta})$ を次のように定める。

(2) $p(\boldsymbol{z}|\boldsymbol{x};\boldsymbol{\phi})$

- 深層学習によって、$\boldsymbol{x}$ から正規分布の平均ベクトルを生成

$$G_\mu : \mathbb{R}^n \to \mathbb{R}^k,$$
$$\boldsymbol{\mu}_E := G_\mu(\boldsymbol{x};\boldsymbol{\phi})$$

※ G_μ は深層学習の回帰関数で ϕ はパラメタ

- $\boldsymbol{z}$ は正規分布 $N(\boldsymbol{\mu}_E, \boldsymbol{I})$ に従う。

注意

エンコーダとデコーダの分散共分散行列として対角行列を考え、その対角成分も深層学習で学習することが多い。
本スライドでは、話を簡単にするため単位行列で固定する場合を考える。対角成分を考える場合でも、式が複雑になるだけで、以下の流れは同じ。

目的関数

正規分布の代入

$d \in [D]$ において

$$LB_d(q(\boldsymbol{z}|\boldsymbol{x}_d;\boldsymbol{\phi}),\boldsymbol{\theta}) = -\frac{n}{2}\log 2\pi + LB_1 + LB_2$$

$$LB_1 := -\frac{1}{2}\mathbb{E}_{\boldsymbol{\epsilon}\sim N(\mathbf{0},\boldsymbol{I})}\left[(\boldsymbol{x}_d - \boldsymbol{\mu}_D)^T(\boldsymbol{x}_d - \boldsymbol{\mu}_D)\right]$$

$$LB_2 := -\frac{1}{2}\boldsymbol{\mu}_E^T\boldsymbol{\mu}_E$$

ただし、 $\boldsymbol{\mu}_E := G_\mu(\boldsymbol{x}_d;\boldsymbol{\phi}), \quad \boldsymbol{z} := \boldsymbol{\mu}_E + \boldsymbol{\epsilon}, \quad \boldsymbol{\mu}_D := F_\mu(\boldsymbol{z};\boldsymbol{\theta})$

式変形１

LB に関する性質２より、

$$LB_d(q(\boldsymbol{z}|\boldsymbol{x}_d;\boldsymbol{\phi}),\boldsymbol{\theta}) = \mathbb{E}_{q(\boldsymbol{z}|\boldsymbol{x}_d;\boldsymbol{\phi})}[\log p(\boldsymbol{x}|\boldsymbol{z};\boldsymbol{\theta}_2)] - KL(q(\boldsymbol{z}|\boldsymbol{x}_d;\boldsymbol{\phi}), p(\boldsymbol{z};\boldsymbol{\theta}_1))$$

第１項と第２項で分けて考える。

$\log p(\boldsymbol{x}|\boldsymbol{z};\boldsymbol{\theta}) = -\frac{n}{2}\log 2\pi - \frac{1}{2}(\boldsymbol{x}_d - \boldsymbol{\mu}_D)^T(\boldsymbol{x}_d - \boldsymbol{\mu}_D)$ より、

$$\begin{aligned}
&\mathbb{E}_{q(\boldsymbol{z}|\boldsymbol{x}_d;\boldsymbol{\phi})}[\log p(\boldsymbol{x}|\boldsymbol{z};\boldsymbol{\theta})] \\
&= -\frac{n}{2}\log 2\pi - \mathbb{E}_{\boldsymbol{z}\sim N(\boldsymbol{\mu}_E,\boldsymbol{I})}\left[\frac{1}{2}(\boldsymbol{x}_d - \boldsymbol{\mu}_D)^T(\boldsymbol{x}_d - \boldsymbol{\mu}_D)\right]
\end{aligned}$$

ただし、$\boldsymbol{\mu}_D := F_\mu(\boldsymbol{z};\boldsymbol{\theta})$

$$= -\frac{n}{2}\log 2\pi - \mathbb{E}_{\boldsymbol{\epsilon}\sim N(\mathbf{0},\boldsymbol{I})}\left[\frac{1}{2}(\boldsymbol{x}_d - \boldsymbol{\mu}_D)^T(\boldsymbol{x}_d - \boldsymbol{\mu}_D)\right]$$

ただし、$\boldsymbol{z} = \boldsymbol{\mu}_E + \boldsymbol{\epsilon}$,　$\boldsymbol{\mu}_D := F_\mu(\boldsymbol{z};\boldsymbol{\theta})$

式変形 2

$$
\begin{aligned}
& -KL(q(\boldsymbol{z}|\boldsymbol{x}_d;\boldsymbol{\phi}), p(\boldsymbol{z};\boldsymbol{\theta}_1)) \\
&= -\mathbb{E}_{q(\boldsymbol{z}|\boldsymbol{x}_d;\boldsymbol{\phi})}\left[\log \frac{q(\boldsymbol{z}|\boldsymbol{x}_d;\boldsymbol{\phi})}{p(\boldsymbol{z})}\right] \\
&= \frac{1}{2}\mathbb{E}_{\boldsymbol{z}}\left[(\boldsymbol{z}-\boldsymbol{\mu}_E)^T(\boldsymbol{z}-\boldsymbol{\mu}_E) - \boldsymbol{z}^T\boldsymbol{z}\right]
\end{aligned}
$$

ただし　$\boldsymbol{z} \sim N(\boldsymbol{\mu}_E, \boldsymbol{I})$

$$
\begin{aligned}
&= \frac{1}{2}\mathbb{E}_{\boldsymbol{z}}\left[-2\boldsymbol{\mu}_E^T\boldsymbol{z} + \boldsymbol{\mu}_E\boldsymbol{\mu}_E^T\right] \\
&= \frac{1}{2}\left(-2\boldsymbol{\mu}_E^T\mathbb{E}_{\boldsymbol{z}}\left[\boldsymbol{z}\right] + \boldsymbol{\mu}_E^T\boldsymbol{\mu}_E\right) \\
&= \frac{1}{2}\left(-2\boldsymbol{\mu}_E^T\boldsymbol{\mu}_E + \boldsymbol{\mu}_E^T\boldsymbol{\mu}_E\right) \\
&= -\frac{1}{2}\boldsymbol{\mu}_E^T\boldsymbol{\mu}_E
\end{aligned}
$$

確率的勾配降下法

確率的勾配降下法

ステップ０　初期点 $\boldsymbol{\theta}$, $\boldsymbol{\phi}$ を決める。

ステップ１　バッチ $S \subset [D]$ を作る

ステップ２　探索方向の計算

$$\boldsymbol{d}_\theta := \frac{1}{|S|}\sum_{d\in S}\nabla_{\boldsymbol{\theta}} LB_d(q(\boldsymbol{z}|\boldsymbol{x}_d;\boldsymbol{\phi}),\boldsymbol{\theta})$$

$$\boldsymbol{d}_\phi := \frac{1}{|S|}\sum_{d\in S}\nabla_{\boldsymbol{\phi}} LB_d(q(\boldsymbol{z}|\boldsymbol{x}_d;\boldsymbol{\phi}),\boldsymbol{\theta})$$

ステップ３　ステップサイズ α を計算

ステップ４　パラメタの更新

$$\boldsymbol{\theta}_{k+1} := \boldsymbol{\theta}_k - \alpha\boldsymbol{d}_\theta$$

$$\boldsymbol{\phi}_{k+1} := \boldsymbol{\phi}_k - \alpha\boldsymbol{d}_\phi$$

ステップ５　終了条件を満たしていなければ、ステップ１に戻る

以下では、$\nabla_{\boldsymbol{\theta}} LB_d(q(\boldsymbol{z}|\boldsymbol{x}_d;\boldsymbol{\phi}),\boldsymbol{\theta})$ と $\nabla_{\boldsymbol{\phi}} LB_d(q(\boldsymbol{z}|\boldsymbol{x}_d;\boldsymbol{\phi}),\boldsymbol{\theta})$ の計算法を考える

モンテカルロ法

$LB_1 := \mathbb{E}_{\boldsymbol{\epsilon}\sim N(\mathbf{0},\boldsymbol{I})}\left[-\frac{1}{2}(\boldsymbol{x}_d-\boldsymbol{\mu}_D)^T(\boldsymbol{x}_d-\boldsymbol{\mu}_D)\right]$ は複雑な分布の期待値計算

$\rightarrow$ モンテカルロ法で近似計算を行う

$$LB_1 = \mathbb{E}_{\boldsymbol{\epsilon}\sim N(\mathbf{0},\boldsymbol{I})}\left[-\frac{1}{2}(\boldsymbol{x}_d-\boldsymbol{\mu}_D)^T(\boldsymbol{x}_d-\boldsymbol{\mu}_D)\right]$$

ただし、$\boldsymbol{\mu}_E := G_\mu(\boldsymbol{x}_d;\boldsymbol{\phi})$, $\boldsymbol{z} = \boldsymbol{\mu}_E + \boldsymbol{\epsilon}$, $\boldsymbol{\mu}_D := F_\mu(\boldsymbol{z};\boldsymbol{\theta})$

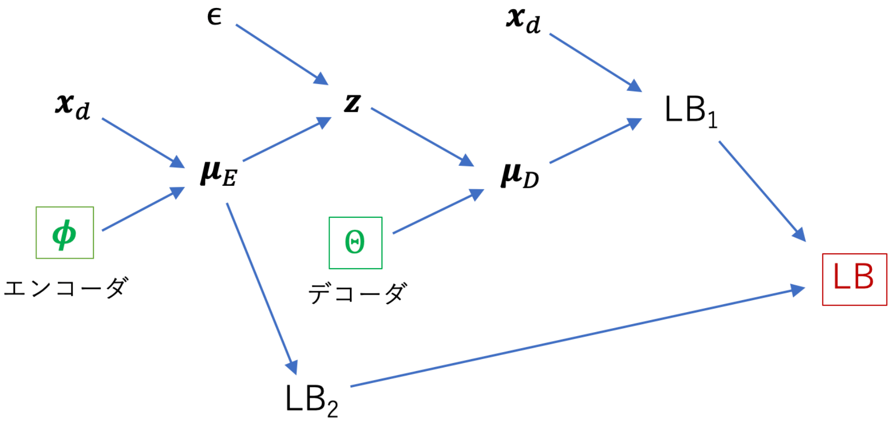
$$\simeq -\frac{1}{2C}\sum_{c\in[C]}(\boldsymbol{x}_d-\boldsymbol{\mu}_{Dc})^T(\boldsymbol{x}_d-\boldsymbol{\mu}_{Dc})$$

ただし、$\boldsymbol{\mu}_E := G_\mu(\boldsymbol{x}_d;\boldsymbol{\phi})$, $\boldsymbol{z}_c = \boldsymbol{\mu}_E + \boldsymbol{\epsilon}_c$, $\boldsymbol{\mu}_{Dc} := F_\mu(\boldsymbol{z}_c;\boldsymbol{\theta})$

$\boldsymbol{\epsilon}_c$ は $N(\mathbf{0},\boldsymbol{I})$ からのランダムサンプリング

以下では、$LB_{1c} := -\frac{1}{2}(\boldsymbol{x}_d-\boldsymbol{\mu}_{Dc})^T(\boldsymbol{x}_d-\boldsymbol{\mu}_{Dc})$ について考える。

変数の依存関係



デコーダーの勾配

$$\frac{\partial LB_d(q(\boldsymbol{z}|\boldsymbol{x}_d;\boldsymbol{\phi}),\boldsymbol{\theta})}{\partial\boldsymbol{\theta}} = \frac{\partial LB_1}{\partial\boldsymbol{\theta}} + \frac{\partial LB_2}{\partial\boldsymbol{\theta}}$$
$$\simeq \frac{1}{C}\sum_{c\in[C]}\frac{\partial LB_{1c}}{\partial\boldsymbol{\theta}}$$

$c\in[C]$ に対して、 $\dfrac{\partial LB_{1c}}{\partial\boldsymbol{\theta}} = \dfrac{\partial LB_{1c}}{\partial\boldsymbol{\mu}_{Dc}}\dfrac{\partial\boldsymbol{\mu}_{Dc}}{\partial\boldsymbol{\theta}}$

- $\dfrac{\partial LB_{1c}}{\partial\boldsymbol{\mu}_{Dc}} = (\boldsymbol{x}_d - \boldsymbol{\mu}_{Dc})^T$
- $\dfrac{\partial\boldsymbol{\mu}_{Dc}}{\partial\boldsymbol{\theta}} := \dfrac{\partial F_\mu}{\partial\boldsymbol{\theta}}$ は深層学習の誤差逆伝播法で効率よく計算できる

エンコーダの勾配

$$\frac{\partial LB_d(q(\boldsymbol{z}|\boldsymbol{x}_d;\boldsymbol{\phi}),\boldsymbol{\theta})}{\partial\boldsymbol{\phi}} = \frac{\partial LB_1}{\partial\boldsymbol{\phi}} + \frac{\partial LB_2}{\partial\boldsymbol{\phi}} \simeq \frac{1}{C}\sum_{c\in[C]}\frac{\partial LB_{1c}}{\partial\boldsymbol{\phi}} + \frac{\partial LB_2}{\partial\boldsymbol{\phi}}$$

第１項： $\dfrac{\partial LB_{1c}}{\partial\boldsymbol{\phi}} = \dfrac{\partial LB_{1c}}{\partial\boldsymbol{\mu}_{Dc}}\dfrac{\partial\boldsymbol{\mu}_{Dc}}{\partial\boldsymbol{z}_c}\dfrac{\partial\boldsymbol{z}_c}{\partial\boldsymbol{\mu}_E}\dfrac{\partial\boldsymbol{\mu}_E}{\partial\boldsymbol{\phi}}$ について

$$\frac{\partial LB_{1c}}{\partial\boldsymbol{\mu}_{Dc}} = (\boldsymbol{x}_d - \boldsymbol{\mu}_{Dc})^T$$

$\dfrac{\partial\boldsymbol{\mu}_{Dc}}{\partial\boldsymbol{z}_c} = \dfrac{\partial F_\mu}{\partial\boldsymbol{z}_c}$ 誤差逆伝播で計算

$$\frac{\partial\boldsymbol{z}_c}{\partial\boldsymbol{\mu}_E} = \boldsymbol{I}$$

$\dfrac{\partial\boldsymbol{\mu}_E}{\partial\boldsymbol{\phi}} = \dfrac{\partial G_\mu}{\partial\boldsymbol{\phi}}$ 誤差逆伝播で計算

第２項： $\dfrac{\partial LB_2}{\partial\boldsymbol{\phi}} = \dfrac{\partial LB_2}{\partial\boldsymbol{\mu}_E}\dfrac{\partial\boldsymbol{\mu}_E}{\partial\boldsymbol{\phi}}$ について

$$\frac{\partial LB_2}{\partial\boldsymbol{\mu}_E} = -\boldsymbol{\mu}_E^T$$

$\dfrac{\partial\boldsymbol{\mu}_E}{\partial\boldsymbol{\phi}} = \dfrac{\partial G_\mu}{\partial\boldsymbol{\phi}}$ 誤差逆伝播で計算