ベイジアンネットワーク

白坂貴規

平成 32 年 3 月 17 日

1　ベイズ推論の基礎

ベイズ推論では学習や予測, モデル選択などを全て確率分布上の計算問題として取り扱う.

1.1　確率推論

M 次元ベクトル $\boldsymbol{x} = (x_1, \ldots, x_M)^T \in \mathbb{R}^M$ の実数関数 $p(\boldsymbol{x})$ が次の条件を満たす時, $p(\boldsymbol{x})$ を**確率密度関数 (probability density function)** と呼ぶ.

$$p(\boldsymbol{x}) \geq 0 \int p(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x} = 1 \tag{1}$$

また, 各要素が離散値である時, $p(\boldsymbol{x})$ を**確率質量関数 (probability mass function)** と呼ぶ. 以降, 確率密度関数や確率質量関数で決められる $\boldsymbol{x}$ の分布を**確率分布 (probabilistic distribution)** あるいは**確率モデル (probabilistic model)** と呼ぶ. ある 2 つの変数 x と y に関する確率分布 $p(x, y)$ を**同時分布 (joint distribution)** とよび,

$$p(y) = \int p(x, y)\,\mathrm{d}x \tag{2}$$

のように一方の変数 x を積分により除去する操作を**周辺化 (marginalization)** とよび, 結果として得られる確率分布 $p(y)$ を y の**周辺分布**と呼ぶ. また, 同時分布 $p(x, y)$ において, y に対して特定の値が決められた時の x の確率分布を**条件付き分布 (conditional distribution)** とよび, 次のように定義する.

$$p(x|y) = \frac{p(x, y)}{p(y)} \tag{3}$$

条件付き分布 $p(x|y)$ は x の確率分布であり, y はこの分布の特性を決めるパラメータのようなものであると解釈できる.

$$\int p(x|y)\,\mathrm{d}x = \frac{\int p(x, y)\,\mathrm{d}x}{p(y)} = \frac{p(y)}{p(y)} = 1 \tag{4}$$

であり, $p(x, y)$ と $p(y)$ が共に非負であることも考慮すれば, 条件付き分布 $p(x|y)$ は確率分布の用件を満たす. さらに, 同時分布を考える際に重要となるのが**独立 (independence)** という概念である. 同時分布が

$$p(x, y) = p(x)p(y) \tag{5}$$

を満たす時, x と y は独立であるという. ある同時分布が与えられた時, そこから興味の対象となる条件付き分布や周辺分布を算出することをここでは**ベイズ推論 (Baysian inference)**, あるいは単に**推論 (inference)** と呼ぶ. **期待値 (expectation)** は確率分布の特徴を定量的に表すことに使われる. $\boldsymbol{x}$ をベクトルとした時に, 確率分布 $p(\boldsymbol{x})$ に対して, ある関数 $f(\boldsymbol{x})$ の期待値 $\mathbb{E}_{p(\boldsymbol{x})}[f(\boldsymbol{x})]$ は次のように計算される.

$$\mathbb{E}_{p(\boldsymbol{x})}[f(\boldsymbol{x})] = \int f(\boldsymbol{x}p(\boldsymbol{x}))\,\mathrm{d}\boldsymbol{x} \tag{6}$$

2 つの確率分布 $p(\boldsymbol{x})$ 及び $q(\boldsymbol{x})$ に対して次のような期待値を **KL ダイバージェンス (Kullback-Leibler divergence)** と呼ぶ.

$$\begin{aligned} D_{KL}[q(\boldsymbol{x})][p(\boldsymbol{x})] &= -\int q(\boldsymbol{x}) \ln \frac{p(\boldsymbol{x})}{q(\boldsymbol{x})}\,\mathrm{d}\boldsymbol{x} && (7) \\ &= \mathbb{E}_{q(\boldsymbol{x})[\ln q(\boldsymbol{x})]} - \mathbb{E}_{q(\boldsymbol{x})}[\ln p(\boldsymbol{x})] && (8)(9) \end{aligned}$$

KL ダイバージェンスは任意の確率分布のくみに対して非負であり, 0 となるのは 2 つの分布が完全に一致する場合 $p(\boldsymbol{x}]) = q(\boldsymbol{x})$ に限られる. また, KL ダイバージェンすは 2 つの確率分布の”距離”を表していると解釈されるが, $p(\boldsymbol{x})$ と $q(\boldsymbol{x})$ が対称ではない為, 数学的な距離の公理は満たしていない.

1.2　変数変換

既知の確率密度関数に対して変数関数を行うことで新たな確率密度関数を導出することを考える. 全単射の関数 $f : \mathbb{R}^M \to \mathbb{R}^M$ によって変数を $\boldsymbol{y} = f(\boldsymbol{x})$ のように一対一に変換する操作は, 既知の確率密度関数を $p_x(\boldsymbol{x})$ とすれば, 変換によって得られる $\boldsymbol{y}$ の確率密度関数は

$$p_y(\boldsymbol{y}) = p_x(g(\boldsymbol{y}))\,|\det(J_g)| \tag{9}$$

とかける．ただし，J は f の逆関数 g の**ヤコビ行列 (Jacobian matrix)**

$$J_g = \begin{pmatrix} \frac{\partial x_1}{\partial y_1} & \cdots & \frac{\partial x_1}{\partial y_M} \\ \vdots & \ddots & \vdots \\ \frac{\partial x_M}{\partial y_1} & \cdots & \frac{\partial x_M}{\partial y_M} \end{pmatrix} \tag{10}$$

であり，$\det(J_g)$ は J_g の行列式である．

1.3 グラフィカルモデル

グラフィカルモデル (graohical model) は，確率モデルに存在する複数の変換の関係性をノードと矢印を使って表現する記法である．例として，次のようにあるパラメータ θ に依存して N 個の変数 $\boldsymbol{X} = x_1, \ldots, x_N$ が発生するような確率モデルを考える．グラフィカルモデルは次図のようになる．

$$p(\boldsymbol{X}, \theta) = p(\theta)p(\boldsymbol{X}|\theta) = p(\theta)\prod_{n=1}^{N} p(x_n|\theta) \tag{11}$$

なお，上式における $p(\boldsymbol{X}|\theta)$ を**尤度関数 (likelihood function)**，$p(\theta)$ をパラメータ θ の**事前分布**と呼ぶ．

1.4 指数分布族

ガウス分布やディクレ分布など，ベイズ推論で用いられる多くの実用的な確率分布は，**指数分布族 (exponential family)** と呼ばれるある形式をもつクラスに属する．その前に，代表的な確率分布をいくつか紹介する．一次元の**ガウス分布 (Gaussian distribution)** または**正規分布 (normal distribution)** は次のような $x \in \mathbb{R}$ の確率密度関数をもつ分布である．

$$\mathcal{N}(x|\mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp(-\frac{(x-\mu)^2}{2\sigma^2}) \tag{12}$$

$\mu \in \mathbb{R}$ は平均パラメータで，$\sigma^2 > 0$ は分散パラメータである．ガウス分布は次のように M 次元の多変量に拡張される．

$$\mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu}, \sum) = \frac{1}{\sqrt{(2\pi)^D |\sum|}} \exp\left(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^T \sum^{-1} (\boldsymbol{x}-\boldsymbol{\mu})\right) \tag{13}$$

ここで，$\boldsymbol{\mu} \in \mathbb{R}^M$ は M 次元の平均パラメータで，$\sum$ はサイズが $M \times M$ の**共分散行列 (covariance matrix)** である． 1 次元ガウス分布の分散が正であったように，共分散行列 $\sum$ は**正定値行列 (positive definite matrix)** である必要がある． **ベルヌーイ分布 (Bernouli distribution)** はいわゆるコイン投げの分布である．2 値をとる変数 $x \in \{0, 1\}$ を生成するための確率分布で，単一のパラメータ $\mu \in 0, 1$によって分布の性質が決まる.確率質量関数は $(x|\mu) = \mu^x(1-\mu)^{1-x}(14)$ と定義される．**カテゴリ分布 (categorical distribution)** は，ベルヌーイ分布を任意の D 値をとるように拡張したもので，$\boldsymbol{s} \in \{0, 1\}^D$ かつ，各要素 s_d が $\sum_{d=1}^{D} s_d = 1$ となるような確率変数 $\boldsymbol{s}$ を生成する分布である．

$$(\boldsymbol{s}|\boldsymbol{\pi}) = \prod_{d=1}^{D} \pi s_d \tag{15}$$

ここで，$\boldsymbol{\pi} = (\pi_1, \ldots, \pi_D)^T$ は分布を決める D 次元のパラメータで，$\pi_d \in (0, 1)$ かつ $\sum_{d=1}^{D} \pi_d = 1$ を満たすように設定する必要がある．**ガンマ分布 (gamma distribution)** は正の実数 $\lambda > 0$ を生成してくれるような確率分布で，次のように定義される．

$$Gam(\lambda|a, b) = C_G(a, b)\lambda^{a-1}e^{-b\lambda} \tag{16}$$

$$C_G(a, b) = \frac{b^a}{\Gamma(a)} \tag{17}$$

パラメータ a, bh 共に正の実数値として与える必要がある．$\Gamma(\cdot)$ は**ガンマ関数 (gamma function)** で

$$\Gamma(x) = \int t^{x-1}e^{-t}\,\mathrm{d}t \tag{18}$$

で定義される．また，

$$\Gamma(x+1) = x\Gamma(x) \tag{19}$$

という性質をもつ．

1.5

2 自然言語と単語の表現

参考文献

[1] 斎藤康毅. ゼロから作る Deep Learning 2. p57-129

[2] 【機械学習】誤差逆伝播法による速度改善（その2） https://qiita.com/m-hayashi/items/fa4749f8080e542787d2

[3] 【機械学習 誤差逆伝播法】word2vec メモ (1) https://qiita.com/sand/items/85ea76f9c26aabb849e7

[4] Improving Distributional Similarity with Lossons Learned from Word Embeddings https://www.aclweb.org/anthology/Q15-1016/