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.

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

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

$$p_y(\boldsymbol{y}) = p_x(g(\boldsymbol{y}))\,|\det(J_g)| \tag{8}$$

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

$$J_g = \begin{pmatrix} \frac{\partial x_1}{\partial y_1} & \cdots & \frac{\partial x_1}{\partial y_M} \\ \vdots & \ddots & \vdots \\ \frac{\partial x_M}{\partial y_1} & \cdots & \frac{\partial x_M}{\partial y_M} \end{pmatrix} \tag{9}$$

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．

$$p(\boldsymbol{X}, \theta) = p(\theta)p(\boldsymbol{X}|\theta) = p(\theta)\prod_{n=1}^{N} p(x_n|\theta) \tag{10}$$

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．

$$\mathcal{N}(x|\mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp(-\frac{(x-\mu)^2}{2\sigma^2}) \tag{11}$$

$\mu \in \mathbb{R}$ 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．

$$\mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu}, \sum) = \frac{1}{\sqrt{(2\pi)^D |\sum|}} \exp\left(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^T \sum^{-1} (\boldsymbol{x}-\boldsymbol{\mu})\right) \tag{12}$$

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

$$Bern(x|\mu) = \mu^x(1-\mu)^{1-x} \tag{13}$$

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．

$$Cat(\boldsymbol{s}|\boldsymbol{\pi}) = \prod_{d=1}^{D} \pi s_d \tag{14}$$

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．

$$Gam(\lambda|a, b) = C_G(a, b)\lambda^{a-1}e^{-b\lambda} \tag{15}$$

$$C_G(a, b) = \frac{b^a}{\Gamma(a)} \tag{16}$$

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

$$\Gamma(x) = \int t^{x-1}e^{-t}\,\mathrm{d}t \tag{17}$$

で定義される．また，

$$\Gamma(x+1) = x\Gamma(x) \tag{18}$$

という性質をもつ．

1.5 定義

指数分布族は次のような形式でかける確率分布の族である．

$$p(\boldsymbol{x}|\boldsymbol{\eta}) = h(\boldsymbol{x})\exp(\boldsymbol{\eta}^{\boldsymbol{T}}\boldsymbol{t}(\boldsymbol{x}) - a(\boldsymbol{\eta})) \tag{19}$$

それぞれ $\boldsymbol{\eta}$ を自然パラメータ，$t(x)$ を十分統計量，$h(\boldsymbol{x})$ を基底測度，$a(\boldsymbol{x})$ を対数分配関数とよぶ．ただし，

$$a(\boldsymbol{x}) = \ln \int h(\boldsymbol{x})\exp(\boldsymbol{\eta}^{\boldsymbol{T}}\boldsymbol{t}(\boldsymbol{x}))\,\mathrm{d}\boldsymbol{x} \tag{20}$$

であり，上式が積分して 1 になることを保証する．また，ガウス分布，ポアソン分布，多項分布，ベルヌーイ分布など多くの分布が指数分布族として表せる．

1.6 分布の共役性

指数分布族に対して次のような共役事前分布と呼ばれる分布族が存在する.

$$p_\lambda(\boldsymbol{\eta}) = h_c(\boldsymbol{\eta}) \exp(\boldsymbol{\eta^T \lambda_1} - a(\eta)\lambda_2 - a_c(\boldsymbol{\lambda})) \quad (21)$$

共役事前分布の重要な性質は, 次のような指数型分布族による尤度関数に対して, 事後分布も事前分布と同じような形式になることである. いま, N 個のデータ $\boldsymbol{X} = \boldsymbol{x_1}, \ldots, \boldsymbol{x_N}$ を観測したとすると事後分布は

$$\begin{aligned} p(\boldsymbol{\eta}|\boldsymbol{X}) &\propto p_\lambda(\boldsymbol{\eta}) \prod_{n=1}^{N} p(\boldsymbol{x_n}|\boldsymbol{\eta}) \\ &= h_c(\boldsymbol{\eta}) \exp(\boldsymbol{\eta^T \lambda_1} - a(\eta)\lambda_2 - a_c(\lambda)) \\ &\quad \left\{ \prod_{n=1}^{N} h(\boldsymbol{x_n}) \right\} \exp\left(\boldsymbol{\eta^T} \sum_{\boldsymbol{n=1}}^{\boldsymbol{N}} \boldsymbol{t(x_n)} - Na(\boldsymbol{\eta}) \right) \\ &\propto h_c(\boldsymbol{\eta}) \exp\left(\boldsymbol{\eta^T} \left(\boldsymbol{\lambda_1} + \sum_{\boldsymbol{n=1}}^{\boldsymbol{N}} \boldsymbol{t(x_n)} \right) - a(\boldsymbol{\eta})(\boldsymbol{\lambda_2} + \boldsymbol{N}) \right) \end{aligned}$$

となる. つまり, 事後分布のパラメータを $\hat{\boldsymbol{\lambda_1}}, \hat{\lambda_2}$ とすれば,

$$\hat{\boldsymbol{\lambda_1}} = \boldsymbol{\lambda_1} + \sum_{n=1}^{N} \boldsymbol{t(x_n)}, \hat{\lambda_2} = \lambda_2 + N \quad (23)$$

となっている. よって, 指数分布族による尤度関数に対して共役事前分布を用いると事前分布が解析的に求められることがわかる. また, 共役性を利用すると事後分布を使って未観測のデータ $\boldsymbol{x_*}$ の予測分布を次のように求められる.

$$\begin{aligned} p(\boldsymbol{x_*}|\boldsymbol{X}) &= \int p(\boldsymbol{x_*}|\boldsymbol{\eta}) p(\boldsymbol{\eta}|\boldsymbol{X}) \, \mathrm{d}\boldsymbol{\eta} \\ &= \int h(\boldsymbol{x_*}) \exp(\boldsymbol{\eta^T x_*} - a(\boldsymbol{\eta})) h_c(\boldsymbol{\eta}) \exp(\boldsymbol{\eta^T \hat{\lambda}_1} - a(\boldsymbol{\eta})\hat{\lambda_2} - a_c(\boldsymbol{\hat{\lambda}})) \, \mathrm{d}\boldsymbol{\eta} \\ &= h(\boldsymbol{x_*}) \frac{\exp(a_c(\hat{\boldsymbol{\lambda_1}} + \boldsymbol{t(x_*)}, \hat{\lambda_2} + 1))}{\exp(a_c(\hat{bm}\lambda_1, \hat{lam}\hat{b}da_2))} \end{aligned}$$

結果の確率分布は一般的には指数分布族にはならない.

1.7 逐次学習

ベイズ推論によるモデルの学習では, 事後分布によって学習結果を保存することにより, 新規に入ってくる学習データに適応的に学習を進めることができる. これを**逐次学習 (sequential learning)** あるいは**オンライン学習 (online learning)** という. 特に, 共役事前分布を使った解析的な学習では, データの生成過程に順序の依存性を仮定しない場合, データを逐次的に与えた場合と一度にすべて与えた場合とで最終的に得られる事後分布が一致する.

2 マルコフ連鎖

確率変数の系列 $\boldsymbol{z^{(1)}}, \boldsymbol{z^{(2)}}, \ldots$ に対して

$$p(\boldsymbol{z^{(t)}}|\boldsymbol{z^{(1)}}, \boldsymbol{z^{(2)}}, \ldots, \boldsymbol{z^{(t-1)}}) = p(\boldsymbol{z^{(t)}}|\boldsymbol{z^{(t-1)}}) \quad (25)$$

が成り立つとき, 系列 $\boldsymbol{z^{(1)}}, \boldsymbol{z^{(2)}}, \ldots$ を**一次マルコフ連鎖 (first-order Marcov chain)** と呼ぶ. **遷移確率 (transition probability)** を $\mathcal{T}(\boldsymbol{z^{(t-1)}}, \boldsymbol{z^{(t)}})$ とおく. 次の式が成り立つとき, 分布 $p_*(\boldsymbol{z})$ は**定常分布 (statinary distribution)** であるという.

$$p_*(\boldsymbol{z}) = \int \mathcal{T}(\boldsymbol{z'}, \boldsymbol{z}) p_*(\boldsymbol{z'}) \, \mathrm{d}\boldsymbol{z'} \quad (26)$$

(定常状態 (22)) $p_*(\boldsymbol{z})$ がサンプルを取り出したい事後分布だとして, $p_*(\boldsymbol{z})$ に分布収束するような遷移確率 $\mathcal{T}(\boldsymbol{z^{(t-1)}}, \boldsymbol{z^{(t)}})$ を設計するのがマルコフ連鎖モンテカルロ法のアイデアである. $p_*(\boldsymbol{z})$ が定常分布となるための十分条件として**詳細釣り合い条件 (detailed balance condition)** がある.

$$p_*(\boldsymbol{z}) \mathcal{T}(\boldsymbol{z^{(t-1)}}, \boldsymbol{z^{(t)}}) = p_*(\boldsymbol{z'}) \mathcal{T}(\boldsymbol{z^{(t-1)}}, \boldsymbol{z^{(t)}}) \quad (27)$$

詳細釣り合い条件に加え, サンプルサイズを無限大にしたとき, 遷移確率よって任意の初期状態 $p(\boldsymbol{z_0})$ から定常分布 $p_*(\boldsymbol{z})$ に収束できなければならない. この特性をエルゴード性と呼ぶ. 具体的には, マルコフ連鎖において任意の状態から任意の状態へ有限回数で遷移できること (既約性), すべての状態が固定の周期性をもたないこと (非周期性), さらに同じ状態に有限回で戻ることができること (正再帰性) が求められる. (24)

2.1 最適化に基づく推論手法

マルコフ連鎖モンテカルロ法は, 無限に計算を続ければ得られるサンプルが真の分布から得られたものと同一視できるという理論面で優れた特性がある. しかし, 実用上は必要なサンプルサイズが明確に決めにくいことや, 計算コストが膨大になるなどの欠点をもっています. 一方で, 機械学習の分野で主流となっている方法は勾配情報を用いた数値最適化に基づく手法で, 実験的には非常に速い収束測度を持つことが示されている. 最適化による近似推論アルゴリズ

ムの中で, 現在最も広く用いられている手法が**変分推論法 (variational inference method)** である. この方法では, 事前分布を計算する際に登場する解析不可能な積分を, 最適化の問題に置き換えることによって近似的に数値計算する. 周辺尤度 $p(\boldsymbol{X})$ の計算には潜在変数 Z の積分除去 $p(\boldsymbol{X}) = \int p(\boldsymbol{X}, \boldsymbol{Z})\,\mathrm{d}\boldsymbol{Z}$ が必要になるが, モデルが複雑になるとこの積分は解析的に解析的に実行できない. 変分推論法では, **エビデンス下界 (evidence lower bound, ELBO)** と呼ばれる対数周辺尤度 $\ln p(\boldsymbol{X})$ の下界 $\mathcal{L}(\boldsymbol{\xi})$ を考える.

$$\ln p(\boldsymbol{X}) \geq \mathcal{L}(\boldsymbol{\xi}) \tag{28}$$

ここで,$\boldsymbol{\xi}$ **変分パラメータ (variational parameter)** と呼ばれてるもので, 変分推論報における近似分布の平均や分散などを指す. ちなみに, ELBO を負にした $\mathcal{F} = -\mathcal{L}(\boldsymbol{xi})$ は**変分エネルギー (variational energy)** と呼ばれる. 勾配降下法などの一般的な最適化手法を用いて $\mathcal{L}(\boldsymbol{\xi})$ を $\boldsymbol{\xi}$ によって最大化すれば, 対数周辺尤度 $\ln p(\boldsymbol{X})$ の近似解が得られることになる. ELBO の設計の仕方はいくつかあり, モデルや目的に応じて使い分ける. 最もよく使われる手法は, 事後分布 $p(\boldsymbol{Z}|\boldsymbol{X})$ をパラメータ $\boldsymbol{\xi}$ で定められるある分布 $q(\boldsymbol{Z};\boldsymbol{\xi})$ によって近似することである. 近似の良さを測る手法にもいくつかあるが, 変分推論法では次のような KL ダイバージェンスを使い, 変分パラメータ $\boldsymbol{\xi}$ に関して最小化することによって近似分布 $q(\boldsymbol{Z};\boldsymbol{\xi})$ を得る.

$$q(\boldsymbol{Z};\boldsymbol{\xi_{opt}}) = \arg\min_{\boldsymbol{opt}} D_{KL}[q(\boldsymbol{Z};\boldsymbol{\xi})||p(\boldsymbol{Z}|\boldsymbol{X})] \tag{29}$$

また, 対数周辺尤度 $\ln p(\boldsymbol{X})$ は次のように ELBO と上式の KL ダイバージェンスに分解できる.

$$\ln p(\boldsymbol{X}) = \mathcal{L}(\boldsymbol{\xi}) + D_{KL}[q(\boldsymbol{Z};\boldsymbol{\xi})||p(\boldsymbol{Z}|\boldsymbol{X})] \tag{30}$$

ただし,

$$\mathcal{L}(\boldsymbol{\xi}) = \int q(\boldsymbol{Z};\boldsymbol{\xi} \ln \frac{p(\boldsymbol{Z}|\boldsymbol{X})}{q(\boldsymbol{Z};\boldsymbol{\xi})}\,\mathrm{d}\boldsymbol{Z} \tag{31}$$

上式からわかるように, 対数周辺尤度 $\ln p(\boldsymbol{X})$ 自体は $\boldsymbol{\xi}$ の値にかかわらず一定なので,$D_{KL}[q(\boldsymbol{Z};\boldsymbol{\xi})||p(\boldsymbol{Z}|\boldsymbol{X})]$ を $\boldsymbol{\xi}$ に関して最小化する問題は, $\mathcal{L}\boldsymbol{\xi}$ を $\boldsymbol{\xi}$ に関して最小化する問題と等価になる. 近似分布 q の置き方には様々な選択があり, 潜在変数の集合 $\boldsymbol{Z}$ が $\boldsymbol{Z} = \{\boldsymbol{Z_1}, \ldots, \boldsymbol{Z_M}\}$ のように M 個に分割できるとする. 複雑なモデルに対しては, 事前分布に独立性を仮定して近似する方法がよく使われている.

$$q(\boldsymbol{Z}) = \prod_{i=1}^{M} q(\boldsymbol{Z_i}) \tag{32}$$

この手法は特に**平均場近似 (mean field approximation)** と呼ばれている. 平均場近似では, 各近似分布 $q(\boldsymbol{Z_1}, \ldots, \boldsymbol{Z_M})$ を交互に更新していく手続きを繰り返すため, アルゴリズムの特性はギブスサンプリングと非常に似たものになっている.

3 ニューラルネットワークのベイズ推論

バッチ学習によるニューラルネットワーク (以下, NN) の基本的な学習方法を示す. 順伝播型 NN をベイズ化する方法は, ネットワークの挙動を支配するパラメータに事前分布を設定することによって確率的な学習や予測が行えるようにするものである. 入力データ $\boldsymbol{X} = \{x_1, \ldots, x_N\}$ が与えられたもとでの, 観測データ $\boldsymbol{Y} = \{\boldsymbol{y_1}, \ldots, \boldsymbol{y_N}\}$ 及びパラメータの同時分布を

$$p(\boldsymbol{Y}, \boldsymbol{W}|\boldsymbol{X}) = p(\boldsymbol{W}) \prod_{n=1}^{N} p(\boldsymbol{y_n}|\boldsymbol{x_n}, \boldsymbol{W}) \tag{33}$$

とおく. ここでは, $\boldsymbol{x_n} \in \mathbb{R}^{H_0}$ から $\boldsymbol{y_n} \in \mathbb{R}^D$ を予測する回帰問題であるとし, 観測モデルには次のようなガウス分布を使う.

$$p(\boldsymbol{y_n}|\boldsymbol{x_n}, \boldsymbol{W}) = \mathcal{N}(\boldsymbol{y_n}|\boldsymbol{f}(\boldsymbol{X_n};\boldsymbol{W}), \sigma_y^2\boldsymbol{I}) \tag{34}$$

ここで σ_y^2 は固定のノイズパラメータで, $\boldsymbol{f}(\boldsymbol{X_n};\boldsymbol{W})$ は出力次元が D である NN である. ベイズ推論の枠組みでは, 学習データが与えられたあとのパラメータの事後分布を計算する. したがって, NN のパラメータには事前分布を明示的に設定する必要がある. ここでは簡単のため各重みパラメータを $w \in \boldsymbol{W}$ とし, 次のような独立なガウス分布を考える.

$$p(w) = \mathcal{N}(w|0, \sigma_w^2) \tag{35}$$

次の図には活性化関数に双曲線正接関数, 入力ベクトルを $(x, 1)^T$ とし, 隠れ層の数 H_1 及び重みパラメータに仮定するノイズ σ_w^2 を変えた場合の関数のサンプル例を示す. H_1 が大きくなる程事前分布によって生成される関数が複雑化し, また σ_w^2 が大きくなるほど急激な変化を持つ関数が生成されていることがわかる.

3.1 ラプラス近似による学習

ベイズニューラルネットワーク (以下, BNN) に対するラプラス近似による学習と予測を導出する. 簡

単のため, 出力次元は $D=1$ とする. はじめにモデルの事後分布の MAP 推定値を最適化により求めたあと, その周辺をガウス分布によって近似する. BNN においては重みパラメータ $\boldsymbol{W}$ の事後分布の MAP 推定値を求めることが最初のステップになる. 事後分布は

$$\begin{aligned} p(\boldsymbol{W}|\boldsymbol{X},\boldsymbol{Y}) &= \frac{p(\boldsymbol{W})p(\boldsymbol{Y}|\boldsymbol{X},\boldsymbol{W})}{p(\boldsymbol{Y}|\boldsymbol{X})} \\ &\propto p(\boldsymbol{W})p(\boldsymbol{Y}|\boldsymbol{X},\boldsymbol{W}) \end{aligned} \tag{36}$$

と書ける. 分母には $\boldsymbol{W}$ がないので, $\boldsymbol{W}$ の最適化の過程では無視できる. 局所最適解 $\boldsymbol{W_{MAP}}$ は, 対数事後分布の勾配を利用して

$$\boldsymbol{W_{new}} = \boldsymbol{W_{old}} + \alpha \nabla_{\boldsymbol{W}} \ln p(\boldsymbol{W}|\boldsymbol{Y},\boldsymbol{X})|_{\boldsymbol{W}=\boldsymbol{W_{old}}} \tag{37}$$

のように繰り返し更新すれば求まる. ここで, $\alpha > 0$ は学習率である. 対数事後分布は,

$$\begin{aligned} \ln p(\boldsymbol{W}|\boldsymbol{Y},\boldsymbol{X}) &= \ln p(\boldsymbol{W}|\boldsymbol{Y},\boldsymbol{X}) + \ln p(\boldsymbol{W}) + c \\ &= \sum_{n=1}^{N} \ln p(y_n|\boldsymbol{x_n},\boldsymbol{W}) + \sum_{w\in\boldsymbol{W}} \ln p(x) + c \end{aligned} \tag{38}$$

と書ける. あるパラメータ $w \in \boldsymbol{W}$ の偏微分を計算すると

$$\frac{\partial}{\partial w} \ln p(\boldsymbol{W}|\boldsymbol{Y},\boldsymbol{X}) = -\{\frac{1}{\sigma_y^2}\frac{\partial}{\partial w}E(\boldsymbol{W}) + \frac{1}{\sigma_w^2}\frac{\partial}{\partial w}\Omega_{L2}(\boldsymbol{W})\} \tag{39}$$

となる. 上式において $\Omega_{L2}(\boldsymbol{W})$ は各パラメータ w に関するガウス事前分布 $p(w)$ に由来するもびであるが, こちらの方が微分が容易である. また, $E(\boldsymbol{W})$ は NN の誤差関数であるが, これは通常の誤差逆伝播を用いて微分を評価する.

$$q(\boldsymbol{W}) = \mathcal{N}(\boldsymbol{W}|\boldsymbol{W_{MAP}}, \{\Lambda(\boldsymbol{W_{MAP}})\}^{-1}) \tag{40}$$

ここで,*Lambda* は精度行列であり,

$$\begin{aligned} \Lambda &= -\nabla_{\boldsymbol{W}}^2 \ln p(\boldsymbol{W}|\boldsymbol{Y},\boldsymbol{X}) \\ &= \frac{1}{\sigma_w^2}\boldsymbol{I} + \frac{1}{\sigma_y^2}\boldsymbol{H} \end{aligned} \tag{41}$$

である. また, $\boldsymbol{H}$ は NN の誤差関数に対するヘッセ行列である.

3.2 予測分布の近似

パラメータの事後分布の近似を得た後は, テストの入力 $\boldsymbol{x_*}$ に対する出力 y_* の予測分布を

$$p(y_*|\boldsymbol{x_a st},\boldsymbol{Y},\boldsymbol{X}) \approx \int p(y_a st|\boldsymbol{x_*},\boldsymbol{W}) q(\boldsymbol{W})\,\mathrm{d}\boldsymbol{W} \tag{42}$$

として近似する. パラメータの事後分布を簡単なガウス分布 $q(\boldsymbol{W})$ によって近似したものの, $p(y_*|\boldsymbol{x_*},\boldsymbol{W})$ の中には NN が含まれているため, 依然として一般的に予測分布の計算は解析的に行えない. ここでは予測分布を計算するために, NN の関数の線形近似を行う. この近似では, パラメータの事後分布が MAP 推定値の周辺に集中しており, かつその小さな範囲においては NN の関数値 $f(\boldsymbol{x_*};\boldsymbol{W})$ が $\boldsymbol{W}$ の線形関数でよく近似できるという仮定をおく. テイラー展開で $\boldsymbol{W}$ の関数 $f(\boldsymbol{x_*};\boldsymbol{W})$ を $\boldsymbol{W_{MAP}}$ まわりで一次近似すれば

$$f(\boldsymbol{x_*};\boldsymbol{W}) \approx f(\boldsymbol{x_*};\boldsymbol{W_{MAP}}) + \boldsymbol{g^T}(\boldsymbol{W}-\boldsymbol{W_{MAP}}) \tag{43}$$

ただし, $\boldsymbol{g}$ は次のように関数の勾配を $\boldsymbol{W_{MAP}}$ で評価したものである.

$$\boldsymbol{g} = \nabla_{\boldsymbol{W}} f(\boldsymbol{x_*};\boldsymbol{W})|_{\boldsymbol{W}=\boldsymbol{W_{MAP}}} \tag{44}$$

この近似を用いれば, 予測分布の計算に NN 特有の非線形関数がなくなるので, 解析的に計算することが可能になる. したがって, 求めたい予測分布の近似は,

$$\begin{aligned} p(y_*|\boldsymbol{x_*},\boldsymbol{Y},\boldsymbol{X}) &\approx \int p(y_*|\boldsymbol{x_*},\boldsymbol{W}) q(\boldsymbol{W})\,\mathrm{d}\boldsymbol{W} \\ &\approx \int \mathcal{N}(y_*|f(\boldsymbol{x_*};\boldsymbol{W_{MAP}})) + \boldsymbol{g^T}(\boldsymbol{W}-\boldsymbol{W_{MAP}}, \\ &\quad \mathcal{N}(\boldsymbol{W}|\boldsymbol{W_{MAP}},\{\boldsymbol{\Lambda}(\boldsymbol{W_{MAP}})\}^{-1})\,\mathrm{d}\boldsymbol{W} \\ &= \mathcal{N}(y_*|f(\boldsymbol{x_*};\boldsymbol{W_{MAP}}),\sigma^2(\boldsymbol{x_*})) \end{aligned} \tag{45}$$

と計算できる. ただし,

$$\sigma^2(\boldsymbol{x_*}) = \sigma_y^2 + \boldsymbol{g^T} + \{\boldsymbol{\Lambda}(\boldsymbol{W_{MAP}})^{-1}\}\boldsymbol{g} \tag{46}$$

である.

3.3 重みパラメータの推論

BNN への正規化されていない事後分布を利用すると対応するポテンシャルエネルギーは

$$\mathcal{U}(\boldsymbol{W}) = -\{\ln p(\boldsymbol{Y}|\boldsymbol{X},\boldsymbol{W}) + \ln p(\boldsymbol{W})\} \tag{47}$$

となる. リープフロッグ法を使うためにはポテンシャルエネルギーの微分が必要になるが, 正則化項をもつ基本的な NN のコスト関数の微分と等価であるから, 誤差逆伝播による勾配計算が利用できる.

3.4 ハイパーパラメータの推論

重みパラメータ $\boldsymbol{W}$ の事前分布を支配する σ_w^2 や観測モデルのノイズパラメータ σ_y^2 はハイパーパラメータとして扱われ, 通常は学習を実行する前に適切なものを固定値として与えておく必要がある. これらのハイパーパラメータはデータの大まかな"スケール感"を反映していると考えられるが, もし直観的に値を指定するのが困難である場合や, 学習データに対して当てはまりの良い学習結果を得たい場合は, ハイパーパラメータに対しても事前分布を設定することによって, サンプリングの枠組みの中で重みパラメータ $\boldsymbol{W}$ と同時に推論することもできる. ハイパーパラメータも同時推論するために, これらに事前分布を与える. パラメータ $\boldsymbol{W}$ は分散 σ_w^2 をもつガウス分布にしたがって決定される. 表記を簡単にするため, 精度パラメータ $\gamma_w = \sigma_w^{-2}$ を導入し, 事前分布として,

$$p(\gamma_w) = Gam(\gamma_w | a_w, b_w) \tag{48}$$

を考える. ここで, $a_w > 0, b_w > 0$ は固定値として与える. 観測ノイズに対する精度パラメータ $\gamma_y = \sigma_y^{-2}$ の事前分布も同様にして

$$p(\gamma_y) = Gam(\gamma_y | a_y, b_y) \tag{49}$$

と設定する. ただし, $a_y > 0, b_y > 0$ である. これらの精度パラメータの事前分布を導入した場合のモデルを改めて書き下すと

$$\begin{aligned} p(\boldsymbol{Y}, \boldsymbol{W}, \gamma_w, \gamma_y | \boldsymbol{X}) &= p(\gamma_w) p(\gamma_y) p(\boldsymbol{W} | \gamma_w) \\ &\quad \prod_{n=1}^{N} p(y_n | \boldsymbol{x_n}, \boldsymbol{W}, \gamma_y) \end{aligned} \tag{50}$$

となる. したがって事後分布全体は

$$p(\boldsymbol{W}, \gamma_w, \gamma_y | \boldsymbol{Y}, \boldsymbol{X}) \tag{51}$$

となる. ここではギブスサンプリングを用いて各確率変数のサンプルを得ることを考える. 各確率変数 $\boldsymbol{W}, \gamma_w, \gamma_y$ を条件付き確率を用いて別々にサンプリングすることである. γ_w, γ_y がサンプルされた値で条件付けされた下での $\boldsymbol{W}$ の分布は

$$p(\boldsymbol{W} | \boldsymbol{Y}, \boldsymbol{X}, \gamma_w, \gamma_y) \tag{52}$$

となるが, これはパラメータの事後分布そのものであるため, 通常通りハミルトニアンモンテカルロ法を実行することによって $\boldsymbol{W}$ のサンプルを得る. $\boldsymbol{W}, \gamma_y$ が与えられた下での γ_w の分布は, γ_w に関わらない部分を無視すると

$$p(\gamma_w | \boldsymbol{Y}, \boldsymbol{X}, \boldsymbol{W}, \boldsymbol{\gamma_y}) \propto p(\boldsymbol{W} | \gamma_w) p(\gamma_w) \tag{53}$$

とかける. $p(\boldsymbol{W}|\gamma_w)$ がガウス分布であり, 精度 γ_w の事前分布 $p(\gamma_w)$ には共役事前分布であるガンマ分布を用いているので, この条件付き分布もガンマ分布として解析的に以下のようにもとまる.

$$\begin{aligned} \gamma_w &\backsim Gam(\hat{a_w}, \hat{b_w}) && (54) \\ \hat{a_w} &= a_w + \frac{K_w}{2} && (55) \\ \hat{b_w} &= b_w + \frac{1}{2} \sum_{w \in \boldsymbol{W}} w^2 && (56) \end{aligned}$$

のようにして γ_w のサンプルが得られる. ただし, K_w は重みパラメータ $\boldsymbol{W}$ の総数である. 次に, $\boldsymbol{W}, \gamma_w$ が与えられた下での γ_y の分布も

$$p(\gamma_y | \boldsymbol{Y}, \boldsymbol{X}, \boldsymbol{W}, \boldsymbol{\gamma_w}) \propto p(\gamma_y) \prod_{n=1}^{N} p(y_n | \boldsymbol{x_n}, \boldsymbol{W}, \gamma_y) \tag{57}$$

と計算できる. 観測モデル $p(\boldsymbol{Y}|\boldsymbol{X}, \boldsymbol{W}, \boldsymbol{\gamma_y})$ がガウス分布であり, 精度 γ_y の事前分布 $p(\gamma_y)$ には共役なガンマ分布を用いたので, こちらも解析的に分布を計算でき,

$$\begin{aligned} \gamma_y &= Gam(\hat{a_y}, \hat{b_y}) && (58) \\ \hat{a_y} &= a_y + \frac{N}{2} && (59) \\ \hat{b_y} &= b_y + \frac{1}{2} \sum_{n=1}^{N} \{y_n - f(\boldsymbol{x_n}; \boldsymbol{W})\}^2 && (60) \end{aligned}$$

のようにして γ_y のサンプルが得られる. 直観的に上式を解釈すると, γ_y は NN の関数 $f(\boldsymbol{x}; \boldsymbol{W})$ で表現できない誤差を学習していると言える. ガンマ分布の平均は $\hat{a_w}/\hat{b_w}$ であるため, $\hat{b_w}$ が大きいほど関数 $f(\boldsymbol{x}; \boldsymbol{W})$ による y の推定の精度が低く, 観測に対する分散が大きくなるように学習される. なお, ここでは簡単のため共通なハイパーパラメータ γ_w が与えられていると仮定したが, ハイパーパラメータを複数のグループに分割して与えることもできる.

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/