電子情報通信学会論文誌 2000/6 Vol. J83–D–II No. 6

非ベイズ的付加項を用いた多層パーセプトロンの学習

内田　真人†（学生員）　　塩谷　浩之†（正員）
伊達　　悖†（正員）

Training Multilayer Perceptrons Using Non-Bayesian Additive Term

Masato UCHIDA†, *Student Member*, Hiroyuki SHIOYA†, and Tsutomu DA-TE†, *Regular Members*

† 北海道大学大学院工学研究科，札幌市
Graduate School of Engineering, Hokkaido University, Sapporo-shi, 060–8628 Japan

あらまし　本論文では，ベイズ推定の枠組みを用いることなく，正則化項に類似した付加項をもつ学習アルゴリズムの導出を試みる．ある条件のもとで導出されたアルゴリズムは，正則化パラメータに相当する部分が正定数とはならないという点で通常の正則化項付き学習アルゴリズムと異なる．また，特別な場合には付加項の意味付けに関する結果を得た．

キーワード　多層パーセプトロン，正則化項

1. まえがき

多層パーセプトロン（以下，MLP と略す）の学習において，過学習の回避，汎化能力の向上などの目的から最小化すべき損失関数にネットワーク構造に関するペナルティ項，いわゆる正則化項を付与する学習法が提案されている [1]．正則化項の理論的説明のために用いられる手段としては，ベイズ推定の枠組みを利用したものが挙げられる [2]．これは，正則化項をパラメータの事前分布としてとらえるものであり，そのハイパパラメータは正則化パラメータに相当する．

本論文の目的は，ベイズ推定の立場をとらずに，正則化項に類似した付加項を用いる学習アルゴリズムを導出し，別の視点から付加項の意味付けを行うことにある．適当な確率分布モデルによるパラメータ推定を議論の出発点とした結果，ある条件のもとで，正則化パラメータに相当する部分が正定数とはならないという点を除いて通常の正則化項付き学習アルゴリズムと形式的に一致する学習アルゴリズムを得た．また特別な場合には，付加項を推定モデルの分散として意味づけることが可能となった．

2. 準　　備

MLP は，入力層・中間層・出力層から構成される階層型ニューラルネットワークであり，結合荷重・しきい値からなる修正可能なパラメータ $\boldsymbol{\theta} = (\theta_1, \cdots, \theta_k)^T$ $(\in \boldsymbol{R}^k)$ をもつ（T はベクトルの転置を表す）．本論文では，出力層の素子数を 1 とし，MLP は入力 $\boldsymbol{x} = (x_1, \cdots, x_m)^T$ $(\in \boldsymbol{R}^m)$ に対し $f(\boldsymbol{x};\boldsymbol{\theta})$ を決定論的に出力するものとする．ただし，$f(\boldsymbol{x};\boldsymbol{\theta})$ は $\boldsymbol{\theta}$ に関して高階微分可能な関数であるとする．ここで，入力 $\boldsymbol{x}$ と教師信号 y^* からなるサンプルの組が，ある確率分布 $p_*(\boldsymbol{x}, y)(= q(\boldsymbol{x})p_*(y|\boldsymbol{x}))$ に従って互いに独立に n 個観測されたとする．このときの学習サンプルを $D_n = \{(\boldsymbol{x}_1, y_1^*), \cdots, (\boldsymbol{x}_n, y_n^*)\}$ と書き，MLP の学習を考える．学習の目的は，事前に設定した損失関数 $l(y|\boldsymbol{x};\boldsymbol{\theta})$ について $\boldsymbol{\theta}_l^* = \arg\min_\theta E_*[l(y|\boldsymbol{x};\boldsymbol{\theta})]$ となるパラメータ $\boldsymbol{\theta}_l^*$ を得ることであるが，そのための手続きとして $\hat{\boldsymbol{\theta}}_l = \arg\min_\theta E_e[l(y|\boldsymbol{x};\boldsymbol{\theta})]$ となるパラメータ $\hat{\boldsymbol{\theta}}_l$ を得ることを考える．ただし，E_* は $p_*(\boldsymbol{x}, y)$ による期待値を表し，E_e は経験分布 $p_e(\boldsymbol{x}, y)$ $(= q_e(\boldsymbol{x})p_e(y|\boldsymbol{x}))$ による期待値を表す．代表的な損失関数の例には，以下に示す 2 乗誤差関数がある．

$$l_2(y|\boldsymbol{x};\boldsymbol{\theta}) = \frac{1}{2}(y - f(\boldsymbol{x};\boldsymbol{\theta}))^2$$

ここで

$$L(\boldsymbol{x}, y;\boldsymbol{\theta}) = -\log p(\boldsymbol{x}, y;\boldsymbol{\theta}) = -\log q(\boldsymbol{x})p(y|\boldsymbol{x};\boldsymbol{\theta})$$

で定義される負の対数ゆう度関数について $\hat{\boldsymbol{\theta}}_L = \arg\min_\theta E_e[L(\boldsymbol{x}, y;\boldsymbol{\theta})]$ となるパラメータ $\hat{\boldsymbol{\theta}}_L$ を考える．特に，$\boldsymbol{x}$ に対する y の条件付き確率分布を

$$p(y|\boldsymbol{x};\boldsymbol{\theta}) = \frac{1}{\sqrt{2\pi}\sigma}\exp\left\{-\frac{1}{2\sigma^2}(y - f(\boldsymbol{x};\boldsymbol{\theta}))^2\right\} \tag{1}$$

で与えられるモデルで推定する場合

$$L(\boldsymbol{x}, y;\boldsymbol{\theta}) = -\log\frac{q(\boldsymbol{x})}{\sqrt{2\pi}\sigma} + \frac{l_2(y|\boldsymbol{x};\boldsymbol{\theta})}{\sigma^2} \tag{2}$$

となり，$E_e[l_2(y|\boldsymbol{x};\boldsymbol{\theta})]$ の最小化と，$E_e[L(\boldsymbol{x}, y;\boldsymbol{\theta})]$ の最小化とは全く同等の意味をもつ．

最急降下法によるオンライン学習アルゴリズムとバッチ学習アルゴリズムを，それぞれ

$$\boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_t - \eta\nabla l(y_t^*|\boldsymbol{x}_t;\boldsymbol{\theta}_t)$$
$$\boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_t - \eta E_e[\nabla l(y|\boldsymbol{x};\boldsymbol{\theta}_t)]$$

で与える．ただし，η は十分小さい正定数であり，$\boldsymbol{\theta}_t$ は t 回目の更新で得られたパラメータを表す．また

$$\nabla = \left(\frac{\partial}{\partial\theta_1}, \cdots, \frac{\partial}{\partial\theta_k}\right)^T$$

である．特に，式 (2) の場合

$$\sigma^2 \nabla L(\boldsymbol{x}, y; \boldsymbol{\theta}) = \nabla l_2(y|\boldsymbol{x}; \boldsymbol{\theta}) \tag{3}$$

となり，σ^2 が正値であるということから $L(\boldsymbol{x}, y; \boldsymbol{\theta})$ と $l_2(y|\boldsymbol{x}; \boldsymbol{\theta})$ のこう配の方向は一致する．

3. 定 式 化

$\boldsymbol{x}$ に対する y の条件付き確率分布を式 (1) で与えられるモデルで推定する場合，$E_e[l_2(y|\boldsymbol{x}; \boldsymbol{\theta})]$ の最小化と，$E_e[L(\boldsymbol{x}, y; \boldsymbol{\theta})]$ の最小化が全く同等の意味をもつという特徴に注目する．このことは，分散を定数と仮定したことに起因する．したがって，その仮定を外した場合，$l_2(y|\boldsymbol{x}; \boldsymbol{\theta})$ と $L(\boldsymbol{x}, y; \boldsymbol{\theta})$ の性質が異なることがあると考えられる．本論文では分散の推定のために新たなパラメータを設定せずに，既存のパラメータを用いて新たな関数を作り，それを分散の推定のために用いる．この結果，通常の正則化項に類似した付加項をもつ学習アルゴリズムを得ることができる．以下では，損失関数が 2 乗誤差関数の場合に限定せず，より一般的な立場からモデルを設定し，それを用いた学習アルゴリズムを導出する．

3.1 モデルの設定

式 (1) を拡張し，以下に示すモデルを設定する．

$$p(y|\boldsymbol{x}; \boldsymbol{\theta}) = \frac{1}{C(\boldsymbol{\theta})} \exp\left\{ -\frac{l(y|\boldsymbol{x}; \boldsymbol{\theta})}{r(\boldsymbol{\theta})} \right\} \tag{4}$$

ただし

$$C(\boldsymbol{\theta}) = \int_R \exp\left\{ -\frac{l(y|\boldsymbol{x}; \boldsymbol{\theta})}{r(\boldsymbol{\theta})} \right\} dy < \infty, \quad \forall \boldsymbol{x} \in \boldsymbol{R}^m \tag{5}$$

$$l(y|\boldsymbol{x}; \boldsymbol{\theta}) \geqq 0, \quad \forall \boldsymbol{x} \in \boldsymbol{R}^m, \forall y \in \boldsymbol{R}, \forall \boldsymbol{\theta} \in \boldsymbol{R}^k$$

$$r(\boldsymbol{\theta}) > 0, \quad \forall \boldsymbol{\theta} \in \boldsymbol{R}^k$$

とする．また，$l(y|\boldsymbol{x}; \boldsymbol{\theta})$ と $r(\boldsymbol{\theta})$ は $\boldsymbol{\theta}$ に関して高階微分可能な関数であるとする．

3.2 学習アルゴリズムの設定

$\boldsymbol{x}$ に対する y の条件付き確率分布を式 (4) で与えたときの $L(\boldsymbol{x}, y; \boldsymbol{\theta})$ について

$$\begin{aligned} & r(\boldsymbol{\theta}) E_e[\nabla L(\boldsymbol{x}, y; \boldsymbol{\theta})] \\ &= E_e[\nabla l(y|\boldsymbol{x}; \boldsymbol{\theta})] + K_e(\boldsymbol{\theta}) \nabla r(\boldsymbol{\theta}) \\ &\quad - E_{q_e(x)}[E_{p(y|x;\theta)}[\nabla l(y|\boldsymbol{x}; \boldsymbol{\theta})]] \end{aligned} \tag{6}$$

$$\begin{aligned} & r(\boldsymbol{\theta}) E_*[\nabla L(\boldsymbol{x}, y; \boldsymbol{\theta})] \\ &= E_*[\nabla l(y|\boldsymbol{x}; \boldsymbol{\theta})] + K_*(\boldsymbol{\theta}) \nabla r(\boldsymbol{\theta}) \\ &\quad - E_{q(x)}[E_{p(y|x;\theta)}[\nabla l(y|\boldsymbol{x}; \boldsymbol{\theta})]] \end{aligned} \tag{7}$$

が成り立つ（付録 **1.** 参照）．ここで，$r(\boldsymbol{\theta}) > 0$ の仮定と，式 (3) との整合性から，オンライン学習アルゴリズムを

$$\boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_t - \eta r(\boldsymbol{\theta}_t) \nabla L(\boldsymbol{x}, y; \boldsymbol{\theta}_t) \tag{8}$$

と定め，$n < \infty$, $n \to \infty$ のときのバッチ学習アルゴリズムをそれぞれ

$$\boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_t - \eta r(\boldsymbol{\theta}_t) E_e[\nabla L(\boldsymbol{x}, y; \boldsymbol{\theta}_t)] \tag{9}$$

$$\boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_t - \eta r(\boldsymbol{\theta}_t) E_*[\nabla L(\boldsymbol{x}, y; \boldsymbol{\theta}_t)] \tag{10}$$

と定める．

4. 考 察

過学習の回避，汎化能力の向上，あるいは構造化のために，本来の損失関数 $l(y|\boldsymbol{x}; \boldsymbol{\theta})$ に正則化項（あるいは，ペナルティ項）と呼ばれる $\boldsymbol{\theta}$ の関数 $r(\boldsymbol{\theta})$ を加えた

$$\sum_{i=1}^{n} l(y_i|\boldsymbol{x}_i; \boldsymbol{\theta}) + \lambda r(\boldsymbol{\theta}), \quad \lambda > 0 \tag{11}$$

を新たな損失関数として MLP の学習に用いることがある．この形式は，ベイズ推定の立場から容易に導出される（付録 **2.** 参照）．したがって，最急降下法によるバッチ学習アルゴリズムは

$$\boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_t - \eta \left\{ E_e[\nabla l(y|\boldsymbol{x}; \boldsymbol{\theta}_t)] + \frac{\lambda}{n} \nabla r(\boldsymbol{\theta}_t) \right\} \tag{12}$$

と与えられる．

これに対し本論文で得られた学習アルゴリズム（式 (6), (9)）は，パラメータの事前分布を全く考慮せずに導出されたものであるが，$-E_{q_e(x)}[E_{p(y|x;\theta)}[\nabla l(y|\boldsymbol{x}; \boldsymbol{\theta})]]$ が加算されているという点と，$\nabla r(\boldsymbol{\theta})$ の係数が正定数とはならないという点で異なる．前者については，**4.1** で考察し，後者については **4.2** で考察する．

4.1 学習アルゴリズム

式 (12) と式 (6), (9) の形式を $\nabla r(\boldsymbol{\theta})$ の係数が正定数とはならないという点を除いて一致させるためには

$$E_{p(y|x;\theta)}[\nabla l(y|\boldsymbol{x}; \boldsymbol{\theta})] = 0$$

が成立すれば十分である．上式を成立させる例には

$$l_{2d}(y|\boldsymbol{x}; \boldsymbol{\theta}) = \frac{1}{2d} (y - f(\boldsymbol{x}; \boldsymbol{\theta}))^{2d}, \quad d \in \boldsymbol{N}$$

がある．これは

$$l_{2d}(f(\boldsymbol{x};\boldsymbol{\theta})+z|\boldsymbol{x};\boldsymbol{\theta})=l_{2d}(f(\boldsymbol{x};\boldsymbol{\theta})-z|\boldsymbol{x};\boldsymbol{\theta})$$
$$\nabla l_{2d}(f(\boldsymbol{x};\boldsymbol{\theta})+z|\boldsymbol{x};\boldsymbol{\theta})=-\nabla l_{2d}(f(\boldsymbol{x};\boldsymbol{\theta})-z|\boldsymbol{x};\boldsymbol{\theta})$$

が成立するためである．特に $d=1$ のときには2乗誤差関数に相当する．また，式 (5) より

$$C(\boldsymbol{\theta})=\sqrt{2\pi r(\boldsymbol{\theta})}$$

であることから

$$K_e(\boldsymbol{\theta})=E_e\left[\frac{1}{2}-\frac{l_2(y|\boldsymbol{x};\boldsymbol{\theta})}{r(\boldsymbol{\theta})}\right]$$
$$K_*(\boldsymbol{\theta})=E_*\left[\frac{1}{2}-\frac{l_2(y|\boldsymbol{x};\boldsymbol{\theta})}{r(\boldsymbol{\theta})}\right]$$

よりバッチ学習アルゴリズムが明示的に定まる．式 (8)，(A·2) より，オンライン学習アルゴリズムも同様に定まることがわかる．

4.2 正則化項との関連性

$n\to\infty$ におけるバッチ学習アルゴリズム（式 (7)，(10)）を $l(y|\boldsymbol{x};\boldsymbol{\theta})=l_2(y|\boldsymbol{x};\boldsymbol{\theta})$ の場合に限って考察する．

教師信号と $\boldsymbol{\theta}^*_{l_2}$ をパラメータとしてもつ MLP の出力との差が，平均 0，分散 τ^2 の確率分布に従うものとし，それが $p_*(\boldsymbol{x},y)$ である場合を考える．このとき式 (7) の右辺第1項は

$$E_*[\nabla l_2(y|\boldsymbol{x};\boldsymbol{\theta})]=-E_{q(x)}[(f(\boldsymbol{x};\boldsymbol{\theta}^*_{l_2})-f(\boldsymbol{x};\boldsymbol{\theta}))\nabla f(\boldsymbol{x};\boldsymbol{\theta})] \quad (13)$$

となり，第2項は

$$E_*[l_2(y|\boldsymbol{x};\boldsymbol{\theta})]=E_*\left[\frac{(y^*-f(\boldsymbol{x};\boldsymbol{\theta}^*_{l_2})+f(\boldsymbol{x};\boldsymbol{\theta}^*_{l_2})-f(\boldsymbol{x};\boldsymbol{\theta}))^2}{2}\right]=\frac{\tau^2+E_{q(x)}[(f(\boldsymbol{x};\boldsymbol{\theta}^*_{l_2})-f(\boldsymbol{x};\boldsymbol{\theta}))^2]}{2}$$

より

$$K_*(\boldsymbol{\theta})\nabla r(\boldsymbol{\theta})=E_*\left[\frac{1}{2}-\frac{l_2(y|\boldsymbol{x};\boldsymbol{\theta})}{r(\boldsymbol{\theta})}\right]\nabla r(\boldsymbol{\theta})$$
$$=\left\{\frac{1}{2}-\frac{\tau^2}{2r(\boldsymbol{\theta})}-\frac{E_{q(x)}[(f(\boldsymbol{x};\boldsymbol{\theta}^*_{l_2})-f(\boldsymbol{x};\boldsymbol{\theta}))^2]}{2r(\boldsymbol{\theta})}\right\}\nabla r(\boldsymbol{\theta})$$
$$=\{(r(\boldsymbol{\theta})-\tau^2)-E_{q(x)}[(f(\boldsymbol{x};\boldsymbol{\theta}^*_{l_2})-f(\boldsymbol{x};\boldsymbol{\theta}))^2]\}\frac{\nabla r(\boldsymbol{\theta})}{2r(\boldsymbol{\theta})} \quad (14)$$

となる．ただし，$E_{q(x)}$ は入力の確率分布による期待値を表す．これらから，次のような見解が得られる．

式 (13) は推定モデルの分布の中心を移動させる働きをもち，それ自体は本来の目的である $E_*[l_2(y|\boldsymbol{x};\boldsymbol{\theta})]$ の最小化を行っていることにほかならない．したがって，$r(\boldsymbol{\theta})$ が定数であれば通常の最小2乗法になることは自明である．

式 (14) は $r(\boldsymbol{\theta})=\tau^2$ のとき $E_{q(x)}[(f(\boldsymbol{x};\boldsymbol{\theta}^*_{l_2})-f(\boldsymbol{x};\boldsymbol{\theta}))^2]\geqq 0$ より $K_*(\boldsymbol{\theta})\leqq 0$ であるから分散 $(r(\boldsymbol{\theta}))$ を増加させる効果がある．また，$E_{q(x)}[(f(\boldsymbol{x};\boldsymbol{\theta}^*_{l_2})-f(\boldsymbol{x};\boldsymbol{\theta}))^2]=0$ の場合には，$r(\boldsymbol{\theta})>\tau^2$ ならば $K_*(\boldsymbol{\theta})>0$ より分散を減少させ，$r(\boldsymbol{\theta})<\tau^2$ ならば $K_*(\boldsymbol{\theta})<0$ より分散を増加させる働きをもつ．以上のように，$K_*(\boldsymbol{\theta})$ の符号はデータやモデルに依存して変化し，必ずしも正値をとらない．そのため，本論文で導出した学習アルゴリズムと通常の正則化項付き学習アルゴリズムとの関連性は形式的類似性にとどまる．これは，本論文の立場とベイズ推定の立場を明確に区別する特徴である．

式 (13)，(14) の以上のような挙動は $r(\boldsymbol{\theta})$ と $f(\boldsymbol{x};\boldsymbol{\theta})$ がともに $\boldsymbol{\theta}$ を含んでいるということからそれぞれ独立には行われない．ここで，$r(\boldsymbol{\theta})$ を $\boldsymbol{\theta}$ の関数ではない新たなパラメータで置き換えれば独立な挙動が可能である．しかしこれは，本論文の立場とは本質的に異なる．

5. むすび

本論文では，2乗誤差関数が自明に満たす条件を制限とする損失関数を用いた場合の非ベイズ的付加項をもつ MLP の学習アルゴリズムを導出した．ここで導かれる付加項付き学習アルゴリズムにおける付加項の係数の正負は，データやモデルに依存して変化し，正定数ではないので，通常の正則化項付き学習との関連は形式的類似性にとどまる．その上で，2乗誤差関数に限定した場合は，学習アルゴリズムも簡易に定まることがわかった．更にその付加項は，モデルの分散に関連をもつ項となることがわかった．今後は，付加項の解析や数値実験による有効性の検証を予定している．

文　献

[1] M. Ishikawa, “Structural learning with forgetting,”

Neural Networks, vol.9, no.3, pp.509–521, 1996.
[2] D.J.C. Mackay, “Bayesian interpolation,” Neural Computation, vol.4, pp.415–447, 1992.

付　　録

1. 式 (6)，(7) の導出

$$L(\boldsymbol{x},y;\boldsymbol{\theta})=\log C(\boldsymbol{\theta})+\frac{l(y|\boldsymbol{x};\boldsymbol{\theta})}{r(\boldsymbol{\theta})}-\log q(\boldsymbol{x})$$

より

$$\begin{aligned}&\nabla L(\boldsymbol{x},y;\boldsymbol{\theta})\\&\quad=\nabla\log C(\boldsymbol{\theta})+\frac{\nabla l(y|\boldsymbol{x};\boldsymbol{\theta})}{r(\boldsymbol{\theta})}-\frac{l(y|\boldsymbol{x};\boldsymbol{\theta})\nabla r(\boldsymbol{\theta})}{r(\boldsymbol{\theta})^2}\end{aligned}$$

である．また，

$$\begin{aligned}&\nabla\log C(\boldsymbol{\theta})\\&\quad=\frac{1}{C(\boldsymbol{\theta})}\nabla\int_R\exp\left\{-\frac{l(y|\boldsymbol{x};\boldsymbol{\theta})}{r(\boldsymbol{\theta})}\right\}dy\\&\quad=\frac{1}{C(\boldsymbol{\theta})}\int_R\nabla\exp\left\{-\frac{l(y|\boldsymbol{x};\boldsymbol{\theta})}{r(\boldsymbol{\theta})}\right\}dy\\&\quad=-\frac{1}{C(\boldsymbol{\theta})}\int_R\frac{\nabla l(y|\boldsymbol{x};\boldsymbol{\theta})}{r(\boldsymbol{\theta})}\exp\left\{-\frac{l(y|\boldsymbol{x};\boldsymbol{\theta})}{r(\boldsymbol{\theta})}\right\}dy\\&\qquad+\frac{\nabla r(\boldsymbol{\theta})}{C(\boldsymbol{\theta})}\int_R\frac{l(y|\boldsymbol{x};\boldsymbol{\theta})}{r(\boldsymbol{\theta})^2}\exp\left\{-\frac{l(y|\boldsymbol{x};\boldsymbol{\theta})}{r(\boldsymbol{\theta})}\right\}dy\\&\quad=-\frac{1}{r(\boldsymbol{\theta})}\int_R\nabla l(y|\boldsymbol{x};\boldsymbol{\theta})p(y|\boldsymbol{x};\boldsymbol{\theta})dy\\&\qquad+\nabla r(\boldsymbol{\theta})\times\\&\qquad\quad\int_R\frac{-\log p(y|\boldsymbol{x};\boldsymbol{\theta})-\log C(\boldsymbol{\theta})}{r(\boldsymbol{\theta})}p(y|\boldsymbol{x};\boldsymbol{\theta})dy\end{aligned}\qquad\text{(A·1)}$$

である．ただし，式 (A·1) では微分と積分の順序を適当に入れ替えることができることを仮定した．以上より

$$\begin{aligned}&r(\boldsymbol{\theta})\nabla L(\boldsymbol{x},y;\boldsymbol{\theta})\\&\quad=\nabla l(y|\boldsymbol{x};\boldsymbol{\theta})\\&\qquad+\{H(p(Y|\boldsymbol{x};\boldsymbol{\theta}))+\log p(y|\boldsymbol{x};\boldsymbol{\theta})\}\nabla r(\boldsymbol{\theta})\\&\qquad-E_{p(y|x;\theta)}[\nabla l(y|\boldsymbol{x};\boldsymbol{\theta})]\end{aligned}\qquad\text{(A·2)}$$

が導かれる．ただし，H は（条件付き）エントロピー，$E_{p(y|x;\theta)}$ は $p(y|\boldsymbol{x};\boldsymbol{\theta})$ に関する条件付き期待値を表す．ここで

$$\begin{aligned}&K_e(\boldsymbol{\theta})\\&\quad\stackrel{\text{def}}{=}\frac{1}{n}\sum_{i=1}^n\{H(p(Y|\boldsymbol{x}_i;\boldsymbol{\theta}))+\log p(y_i^*|\boldsymbol{x}_i;\boldsymbol{\theta})\}\\&\quad=\frac{1}{n}\sum_{i=1}^n H(p(Y|\boldsymbol{x}_i;\boldsymbol{\theta}))-\frac{1}{n}\sum_{i=1}^n\log q_e(\boldsymbol{x}_i)\\&\qquad+\frac{1}{n}\sum_{i=1}^n\log p_e(y_i^*|\boldsymbol{x}_i)+\frac{1}{n}\sum_{i=1}^n\log q_e(\boldsymbol{x}_i)\\&\qquad-\frac{1}{n}\sum_{i=1}^n\log\frac{p_e(y_i^*|\boldsymbol{x}_i)q_e(\boldsymbol{x}_i)}{p(y_i^*|\boldsymbol{x}_i;\boldsymbol{\theta})q_e(\boldsymbol{x}_i)}\end{aligned}$$

とおけば

$$\begin{aligned}&r(\boldsymbol{\theta})E_e[\nabla L(\boldsymbol{x},y;\boldsymbol{\theta})]\\&\quad=E_e[\nabla l(y|\boldsymbol{x};\boldsymbol{\theta})]+K_e(\boldsymbol{\theta})\nabla r(\boldsymbol{\theta})\\&\qquad-E_{q_e(x)}[E_{p(y|x;\theta)}[\nabla l(y|\boldsymbol{x};\boldsymbol{\theta})]]\end{aligned}$$

と書ける．ここで，$E_{q_e(x)}$ は $q_e(\boldsymbol{x})$ に関する期待値を表す．このとき，大数の法則より

$$\begin{aligned}\lim_{n\to\infty}K_e(\boldsymbol{\theta})&=H(p(\boldsymbol{X},Y;\boldsymbol{\theta}))-H(p_*(\boldsymbol{X},Y))\\&\qquad-D(p_*(\boldsymbol{X},Y)\|p(\boldsymbol{X},Y;\boldsymbol{\theta}))\\&\stackrel{\text{def}}{=}K_*(\boldsymbol{\theta})\end{aligned}$$

が成り立つ．ここで，学習サンプル $(\boldsymbol{x}_i,y_i^*)$ が互いに独立に同一の確率分布 $p_*(\boldsymbol{x},y)$ に従うという仮定を用いた．また，$D(\cdot\|\cdot)$ はカルバック情報量を表す．

2. 式 (11) の導出

未知のパラメータ $\boldsymbol{\theta}$ を確率変数とみなし，ハイパパラメータ λ をもつ事前確率を $\omega(\boldsymbol{\theta},\lambda)$ とおく．このとき，サンプル D_n に対する $\boldsymbol{\theta}$ の事後確率

$$p(\boldsymbol{\theta}|D_n,\lambda)=\frac{p(D_n,\boldsymbol{\theta},\lambda)}{p(D_n,\lambda)}$$

を最大化することを考える．ただし

$$p(D_n,\boldsymbol{\theta},\lambda)=\omega(\boldsymbol{\theta},\lambda)\prod_{i=1}^n q_e(\boldsymbol{x}_i)p(y_i|\boldsymbol{x}_i;\boldsymbol{\theta})$$

である．このとき $-\log p(D_n,\boldsymbol{\theta},\lambda)$ の最小化により，事後確率の最大化が実現される．ここで

$$\omega(\boldsymbol{\theta},\lambda)=\frac{\exp\{-\lambda r(\boldsymbol{\theta})\}}{\gamma(\lambda)}$$

$$\gamma(\lambda)=\int_R\exp\{-\lambda r(\boldsymbol{\theta})\}d\boldsymbol{\theta}$$

$$l(y|\boldsymbol{x};\boldsymbol{\theta}) = -\log p(y|\boldsymbol{x};\boldsymbol{\theta})$$

とおけば

$$-\log p(D_n,\boldsymbol{\theta},\lambda) = \sum_{i=1}^{n} l(y_i|\boldsymbol{x}_i;\boldsymbol{\theta}) + \lambda r(\boldsymbol{\theta}) + \text{const}$$

となり，定数を除いた部分は式 (11) そのものである．

（平成 11 年 11 月 29 日受付，12 年 1 月 24 日再受付）