4Q-4 評価関数について*

4Q－4

永井隆弘†
大阪大学 基礎工学部 制御工学科‡
tnagai@is.toyo-eng.co.jp§

1 はじめに

人工神経回路網（以下、ニューラルネット）最適化の尺度に用いる評価関数に関する考察を与える.

与えられたデータに対するモデル当てはめの尺度となる評価関数には、通常、何の注意も払われないままに、2乗誤差が用いられる. しかし、その評価関数はそもそもデータの確率分布に対するある仮説から導かれたものであり、実際上、その仮説が正しくない場合も多く、その尺度に関して最適化することが必ずしも最良とはいえない.

小稿は、評価関数本来の起源に立ち返って、その微分構造を解析し、その変形を考察することで、評価関数を統一的に捉えることを試みる [1]. これは、ニューラルネットを産業分野に適用する際に行なうべき評価関数設計の指針となる意義のあるものである.

2 最小2乗法

N 個のデータ点 $(x_i, y_i)_{i=1,\ldots,N}$ をシグモイド関数などの適当な素子関数の類を用いて、重みとオフセットからなるパラメータ W で当てはめることを考える. ニューラルネットでは、x_i が入力データ、y_i が教師データ、$y(x_i)$ が教師データに対する近似出力データとなる. パラメータ W のこのモデルは独立変数 x と従属変数 y の間の関数関係を与える.

$$y(x) = y(x; W) \tag{1}$$

さて、W についての当てはめ値を得るのにどんな評価関数を用いれば良いだろうか？多くは最小2乗誤差、

$$\min_W \sum_i^N [y_i - y(x_i; W)]^2 \tag{2}$$

を用いている. 以下でその起源に立ち返る.

モデルを評価するにはまず、データの確率分布に関する仮説が必要である. そして、y_i が連続値を取り得る時には、各データが $\pm\Delta y$ に入る確率を考える. パラメータ W からそのデータが得られる確率（尤度）を最大にすることによってパラメータ W を当てはめる（最尤推定）. 一般には、データ点 y_i が「真の」モデル $y(x)$ のまわりにそれぞれ独立でランダムな正規分布（Gauss 分布）の測定誤差を持っていると仮定する. さらに各々の正規分布の標準偏差 σ はすべての測定成分で同一とする. このときのデータ組の現われる確率は各々の測定成分発生の確率の積として書ける.

$$P = \prod_{i=1}^N \{\exp[-\frac{(y_i - y(x_i))^2}{2\sigma^2}]\Delta y\} \tag{3}$$

これを最大にするときに、その対数をとって負号をつけたもの

$$[\sum_{i=1}^N \frac{(y_i - y(x_i))^2}{2\sigma^2}] - N \log \Delta y \tag{4}$$

は最小になる. ここで、$N, \sigma, \Delta y$ はすべて定数であるので、この式を最小にすることは (2) と同値である. これが最小2乗法である [2].

3 評価関数の一般形

データがある分布 ρ に従うと仮定すると、

$$P = \prod_{i=1}^N \{\rho(y_i, y(x_i; W)\Delta y\} \tag{5}$$

は、式 (3) の一般形となる. 式 (4) と同様に、対数をとって負号をつけ、

$$E = -\sum_{i=1}^N \log \rho(y_i, y(x_i; W)) \tag{6}$$

を最小化することによって、W の最尤推定量が求められる. したがって式 (6) が評価関数の一般形となる.

評価関数 E の $y(x)$ に関する勾配は、式 (6) より、

$$E_y = -\sum_{i=1}^N \frac{1}{\rho}\frac{\partial \rho}{\partial y} \tag{7}$$

となり、φ 関数を、

$$\varphi(y_i, y(x_i)) = \frac{1}{\rho}\frac{\partial \rho}{\partial y} \tag{8}$$

とおく. 評価関数 E のパラメータ W に関する勾配は、変分原理に基づいて、

$$\begin{aligned} &\frac{\partial E}{\partial W} \\ &= -\sum_{i=1}^N \varphi(y_i, y(x_i)) \star y_x \star x_W \end{aligned} \tag{9}$$

---

*A Study On Estimators.
†Takahiro NAGAI
‡Saburo TSUJI Group, Lab. of Image Information Science and Technology, Dep. of Control Eng., Faculty of Engineering Science, Osaka Univ.
§Toyo Engineering, 2-8-1 Akanehama, Narashino-shi, Chiba, 275

となる. 式(9) の積項のうち$\varphi(y_i, y(x_i))$ はデータ y_iと近似データ $y(x_i)$ の違いの程度（以下、相関）を表す一変数$y(x_i)$の関数である. 評価関数が最小のとき、その点での微分可能性を仮定するとその微分値は0となる. 一般にデータ y_iとモデル $y(x_i)$ が一致する、すなわち相関 = 0 のとき相関関数 $\varphi = 0$ とする（以下、収束条件）.

式(9) に比例係数をかけ、修正量とする方法が従来よく使用される最急降下法である.

4 差分型評価関数

データ y_iと近似データ $y(x_i)$ の相関がそれらの差分（誤差）で決定される差分型評価関数を考える. このとき相関関数φ は、

$$\varphi(y_i, y(x_i)) = \varphi(y_i - y(x_i)) \tag{10}$$

データがp次の指数分布に従うと仮定すると、式(6) より評価関数はいわゆるL_pノルムとなり、相関関数φ は、重みα係数をかけて、

$$\varphi = \alpha |y_i - y(x_i)|^{p-1} sgn(y_i - y(x_i)) \tag{11}$$

となる. 収束条件より実数pの範囲は$p \geq 1$となる. データが正規分布、両側指数分布のときそれぞれ$p = 2, 1$である. 相関関数φ はそれぞれ、

$$\varphi = y_i - y(x_i) \tag{12}$$

$$\varphi = sgn(y_i - y(x_i)) \tag{13}$$

となる. 一般に分布の裾ほど、たくさんのデータを集めないと正規分布に近づかず、たいていそれよりは漸近的にずっと大きい. 以上のものより、さらに広い、したがって時にはさらに現実的な裾を持つものにコーシー分布があり、その場合は、

$$\varphi = \alpha \frac{y_i - y(x_i)}{1 + \beta(y_i - y(x_i))^2/2} \tag{14}$$

となる. これは次節に紹介する 3) の smooth rejection 推定量としても有用であろう.

5 差分型評価関数の微分関数の変形

前節までは、データの確率分布に対する仮説を元に最尤推定となる相関関数を導いた. しかし、これが最良とならない場合がある. ある一定の確率分布にほとんどのデータが従う一方で、ノイズを含んだ外れ値もデータに混入する場合である. 3節で示したように、相関関数φ は相関量に対する一種の重み関数となり、相関の程度によるパラメータ Wの修正量を制御できる関数である. したがって、そういったデータに対するロバスト（頑健）性もたせるには、その外れ値の影響を抑える、もしくは排除するように相関関数を変形すればよい. これはまさにロバスト推定の考え方である. 例えば、正規分布に従うデータの中に外れ値も混入している場合、相関の大きさ $|y_i - y(x_i)|$ に単調に比例して相関量φ を定義するのでなく、ある一定以上（カット位置）の相関 $|y_i - y(x_i)|$ に対し

1. 一定の相関量を与える (hard bounding)
2. 相関量を0とする (hard rejection)
3. 徐々に減少する相関量を与える (smooth rejection)

いずれかのように相関関数$\varphi(z)$ を変形すればよい. 良く知られたこれらの方法以外にも、徐々に飽和するいわば smooth bounding となる関数

$$\varphi = \alpha \frac{1 - \exp(-\beta(y_i - y(x_i))}{1 + \exp(-\beta(y_i - y(x_i))} \tag{15}$$

も可能である.

6 比型評価関数

データ y_iと近似データ $y(x_i)$ の相関は、4, 5節のように一般には差分で決定されるが、それが唯一ではない. 相関としてデータの比

$$\varphi(y_i, y(x_i)) = \varphi(\frac{y_i}{y(x_i)}, \frac{1 - y_i}{1 - y(x_i)}) \tag{16}$$

で決定される比型評価関数を考えることも可能である. ここでは、あらかじめ、データ y_iの有界性 [0,1] を仮定する. 収束条件を考慮して、

$$\varphi = \frac{1 - y_i}{1 - y(x_i)} - \frac{y_i}{y(x_i)} \tag{17}$$

$$\varphi = \log \frac{1 - y_i}{1 - y(x_i)} - \log \frac{y_i}{y(x_i)} \tag{18}$$

が考えられよう. 比型の特徴は、データ y_iの存在限界 0, 1 において無限大の量を持つことである. これは、出力層の素子に存在限界においてその微分値が0となる関数を選んだ時の学習速度減速の欠点を補うことになる.

良く知られているものは、出力層の素子関数にシグモイド関数、相関関数に式(17) を用いた場合であり、特異点が互いに相殺されて、式(9) の積項二項目までは

$$\varphi y_x = y(x) - y \tag{19}$$

となる.

7 おわりに

モデルの近似程度を示す相関関係によって、評価関数は差分型、比型の2種類に分類できる. いずれの場合も、評価関数の設計方針は、データの確率分布に対する仮説、収束条件を満足する相関関数φ の形状、さらにその相関関数と素子の微分関数とのコンビネーション関係からなることを示した.

最後に、小稿であまり触れなかったが注意すべき点を一つあげる. それは、ニューラルネットの特殊性に由来する. ニューラルネットの出力層の素子関数が有界の時、その存在範囲内で、裾の幅の減衰率やカット位置を決定しなければならないことである. 式でいえば$\alpha(y_i, y(x_i))$ と$\beta(y_i, y(x_i))$ の調整である. 外れ値の分布や大きさは、通常、未知であるため、これらは試行錯誤によらざるを得ない. ただ、それら外れ値の少々の違いに敏感とならないような Fuzzy 量を定義するように注意を払うことは重要であろう.

参考文献

[1] 永井隆弘:フィードフォワード型ニューラルネットの動作原理に関する考察. 大阪大学修士学位論文, 1995

[2] NUMERICAL RECIPES in C[日本語版]. 技術評論社, 1993