ニューラルネットワークにおけるジェフリーズの事前分布

本村 陽一 $^{(PY)}$ (motomura@etl.go.jp)
電子技術総合研究所 情報科学部 情報数理研究室

Jeffreys' Prior for Layered Neural Networks

Yoichi MOTOMURA (Electrotechnical Laboratory)

ABSTRACT: We study Jeffreys' Prior for layered neural networks. With this prior for learning and generalization, we can get the parameter transformation under which a shape of a log likelihood function is invariant. We also give the numerical experiments and the transformation formula explicitly based on the numerical results.

1 はじめに

ニューラルネットの学習を確率分布の推定としてとらることがしばしば行われている。この場合ネットワークの結合などが確率分布のパラメータに対応づけられ, 与えられたデータからパラメータ推定を行うことになる。このパラメータ推定を Bayes 的に行う場合にはパラメータの事前分布の選択が一つの問題となる [1][2]。この問題に対して, Bayes 統計学では真のパラメータの変化に対して尤度関数が一様になることが望ましいという考え方にもとづいた Jeffreys' Prior を用いることが提案されている [3]。

本稿では階層型ニューラルネットに対して Jeffreys' Prior を近似的に求めた結果について報告する。

2 定義

ここでは線形の入出力素子がそれぞれ 1 個の 3 層のニューラルネットを考える。ネットワークの出力関数を $f(x)$ として次のように書く。

$$f(x) = \sum_i v_i \frac{1}{\exp(-(w_i x + b_i))} + b_0 \qquad (1)$$

(ここで v_i, b_0は出力層の重みとしきい値,w_i, b_iは中間層の重みとしきい値パラメータ。) このニューラルネット中の可変パラメータをθと書くことにする。このネットワークを次のように定義する x, y空間の上の同時確率分布に対応づける。

$$\begin{aligned} P(x, y|\theta) &= P(y|x, \theta)P(x) \\ &= \frac{1}{\sqrt{2\pi}\sigma} \exp(-\frac{(y - f(x;\theta))^2}{2\sigma^2}) P(x) \qquad (2) \end{aligned}$$

ただし分散パラメータσ^2は既知とし, 以下では $P(x)$ は局所一様分布であるとする。真のパラメータをθ^*とし, $P(x, y|\theta^*)$ からの独立なサンプリングで得られた学習用データを $D_{\theta^*} = \{(x_i, y_i)|i = 1 \cdots n\}$ とする。D_{θ^*}からθを推定することを考え, 通常用いられる対数尤度をとれば,

$$\begin{aligned} L_{\theta^*}(\theta) &= \sum_i \log P(x, y|\theta) \\ &= \sum_i -(y_i - f(x_i;\theta))^2 + C_1 \\ &= \sum_{i=1}^{n} -\{f(x_i;\theta^*) - f(x_i;\theta)\}^2 + C_2 \qquad (3) \end{aligned}$$

となる。(C_2はθによらない定数) したがって真のパラメータがθ^*のもとでの対数尤度 $L_{\theta^*}(\theta)$ は$\theta = \theta^*$で最大となる関数である。$L_{\theta^*}(\theta)$ が真のパラメータとの差 $(\theta - \theta^*)$ だけに依存する関数の場合を尤度関数がθに関して一様であると言う。すなわち $L_{\theta^*}(\theta)$ がθ^*の値にも依存し, データによって関数の形状 ・サイズが大きく変わってしまう場合には一様でない。

3 Jeffreys' Prior

ベイズ統計学において特定の値に重みを置かないという意味で客観的な事前分布を選ぶ手法として Jeffreys のルールが提案されており, これにしたがって決められる事前分布が Jeffreys' Prior と呼ばれている。例えばあるパラメータに関して尤度が一様である場合には局所的一様事前分布が客観的な事前分布になっている。しかし, 一般には必ずしも任意のパラメータについて尤度が一様になるとは限らない。そこでパラメータの一対一の変換 $\theta = g(\phi)$ を導入し, パラメータϕ については対数尤度関数 $L_{\phi^*}(\phi)$ が一様であるようにすることを考える。そのためにまず $L_{\phi^*}(\phi)$ を$\phi^*(= g^{-1}(\theta^*))$ で二次までの Taylor 展開で近似する。明らかに $L_{\phi^*}(\phi^*) = C_2$, $\frac{d}{d\phi} L_{\phi^*}(\phi^*) = 0$。Fisher 情報量を$-\frac{1}{n}\frac{d^2}{d\phi^2} L_{\phi^*}(\phi)|_{\phi=\phi^*} = J_{\phi^*}$と表すことにすると,

$$L_{\phi^*}(\phi) \approx C_2 - J_{\phi^*}(\phi^* - \phi)^2. \qquad (4)$$

したがって $L_{\phi^*}(\phi)$ が一様であるためには, J_{ϕ^*}が定数となれば十分。ここで, $-\frac{1}{n}\frac{d^2}{d\theta^2} L_{\theta^*}(\theta)|_{\theta=\theta^*} = J_{\theta^*}$として,

$$J_{\phi^*} = J_{\theta^*} \cdot \left(\frac{d\theta}{d\phi}\right)^2_{\theta=\theta^*} \qquad (5)$$

だから,

$$\frac{d\theta}{d\phi}_{\theta=\theta^*} \propto \frac{1}{\sqrt{J_{\theta^*}}} \qquad (6)$$

を満たせば J_{ϕ^*}は定数になるので, 尤度関数はパラメータϕに関して一様となる。したがってϕの事前分布 $P(\phi)$

Key Words :学習 ・汎化, 統計的学習, Jeffreys' Prior

の無情報事前分布として一様分布を考えることは妥当である。これをθについて見直すことでθの事前分布を

$$P_j(\theta) \propto P(\phi)\frac{d\phi}{d\theta}|_\theta \propto \sqrt{J_{\theta^*}}|_{\theta^*=\theta} \tag{7}$$

と置けばよい。この $P_j(\theta)$ が Jeffreys' Prior である。

4 ニューラルネットにおける Jeffreys' Prior

ここでは $x \in [-10, 10]$ を一様等間隔に固定して考える。この条件のもとでθ^*を定めると, $L_{\theta^*}(\theta)$, $P_j(\theta)$ を計算することができる。以下ではθとして一つのパラメータをとり, 他のパラメータを固定した一変数のモデルについて考える。しきい値, 出力層の結合をθとした場合は, これについて尤度関数が一様になる。しかし中間層の重み結合については一様にならない。これを計算した結果を図 1に示す。これからθ^*について尤度関数は一様でなく, Jeffreys' Prior も比較的偏った分布になっている。したがって, 初期値として一様ランダムを与えるような場合は wについては偏りを与えていることになる。また一般に BP 学習などの学習では初期値から尤度関数の上を山登り法によって探索するが, 真のパラメータθ^*の値によってこの関数の形状が異なるために収束性がばらつくことが予想される。Jeffreys' Prior を用いることによりこうした問題が解消されることが期待される。しかしニューラルネットの場合, これを解析的に求めることは難しいので, 変換$\theta = g(\phi)$ を数値的に近似して構成する。ニューラルネットの符号対称性 [6] から$\theta \geq 0$ についてのみ考えればよい。$P_j(\theta)$ のプロットからθの定義域を適当に ($\theta \in [0, 3]$) 定め, $1/\sqrt{J_\theta}$をこの範囲で一次近似すると,

$$\frac{1}{\sqrt{J_\theta}} \approx 10.0\theta - 1.0$$

のようにできる。これと 式 (6) から$\frac{d\theta}{d\phi} \approx \alpha(\theta - 0.1)$ と書け (αは任意の定数), この微分方程式を解くと, 定義域の範囲においては任意定数α, βを使って次式を得る。

$$\theta = g(\phi) \approx \beta(e^\alpha)^\phi \tag{8}$$

この変換を用いて尤度関数 $L_{\phi^*}(\phi)$ と $P_j(\phi)$ をプロットしたものが図 2 である。これを図 1と比較すると, 尤度関数は真のパラメータθ^* の値によらずどれも似た形状であり, $P_j(\phi)$ も局所的な領域では幅広く分布している。

5 議論

従来の学習では初期値や事前分布が必要であり一様分布などを仮定することが多い。しかし尤度関数を一様にする意味では Jeffreys' Prior が最適であり, これは中間層パラメータの場合には一様分布には一致していないことがわかった。逆にパラメータを一様に分布させる場合には式 (8) の変換を用いれば Jeffreys' Prior を近似する。例えばこの変換は [6] においてウェーブレット展開から導出した初期値設定 $w = 2^\phi$に一致しており, この場合には BP 法による従来の乱数初期値と比較した実験で学習対象が変わっても比較的安定して収束することを確認している。これは Jeffreys' Prior の良い近似になっているためと考えられる。また最近 Jeffreys' Prior によるベイズ符号化が Min-Max で最適になることが示され [4] 注目されている [5] が, この結果は一般の確率分布についてのベイズリスクを評価することで得られているので, 事前分布を必要とする学習 (Bayes アルゴリズム, Gibbs アルゴリズム [1] [2] など) に Jeffreys' Prior を適用することで汎化能力などが向上することが期待される。これらについての詳細な検討, 実験評価などは今後の課題である。

謝辞

Jeffreys'Prior について教えて頂いた川端電通大助教授, また御助言を頂いた情報数理研究室 梅山室長, 麻生主任研究官, 赤穂, 田中両研究官に感謝いたします。

図 1: Left:$L_{\theta^*}(\theta)$, Right:$P_j(\theta)$

図 2: Left:$L_{\phi^*}(\phi)$, Right:$P_j(\phi)$

参考文献

[1] Haussler D., Kearns M. and Schapire R.:"Bounds on the Sample Complexity of Bayesian Learning Using Information Theory and the VC Dimension", Proceeding of the COLT 4, pp.61-73,(1991).

[2] 麻生 :" 汎化能力の高い学習アルゴリズムについて", 電気学会光応用 ・視覚研究会資料 LAV-93-3 (1993).

[3] Box G. and Tiao G.:"Bayesian Inference in Statistical Analysis", Addison Wesley,(1973).

[4] Clarke B.S. and Barron A.:"Jeffreys' Prior is asymtotically least favorable under entropy risk", submitted to the JSPI, (1992).

[5] 川端 :" ベイズ符号と文脈重みづけ法", 信学技報 IT93-121, (1994).

[6] 本村:" 階層型ニューラルネットワークにおける多重解像度構造による学習能力の改善", 信学技報 NC93-50(1993).