Jeffreys' prior を用いた min-max 最適なランダム結合の神経回路網

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

Min-max optimum Neural Networks using Jeffreys' prior

Yoichi MOTOMURA (Electrotechnical Laboratory)

ABSTRACT: In order to improve reliability and robustness, min-max optimum is hopeful. In statistical modeling, we can use Bayesian mixture model. Recently, it has been proved that Jeffreys' prior on the Bayesian mixture can give min-max optimum in the context of Bayesian coding. In the case of neural networks, we can derive this model also naturally, and some theoretical result can be applied. In this paper, some examples to suit this framework are shown. Finally, several prior are evaluated as experimental result.

1 はじめに

人工神経回路網モデルを現実問題に応用する場合, しばしば対象の複雑性 (モデルのパラメータ数など) が非常に大きいにもかかわらずサンプルを十分採取することができないといった現実的な制約の存在のため, とても少ない情報から無理矢理学習を行わなければいけないことが起こり得る。このような場合には限られた学習サンプルだけから解を探索すると該当する候補が複数存在し, どれが選ばれるかは不定であるため学習結果が不適切なものになる危険性がある。このような時に解の信頼性, ロバスト性を向上するためには損失が最大となる未学習の入力に対する答を最良にする min-max 最適解 (最大損失の最小化) が有効であると考えられる。本研究ではこのような min-max 最適性の評価を目的とする。

単純な確率分布の系列をある事前確率で重みづけて混合 (Mixture) するモデルとして Bayesain mixture がある。最近, 符号化理論においてこの Bayesian mixture で Jeffreys' prior を用いたものが min-max 最適となることが明らかになった [1][2]。

本研究ではまず神経素子を確率分布とした場合の Mixture について KL ダイバージェンスの最悪値を他の事前分布と実験により比較し Jeffreys' prior の min-max 最適性を実験的に評価する。さらにより一般的なモデルに対する min-max 最適性を実現することを目的として Jeffreys' prior をランダム結合の神経回路網 [3] の重み決定に適用しこのネットワークの各素子の出力の平均値を出力とした場合の真のモデルとの間の最小 2 乗誤差の最悪値が他の分布を用いた場合よりも改善されることを実験評価により示す。

2 min-max 最適性

2.1 Bayesian mixture

$p(x;\theta)$ を $\theta \in \Theta$ をパラメータとする x の確率分布とする。この時, θ に関する確率分布を $\pi(\theta)$ として, $p(x;\theta)$ を混合した次のような分布を Bayesian mixture と呼ぶ。

$$m(x;\pi) = \int_{\Theta} p(x;\theta)\pi(\theta)d\theta$$

Key Words :Bayesian mixture, Jeffreys' Prior, ランダム結合の神経回路網

2.2 Jeffreys' prior

真のパラメータ θ^* の分布 $p(x;\theta^*)$ と混合分布 $m(x;\pi)$ の間の距離を KL ダイバージェンス:

$$D(p(x;\theta^*)||m(x;\pi)) = \sum_{i=1}^{n} p(x_i;\theta^*)\log\frac{p(x_i;\theta^*)}{m(x_i;\pi)}$$

により評価する。このとき D が最大となるように θ^* を選んだ場合の $max_{\theta^*}D$(最悪評価) を最小にするような $\pi(\theta)$ の選び方が最近ベイズ符号化の理論によって明らかになった [1][2]。KL ダイバージェンスの漸近評価を θ^* に対する $m(x;\pi)$ の損失として $R_m(\theta^*,\pi)(\equiv \lim_{n\to\infty} D(p(x;\theta^*)||m(x;\pi)))$ を考えると, これの min-max 最適解 $min_{\pi}max_{\theta^*}R(\theta^*,\pi)$ を実現する $\pi(\theta)$ は,

$$\sqrt{I(\theta)}/C_{\pi} \equiv \pi_J(\theta)$$

であることが証明された。(C_{π} は正規化係数, $I(\theta)$ は Fischer 情報量) この π_J は Jeffreys' prior として既に知られていたものである。これにより, Bayesian mixture においては $m(x;\pi_J(\theta))$ が min-max 最適解 (最悪の θ^* に対する損失を最小にする解) となる。

3 神経回路網への適用と実験評価

さて, 以上の結果を神経回路網に適用し, Jeffreys' prior の min-max 最適性を実験的に評価する。

3.1 神経素子モデルの Bayesian mixture

入力を x, 出力を $y = f(x)$ とした神経素子のモデルを次のように書く。

$$f(x) = \frac{1}{1+\exp(-(wx+b))}$$

ここで w, b は重みとしきい値パラメータ。以後可変パラメータを θ と表す。この素子の出力を確率化して実際の出力 y は $f(x)$ を平均値とした正規分布に従って発生すると考える。すなわち, 入力 x のもとでの y の条件付き確率分布 $p(y|x)$ を次のように定義する。

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

(ただしσ^2は適当に定める分散パラメータ) この場合x, yの同時確率分布は $p(x)$ を与えると $p(x, y; \theta) = p(y|x, \theta)p(x)$ となる。

この x のもとでの yに関する条件付き確率分布を基本となるモデルとしてみて, 事前分布$\pi(\theta)$ を用いて k個のθを発生し, 次のように混合した Bayesian mixture

$$m(y|x; \pi) = \sum_{i=1}^{k} \pi(\theta_i) p(y|x; \theta) \qquad (2)$$

を考える。このモデルに対する Jeffreys' prior を実現するために逆関数法により一様乱数を Jeffreys' prior へ変換する方法 [4] を用いて Bayesian mixture のダイバージェンスの最悪値 $max_{\theta^*} R_m(\theta^*, \pi)$ を実験的に評価する*。

(1) 区間 $[-0.1, 0.1]$ の一様乱数, (2) 区間 $[-10, 10]$ の一様乱数, (3)$[-10, 10]$ の一様格子点列と Jeffreys' prior((4) 一様乱数を変換して生成した場合, (5) 一様格子点列を変換して生成した場合) の各事前分布について, θ^*を 0 から 10 まで変化させた場合の $R(\theta^*, \pi)$ の最悪値を $k = 2$ から $k = 100$ までについて計算し, 比較した (図 1) †。

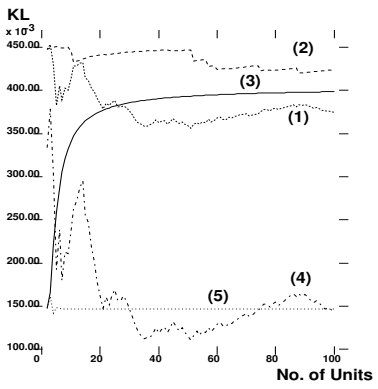

図 1: y:Maximum KL divergence, x:number of units

この結果, 混合分布のサイズ kによらず KL ダイバージェンスの最悪値は Jeffreys' prior を用いた場合 (4),(5) の方が他の事前分布よりも小さなることがわかる。また kが小さい場合, 一様格子点列を用いる方が安定している。

3.2 ランダム結合の神経回路網

N個の素子を並列に接続し, その重みパラメータを確率法則にしたがって決定するモデルとしてランダム結合の神経回路網 [3] がある。このパラメータを事前分布 $\pi(\theta)$ によって定める。各素子の出力の平均値を f_mとすると,

$$f_m(x) = \frac{1}{N} \sum_{i=1}^{N} f(x; \theta_i) \pi(\theta).$$

*ただし b の Fischer 情報量は定数となり Jeffreys' prior は一様分布になるのでここでは特に考慮しない。そこで $b = 0$ に固定し$\theta \equiv w$ についてのみ検討を行ったがここでの結果は任意の b の場合へと自然に拡張できる。またここでは $p(x)$ は局所一様分布に固定して考える

†実験条件:$\sigma = 1, \theta \in [0, 10], x \in [-10, 10]$ として, x, yについては 0.1 刻みで計算を行った。

である。一方, 前述の Mixture の出力の期待値$\bar{y} = \int_y y m(y|x) dy$ であるから, 式 (1),(2) より

$$\bar{y}(x) = \frac{1}{N} \sum_{i=1}^{N} f(x; \theta_i) \pi(\theta_i)$$

となるから f_mの出力は, $\bar{y}(x)$ と一致する。したがって Bayesian mixture の持つ理論的性質はこのモデル f_mに対してもある程度成立することが予想できる。そこで f_mを先の (1)..(5) によって定めた場合の平均 2 乗誤差

$$M.S.E(\theta^*, \pi) \equiv \frac{1}{n} \sum_{i=1}^{n} (f(x_i; \theta^*) - f_m(x_i; \pi))^2$$

について先と同様の条件のもとで実験し確認した (図 2)。

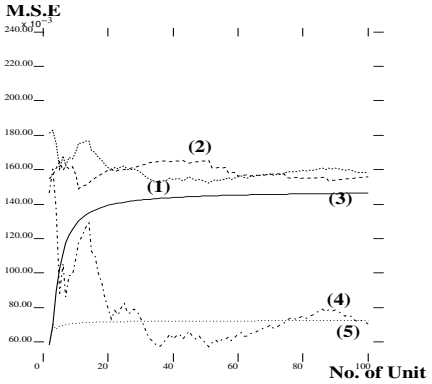

図 2: y:Maximum squared error, x:number of units

この結果, ランダム結合の神経回路網の最小二乗誤差に関しても Jeffreys' prior(4),(5) が最悪評価を小さくできることがわかる。

4 議論

本研究では未学習サンプルに対する損失の最悪値を最小に押えるという観点から学習以前のパラメータの事前分布の問題を扱った。これを正則化の枠組 (Bayesian regularization) によって学習アルゴリズムに適用する方法が考えられる。今後の課題としてこうした手法の有効性を現実的な悪条件のもとで実験的に評価することが必要である。

参考文献

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

[2] 韓, 小林:"情報と符号化の数理", 岩波講座 応用数学 (1994).

[3] 甘利:" 神経回路網の数理", 産業図書,(1978)

[4] Motomura, Y.:"Jeffreys' prior for Layered Neural Networks" ,Proceeding of SPIE, vol.2492,pp.810-819,(1995).