Robustness of Machine Learning Algorithms using Non-Convex Loss Functions

金森 敬文 (名大)　　藤原 秀平 ((株) TOP GATE)　　武田 朗子 (東大)

概要

本研究では，非凸損失を用いたロバストな学習アルゴリズムを提案し，任意の局所解に対して所望の統計的性質が成り立つことを理論的に保証する枠組を提供する．結果の一部は [1] を参照のこと．

1. 非凸損失に基づく学習アルゴリズム

機械学習の分野では，凸損失を用いた計算効率の高い学習法が多く提案されている．例として 2 値判別の学習アルゴリズムを考える．入力 $x \in \mathcal{X}$ に対する 2 値ラベル $y \in \{\pm 1\}$ を予測するとき，データから判別関数 $g : \mathcal{X} \to \mathbb{R}$ を構成し，その符号を予測ラベルとする．C-SVM や ν-SVM などでは，凸損失としてヒンジ損失 $[1 - yg(x)]_+$ を用い，ヒンジ損失の平均値を適当な正則化のもとで最適化することで判別関数を学習する．凸性により効率的なアルゴリズムを設計でき，大規模データにも適応可能になっている．

一方で，非凸関数を用いるほうが統計的に優れた結果が得られる場合もある．外れ値の混入が想定されるデータを扱うとき，凸損失を用いる学習アルゴリズムではロバストな判別器を構成することは一般に困難であり，非凸損失の利用が推奨される．例えばヒンジ損失の代わりにランプ損失 $\min\{[1 - yg(x)]_+, 1\}$ などを用いることで，外れ値にあまり影響を受けない学習アルゴリズムを構成することができる．

しかし非凸損失を用いる学習法では，大域的な最適解を得ることは一般に困難である．実際には局所解が得られることが多いが，理論的な解析では大域的最適解が得られたと仮定することもある．非凸損失を用いる学習法の統計的性質について，数値的には好ましい結果が多く報告されているが，理論的な理解は必ずしも進んでいないのが現状である．

2. 非凸損失とロバスト性

本研究では outlier indicator を用いたロバスト学習法の理論的性質について考察する．まず 2 値判別問題について述べる．2 値ラベルデータ $D = \{(x_i, y_i) | i = 1, \dots, m\} \subset \mathcal{X} \times \{\pm 1\}$ が得られたとき，適切な判別器 $h(x) = \mathrm{sign}(f(x) + b)$ を学習することを考える．ここで f は再生核ヒルベルト空間 (RKHS) $\mathcal{H}$ から選ばれ，$b \in \mathbb{R}$ はバイアス項である．データ (x, y) に対して $y(f(x) + b) > 0$ なら正しく判別していることになる．したがって $y(f(x) + b)$ が平均的に大きな値を取るように学習を進めることで，予測精度の高い判別器が得られると期待される．以上の考察から，ν-SVM では次の凸最適化問題を解くことで判別器を得る：

$$\min_{f \in \mathcal{H}, b, \rho \in \mathbb{R}} \frac{1}{2}\|f\|^2 + \rho\nu + \frac{1}{m}\sum_{i=1}^{m} L_i, \qquad L_i = [\rho - y_i(f(x_i) + b)]_+.$$

ここで ρ はヒンジ損失のしきい値パラメータであり，ν-SVM ではデータに合わせて適応的に調整される．正則化パラメータ $\nu \in (0, 1)$ が大きいほど，自由度の大きなモデルを使うことに対応する．

外れ値が混入しているデータを扱う場合を考える．各データに対して outlier indicator $\eta_i \in \{0, 1\}$ を導入し，外れ値と考えられるデータを無視するような学習法を構成する．ν-SVM に outlier indicator を導入し，非凸最適化問題

$$\min_{f, b, \rho, \eta} \frac{1}{2}\|f\|^2 + \rho\nu + \frac{1}{m}\sum_{i=1}^{m} \eta_i L_i, \quad \text{subject to} \quad (\eta_1, \dots, \eta_m) \in \{0, 1\}^m, \ \sum_{i=1}^{m} \eta_i \geq m(1 - \mu) \tag{1}$$

を解くことで判別器 $\mathrm{sign}(f(x) + b)$ を得る．ここで $\mu \in [0, 1)$ は外れ値の割合の上界である．実際にはパラメータ ν と μ をデータに合わせて適切に設定する必要がある．本研究では，DC(difference of convex functions) 法を用いて上の最適化問題の解を得る．一般には局所解が得られる．学習アルゴリズムのロバスト性について，finite-sample breakdown ponit (FBP) を用いて解析する．

定理 1. DC 法によって得られる (1) の任意の局所解を $f + b \in \mathcal{H} + \mathbb{R}$ とする．2 値ラベルデータ D において，少ないほうのラベルの比率を r_D とする．このとき $0 \leq 2\mu < \nu < 2(r_D - 2\mu)$ なら $f + b$ の FBP は μ 以上，さらに RKHS $\mathcal{H}$ のカーネル関数が有界なら $0 < \nu < 2(r_D - 2\mu)$ に対して $f + b$ の FBP は μ 以上になる．$f + b$ が大域的最適解なら FBP について等号が成り立つ．

上の結果は，観測データから $m\mu$ 個のデータを除いて学習するとき，FBP の下限が μ になるための十分条件を示している．関数部分 f の大域解については必要十分条件になることも導出できる．また，大域的最適解よりも局所解のほうが FBP の意味でロバストになることを示している．定理 1 の結果を使うと，交差検証法を用いてグリッド探索で ν, μ を決めるとき，探索範囲を適切に制約することができる．既存の結果として，位置パラメータに対する並進共変的推定量では FBP の上限は 1/2 となることが知られているが，定理の条件 ($\mu < r_D/2$) は判別問題での対応する結果と言える．2 値判別だけでなく 1-クラス，多値判別でも局所解に対して同様の結果が得られる．ロバスト化したサポートベクトル回帰では，大域的最適解の FBP に対して同様の結果が得られるが，局所解に関しては今後の課題である．

References

[1] T. Kanamori, S. Fujiwara, & A. Takeda, Breakdown Point of Robust Support Vector Machines, arXiv:1409.0934.