教師なしデータ学習のためのブースティング

予測発見戦略研究センター 遺伝子多様性グループ
江口 真透

1. はじめに

ゲノム・オミックスデータの統計解析において“$p \gg n$”問題が課題となり幾つかの新しい展開がなされつつある．パターン認識の予測問題を中心に考えてきたが，この発表では教師なしデータの解析について焦点を当て機械学習のアプローチを考える．特にブースティングの枠組みで密度推定，正規混合モデル，独立成分分析について高次元データの場合でも解析可能な方法を提案したい．

単調増加で凸の実関数 $U(t)$ を生成関数として採用する．この $U(t)$ の導関数を $u(t)$，$u(t)$ の逆関数を $\xi(s)$ とし，以後この 3 つ組 (U, u, ξ) を混同して使うことにする [1]．p 次元ユークリッド空間上の有限密度関数の全体を $\mathcal{M} = \{f(x) : f(x) \geq 0, \int f(x)dx < \infty\}$ と表す．このとき，$\mathcal{M}$ の関数 f の関数 g に対する U クロスエントロピーは

$$C_U(g, f) = -\mathrm{E}_g\xi(f(x)) + \int U(\xi(f(x)))dx \tag{1.1}$$

と定義される．ここで E_g は g に関する期待値を表す．

2. ブースト学習アルゴリズム

ブースト学習に使う辞書を $\mathcal{D} = \{g_\theta(x) \in \mathcal{M} : \theta \in \Theta\}$ としよう．辞書 $\mathcal{D}$ の元 g_j たちを使って

$$f^*(x) = \xi^{-1}\Big(\sum \pi_j\xi\big(g_j(x)\big)\Big) \tag{2.1}$$

を考えよう．ここで π_j は非負で $\sum_j \pi_j = 1$ を満たすとする．自明な関係 $\xi^{-1}(\xi(\mathcal{D})) = \mathcal{D}$ に対して，(2.1) は $\mathcal{D}$ を拡大した空間 $\xi^{-1}(\mathrm{co}(\xi(\mathcal{D})))$ を考えていることになる．ここで co は凸包を表す．

データ $\{x_1, \cdots, x_n\}$ が与えられたとき (1.1) の経験版として与えられる U ロス (汎) 関数

$$\bar{L}_U(f) = -\frac{1}{n}\sum_{i=1}^{n}\xi(f(x_i)) + \int U(\xi(f(x)))dx$$

を使って次のアルゴリズムを考える．例えば，モデル $\{g_\theta(x)\}$ に対して生成関数 $U(t) = \exp(t)$ と選べば U ロス関数 $\bar{L}_U(g_\theta)$ は負の対数尤度関数に他ならない．

(A). 初期推定量を $f_1 = \mathrm{argmin}_{g\in\mathcal{D}}\,\bar{L}_U(g)$ とする．

(B). 自然数 $k = 1, \cdots$ に対して f_k が与えられたとき

$$(\alpha_{k+1}, g_{k+1}) = \underset{(\alpha,g)\in(0,1)\times\mathcal{D}}{\mathrm{argmin}}\ \bar{L}_U(\xi^{-1}((1-\alpha)\xi(f_k) + \alpha\xi(g))) \tag{2.2}$$

を求めて $f_{k+1} = \xi^{-1}\big((1-\alpha_{k+1})\xi(f_k) + \alpha_k\xi(g_{k+1})\big)$ と更新する．

(C). 適当に決められた K に対して，$\widehat{f}(x) = \xi^{-1}((1-\alpha_K)\xi(f_{K-1}(x)) + \alpha_K\xi(g_K(x)))$ と定義する．

注意 1. 最終的に (C) で得られた $\widehat{f}(x)$ は (2.1) の形で表される．実際，

$$\widehat{f}(x) = \xi^{-1}\Big(\sum_{k=1}^{K}\pi_k\xi(g_k(x))\Big)$$

となる．ここで $\pi_k = \alpha_k \prod_{\ell=k+1}^{K}(1-\alpha_\ell)$．ただし $\alpha_1 = 1$ とする．

注意 2. ステップ (B) の式 (2.2) の右辺の関数は次のように分解される．

$$\bar{L}_U(\xi^{-1}((1-\alpha)\xi(f_k)+\alpha\xi(g))) = (1-\alpha)\bar{L}_U(f_k)+\alpha\bar{L}_U(g)-\Delta_U(f_k,g,\alpha). \tag{2.3}$$

ここで

$$\Delta_U(f,g,\alpha) = \int \big\{(1-\alpha)U(\xi(f))+\alpha U(\xi(g)) - U\big((1-\alpha)\xi(f)+\alpha\xi(g)\big)\big\}$$

と定義する．実関数 U の凸性から任意の $\alpha \in (0,1)$ に対して $\Delta_U(f,g,\alpha) \geq 0$ となり，等号は $f = g$ のときにのみ成立する．ステップ (B) の逐次更新式は (2.3) 式の分解から次のような理解が得られる．辞書 $\mathcal{D}$ の元 g は単にロス関数 $\bar{L}_U(\cdot)$ を小さくするだけでなくて，同時に $\Delta(f_k,\cdot,\alpha)$ を大きくすることで現ステップ f_k と隔るように選ばれる．このように $\mathcal{D}$ の中で選ばれた集合 $\{g_j\}_{j=1}^{k}$ に対応して新たな候補 g_{k+1} が選ばれる．

3. ガウシアン・ミクスチュアー

平均ベクトル μ, 分散行列 V を持つ p 次元ガウス密度関数を $g(x,\mu,V)$ と書き，辞書 $\mathcal{D}$ を $\{g(x,\mu,V) : (\mu,V) \in \mathbb{R}^p \times \mathbb{S}^p\}$ とする．ここで $\mathbb{S}^p$ を p 次半正定値行列全体とする．生成関数 $U(t) = \frac{1}{2}t^2$ による上記のアルゴリズムは k ミクスチュアー $f_k(x)$ に対して $(\alpha_{k+1},\mu_{k+1},V_{k+1})$ を

$$\underset{(\alpha,s\mu,sV)}{\operatorname{argmin}} \Big\{ -\frac{1}{n}\sum_{i=1}^{n}\{(1-\alpha)f_k(x_i)+\alpha g(x_i,\mu,V)\} + \frac{1}{2}\int\{(1-\alpha)f_k+\alpha g(\cdot,\mu,V)\}^2 \Big\} \tag{3.1}$$

と定めて, $(k+1)$ ミクスチュアー $f_{k+1}(x) = (1-\alpha_{k+1})f_k(x)+\alpha_{k+1}g(x,\mu_{k+1},V_{k+1})$ に更新する．(3.1) 式の括弧内の第 2 項の働きで多くの場合 (μ_{k+1},V_{k+1}) は $\{(\mu_j,V_j)\}_{j=1}^{k}$ の凸包の外で求まることになる．このように各々のステップで新たなガウス密度成分が探索される．データの分布が無限ガウシアン・ミクスチュアーで書き下されるならば十分大きなステップ数 K を選べば $f_K(x)$ はデータの分布を十分近似していることが示される [2,5]．生成関数 $U(t) = (1+\beta t)^{\frac{\beta+1}{\beta}}$ に対しては次の β バージョンのガウシアン・ミクスチュアーの議論が可能となる．

$$f^*(x) = \Big(\sum_j \pi_j g(x,\mu_j,V_j)^\beta\Big)^{\frac{1}{\beta}}.$$

4. 検討すべき課題

このような方向に関連する研究として [3,4] があったがブースティングの考えは取り入れられてなく，単に U ロス関数の局所最小化を逐次的に初期値を変えることによって局所構造を求めていた．今回の考察ではブースティング，または forward stagewise な方法で大域的な U ロス関数の最小化が可能になっている．今後，検討すべき課題として，U ロス関数でどの U が良いか？ 過学習を防ぐための K を決め方，適切な辞書 $\mathcal{D}$ の選択，ステップサイズ α_k の他の決め方などが挙げられる．

参考文献

[1] Eguchi, S. (2008). Chapter 13, Springer.
[2] Klemela, J. (2007). Machine Learning 67, 3, 169-195
[3] Mollah, N. H., Sultana N., Minami M. and Eguchi, S. (2010). Neural Networks 23, 2, 226-238.
[4] Mollah, N. H., Minami, M. and Eguchi, S. (2006). Neural Computation, 18, 1, 166-190.
[5] Naito, K. and Eguchi, S. In preparation.