2017年6月16日　統計数理研究所　オープンハウス

最適な半教師付き学習

藤澤　洋徳　　数理・推論研究系　教授

はじめに

半教師付き学習において，教師なしデータの有効利用が幾つか提案されている．本発表では，最適性の観点から，教師なしデータの有効利用を考える．得られた結果から自然に，過去に提案されていない，教師なしデータの新しい有効利用法が見つかる．

問題設定

教師付きデータ

$$\begin{pmatrix} x_1 \\ y_1 \end{pmatrix}, \ldots, \begin{pmatrix} x_1 \\ y_1 \end{pmatrix} \sim g_{x,y}(x,y)$$

教師なしデータ

$$x_{n+1}, \ldots, x_N \sim g_x(x)$$

半教師付きデータ

上記を合わせたもの

教師付き学習　（u_βの代表例はスコア）

$$\hat{\beta}_0 = \arg\operatorname{solve}\left\{ \beta : \sum_{i=1}^{n} u_\beta(x_i, y_i; \beta) = 0 \right\}$$

目的：教師なしデータを上手く利用して，推定量$\hat{\beta}$を改良する．

半教師付き学習

Kawakita (unpublished, DRESS I)　$s_\eta(x;\eta) = (\partial/\partial\eta)\log g_x(x;\eta)$

$$\hat{\beta}_K = \arg\operatorname*{solve}_\beta \left\{ \beta : \sum_{i=1}^{n} \frac{g_x(x_i;\hat{\eta}')}{g_x(x_i;\hat{\eta})} u_\beta(x_i, y_i; \beta) = 0 \right\}$$

$$\hat{\eta} = \arg\operatorname{solve}\left\{ \eta : \sum_{i=1}^{n} s_\eta(x_i;\eta) = 0 \right\}$$

$$\hat{\eta}' = \arg\operatorname{solve}\left\{ \eta' : \sum_{i=n+1}^{N} s_\eta(x_i;\eta') = 0 \right\}$$

REMARK：密度比の極限は1である．意味がないように見える．しかし実際には漸近分散の意味で$\hat{\beta}_K$は$\hat{\beta}_0$を改良している．

Kawakita and Takeuchi (2014, DRESS II)

$$\hat{\beta}_{KT} = \arg\operatorname{solve}\left\{ \beta : \sum_{i=1}^{n} \frac{g_x(x_i;\hat{\eta}'_N)}{g_x(x_i;\hat{\eta})} u_\beta(x_i, y_i; \beta) = 0 \right\}$$

$$\hat{\eta}'_N = \arg\operatorname{solve}\left\{ \eta : \sum_{i=1}^{N} s_\eta(x_i;\eta) = 0 \right\}$$

Kawakita and Kanamori (2013) (nDRESS I)

密度比を$w(x;\theta) = \exp\{\theta^T\psi(x)\}$によって適当に推定.（推定方法は省略）

$$\hat{\beta}_{KK} = \arg\operatorname*{solve}_\beta \left\{ \beta : \sum_{i=1}^{n} w(x_i;\hat{\theta}) u_\beta(x_i, y_i; \beta) = 0 \right\}$$

目的：過去の提案手法は密度比に主眼を置いている．本発表では，密度比という観点からではなく，ある種の最適性から手法を提案する．

Kawakita and Fujisawa (2017) (nDRESS II)　$w(x;\theta) = \exp\{\theta^T\psi(x)\}$

$$\hat{\beta}_{KF} = \arg\operatorname{solve}\left\{ \beta : \sum_{i=1}^{n} w(x_i;\hat{\theta}) u_\beta(x_i, y_i; \beta) = 0 \right\}$$

$$\hat{\theta} = \arg\operatorname{solve}\left\{ \theta : \frac{1}{n}\sum_{i=1}^{n} \psi(x_i) w(x_i;\theta) - \frac{1}{N}\sum_{i=1}^{N} \psi(x_i) = 0 \right\}$$

最適な半教師付き学習

仮定：$\hat{\beta}$は漸近線形展開をもち正則であるとする．

漸近線形展開：

$$\sqrt{n}(\hat{\beta} - \beta_0) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n} \phi_1(x_i, y_i) + \frac{1}{\sqrt{N-n}}\sum_{i=n+1}^{N} \phi_2(x_i) + o_p(1).$$

ただし$\beta_0 = \arg\operatorname{solve}\left\{\beta : \mathrm{E}_{g_{x,y}}[u_\beta(x,y;\beta)] = 0\right\}$とする．ただし核関数$\phi_1(x,y)$と$\phi_2(x)$は適当な性質を満たすとする．

正則性：適当なセミパラメトリック理論の意味で正則 (Tsiatis 2006).（超有効推定量を排除するための性質．詳細は省略）.

定理．　核関数は次の性質を満たす：

$$\phi_1(x,y) = J_{\beta\beta}^{-1} u_\beta(x,y;\beta_0) - \frac{1}{\sqrt{r}}\phi_2(x).$$

ただし

$$r = \frac{N-n}{n}, \qquad J_{\beta\beta} = \mathrm{E}_{g_{x,y}}\left[-\frac{\partial u_\beta}{\partial\beta}(x,y;\beta_0)\right],$$

とする．

定理．　$\hat{\beta}$の漸近分散を最小にするのは次のときである：

$$\phi_2(x) = \frac{\sqrt{r}}{1+r} J_{\beta\beta}^{-1} \mathrm{E}_{g_{y|x}}[u_\beta(X,Y;\beta_0)|X=x].$$

REMARK: この定理から，最適な推定量を作るためには，条件付密度関数$g(y|x)$に関する期待値を得る必要がある．しかし，これは，半教師付きデータからは推定しにくい．

定理．　$\phi_2(x) = B\psi(x)$と制限されているとする．このクラスの中で漸近分散を最小にする$\hat{\beta}$では以下が成り立つ：

$$\phi_2(x) = \frac{\sqrt{r}}{1+r} J_{\beta\beta}^{-1} \mathrm{E}_{g_{x,y}}\left[u_\beta(x,y,;\beta_0)\psi(x)^T\right] \mathrm{E}_{g_x}\left[\psi(x)\psi(x)^T\right]^{-1} \psi(x).$$

REMARK: 上記の定理の中にある$\phi_2(x)$は半教師付きデータから推定可能である．そのような推定量$\hat{\beta}$を与える推定方程式は，自然に導出できる．その一つが nDRESS II である．

REMARK: 上記の定理の性質をもつ推定量$\hat{\beta}$を与える推定方程式は，その他にもたくさん作ることができる．