Minimalizacja ryzyka strukturalnego, podejście Vapnika

Wykład IV

Wisła, grudzień 2009

Minimalizacja ryzyka strukturalego

Problem klasyfikacji dla dwóch klas, $g = 2$.

$$L(f(\mathbf{x}), y) = I\{f(\mathbf{x}) \neq y\}.$$

Załóżmy, że dana jest rodzina $\mathcal{C}$ potencjalnych klasyfikatorów

$$\phi \in R^p \qquad \phi : R^p \to \{-1, 1\}.$$

(indeks klasy $Y = \pm 1$). Na przykład $\mathcal{C} = \{\phi_{\mathbf{a}} : \ \phi_{\mathbf{a}}(\mathbf{x}) = sign(\mathbf{a}'\mathbf{x})\}$

Problem: Ocena błędu predykcji dla reguły minimalizującej ryzyko empiryczne (estymator błędu przez powtórne podstawienie).

Przykład

$p(x|-1) \sim U[0,1]$ i $p(x|1) \sim U[1,2]$

Minimalne ryzyko reguły (ryzyko bayesowskie równa) się 0.

$A_1, \ldots, A_n$ obserwacje z populacji pierwszej

$B_1, \ldots, B_n$ obserwacje z populacji drugiej.

LDA Fishera w tym przypadku redukuje się do:

klasyfikuj do drugiej populacji gdy $X > T = (\bar{A}_n + \bar{B}_n)/2$

$$Err_{\mathcal{U}} = \frac{1}{2}(P(X > T|Y = -1) + P(X < T|Y = 1) = |T - 1|/2$$

$$Err = E|T - 1|/2 \sim E|N(0, \frac{1}{24n})| = \frac{1}{4\sqrt{\pi \times n}}$$

Szacujemy blad, jaki popełnimy, wybierając klasyfikator ϕ_n^* minimalizujący estymator błędu przez powtórne podstawienie (ryzyko empiryczne)

$$\bar{err}(\phi) = n^{-1} \sum_{i=1}^{n} I\{\phi(\mathbf{X}_i) \neq Y_i\}.$$

$$\phi_n^* = \operatorname{argmin}_{\phi \in \mathcal{C}} \bar{err}(\phi).$$

Ile różni się

$$Err_{\mathcal{U}}(\phi_n^*) = P(\phi_n^*(\mathbf{X}^0) \neq Y^0 | \mathcal{U})$$

od minimalnego błędu w klasie wszystkich klasyfikatorów $\inf_\phi Err(\phi)$.

$$Err_{\mathcal{U}}(\phi_n{}^*) - \inf_{\phi} Err(\phi) =$$

$$\underbrace{(Err_{\mathcal{U}}(\phi_n^*) - \inf_{\phi \in \mathcal{C}} Err(\phi))}_{\text{błąd estymacji}} + \underbrace{(inf_{\phi \in \mathcal{C}} Err(\phi) - \inf_{\phi} Err(\phi))}_{\text{błąd aproksymacji}}$$

Zwiększając rodzinę $\mathcal{C}$ zwiększamy (zmniejszamy) błąd estymacji (aproksymacji).
Zmniejszając rodzinę $\mathcal{C}$ zmniejszamy (zwiększamy) błąd estymacji (aproksymacji).

Błąd estymacji $\tilde{\phi} = \operatorname{argmin}_{\phi \in \mathcal{C}} Err(\phi)$

$$Err_{\mathcal{U}}(\phi_n^*) - Err(\tilde{\phi}) \leqslant Err_{\mathcal{U}}(\phi_n^*) \underbrace{-\bar{err}(\phi_n^*) + \bar{err}(\tilde{\phi})}_{\geqslant 0} - Err(\tilde{\phi}) \leqslant$$

$$\underbrace{Err_{\mathcal{U}}(\phi_n^*) - \bar{err}(\phi_n^*)} + \underbrace{\bar{err}(\tilde{\phi}) - Err(\tilde{\phi})} \leqslant 2 \sup_{\phi \in \mathcal{C}} |Err(\phi) - \bar{err}(\phi)| \qquad (*)$$

Ponadto

$$Err_{\mathcal{U}}(\phi_n^*) \leqslant \bar{err}(\phi_n^*) + \sup_{\phi \in \mathcal{C}} (Err(\phi) - \bar{err}(\phi)) \qquad (**)$$

Problem $(**)$: szacowanie błędu predykcji $Err_{\mathcal{U}}(\phi_n^*)$ w terminach $\bar{err}(\phi_n^*)$.

$$\sup_{\phi \in \mathcal{C}} (Err(\phi) - \bar{err}(\phi)) = \sup_{f \in \mathcal{F}} (Pf - P_n f)$$

$$\mathbf{Z} = (\mathbf{X}, Y),\ \mathbf{Z}_i = (\mathbf{X}_i, Y_i)$$

$f(\mathbf{z}) = I\{\phi(\mathbf{x}) \neq y\}$, $Pf = Ef(\mathbf{Z})$, $P_n f = n^{-1} \sum_{i=1}^{n} f(\mathbf{Z}_i)$.

$\mathcal{F}$: rodzina funkcji f utworzona na podstawie $\mathcal{C}$.

Nierówność koncentracyjna McDiarmida

$g : X^n \to R$:

$$\sup_{x_1,\ldots,x_n,x_i' \in X} |g(x_1,\ldots,x_n) - g(x_1,\ldots,x_{i-1},x_i',x_{i+1},\ldots,x_n)| \leqslant c \quad 1 \leqslant i \leqslant n$$

to dla $Z = g(X_1,\ldots,X_n)$, gdzie $X_1,\ldots,X_n$- niezależne,

$$P(Z - EZ > t) \leqslant exp(-2t^2/nc^2) \qquad (!)$$

Dla $Z = \sup_{f\in\mathcal{F}} Pf - P_n f$, $c = 2/n$ i (!) implikuje

Z prawdopodobieństwem $> 1 - \delta$

$$\sup_{f \in \mathcal{F}} [Pf - P_n f] \leqslant E\Big(\sup_{f \in \mathcal{F}} [Pf - P_n f] \Big) + \sqrt{\frac{2 \log(1/\delta)}{n}}$$

Średnie Rademachera zbioru $A \subset R^n$

$$R_n(A) = E\left\{ \sup_{\mathbf{a}} \frac{1}{n} \sum_{i=1}^{n} \sigma_i a_i \right\},$$

$\mathbf{a} = (a_1, \ldots, a_n)$ i σ_i nzl Rademachera.

Własności średnich Rademachera

- (i)
$$R_n(A) = R_n(\mathrm{absconv}(A)),$$
gdzie $\mathrm{absconv}(A) = \{\sum_{j=1}^{N} c_j \mathbf{a}^{(j)}, \sum_{j=1}^{N} |c_j| \leqslant 1, \mathbf{a}^{(j)} \in A\}$

- (ii)
$$R_n(\phi \circ A) \leqslant L_\phi R_n(A),$$
gdzie $\phi : R \to R$ Lipschitzowska ze stałą L_ϕ, a
$\phi \circ A = \{(\phi(a_1), \ldots, \phi(a_n)), \mathbf{a} = (a_1, \ldots, a_n) \in A\}$

- (iii) $A = \{\mathbf{a}_1, \ldots, \mathbf{a}_m\}$ to $R_n(A) \leqslant \max_{i=1,\ldots,m} ||\mathbf{a}_i|| \frac{\sqrt{2 \log m}}{n}$

$(\mathbf{Z}'_1, \ldots, \mathbf{Z}'_n)$ niezależna kopia $(\mathbf{Z}_1, \ldots, \mathbf{Z}_n)$ i $P'_n f = n^{-1} \sum_{i=1}^n f(\mathbf{Z}_i)$.

Technika sprzęgania daje

$$E\Big(\sup_{f\in\mathcal{F}}[Pf - P_n f]\Big) = E(\sup_{f\in\mathcal{F}} E[P'_n f - P_n f | \mathbf{Z}_1, \ldots, \mathbf{Z}_n]) \leqslant E \sup_{f\in\mathcal{F}} E[P'_n f - P_n f]$$

$$= E \sup_{f\in\mathcal{F}} n^{-1} \sum_{i=1}^n \sigma_i (f(Z'_i) - f(Z_i)) \leqslant 2ER_n(\mathbf{Z}^n)$$

Nasz zbiór A: $\mathcal{F}(\mathbf{z}^n) = \{(f(\mathbf{z}_1), \ldots, f(\mathbf{z}_n))\}_{f\in\mathcal{F}}$, $(iii) \Rightarrow$

$$R_n(\mathbf{Z}^n) \leqslant \sqrt{\frac{2 \log S_{\mathcal{F}}(\mathbf{Z}^n)}{n}},$$

gdzie $S_{\mathcal{F}}(\mathbf{z}^n) = |\mathcal{F}(\mathbf{z}^n)|$, $\mathbf{z}^n = (\mathbf{z}_1, \mathbf{z}_2, \ldots, \mathbf{z}_n)$.

Jeśli $V(\mathbf{z}^n)$ wymiar Vapnika-Czervonenkisa zbioru $S_{\mathcal{F}}(\mathbf{z}^n)$ to

$$\log S_{\mathcal{F}}(\mathbf{z}^n) \leqslant V(\mathbf{z}^n) \log(n+1)$$

i

$$R_n(\mathbf{Z}^n) \leqslant \sqrt{\frac{2V(\mathbf{Z}^n)\log(n+1)}{n}},$$

Jeśli $\sup_{\mathbf{z}^n} V(\mathbf{z}^n) \leqslant V$ to z prawdopodobieństwem $> 1-\delta$

$$\sup_{f\in\mathcal{F}}[Pf - P_n f] \leqslant C\sqrt{\frac{V\log n}{n}} + \sqrt{\frac{2\log(1/\delta)}{n}}$$

Czynnik $\log n$ można pominąć.

Minimalizacja ryzyka strukturalnego

Rodziny klasyfikatorów $\mathcal{C}_1 \subset \mathcal{C}_2 \ldots \subset \mathcal{C}_N$

$$(1) \qquad Err_{\mathcal{U}}(\phi^*_{n,i}) \leqslant \bar{err}(\phi^*_{n,i}) + B(n, \mathcal{F}_i) \qquad i = 1, \ldots, n$$

Minimalizacja prawej strony (1)
Problemy: (i) Klasy z małym wymiarem mogą mieć duży błąd aproksymacji (ii) minimalizacja $\bar{err}(\phi)$ często problem NP-trudny
(ii) nierówność

$$Err_{\mathcal{U}}(\phi^*_n) \leqslant \bar{err}(\phi^*_n) + \sup_{\phi \in \mathcal{C}} (Err(\phi) - \bar{err}(\phi)) \qquad (**)$$

za gruba !

Oszacowania oparte na marginesach

Niech ϕ będzie regułą klasyfikacyjną oparta na funkcji klasyfikacyjnej d:
$\phi(\mathbf{x}) = 1$ jeśli $d(\mathbf{x}) \geqslant 0$ i $\phi(\mathbf{x}) = -1$ w przeciwnym przypadku.
Wtedy

$$P(\phi(\mathbf{X}) \neq Y) = P(\mathrm{sign}(d(\mathbf{X})) \neq Y) \leqslant E\, I\{d(\mathbf{X})Y \leqslant 0\}$$

Niech $\gamma : R \to R_+$ będzie funkcją kosztu taką, że $\gamma(x) \geqslant I\{x \geqslant 0\}$.
Typowe przykłady

$$\gamma(x) = e^x, \quad \gamma(x) = (1+x)_+, \quad \gamma(x) = \log_2(1+x)$$

Funkcjonał kosztu i jego wersja empiryczna

$$A(d) = E(\gamma(-d(\mathbf{X})Y)|\mathcal{U}), \qquad A_n(d) = n^{-1}\sum_{i=1}^{n}\gamma(-d(\mathbf{X}_i)Y_i)$$

Mamy

$$Err(d) \leqslant A(d) \qquad \bar{err}_n(d) \leqslant A_n(d).$$

Niech $B = ||\gamma(-d(\mathbf{x})y))||_\infty$. Wtedy

$$Err_{\mathcal{U}}(d_n) \leqslant A(d_n) = E(\gamma(-d_n(\mathbf{X})Y)|\mathcal{U})) \leqslant$$

$$A_n(d_n) + \sup_{d \in \mathcal{D}}(A(d) - A_n(d)) \leqslant$$

$$A_n(d_n) + 2L_\gamma ER_n(\mathcal{F}(\mathbf{Z}^n)) + B\sqrt{\frac{2\log\frac{1}{\delta}}{n}},$$

L_γ - stała Lipschitza funkcji γ.

Systemy głosowania

Systemy głosowania opierają się na podejmowaniu decyzji na podstawie kombinacji liniowych klasyfikatorów. $\mathcal{C}$- rodzina bazowych klasyfikatorów.

$$\mathcal{D}_\lambda = \left\{ d(\mathbf{x}) = \sum_{i=1}^{m} c_j g_j(\mathbf{x}) : m \in N, \sum_{j=1}^{n} |c_j| \leqslant \lambda, g_1, \ldots, g_n \in \mathcal{C} \right\}$$

Na podstawie własności średnich Rademachera zbioru

$$R_n(\mathcal{D}_\lambda(\mathbf{Z}^n)) = \lambda R_n(\mathcal{C}(\mathbf{Z}_n)) \leqslant \lambda \sqrt{\frac{2 V_\mathcal{C} \log(n+1)}{n}}$$

Stąd

$$Err_{\mathcal{U}}(\mathrm{sign}(d_n^*)) \leqslant A_n(d_n^*) + 2L_\phi \lambda \sqrt{\frac{2V_{\mathcal{C}} \log(n+1)}{n}} + B\sqrt{\frac{2\log\frac{1}{\delta}}{n}},$$

gddzie $\mathrm{sign}(d_n^*)$ minimalizuje ryzyko empiryczne w klasie $\{\mathrm{sign}(d)\}_{d\in\mathcal{D}}$
Oszacowanie błedu predykcji $\mathrm{sign}(d_n^*)$ w terminach wymiaru Vapnika-Chervonenkisa klasy bazowej.

Marginesy

Przykład

$$\phi_\eta(x) = 1 \quad x \leqslant -\eta$$

$$\phi_\eta(x) = 0 \quad x \geqslant 0$$

$$\phi_\eta(x) = 1 + x/\eta \quad \text{w p.p.}$$

W tym przypadku $B = 1$ i $L_\phi = 1/\eta$. Ponadto $A_n(d) \leqslant \bar{err}_n^\gamma(d)$, gdzie

$$\bar{err}_n^\gamma(d) = \frac{1}{n} \sum_{i=1}^{n} I\{d(\mathbf{X}_i) Y_i \leqslant \eta\}.$$

Liczymy nie tylko źle sklasyfikowane elementy, ale również te sklasyfikowane dobrze, ale z małym 'marginesem'

$$Err_{\mathcal{U}}(d_n) \leqslant \bar{err}_n(d_n) + \frac{2\lambda}{\eta}\sqrt{\frac{2V_{\mathcal{C}}\log(n+1)}{n}} + \sqrt{\frac{2\log\frac{1}{\delta}}{n}}.$$

Inne ważne kierunki/perspektywy

- Własności skończeniepróbkowe estymatorów PMS i ich ryzyka (Barron, Birgé, Massart, Barraud ..)
- Metody rzadkiej regresji (Lasso, Least Angle Regression LAR,...). potencjalna liczba predyktorów w modelu liniowym $p \approx n$, $p > n$, jednakże $\#\{i : \beta_i \neq 0\} << p$
- Związki metod selekcji z wielokrotnym testowaniem, w szczególności z procedurami typu Benjaminiego-Hochberga $(M_1 \subset M_2 \Rightarrow \hat{M}_{AIC} = M_2 \Longleftrightarrow LRT_{M_1,M_2} > 2(p_2 - p_1))$
- metody selekcji w modelach regresyjnych oparte na wstępnym uporządkowaniu predyktorów.
- metody predykcji oparte na uśrednianiu modeli

Własności skończeniepróbkowe

$$Y_i = f(x_i) + \varepsilon_i, \qquad i = 1, \ldots, n$$

$\varepsilon_i \sim (0, \sigma^2)$, σ^2-znane, $f; [0,1] \to R$, $x_i \in [0,1]$.
M_m: funkcje stałe na odcinkach $[(j-1)/m, j/m)$.
$f \in Holder(L, \alpha)$, to dla $\hat{f} = \hat{f}_{MNK}(M_m)$

$$R_n = E||f - \hat{f}||_n^2 \leqslant L^2 M^{-2\alpha} + \frac{m\sigma^2}{n}$$

optymalny wybór m:

$$m(\alpha, L) = \left(\frac{L}{\sigma}\right)^{2/(1+\alpha)} m^{1/(2\alpha+1)}$$

Ryzyko $R_n = O(n^{-2\alpha/(2\alpha+1)})$. Konstrukcja estymatorów PMS o takim rzędzie ryzyka, *bez* wykorzystywania wiedzy, że $f \in Holder(L, \alpha)$. Kalibracja stałych w funkcji kryterialnej !!!

Referencje

- Bousquet, O., Boucheron, S., Lugosi, G. (2005) Theory of classification: a survey of recent results, ESAIM: Probability and Statistics, 323-375
- Claeskens, G., Hjort, N. (2008) Model selection and model averaging, Cambridge University Press
- Friedman, J. (1997) On bias, variance, 0/1-loss and the curse of dimensionality, Data mining and Knowledge Discovery, str. 55-77
- Hastie, T., Tibshirani, R., Friedman, J. (2009) The Elements of Statistical Learning, (wydanie drugie) http://www-stat.stanford.edu/ tibs/ElemStatLearn/
- Konishi, S., Kitagawa, G. (2008) Information Criteria and Statistical Modelling, Springer
- Leeb, H., Pötscher, B. (2008) Sparse estimators and the oracle property, or the return of Hodges estimator, J. Econometrics, str. 201-211
- Leeb, H., Pötscher, B. (2008) Model selection,w Handbook of Financial Time Series (wyd Andersen,Davis,Kreiss,Mikosch), Springer,2008

- Nishii, R. (1984) Asymptotic properties of criteria for selection of variables in multiple regression, An. Statist., str. 758-765.
- Schiavo, R., Hand, D.(2000) Ten more years of error rate research, International Statistical Review, str. 295-310.
- Shao, J. (1993) Linear model selection by cross-validation, JASA, str. 486-494.
- Shao, J. (1997) An asymptotic theory for linear model selection, Statistica Sinica, str. 221-264
- Yang, Y. (2005) Can the strengths of AIC and BIC be shared? A conflict between model identification and regression estimation, Biometrika, str. 937-950.