Estymacja błędu predykcji dla logarytmicznej funkcji straty. Predykcja w modelu liniowym

Wykład II

Wisła, grudzień 2009

Plan

- Tw. Takeuchiego: obciążenie $\bar{err}$ w przypadku logarytmicznej funkcji straty.
- Błąd 0-1 a błąd L^2 w problemie klasyfikacji.
- Predykcja w modelu liniowym.

Model (iii)

$Y_1, \ldots, Y_n$ prosta próba losowa $\sim f$, $\hat{\theta}(Y_1, Y_2, \ldots, Y_n)$ estymator NW w modelu $\mathcal{M} = \{f_\theta\}_{\theta \in \Theta}$.

Zauważmy, że ($\mathcal{U} = \{Y_1, \ldots, Y_n\}$ i $L(y, \hat{\theta}) = -2 \log f_{\hat{\theta}}(y)$)

$$-2E_f \log f_{\hat{\theta}}(Y^0) = E_f(L(Y^0, \hat{\theta})|\mathcal{U}) = Err_{\mathcal{U}} = Err_{in}$$

Skonstruujemy asymptotycznie nieobciążone estymatory Err w oparciu o $\bar{err}$.

$$\bar{err} = \frac{1}{n}\sum_{i=1}^{n} L(Y_i, \hat{\theta}_{NW}) = -\frac{2}{n}\sum_{i=1}^{n} \log f_{\hat{\theta}_{NW}}(Y_i) = -2\log \mathcal{L}(\hat{\theta}_{NW})/n$$

Przybliżenie $E(\bar{err})$.

Rzut informacyjny f na $\mathcal{M} = \{f_\theta\}_{\theta\in\Theta}$.

θ^* minimalizuje odległość Kullbacka-Leiblera f od $\mathcal{M} = \{f_\theta\}_{\theta\in\Theta}$.

$$\theta^* = \operatorname{argmin}_{\theta\in\Theta} E_f \log(f/f_\theta) = \operatorname{argmin}_{\theta\in\Theta}\{-2E_f \log f_\theta(Y^0)\} =$$

$$\operatorname{argmin}_{\theta\in\Theta}\{-E_f L(Y^0, \theta)\}.$$

Twierdzenie 2. (Takeuchi (1976)) Przy odpowiednich warunkach regularności na $M = \{f_\theta\}_{\theta \in \Theta}$

$$Err - E(\bar{err}) = \frac{2}{n} tr\{I(\theta^*)J(\theta^*)^{-1}\} + o\left(\frac{1}{n}\right),$$

gdzie

$$I(\theta^*) = E_f\left\{ \frac{\partial \log f_\theta(Y^0)}{\partial \theta} \frac{\partial \log f_\theta(Y^0)}{\partial \theta'}_{|\theta=\theta^*} \right\}$$

$$J(\theta^*) = -E_f\left\{ \frac{\partial^2 \log f_\theta(Y^0)}{\partial \theta \partial \theta'}_{|\theta=\theta^*} \right\}$$

Idea:

$$Err - E \log f_{\theta^*}(Y^0) \approx E \log f_{\theta^*}(Y^0) - E(\bar{err}) \approx \frac{1}{n} tr\{I(\theta^*)J(\theta^*)^{-1}\}.$$

Heurystyka dowodu.

Porównanie

$$nE_{\mathbf{Y}_n}(\bar{err}) = E_{\mathbf{Y}_n} \log f_{\hat{\theta}_n}(\mathbf{Y}_n) = (1)$$

z

$$nErr = nE_{\mathbf{Y}_n}(Err_{\mathcal{U}}) = nE_{\mathbf{Y}_n} E_f \log f_{\hat{\theta}_n}(Y^0) = (2)$$

$$(1) - (2) = E_{\mathbf{Y}_n}\Big\{ \log f_{\hat{\theta}_n}(\mathbf{Y}_n) - \log f_{\theta^*}(\mathbf{Y}_n)\Big\} +$$

$$+ E_{\mathbf{Y}_n}\Big\{ \log f_{\theta^*}(\mathbf{Y}_n) - nE_f \log f_{\theta^*}(Y^0)\Big\}$$

$$+ E_{\mathbf{Y}_n}\Big\{ nE_f \log f_{\theta^*}(Y^0) - E_f \log f_{\hat{\theta}_n}(Y^0)\Big\} = D_1 + D_2 + D_3.$$

$$D_2 = 0, \qquad D_2 = \frac{1}{2} tr\{I(\theta^*)J(\theta^*)^{-1}\} + o(1) = D_3$$

Analiza D_1 i D_3 oparta na rozwinięciu Taylora. Niech $\log \mathcal{L} = \ell$

$$\ell(\theta) - \ell(\hat{\theta}_{NW}) = \frac{1}{2}(\theta - \hat{\theta}_{NW})' \frac{\partial^2 \ell(\tilde{\theta})}{\partial\theta\partial\theta'}(\theta - \hat{\theta}_{NW})$$

Stąd

$$\ell(\theta^*) - \ell(\hat{\theta}_{NW}) \approx -\frac{n}{2}(\theta^* - \hat{\theta}_{NW})' J(\theta^*)(\theta^* - \hat{\theta}_{NW})$$

$$D_1 \approx \frac{n}{2} E_{\mathbf{Y}_n}(\theta^* - \hat{\theta}_{NW})' J(\theta^*)(\theta^* - \hat{\theta}_{NW}) =$$

$$\frac{n}{2} E_{\mathbf{Y}_n}(tr(J(\theta^*)(\theta^* - \hat{\theta}_{NW})(\theta^* - \hat{\theta}_{NW})') \approx \frac{1}{2} tr(I(\theta^*)J(\theta^*)^{-1})$$

$$(nE(\hat{\theta}_{NW} - \hat{\theta}^*)(\hat{\theta}_{NW} - \theta^*)' \to J(\theta^*)^{-1} I(\theta^*) J(\theta^*)^{-1}).$$

Podobnie $D_3 = \frac{1}{2} tr(I(\theta^*)J(\theta^*)^{-1}) + o(1)$.

Jeśli $f \in M$ $(f = f_{\theta^*})$ to przy warunku

$$\int \frac{\partial^2 f_\theta(x)}{\partial\theta\partial\theta'}_{|\theta=\theta^*} dx = \frac{\partial^2}{\partial\theta\partial\theta'}\left\{ \int f_\theta(x)\, dx \right\}_{|\theta=\theta^*} = 0$$

$$I(\theta^*) = J(\theta^*) \qquad (*).$$

W przypadku błędnej specyfikacji modelu parametrycznego niekoniecznie równość !
W przypadku $(*)$ as. nieobciążony estymator Err ma postać

$$\bar{err} + \frac{2p}{n} = -2\log L(\hat{\theta}_{NW}) + \frac{2p}{n}$$

(kryterium Akaikego (AIC)).
W ogólnym przypadku (Kryterium Takeuchiego (TIC))

$$\bar{err} + \frac{2tr\, I(\hat{\theta}_{NW}) J(\hat{\theta}_{NW})^{-1}}{n}$$

Model (iv)

Analogiczna własność do tezy tw. Takeuchiego jest często używana w przypadku modelu (iv)

$$Err_{in} = E_{\mathbf{Y}^0}\left(\frac{1}{n}\sum_{i=1}^{n} -2\log f_{\hat{\theta}(\mathbf{X}_i)}(Y_i^0)\right),$$

gdzie Y_i^0 niezależnie generowane z rozkładu $P_{Y|\mathbf{X}=\mathbf{X}_i}$. Wtedy

$$E_{\mathbf{Y}}(Err_{in}) \approx E_{\mathbf{Y}}\left(\frac{1}{n}\sum_{i=1}^{n} -2\log f_{\hat{\theta}(\mathbf{X}_i)}(Y_i)\right) + \frac{2p}{n}$$

brak formalnego dowodu ..

$$E_{\mathbf{Y}}(Err_{in}|\mathbf{X}) \approx \frac{-2}{n} E_{\mathbf{Y}}\Big(\sum_{i=1}^{n} \log f_{\hat{\theta}(\mathbf{X}_i)}(Y_i)\Big) + \frac{2p}{n} \quad (*)$$

W przypadku modelu liniowego ze znanym σ^2 nieobciążony estymator prawej strony $(*)$ ma postać

$$\frac{RSS}{n\sigma^2} + \frac{2p}{n} + C$$

(σ^2 zastępowane przez nieobciążony estymator)
W przypadku modelu liniowego z nieznanym σ^2 nieobciążony estymator prawej strony $(*)$ ma postać

$$\log\Big(\frac{RSS}{n}\Big) + \frac{2p}{n} + C$$

.

Przykład Zbiór `mjap` zawiera rekordy dotyczące temperatury (minimalna temperatura stycznia uśredniona z lat 1971–2000), długości, szerokości geog. oraz wysokości 26 miast japońskich. Interesuje nas wybór zbioru zmiennych prognozującego temperaturę.

```
 Temp Latitude Longitude Altitude

1   -7.6   45.413   141.683      2.8
2   -7.7   43.057   141.332     17.2
...............
26  14.3   26.203   127.688     28.1
```

```
g=lm(Temp~1,data=mjap)  # model początkowy

wprzod=step(g,direction="forward",scope=
list(upper=.~.+Latitude + Longitude + Altitude))
Start:  AIC=88.51
Temp ~ 1

             Df Sum of Sq    RSS    AIC
+ Latitude    1    576.69 147.72  49.17
+ Longitude   1    518.12 206.29  57.85
+ Altitude    1     93.93 630.48  86.90
<none>                    724.41  88.51

Step:  AIC=49.17
Temp ~ Latitude
```

Wartość AIC wynosi dla modelu ze stałą 88.5, największa redukcja AIC przy dodaniu `Latitude`. Ją wybieramy. Zauważmy, że dołączenie zmiennej `Altitude` spowodowałoby minimalną zmianę AIC (z 88.51 do 86.90).

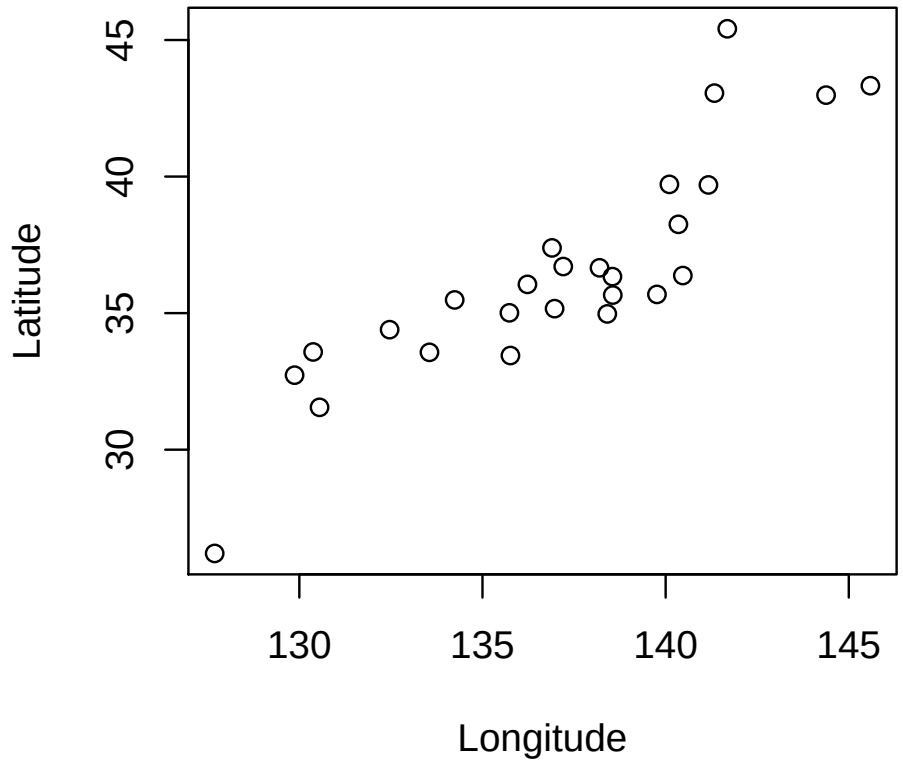
```
             Df Sum of Sq     RSS     AIC
+ Altitude    1    95.098  52.625  24.332
+ Longitude   1    12.791 134.931  48.813
<none>                    147.722  49.168

Step:  AIC=24.33
Temp ~ Latitude + Altitude

             Df Sum of Sq    RSS    AIC
<none>                    52.625 24.332
+ Longitude   1     0.840 51.785 25.914
```

W drugim kroku wybieramy jednak `Altitude`, mimo jej pozornie małej przydatności. Model `Temp` $\sim$ `Latitude + Altitude` daje redukcję AIC do 24.3, dodanie `Longitude` zwiększa AIC więc nie dokładamy jej do modelu.

Dlaczego na drugim kroku wybrana została `Altitude` zamiast `Longitude`? Współliniowość `Longitude` i `Latitude`... Kierunek wysp NE – SW.



Przeszukanie wszystkich podzbiorów i użycie BIC daje w tym przypadku ten sam wybór zmiennych.

Ogólna postać funkcji kryterialnej z karą w modelu regresyjnym

$$-2\sum_{i=1}^{n}\log f_{\hat{\theta}_{NW}}(Y_i|\mathbf{X}_i) + pC_n$$

Kryterium Akaike: $C_n = 2$

Kryterium Schwarza $C_n = \log n$.

Kryterium Schwarza: standardowa motywacja oparta o podejście bayesowskie:
$M_1, \ldots, M_M$, parameter θ_m modelu M_m o gęstości $p(\theta_m|M_m)$. Wtedy

$$P(M_m|\mathcal{U}) \sim P(M_m)P(\mathcal{U}|M_m) \sim$$

$$P(M_m) \int P(\mathcal{U}|M_m, \theta_m)p(\theta_m|M_m)\, d\theta_m.$$

Dla regularnej gęstości $p(\theta_m|M_m)$

$$\log P(\mathcal{U}|M_m) = \log P(\mathcal{U}|\hat{\theta}_{NW}, M_m) - \frac{p_m}{2} \log n + 0(1)$$

Przy jednostajnym rozkładzie apriori na modelach $P(M_m) = 1/M$ i pominięciu członu $0(1)$ dostajemy kryterium Schwarza. Inne podejście: Pokarowski, JM (09) -kryterium MPVC.

Problem klasyfikacji: strata 0-1 vs strata L^2

Niech $f(\mathbf{x}) = P(Y = 1|\mathbf{X} = \mathbf{x})$ i $d_B(\mathbf{x}) = I\{f(\mathbf{x}) > 1/2\}$,

$$\operatorname{argmin}_{\hat{d}} Err(\mathbf{x}_0) = d_B(\mathbf{x}_0) \quad \text{reguła bayesowska}$$

$$\hat{d}_B(\mathbf{x}) = I\{\hat{f}(\mathbf{x}) > 1/2\}$$

empiryczna reguła bayesowska. Czy dobra estymacja f w sensie średniokwadratowym przekłada się na niską wartość Err ?

$$E(\hat{f}(\mathbf{X^0}) - f(\mathbf{X^0}))^2 = E_{\mathbf{X}^0}\{\operatorname{Var}(\hat{f}(\mathbf{X}^0)) + (E\hat{f}(\mathbf{X}^0) - f(\mathbf{X}^0))^2\}$$

Przykład (ESL, str. 226)

$(Y, \mathbf{X}) \in \{0,1\} \times R^{20}$, $\mathbf{X}$- r. jednostajny na $[0,1]^{20}$

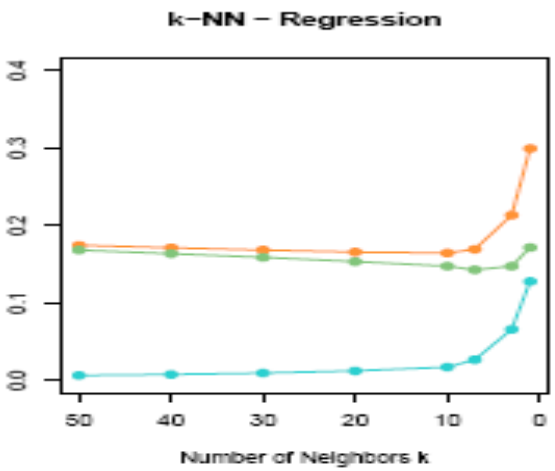
(i) $Y = 0$ jeśli $X_1 \leqslant 1/2$, $Y = 1$ w p.p.;

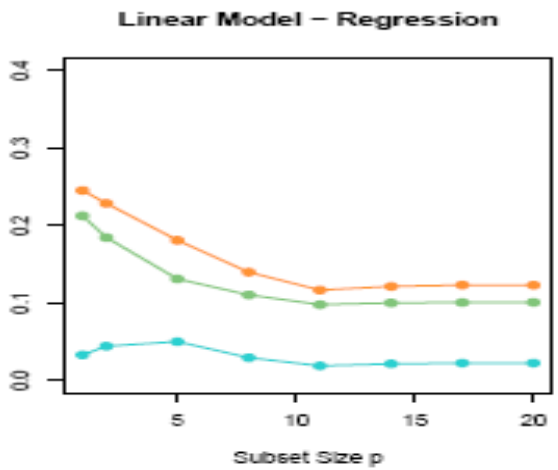
(ii) $Y = 0$ jeśli $\sum_{i=1}^{10} X_i \leqslant 5$, $Y = 1$ w p.p.;
Metody estymacji: regresja liniowa (najlepszy podzbiór predyktorów danej liczności) i metoda kNN.
Związane klasyfikatory: $\hat{f}(\mathbf{x}) > 0.5 \Rightarrow \hat{d}(\mathbf{x}) = 1, = 0$ w p.p.

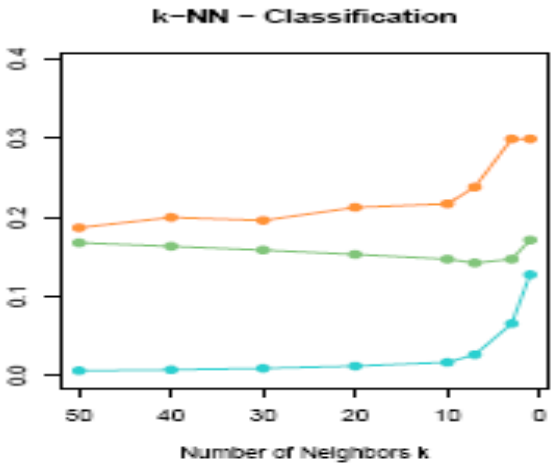
Błąd średniokwadratowy lub blad klasyfikacji

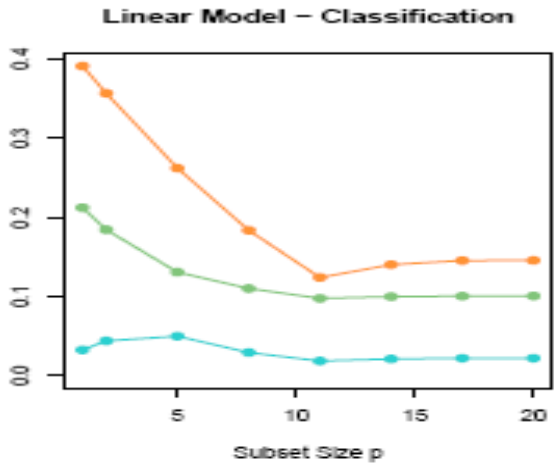
kwadrat obciążenia $\hat{f}(\mathbf{x})$

wariancja $\hat{f}(\mathbf{x})$


k−NN − Regression
0.0
0.1
0.2
0.3
0.4
50
40
30
20
10
0
Number of Neighbors k


Linear Model − Regression
0.0
0.1
0.2
0.3
0.4
5
10
15
20
Subset Size p


k−NN − Classification
0.0
0.1
0.2
0.3
0.4
50
40
30
20
10
0
Number of Neighbors k


Linear Model − Classification
0.0
0.1
0.2
0.3
0.4
5
10
15
20
Subset Size p

Dla dobrej reguły klasyfikacji, ważne, aby $\hat{f}(\mathbf{x}) > 1/2$, gdy $f(\mathbf{x}) > 1/2$,

Wartość $(\hat{f}(\mathbf{x}) - f(\mathbf{x}))^2$ nie ma znaczenia !

$$Err(\mathbf{x}_0) = E(L(Y^0, \hat{d}(\mathbf{X}^0))|\mathbf{X}^0 = \mathbf{x}_0) = P(Y^0 \neq \hat{d}(\mathbf{X}^0)|\mathbf{X}^0 = \mathbf{x}_0)$$

dla funkcji straty 0-1.
Niech $f(\mathbf{x}) = P(Y = 1|\mathbf{X} = \mathbf{x})$ i $d_B(\mathbf{x}) = I\{f(\mathbf{x}) > 1/2\}$

$$\text{argmin}_{\hat{d}} Err(\mathbf{x}_0) = d_B(\mathbf{x}_0) \quad \text{reguła bayesowska}$$

$$\hat{d}_B(\mathbf{x}) = I\{\hat{f}(\mathbf{x}) > 1/2\}$$

empiryczna reguła bayesowska.
Błąd empirycznej reguły bayesowskiej vs reguły bayesowskiej ?

$$Err_{d_B}(\mathbf{x}_0) = (1 - f(\mathbf{x}_0))I\{f(\mathbf{x}_0) > 1/2\} + f(\mathbf{x}_0)I\{f(\mathbf{x}_0) \leqslant 1/2\}$$

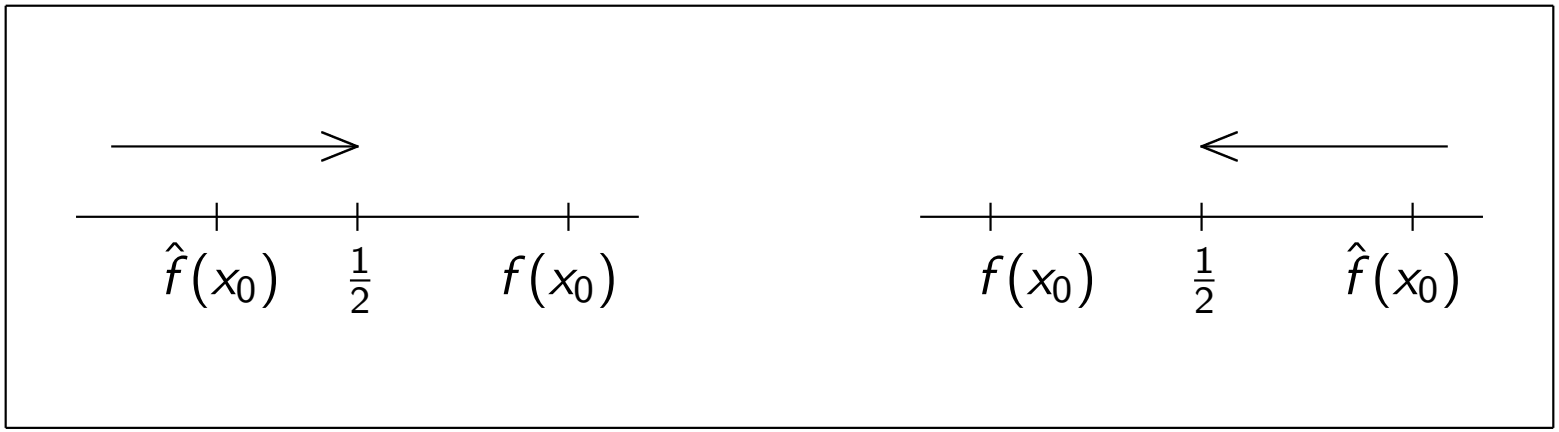
$$Err_{\hat{d}_B}(\mathbf{x}_0) = (1 - f(\mathbf{x}_0))P(\hat{f}(\mathbf{x}_0) > 1/2) + f(\mathbf{x}_0)P(\hat{f}(\mathbf{x}_0) \leqslant 1/2)$$

$$\Downarrow$$

$$Err_{\hat{d}_B}(\mathbf{x}_0) = Err_{d_B}(\mathbf{x}_0) + |2f(\mathbf{x}_0) - 1| P(\hat{d}_B(\mathbf{x}_0) \neq d_B(\mathbf{x}_0) | \mathbf{X^0} = \mathbf{x}_0)$$

Zachowanie $P(\hat{d}_B(\mathbf{x}_0) \neq d_B(\mathbf{x}_0)) \quad (*)$?

Załóżmy, że $\hat{f}(\mathbf{x}_0) \sim N(E\hat{f}(\mathbf{x}_0), \mathrm{Var}(\hat{f}(\mathbf{x}_0)))$.



Wtedy

$$P(\hat{d}(\mathbf{x}_0) \neq d(\mathbf{x}_0)|\mathbf{X}^0 = \mathbf{x}_0) = \Phi\left(\frac{(\text{sign}(1/2 - f(\mathbf{x}_0))(E\hat{f}(\mathbf{x}_0) - 1/2)}{(\text{Var}(f(\mathbf{x}_0))^{1/2}}\right)$$

Jeśli $(\mathrm{sign}(1/2 - f(\mathbf{x}_0))(E\hat{f}(\mathbf{x}_0) - 1/2) > 0$ to zmniejszenie wariancji $(\mathrm{Var}(f\mathbf{x}_0))$ zmniejsza błąd,

natomiast jeśli

$(\mathrm{sign}(1/2 - f(\mathbf{x}_0))(E\hat{f}(\mathbf{x}_0) - 1/2) < 0$ to zwiększenie wariancji $(\mathrm{Var}(f\mathbf{x}_0))$ zmniejsza bład !

Predykcja w modelu liniowym

Do końca wykładu II: $L(y, \hat{f}(\mathbf{x})) = (y - \mathbf{x})^2$.
W modelu liniowym możemy bezpośrednio liczyć $E(Err_{in}|\mathbf{X}))$ dla modelu opartego o wektor predyktorów $\mathbf{x}_\alpha$ będący podwektorem wektora wszystkich predyktorów $\mathbf{x}$.
Zakładamy, że zachodzi $Y = \mathbf{X}'\beta + \varepsilon$

$$E_Y(Err_{in}|\mathbf{X}) = (n + p_\alpha)\sigma^2 + \beta'\mathbf{X}(I - \mathbf{H}_\alpha)\mathbf{X}\beta =: (1) \quad + (2)$$

gdzie p_α to wymiar wektora $\mathbf{X}_\alpha$ i $\mathbf{H}_\alpha = \mathbf{X}_\alpha(\mathbf{X}'_\alpha\mathbf{X}_\alpha)^{-1}\mathbf{X}'_\alpha$.

Jeśli mniejszy model dobrze wyspecyfikowany, to (2)=0.

Model liniowy: kryteria CV i GCV

W sytuacji gdy $\hat{\mathbf{Y}} = \mathbf{S}\mathbf{Y}$, gdzie $\mathbf{S} : R^n \longrightarrow R^n$ jest przekształceniem liniowym.
Dla modelu liniowego ($\mathbf{S} = \mathbf{H}$) (i wielu innych $\mathbf{S}$)

$$Y_i - \hat{Y}_{i(i)} = \frac{Y_i - \hat{Y}_i}{1 - s_{ii}},$$

gdzie $\mathbf{S} = (s_{ij})$. Umożliwia liczenie funkcji CV (n-krotnej kroswalidacji) bez konieczności n-krotnego dopasowywania modelu.

$$CV = n^{-1} \sum_{i=1}^{n} \left(\frac{Y_i - \hat{Y}_i}{1 - s_{ii}} \right)^2$$

GCV: aproksymacja CV: $s_{ii} \leftarrow tr(\mathbf{S}/n)$

$$GCV = n^{-1} \sum_{i=1}^{n} \left(\frac{Y_i - \hat{Y}_i}{1 - tr(\mathbf{S}/n)} \right)^2$$

Związek GCV z AIC

$$GCV = \frac{RSS}{n(1 - tr(\mathbf{S})/n)^2} \approx \frac{RSS}{n} + 2p\frac{RSS}{n} = \frac{RSS}{n} + 2p\hat{\sigma}^2$$

Zamieniając $\hat{\sigma}^2$ przez $\hat{\sigma}_p^2$ (obliczony na podstawie pełnego modelu) dostajemy kryterium AIC (z dokładnością do stałej).

Konserwatywność CV

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \quad \boldsymbol{\beta} \in R^p.$$

2^{p-1} podmodeli odpowiadających wszystkim podwektorom wektora predyktorów

$$\mathbf{Y} = \mathbf{X}_\alpha \boldsymbol{\beta}_\alpha + \boldsymbol{\varepsilon}, \quad \boldsymbol{\beta} \in R^{p_\alpha}.$$

M_{min}: minimalny model liniowy, taki, że $P \in M_{min}$. Jeśli

(i) $\mathbf{X}'\mathbf{X} = 0(n)$ i $\mathbf{X}'\mathbf{X}^{-1} = 0(n)$

(ii) Dla każdego α $\lim_{n\to\infty} \max_{i\leqslant n} (\mathbf{H}_\alpha)_{ii} = 0$

to

$$\lim_{n\infty} P(M_{CV} \supset M_{min}) = 1.$$

Jednocześnie dla $M_\alpha \supset M_{min}$ i $M_\alpha \neq M_{min}$

$$P(M_{CV} = M_{min}) = P(2(p_\alpha - p_{\alpha_{min}})\sigma^2 < \varepsilon'(\mathbf{H}_\alpha - \mathbf{H}_{\alpha_{min}})\varepsilon) + o(1)$$

Estymator $E(Err_{in}|\mathbf{X})$ metodą jacknife

$$\theta = E_{\mathbf{Y}}(Err_{in}|\mathbf{X}) = n^{-1}\sum_{i=1}^{n} E_{\mathbf{Y},\mathbf{Y^0}}(Y_i^0 - \hat{Y}_i)^2|\mathbf{X})$$

estymowane za pomocą $\hat{\theta}_n = \bar{err} = RSS/n$.
Estymator typu jacknife na podstawie $\hat{\theta}_n$. Ogólna konstrukcja dla $\hat{\theta}_n$, takich, że

$$E\hat{\theta}_n - \theta = a_1(\theta)/n + 0(n^{-2}), \qquad (*)$$

($\bar{err}$ spełnia $(*)$ na podstawie twierdzenia podstawowego)

$$\hat{\theta}_{J,n} = n\hat{\theta}_n - (n-1)\sum_{i=1}^{n}\hat{\theta}_{-i}/n$$

$$E\hat{\theta}_{J,n} = \theta + 0(n^{-2}).$$

$$\hat{\theta}_{J,n} = n\hat{\theta}_n - (n-1)\sum_{i=1}^{n} \hat{\theta}_{-i}/n =$$

$$RSS - n^{-1}\sum_{i=1}^{n}\sum_{j\neq i}(Y_j - \hat{Y}_{j(i)})^2$$

Pokarowski (2009)

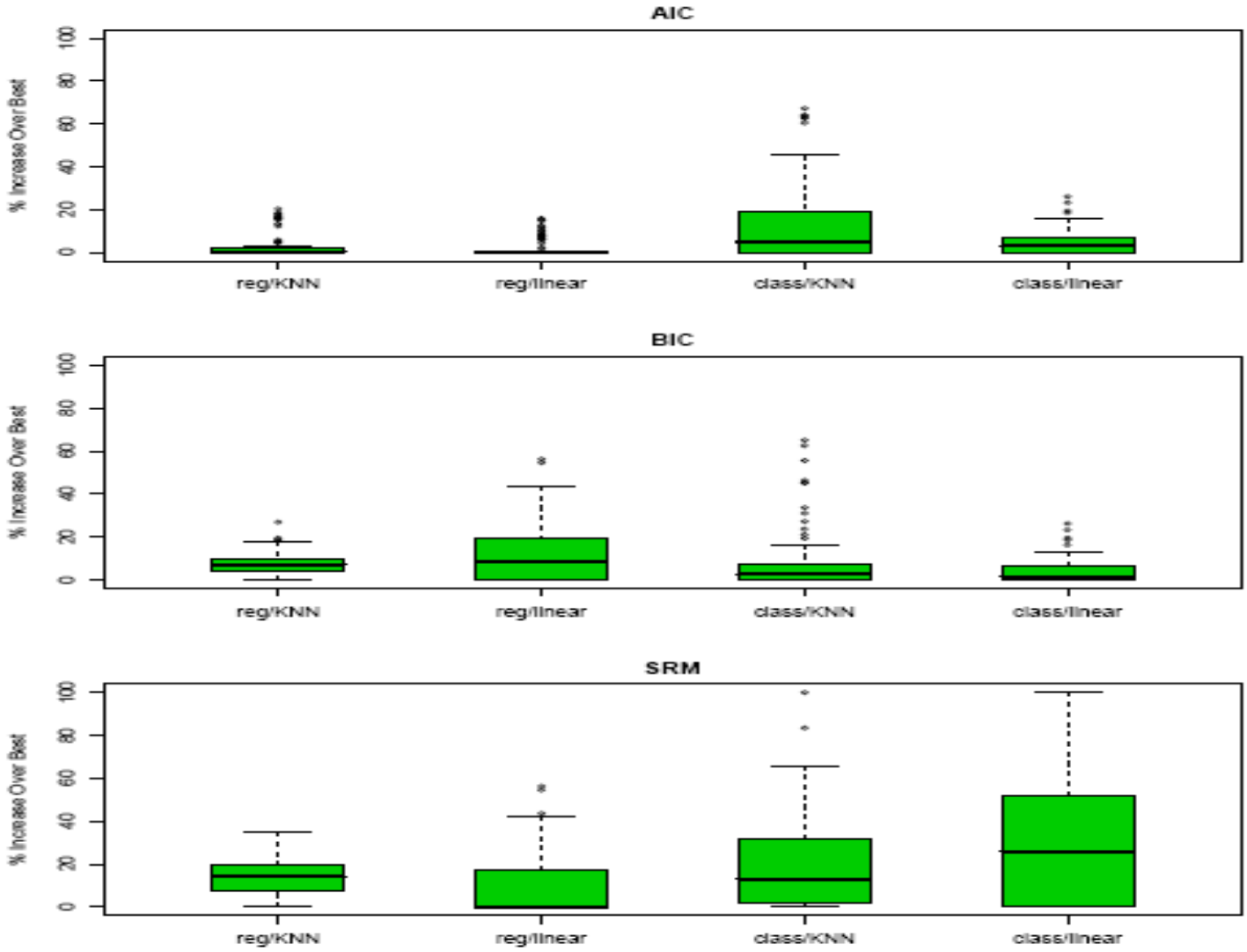
$$\hat{\theta}_{J,n} = n^{-1}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)(Y_i - \hat{Y}_{i(i)}) = n^{-1}\sum_{i=1}^{n}\frac{(Y_i - \hat{Y}_i)^2}{1 - h_{ii}}$$

Porównanie AIC, BIC i SRM

Dwa przykłady rozpatrzone poprzednio (ESL, str. 247):

Liczono $Err_{\mathcal{U}}(\hat{M})$ dla metody selekcji ($\#\mathcal{U} = 80$, próby testowe liczności 10 000). Miara jakości

$$100 \times \frac{Err_{\mathcal{U}}(\hat{M}) - \min_M Err_{\mathcal{U}}(M)}{\max_M Err_{\mathcal{U}}(M) - \min_M Err_{\mathcal{U}}(M)}$$


AIC
% Increase Over Best
0
20
40
60
80
100
reg/KNN
reg/linear
class/KNN
class/linear
BIC
% Increase Over Best
0
20
40
60
80
100
reg/KNN
reg/linear
class/KNN
class/linear
SRM
% Increase Over Best
0
20
40
60
80
100
reg/KNN
reg/linear
class/KNN
class/linear