Błędy predykcji, ocena dopasowania modelu

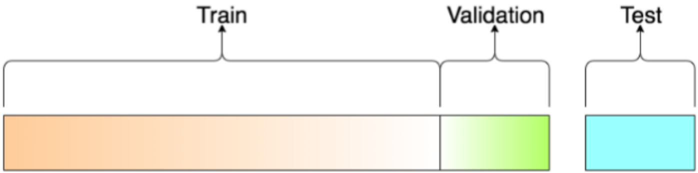
Karolina Sopata

25 kwietnia 2022

Selekcja modelu (*model selection*) - wybór najlepszego spośród rozważanych modeli na podstawie ich dopasowania do danych (do zbioru treningowego)

Ocena dopasowania modelu (*model assessement*) - oszacowanie błędu predykcji wybranego modelu na nowych danych (spoza zbioru treningowego)

Podział danych na zbiory



Przykładowe podziały:

treningowy	walidacyjny	testowy
50%	25%	25%
70%	15%	15%
80%	-	20 %

Metody

- metody wykorzystujące wielokrotnie tą samą próbę
  - bootstrap
  - kroswalidacja
- metody analityczne
  - kryteria informacyjne: AIC, BIC
  - regularyzacja: ridge, LASSO

Błąd predykcji dla obserwacji

$$PE_i = L(y_i, \hat{f}(x_i))$$

y_i - wartość zmiennej objaśnianej dla danej obserwacji

$\hat{f}(x_i)$ - wartość predykcji

L - funkcja straty

$$L(Y, \hat{f}(X)) = \begin{cases} (Y - \hat{f}(X))^2 \\ |Y - \hat{f}(X)| \end{cases}$$

Rodzaje błędów predykcji

- Err - oczekiwany błąd predykcji
- Err_T - oczekiwany błąd predykcji dla ustalonego zbioru treningowego
- $\overline{err}$ - błąd treningowy
- Err_{in} - oczekiwany błąd »wewnątrz próby«

Oczekiwany błąd predykcji

$$Err = E[L(Y, \hat{f}(X))]$$

Y - zmienna objaśniana
$\hat{f}(X)$ - wartość predykcji dla zmiennej/zmiennych objaśniających X

Błąd testowy

test error, generalization error - oczekiwana wartość błędu predykcji dla modelu uczonego na ustalonym zbiorze treningowym τ

$$Err_\tau = E_{X,Y}[L(Y, \hat{f}(X))|\tau]$$

Oczekiwany błąd predykcji (Err) - wartość oczekiwana Err_τ po wszystkich możliwych zbiorach treningowych:

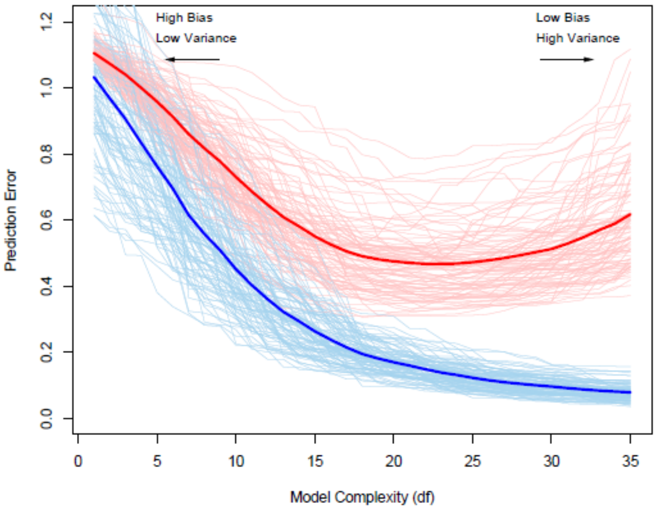
$$Err = E[Err_\tau] = E[E_{X,Y}[L(Y, \hat{f}(X))|\tau]]$$

Błąd treningowy ($\overline{err}$)

$\overline{err}$ - uśrednione błędy predykcji dla wszystkich obserwacji ze zbioru treningowego

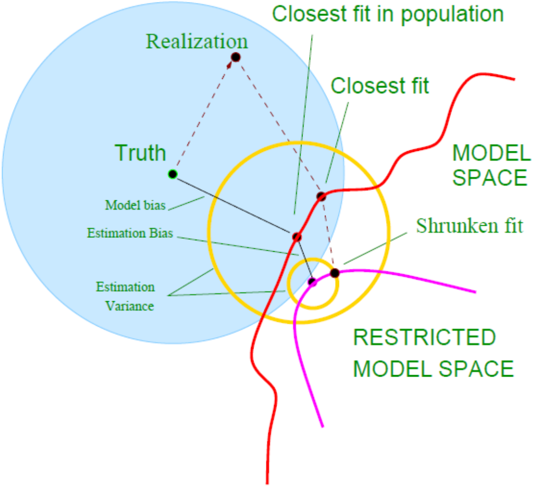
$$\overline{err} = \frac{1}{N}\sum_{1=i}^{N} L(y_i, \hat{f}(x_i))$$

N-liczba obserwacji w próbie



High Bias
Low Variance
Low Bias
High Variance
Prediction Error
0.0
0.2
0.4
0.6
0.8
1.0
1.2
0
5
10
15
20
25
30
35
Model Complexity (df)

Obciążenie, wariancja, złożoność modelu



Obciążenie, wariancja, złożoność modelu

$$Y = f(X) + \epsilon$$

gdzie $E(\epsilon) = 0, Var(\epsilon) = \sigma_\epsilon^2$,

błąd predykcji dla modelu regresji, dla pewnego ustalonego wektora zmiennych objaśniających $X = x_0$:

$$\begin{aligned}
\mathrm{Err}(x_0) &= E[(Y - \hat{f}(x_0))^2 | X = x_0] \\
&= \sigma_\varepsilon^2 + [\mathrm{E}\hat{f}(x_0) - f(x_0)]^2 + E[\hat{f}(x_0) - \mathrm{E}\hat{f}(x_0)]^2 \\
&= \sigma_\varepsilon^2 + \mathrm{Bias}^2(\hat{f}(x_0)) + \mathrm{Var}(\hat{f}(x_0)) \\
&= \text{Irreducible Error} + \mathrm{Bias}^2 + \text{Variance}.
\end{aligned}$$

Regresja liniowa - błąd predykcji

$$\begin{aligned}\mathrm{Err}(x_0) &= E[(Y - \hat{f}_p(x_0))^2 | X = x_0] \\ &= \sigma_\varepsilon^2 + [f(x_0) - \mathrm{E}\hat{f}_p(x_0)]^2 + ||\mathbf{h}(x_0)||^2 \sigma_\varepsilon^2.\end{aligned}$$

gdzie $h(x_0) = X(X^T X)^{-1} x_0$ - wektor wag, które dają następujące dopasowanie:

$$\hat{f}_p(X_0) = x_0^T (X^T X)^{-1} X^T y$$

Stąd $Var[\hat{f}_p(X_0)] = ||h(x_0||^2 \sigma_\epsilon^2$
Błąd "wewnątrz próby"(in-sample error):

$$\frac{1}{N}\sum_{i=1}^{N} \mathrm{Err}(x_i) = \sigma_\varepsilon^2 + \frac{1}{N}\sum_{i=1}^{N}[f(x_i) - \mathrm{E}\hat{f}(x_i)]^2 + \frac{p}{N}\sigma_\varepsilon^2,$$

Regresja grzbietowa - błąd predykcji

β_* - wektor parametrów dla najlepszego dopasowania liniowego

$$\beta_* = \arg\min_\beta \mathrm{E}\left(f(X) - X^T\beta\right)^2.$$

$$\begin{aligned} \mathrm{E}_{x_0}\left[f(x_0) - \mathrm{E}\hat{f}_\alpha(x_0)\right]^2 &= \mathrm{E}_{x_0}\left[f(x_0) - x_0^T\beta_*\right]^2 + \mathrm{E}_{x_0}\left[x_0^T\beta_* - \mathrm{E}x_0^T\hat{\beta}_\alpha\right]^2 \\ &= \mathrm{Ave}[\text{Model Bias}]^2 + \mathrm{Ave}[\text{Estimation Bias}]^2 \end{aligned}$$

Err_{in} - *in-sample error*

$$Err_{in} = \frac{1}{N}\sum_{i=1}^{N} E_{Y^0}[L(Y_i^0, \hat{f}(x_i))|\tau]$$

- $x_i \in \tau$ - wartości zmiennych objaśniających z ustalonego zbioru treningowego τ
- $Y_i^0 \in Y^0$ - nowa wartość zmiennej odpowiedzi dla x_i (różna od y_i danych w zbiorze treningowym)

Optymizm błędu treningowego

Optymizm:

$$op \equiv Err_{in} - \overline{err}$$

Oczekiwany optymizm:

$$\omega \equiv E_Y(op)$$

Generalnie:

$$\omega = \frac{2}{N} \sum_{i=1}^{N} Cov(\hat{y}_i, y_i)$$

Optymizm błędu treningowego c.d.

$$E_y(Err_{in}) = E_y(\overline{err}) + E_Y(op)$$

$$E_y(Err_{in}) = E_y(\overline{err}) + \omega$$

$$E_y(Err_{in}) = E_y(\overline{err}) + \frac{2}{N}\sum_{i=1}^{N} Cov(\hat{y}_i, y_i)$$

Dla modelu addytywnego $(Y = f(X) + \epsilon)$ upraszczamy:

$$Cov(\hat{y}_i, y_i) = d\sigma_\epsilon^2$$

d - liczba zmiennych w modelu
Wówczas:

$$E_y(Err_{in}) = E_y(\overline{err}) + 2\frac{d}{N}\sigma_\epsilon^2$$

Statystyka C_p

Estymator błędu in-sample:

$$\widehat{Err_{in}} = \overline{err} + \hat{\omega}$$

Stosujemy (gdy d parametrów jest dopasowanych metodą najmniejszych kwadratów):

$$\widehat{Err_{in}} = \overline{err} + 2\frac{d}{N}\hat{\sigma}_\epsilon^2$$

Dana wielkość znana jako C_p Mallows'a:

$$C_p = \overline{err} + 2\frac{d}{N}\hat{\sigma}_\epsilon^2$$

Kryterium AIC

Stosowane, gdy funkcją straty jest funkcja log-wiarogodności.

$$\text{loglik} = \sum_{i=1}^{N} \log \Pr_{\hat{\theta}}(y_i)$$

$$\text{AIC} = -\frac{2}{N} \cdot \text{loglik} + 2 \cdot \frac{d}{N}$$

Kryterium BIC

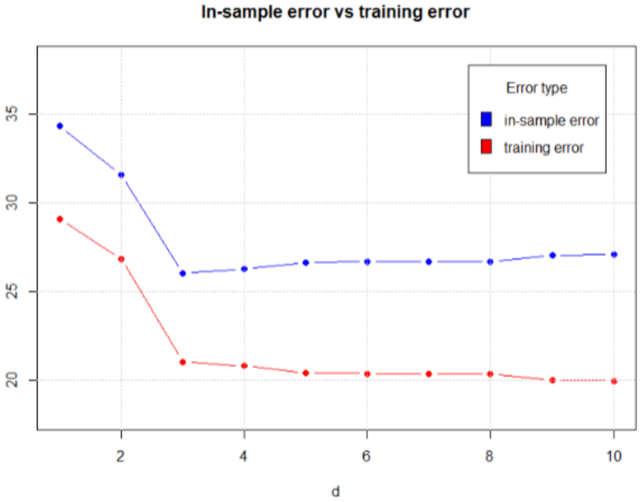
$$\mathrm{BIC} = -2 \cdot \mathrm{loglik} + (\log N) \cdot d$$

Model gaussowski:

$$\mathrm{BIC} = \frac{N}{\sigma_\varepsilon^2}\left[\overline{\mathrm{err}} + (\log N) \cdot \frac{d}{N}\sigma_\varepsilon^2\right]$$

Kryterium BIC jest proporcjonalne do AIC.

Błąd treningowy a błąd in-sample



Dziękuję za uwagę!