Ocena i wybór modelu

Justyna Fijałkowska

Uniwersytet Wrocławski

2 czerwca 2023

Plan prezentacji

1. Obciążenie, wariancja i złożoność modelu
2. Oczekiwany błąd predykcji
3. Optymizm wskaźnika błędów treningowych

Zmienne ilościowe

Niech Y będzie zmienną zależną, X wektorem danych wejściowych, a $\widehat{f}(X)$ modelem predykcyjnym, który oszacowany był na zbiorze uczącym τ.

Przykładowe funkcje straty to:

$$L(Y, \widehat{f}(X)) = \begin{cases} (Y - \widehat{f}(X))^2 & \text{błąd kwadratowy} \\ |Y - \widehat{f}(X)| & \text{błąd całkowity} \end{cases}$$

Oznaczenia:

- błąd testowy

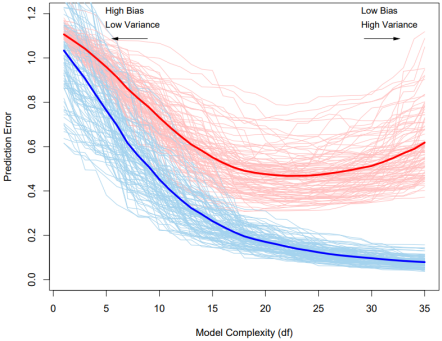
$$Err_\tau = E[L(Y, \widehat{f}(X))|\tau]$$

- oczekiwany błąd testowy

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

Przykład – błąd testowy

Liczba danych wejściowych: 50, liczba powtórzeń eksperymentu: 100.



- Jasnoczerwone krzywe – Err_{τ}.
- Pogrubiona czerwona krzywa – średnia z Err_{τ}.

Przykład – błąd treningowy

Błąd treningowy to średnia strata w próbie treningowej:

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

Uwagi:

- $\overline{err}$ nie jest dobrym oszacowaniem błędu testowego,
- $\overline{err}$ maleje wraz ze zwiększaną złożonością modelu,
- problem overfittingu.

Zmienne jakościowe, kategoryczne

G przyjmuje jedną z K wartości w zbiorze $\mathcal{G}$, etykietowanym $1, 2, \ldots, K$. Modelujemy prawdopodobieństwa $p_k(X) = P(G = k|X)$, a następnie $\widehat{G}(X) = \text{argmax}_k \widehat{p}_k(X)$.

Przykładowe funkcje straty:

$$L(G, \widehat{G}(X)) = \mathbb{1}(G \neq \widehat{G}(X)) \quad \text{0-1 loss},$$

$$L(G, \widehat{p}(X)) = -2 \sum_{k=1}^{K} \mathbb{1}(G = k) \log \widehat{p}_k(X) = -2 \log \widehat{p}_G(X) \quad -2 \times \text{log-wiarogodność}.$$

- Błąd testowy: $Err_\tau = E[L(G, \widehat{G}(X))|\tau]$.
- Oczekiwany błąd testowy – oczekiwany błąd błędnych klasyfikacji.
- Jeżeli funkcja straty jest postaci $-2 \log \widehat{p}_G(X)$, to błąd treningowy jest postaci:

$$\overline{err} = -\frac{2}{N} \sum_{i=1}^{N} \log \widehat{p}_{g_i}(x_i).$$

Log-wiarogodność jako funkcja straty

Niech $P_{\theta(X)}(Y)$ będzie gęstością zmiennej losowej Y, indeksowaną parametrem $\theta(X)$, który zależy od predyktora X. Wówczas funkcja straty może być zdefiniowana za pomocą log-wiarogodności w następujący sposób:

$$L(Y, \theta(X)) = -2 \cdot \log P_{\theta(X)}(Y).$$

W dalszej części zwykle Y będzie zmienna ilościowa z kwadratową funkcją straty.

Podział zbioru danych

Możemy chcieć zajmować się problemami:

- wybór modelu,
- ocena modelu.

Pomysł: Podzielić dane na trzy zbiory: treningowy, walidacyjne i testowe.

Problem: Danych jest za mało.

Pomysł: Przybliżanie etapu walidacji metodami analitycznymi lub efektywne ponowne wykorzystywanie próby.

Model regresji

Niech $Y = f(X) + \epsilon$, gdzie $E(\epsilon) = 0$ i $Var(\epsilon) = \sigma_\epsilon^2$ oraz $X = x_0$ – punkt wejściowy. Wówczas oczekiwany błąd predykcji jest postaci:

$$\begin{aligned} Err(x_0) &= E[(Y - \widehat{f}(x_0))^2 | X = x_0] \\ &= \sigma_\epsilon^2 + [E\widehat{f}(x_0) - f(x_0)]^2 + E[\widehat{f}(x_0) - E\widehat{f}(x_0)]^2 \\ &= \sigma_\epsilon^2 + \text{Bias}^2(\widehat{f}(x_0)) + \text{Var}(\widehat{f}(x_0)) \\ &= \text{błąd nieredukowalny} + \text{obciążenie}^2 + \text{wariancja} \end{aligned}$$

Regresja k najbliższych sąsiadów

Załóżmy, że dane wejściowe x_i są ustalone, a losowość wynika z y_i. Wówczas błąd predykcji jest postaci:

$$\begin{aligned} Err(x_0) &= E[(Y - \widehat{f}_k(x_0))^2 | X = x_0] \\ &= \sigma_\epsilon^2 + \left[f(x_0) - \frac{1}{k} \sum_{l=1}^{k} f(x_{(l)}) \right]^2 + \frac{\sigma_\epsilon^2}{k} \end{aligned}$$

Uwaga: Wraz ze wzrostem k obciążenie będzie zwykle rosło, podczas gdy wariancja będzie maleć.

Model liniowy i regresja grzbietowa

- Załóżmy, że $\widehat{f}_p(x) = x^T\widehat{\beta}$, gdzie β jest wektorem długości p. Wówczas błąd predykcji jest postaci:

$$\begin{aligned} Err(x_0) &= E[(Y - \widehat{f}_p(x_0))^2 | X = x_0] \\ &= \sigma_\epsilon^2 + [f(x_0) - E\widehat{f}_p(x_0)]^2 + \|\boldsymbol{h}(x_0)\|^2\sigma_\epsilon^2, \end{aligned}$$

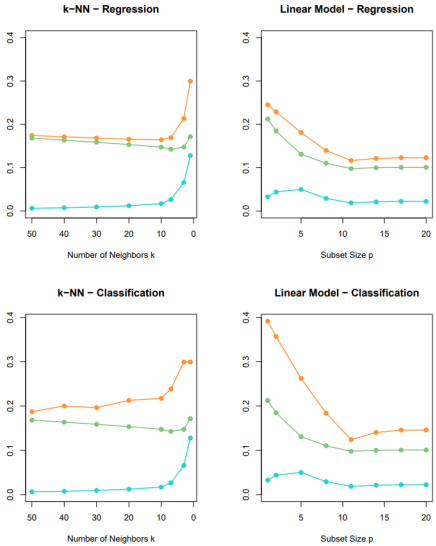
  gdzie $\boldsymbol{h}(x_0) = \boldsymbol{X}(\boldsymbol{X}^t\boldsymbol{X})^{-1}x_0$.
- W przypadku dopasowania za pomocą regresji grzbietowej funkcja $\boldsymbol{h}(x_0)$ jest postaci $\boldsymbol{h}(x_0) = \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X} + \alpha\boldsymbol{I})^T x_0$ oraz zmienia się obciążenie (zwykle jest dodatnie).

Przykład - kompromis obciążenia i wariancji

Zbiór składa się z 80 obserwacji i 20 predyktorów, równomiernie rozłożonych w hipersześcianie $[0,1]^{20}$.

- Obrazki po lewej stronie odpowiadają sytuacji, w której: $Y = 0$, jeżeli $X_1 \leqslant 1/2$ i $Y = 1$, jeżeli $X_1 > 1/2$ oraz stosujemy kNN.
- Obrazki po prawej stronie odpowiadają sytuacji, w której: $Y = 1$, jeżeli $\sum_{j=1}^{10} X_j > 5$ i $Y = 0$, jeżeli $\sum_{j=1}^{10} X_j \leqslant 5$ oraz używamy najlepszego podzbioru regresji liniowej rozmiaru p.
- Górny rząd to regresja z kwadratową funkcją straty.
- Dolny rząd to klasyfikacja ze stratą 0-1.

Przykład - kompromis obciążenia i wariancji



Optymizm wskaźnika błędów treningowych

Oznaczmy zbiór treningowy przez $\tau = \{(x_1, y_1), (x_2, y_2), \ldots, (x_N, y_N)\}$. Wówczas błąd uogólnienia (predykcji) modelu $\widehat{f}$ wynosi:

$$Err_\tau = E_{X^0, Y^0}[L(Y_0, \widehat{f}(X^0))|\tau].$$

Niech (X^0, Y^0) będzie nowym punktem, wtedy błąd oczekiwany wynosi:

$$Err = E_\tau E_{X^0, Y^0}[L(Y^0, \widehat{f}(X^0))|\tau].$$

Zwykle błąd treningowy postaci:

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

jest mniejszy niż prawdziwy błąd Err_τ.

Optymizm wskaźnika błędów treningowych

Zdefiniujmy błąd w próbie jako:

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

Wtedy optymizm definiujemy jako różnicę między Err_{in}, a $\overline{err}$, tzn.:

$$op = Err_{in} - \overline{err}.$$

Wówczas, średni optymizm jest postaci:

$$\omega = E_{\mathbf{y}}(op).$$

W przypadku błędu kwadratowego, 0-1 i innych funkcji straty, można pokazać, że:

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

Optymizm wskaźnika błędów treningowych

Dostajemy ważną zależność:

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

- Dla modelu $Y = f(X) + \epsilon$ zachodzi $\sum_{i=1}^{N} Cov(\hat{y}_i, y_i) = d\sigma_\epsilon^2$, gdzie d – liczba danych wejściowych, a co za tym idzie:

$$E_{\mathbf{y}}(Err_{in}) = E_{\mathbf{y}}(\overline{err}) + 2 \cdot \frac{d}{N}\sigma_\epsilon^2.$$

- Optymizm wzrasta liniowo wraz z liczbą d danych wejściowych i maleje wraz ze wzrostem wielkość próby treningowej.
- Oszacowania błędu predykcji można dokonać szacując optymizm i dodając go do błędu treningowego $\overline{err}$.

Dziękuję za uwagę.