Strukturalne modele regresji. Wybór modelu oraz kompromis obciążenia i wariancji.

Justyna Fijałkowska

Uniwersytet Wrocławski

17 marca 2023

Plan prezentacji

Kara za nieregularność

- Klasa kontrolowana jest poprzez nakładanie kary na *RSS*:

$$\mathrm{PRSS}(f;\lambda) = \mathrm{RSS}(f) + \lambda J(f).$$

Kara za nieregularność

- Klasa kontrolowana jest poprzez nakładanie kary na *RSS*:

$$\mathrm{PRSS}(f;\lambda) = \mathrm{RSS}(f) + \lambda J(f).$$

- $J(f)$ jest duże, jeżeli f zmienia się gwałtownie na małych obszarach przestrzeni wejściowej.

Kara za nieregularność

- Klasa kontrolowana jest poprzez nakładanie kary na *RSS*:

$$\mathrm{PRSS}(f;\lambda) = \mathrm{RSS}(f) + \lambda J(f).$$

- $J(f)$ jest duże, jeżeli f zmienia się gwałtownie na małych obszarach przestrzeni wejściowej.
- Do tworzenia modeli addytywnych z gładkimi funkcjami dla współrzędnych można stosować kary addytywne postaci:

$$J(f) = \sum_{j=1}^{p} J(f_j)$$

w połączeniu z funkcjami addytywnymi:

$$f(X) = \sum_{j=1}^{p} f_j(X_j)$$

Kara za nieregularność

Przykład

Niech $\{(x_i, y_i),\ i = 1, 2, \ldots, N\}$ będzie zbiorem obserwacji modelowanym przez relację $y_i = f(x_i) + \epsilon_i$, gdzie ϵ_i są niezależnymi zmiennymi losowymi o średniej 0 i stałej wariancji.

Jak wybrać $f \in C^2$ najlepiej dopasowaną do danych?

Kara za nieregularność

Przykład

Niech $\{(x_i, y_i),\ i = 1, 2, \ldots, N\}$ będzie zbiorem obserwacji modelowanym przez relację $y_i = f(x_i) + \epsilon_i$, gdzie ϵ_i są niezależnymi zmiennymi losowymi o średniej 0 i stałej wariancji.

Jak wybrać $f \in C^2$ najlepiej dopasowaną do danych?

$$\mathrm{PRSS}(f;\lambda) = \sum_{i=1}^{N} (y_i - f(x_i))^2 + \lambda \int [f''(x)]^2 dx.$$

Kara za nieregularność

Przykład

Niech $\{(x_i, y_i),\ i = 1, 2, \dots, N\}$ będzie zbiorem obserwacji modelowanym przez relację $y_i = f(x_i) + \epsilon_i$, gdzie ϵ_i są niezależnymi zmiennymi losowymi o średniej 0 i stałej wariancji.

Jak wybrać $f \in C^2$ najlepiej dopasowaną do danych?

$$\mathrm{PRSS}(f;\lambda) = \sum_{i=1}^{N} (y_i - f(x_i))^2 + \lambda \int [f''(x)]^2 dx.$$

- Jeżeli $\lambda = 0$, to f jest dowolną funkcją interpolującą.
- Jeżeli $\lambda = \infty$, to f jest funkcją liniową.

Metody jądra i regresja lokalna

- Metody szacują funkcje regresji lub warunkowej wartości oczekiwanej poprzez określenie charakteru sąsiedztwa lokalnego, a także za pomocą klas funkcji dopasowanych lokalnie.

Metody jądra i regresja lokalna

- Metody szacują funkcje regresji lub warunkowej wartości oczekiwanej poprzez określenie charakteru sąsiedztwa lokalnego, a także za pomocą klas funkcji dopasowanych lokalnie.
- Przykład jądra (jądro Gaussa):

$$K_\lambda(x_0, x) = \frac{1}{\lambda} \exp\left[-\frac{||x - x_0||^2}{2\lambda}\right]$$

Metody jądra i regresja lokalna

- Metody szacują funkcje regresji lub warunkowej wartości oczekiwanej poprzez określenie charakteru sąsiedztwa lokalnego, a także za pomocą klas funkcji dopasowanych lokalnie.
- Przykład jądra (jądro Gaussa):

$$K_\lambda(x_0, x) = \frac{1}{\lambda}\exp\left[-\frac{||x - x_0||^2}{2\lambda}\right]$$

- Jądrowy estymator regresji uzyskany za pomocą średniej ważonej Nadaraya–Watson:

$$\hat{f}(x_0) = \frac{\sum_{i=1}^{N} K_\lambda(x_0, x_i) y_i}{\sum_{i=1}^{N} K_\lambda(x_0, x_i)}.$$

Metody jądra i regresja lokalna

Niech $f_{\hat{\theta}}(x_0)$ będzie funkcją przybliżającą $f_\theta(x_0)$, gdzie $\hat{\theta}$ minimalizuje RSS:

$$\mathrm{RSS}(f_\theta, x_0) = \sum_{i=1}^{N} K_\lambda(x_0, x_i)(y_i - f_\theta(x_i))^2,$$

- Jeżeli $f_\theta(x) = \theta_0$, to uzyskamy estymator Nadaraya–Watson.
- Jeżeli $f_\theta(x) = \theta_1 x + \theta_0$, to otrzymamy model lokalnej regresji.

Metody jądra i regresja lokalna

Uwagi:

- Metoda *k* najbliższych sąsiadów może być interpretowana jako metoda jądra z metryką:

$$K_k(x, x_0) = I(||x - x_0|| \leq ||x_{(k)} - x_0||),$$

Metody jądra i regresja lokalna

Uwaga:

- Metoda *k* najbliższych sąsiadów może być interpretowana jako metoda jądra z metryką:

$$K_k(x, x_0) = I(||x - x_0|| \leq ||x_{(k)} - x_0||),$$

- Przekleństwo wymiarowości.

Funkcje bazowe

- Klasa zawiera rozwinięcia liniowe i wielomianowe oraz inne bardziej elastyczne modele.

Funkcje bazowe

- Klasa zawiera rozwinięcia liniowe i wielomianowe oraz inne bardziej elastyczne modele.
- Model dla funkcji f jest postaci:

$$f_\theta(x) = \sum_{m=1}^{M} \theta_m h_m(x),$$

Funkcje bazowe

- Klasa zawiera rozwinięcia liniowe i wielomianowe oraz inne bardziej elastyczne modele.
- Model dla funkcji f jest postaci:

$$f_\theta(x) = \sum_{m=1}^{M} \theta_m h_m(x),$$

- Przykład funkcji bazowych:

$$b_1(x) = 1, \; b_2(x) = x, \text{ and } b_{m+2}(x) = (x - t_m)_+, \; m = 1, \ldots, M-2,$$

Wybór modelu oraz kompromis obciążenia i wariancji

- Problem doboru optymalnego parametru wygładzającego lub złożoności.

Wybór modelu oraz kompromis obciążenia i wariancji

- Problem doboru optymalnego parametru wygładzającego oraz złożoności.
- Używanie RSS do wyboru optymalnego parametru wygładzającego λ to zły pomysł, gdyż dostaniemy dopasowanie interpolacyjne zazwyczaj o słabych własnościach predykcyjnych.

Wybór modelu oraz kompromis obciążenia i wariancji

Przykład

Załóżmy, że dane pochodzą z modelu

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

gdzie $\mathbb{E}(\epsilon) = 0$ i $Var(\epsilon) = \sigma^2$. Dla ułatwienia załóżmy, że wartości x_i w próbie nie są losowe. Ile wynosi oczekiwany błąd predykcji?

Wybór modelu oraz kompromis obciążenia i wariancji

Przykład

Załóżmy, że dane pochodzą z modelu

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

gdzie $\mathbb{E}(\epsilon) = 0$ i $Var(\epsilon) = \sigma^2$. Dla ułatwienia załóżmy, że wartości x_i w próbie nie są losowe. Ile wynosi oczekiwany błąd predykcji?

$$\begin{aligned}
\mathrm{EPE}_k(x_0) &= \mathrm{E}[(Y - \hat{f}_k(x_0))^2 | X = x_0] \\
&= \sigma^2 + [\mathrm{Bias}^2(\hat{f}_k(x_0)) + \mathrm{Var}_{\mathcal{T}}(\hat{f}_k(x_0))] \\
&= \sigma^2 + \left[f(x_0) - \frac{1}{k} \sum_{\ell=1}^{k} f(x_{(\ell)}) \right]^2 + \frac{\sigma^2}{k}.
\end{aligned}$$

Wybór modelu oraz kompromis obciążenia i wariancji

Z czego składa się błąd predykcji?

- Błąd nieredukowalny,
- Błąd obciążenia,
- Błąd wariancji

Wybór modelu oraz kompromis obciążenia i wariancji

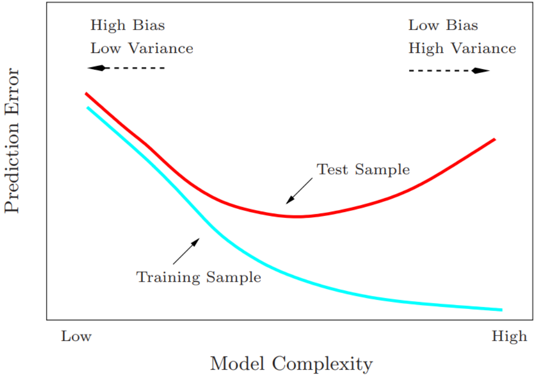

FIGURE 2.11. *Test and training error as a function of model complexity.*

Wybór modelu oraz kompromis obciążenia i wariancji

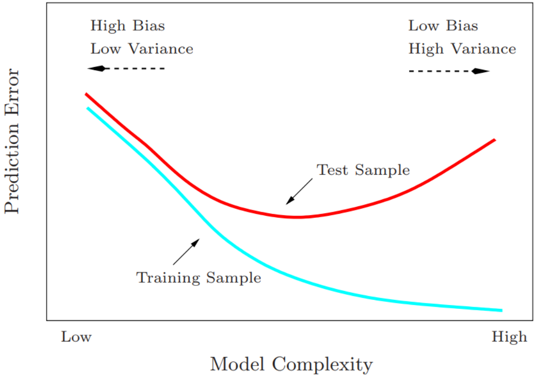

FIGURE 2.11. *Test and training error as a function of model complexity.*

- Jeżeli model jest skomplikowany, to wariancja jest duża, a obciążenie małe.
- Jeżeli model mniej skomplikowany, to wariancja jest mała, a obciążenie duże.

Dziękuję za uwagę.