Metody estymacji błędu prognozy
Bias-variance trade-off

Uogólnianie modelu poza próbę uczącą

- Odnosi się do możliwości prognostycznych modelu w odrębnym i niezależnym od próby uczącej zbiorze danych
- Oszacowanie zdolności uogólniania jest podstawą do wyboru metod, modeli i jest miarą ich jakości
- Zagadnienie: $Y, X, \hat{f}(X)$ bazujące na próbie uczącej T

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

$$L\left(G, \hat{G}(X)\right) = I(G \neq \hat{G}(X))$$

Rodzaje błędu prognozy

- Błąd uczący (treningowy) – *training error*

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

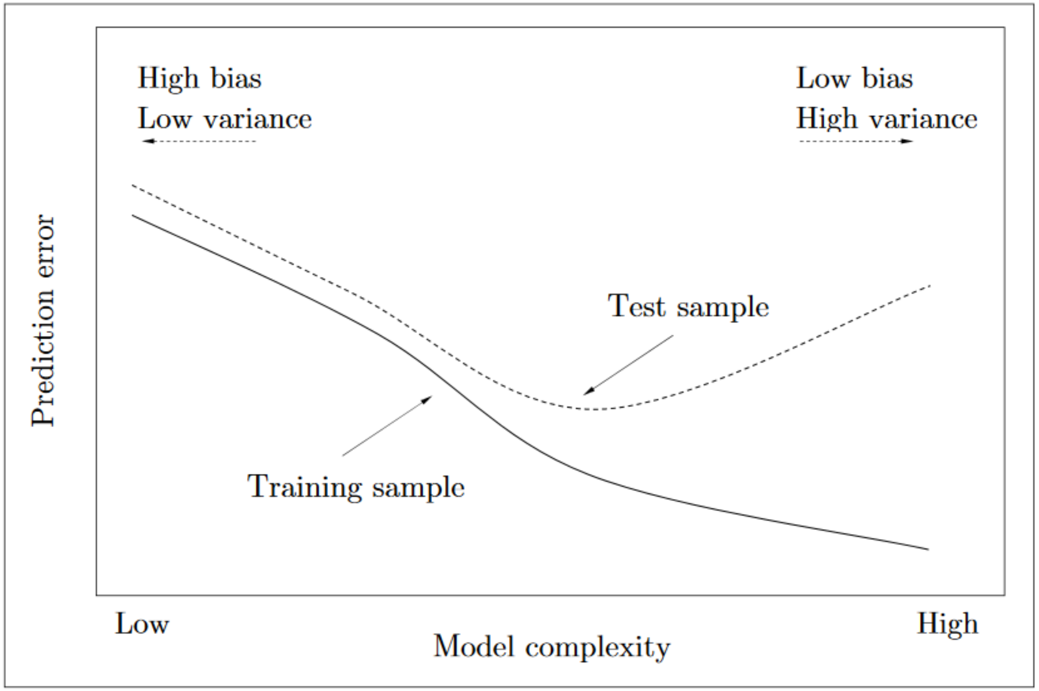
- Błąd testowy (uogólniania) – *test (generalization) error*

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

- Oczekiwany błąd testowy – *expected test error*

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

Bias-variance trade-off



- Istnieje pośredni poziom złożoności modelu dający najmniejszy możliwy Err_T
- Problem: znaleźć α (parametr złożoności) w $f_\alpha(X)$

Bias-variance decomposition

- $Y = f(X) + \varepsilon, \;\; E(\varepsilon) = 0, Var(\varepsilon) = \sigma_\varepsilon^2$

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

- Regresyjny model *k*-najbliższych sąsiadów:

$$Err(x_0) = \sigma_\varepsilon^2 + \left[f(x_0) - \frac{1}{k}\sum_{i=1}^{k} f(x_{(i)})\right]^2 + \frac{\sigma_\varepsilon^2}{k}$$

Wybór modelu i ocena jego jakości

Wybór modelu: wybór najlepszego modelu spośród wielu oszacowań

Ocena modelu: po wybraniu najlepszego modelu należy oszacować spodziewany błąd związany z jego stosowaniem, tj. błąd testowy

Zbiór uczący	Zbiór walidacyjny	Zbiór testowy

Inne podejścia: kroswalidacja, bootstrap

Kroswalidacja

1	2	3	4	5
Treningowy	Treningowy	Walidacyjny	Treningowy	Treningowy

- $k = 2,3, \ldots, K$; K to najczęściej 5 lub 10
- $\rho: \{1, \ldots, N\} \rightarrow \{1, \ldots, K\}$

$$CV(\hat{f}, \alpha) = \frac{1}{N} \sum_{i=1}^{N} L(y_i, \hat{f}^{-\rho(i)}(x_i, \alpha))$$

- kroswalidacja jest silnie zależna od liczby obserwacji

Bootstrap

- Ogólnie metoda statystycznej oceny precyzji z wykorzystaniem wielokrotnego losowania ze zwracaniem ze zbioru
- Z pierwotnego zbioru danych losujemy ze zwracaniem B prób o liczności takiej jak zbioru pierwotnego (np. $B = 100$)
- W każdej z B próbek estymujemy model i sprawdzamy jego jakość

$$\widehat{Err}_{boot} = \frac{1}{B}\frac{1}{N}\sum_{b=1}^{B}\sum_{i=1}^{N} L(y_i, \hat{f}^{*b}(x_i))$$

- Wersja uwzględniająca fakt, że część obserwacji powtarza się w próbkach i pierwotnym zbiorze danych

$$\widehat{Err}_{boot}^{(1)} = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{|C^{-i}|}\sum_{b\in C^{-i}} L(y_i, \hat{f}^{*b}(x_i))$$

C^{-i} to zbiór wskaźników próbek b, które nie zawierają obserwacji i