# Kroswalidacja

## Na podstawie rozdziału 7.10 książki *The Elements of Statistical Learning* (Trevor Hastie, Robert Tibshirani, Jerome Friedman)

Mateusz Łyczko

Instytut Matematyczny UWr

07.06.2023

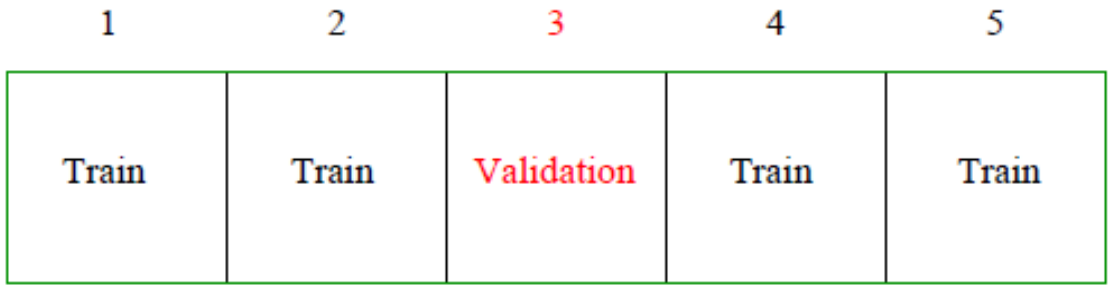
# Wstęp

Prawdopodobnie najprostszą i najczęściej stosowaną metodą szacowania błędu predykcji jest walidacja krzyżowa. Ta metoda bezpośrednio szacuje

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

# Kroswalidacja K-krotna (1)

$K$-krotna walidacja krzyżowa ($K$-fold crossvalidation) wykorzystuje część dostępnych danych do dopasowania modelu, a inną część do jego przetestowania. Dzielimy dane na $K$ mniej więcej równych części. Przykładowo, gdy $K = 5$, to scenariusz może wyglądać następująco:




Dla k-tej części (na obrazku to trzecia część) dopasowujemy model do pozostałych $K - 1$ części danych i obliczamy błąd predykcji dopasowanego modelu podczas przewidywania $k$-tej części danych. Robimy to dla $k = 1, 2, \ldots, K$ i łączymy $K$ oszacowań błędu predykcji.

# Kroswalidacja K-krotna (2)

Niech $\kappa : \{1, \ldots, N\} \to \{1, \ldots, K\}$ będzie funkcją indeksującą wskazującą partycję, do której wpadła $i$-ta obserwacja po zrandomizowaniu, a $\widehat{f}^{-k}(x)$ to niech będzie dopasowana funkcja, która została obliczona, gdy $k$-ta część danych jest usunięta. Wtedy estymacja błędu predykcji na podstawie kroswalidacji to

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

Przypadek z $K = N$ nazywany jest kroswalidacją typu „leave-one-out". Mając zbiór modeli $f(x, \alpha)$, gdzie $\alpha$ to parametr, błąd predykcji będzie równy

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

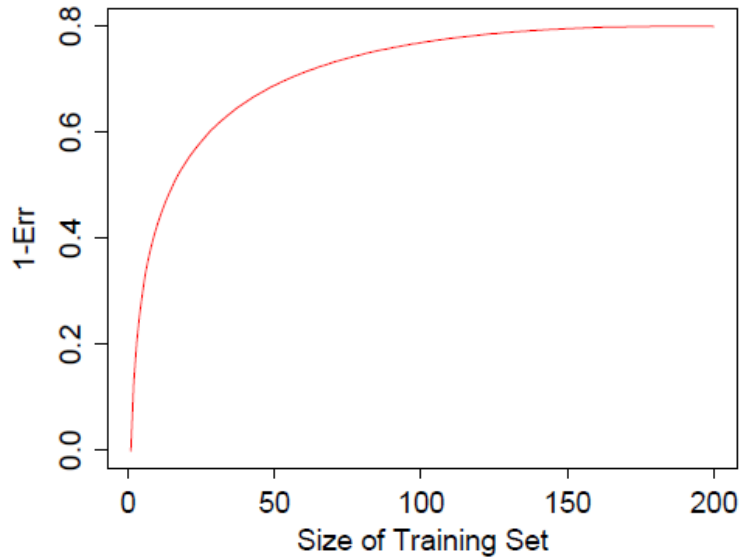
# Kroswalidacja K-krotna (3)

Jaką wartość $K$ powinniśmy wybrać?

Z jednej strony przy $K = N$ estymator kroswalidacji jest w przybliżeniu nieobciążony, ale może mieć dużą wariancję.

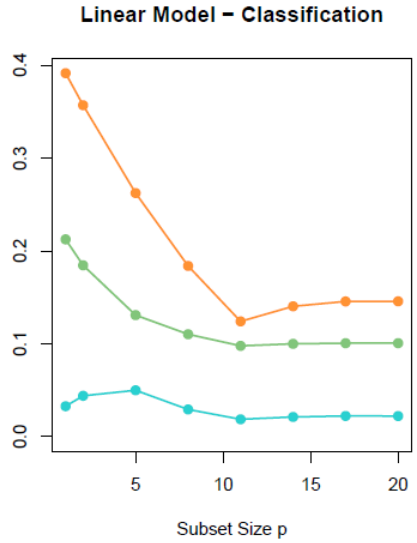
Z drugiej strony, powiedzmy że $K = 5$, kroswalidacja ma mniejszą wariancję, ale może wystąpić obciążenie.

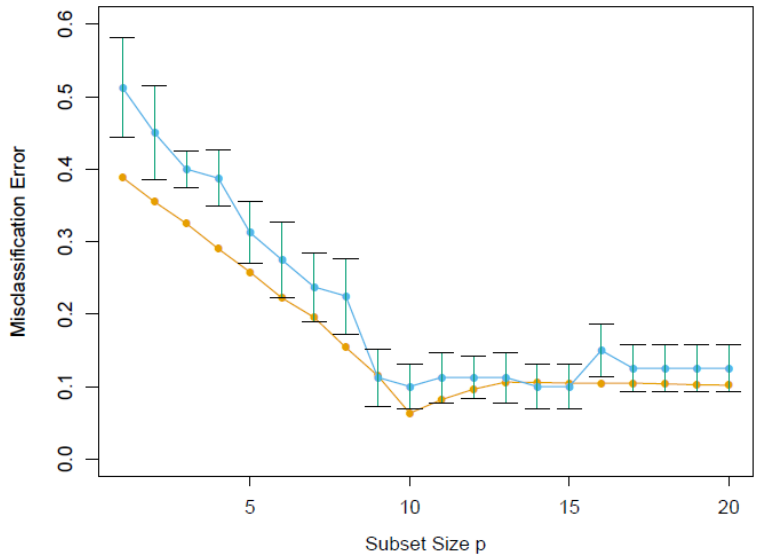
Spójrzmy na hipotetyczną „krzywą uczenia się”:




Odpowiedź zależy od rozmiaru zbioru treningowego, ale ogólnie 5−krotna lub 10−krotna kroswalidacja jest dobrym kompromisem (nie zawsze obciążenie jest dużą wadą).

# Kroswalidacja K-krotna (4)



# Dobry i zły sposób przeprowadzenia kroswalidacji (1)

Rozważmy problem klasyfikacji z dużą liczbą predyktorów, jaki może pojawić się np. w zastosowaniach medycznych.
Typowa strategia może wyglądać następująco:
(1) Sprawdź predyktory: znajdź podzbiór „dobrych" predyktorów, które wykazują dość silną (jednowymiarową) korelację z etykietami klas.
(2) Używając tylko tego podzbioru predyktorów, zbuduj klasyfikator wielowymiarowy.
(3) Użyj walidacji krzyżowej do oszacowania nieznanych parametrów i do oszacowania błędu predykcji ostatecznego modelu.

# Dobry i zły sposób przeprowadzenia kroswalidacji (1)

Rozważmy problem klasyfikacji z dużą liczbą predyktorów, jaki może pojawić się np. w zastosowaniach medycznych.
Typowa strategia może wyglądać następująco:
(1) Sprawdź predyktory: znajdź podzbiór „dobrych" predyktorów, które wykazują dość silną (jednowymiarową) korelację z etykietami klas.
(2) Używając tylko tego podzbioru predyktorów, zbuduj klasyfikator wielowymiarowy.
(3) Użyj walidacji krzyżowej do oszacowania nieznanych parametrów i do oszacowania błędu predykcji ostatecznego modelu.
Czy to jest dobry sposób?

# Dobry i zły sposób przeprowadzenia kroswalidacji (1)

Rozważmy problem klasyfikacji z dużą liczbą predyktorów, jaki może pojawić się np. w zastosowaniach medycznych.
Typowa strategia może wyglądać następująco:
(1) Sprawdź predyktory: znajdź podzbiór „dobrych" predyktorów, które wykazują dość silną (jednowymiarową) korelację z etykietami klas.
(2) Używając tylko tego podzbioru predyktorów, zbuduj klasyfikator wielowymiarowy.
(3) Użyj walidacji krzyżowej do oszacowania nieznanych parametrów i do oszacowania błędu predykcji ostatecznego modelu.
Czy to jest dobry sposób? Nie.

# Dobry i zły sposób przeprowadzenia kroswalidacji (2)

Spójrzmy na inne podejście dla tego przykładu:
(1) Podziel próbki losowo na K grup.
(2) Dla każdego podziału $k = 1, 2, \ldots, K$:
  (A) Znajdź podzbiór „dobrych" predyktorów, które wykazują dość silną (jednowymiarową) korelację z etykietami klas, używając wszystkich próbek z wyjątkiem tych z $k$-tej części.
  (B) Używając tylko tego podzbioru predyktorów, zbuduj wielowymiarowy klasyfikator, używając wszystkich próbek z wyjątkiem tych z $k$-tej części.
  (C) Użyj klasyfikatora, aby dokonać predykcji etykiet klas dla próbek w $k$-tej części.

# Dobry i zły sposób przeprowadzenia kroswalidacji (2)

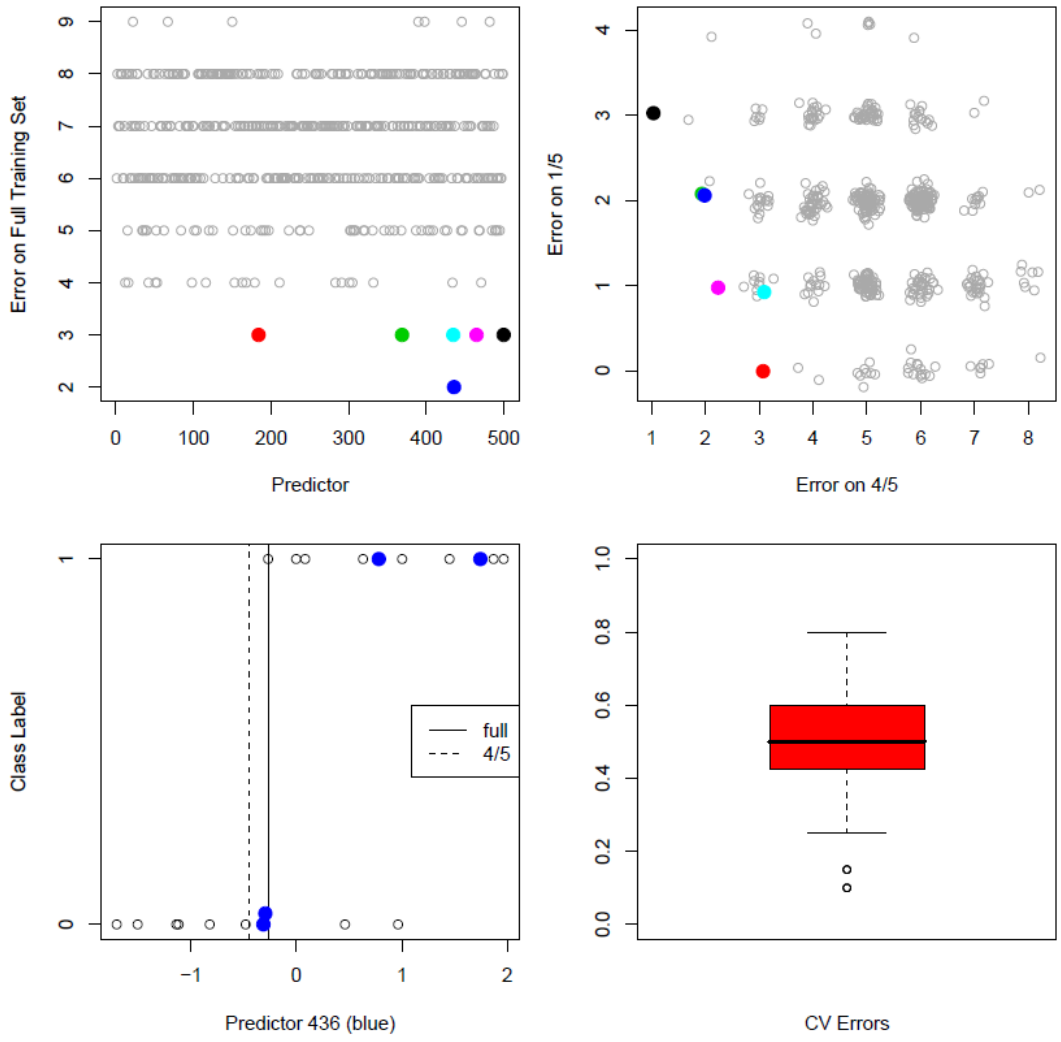
Spójrzmy na inne podejście dla tego przykładu:
(1) Podziel próbki losowo na K grup.
(2) Dla każdego podziału $k = 1, 2, \ldots, K$:
  (A) Znajdź podzbiór „dobrych" predyktorów, które wykazują dość silną (jednowymiarową) korelację z etykietami klas, używając wszystkich próbek z wyjątkiem tych z $k$-tej części.
  (B) Używając tylko tego podzbioru predyktorów, zbuduj wielowymiarowy klasyfikator, używając wszystkich próbek z wyjątkiem tych z $k$-tej części.
  (C) Użyj klasyfikatora, aby dokonać predykcji etykiet klas dla próbek w $k$-tej części.
Czy to jest dobry sposób?

# Dobry i zły sposób przeprowadzenia kroswalidacji (2)

Spójrzmy na inne podejście dla tego przykładu:
(1) Podziel próbki losowo na K grup.
(2) Dla każdego podziału $k = 1, 2, \ldots, K$:
  (A) Znajdź podzbiór „dobrych" predyktorów, które wykazują dość silną (jednowymiarową) korelację z etykietami klas, używając wszystkich próbek z wyjątkiem tych z $k$-tej części.
  (B) Używając tylko tego podzbioru predyktorów, zbuduj wielowymiarowy klasyfikator, używając wszystkich próbek z wyjątkiem tych z $k$-tej części.
  (C) Użyj klasyfikatora, aby dokonać predykcji etykiet klas dla próbek w $k$-tej części.
Czy to jest dobry sposób? Tak.

# Dobry i zły sposób przeprowadzenia kroswalidacji (3)

# Przykład z n=20 i p=500



# Bibliografia

1. Trevor Hastie, Robert Tibshirani, Jerome Friedman, *The Elements of Statistical Learning. Data Mining, Inference, and Prediction*