Metody nieparametryczne w statystyce

Tomasz Wójtowicz

Wydział Zarządzania
AGH Akademia Górniczo-Hutnicza im. S. Staszica w Krakowie

Ocena własności prognostycznych modeli

Ocena modeli

Ocena modelu:

- poprawność estymacji,
- jakość dopasowania,
- jakoś prognoz.

Poprawność estymacji:

- testy, np. homoskedastyczności, autokorelacji, normalności, stabilności

Jakość dopasowania:

- współczynnik determiniacji.

Ocena modeli

Ocena zdolności prognostycznych:

- poprawność i dobre dopasowanie modelu do danych nie gwarantują, że model będzie przydatny do prognozowania,
- model powinien być dobrze dopasowany do danych, ale też dobrze prognozować wartości dla nowych obserwacji,
- *bias – variance*

Ocena modeli

Wartość oczekiwana błędu prognozy MSE można zapisać w postaci:

$$E\left(y_0 - \hat{f}(x_0)\right)^2 = var\left(\hat{f}(x_0)\right) + \left[bias\left(\hat{f}(x_0)\right)\right]^2 + var(\varepsilon)$$

gdzie:

$var(\varepsilon)$ - nieredukowalna wariancja błędu,

$var\left(\hat{f}(x_0)\right)$ - wariancja oszacowań $f(x_0)$ powstająca, gdy model estymujemy na podstawie różnych danych, np. gdy dodajemy nowe dane w celu polepszenia oszacowania,

$bias\left(\hat{f}(x_0)\right)$ – obciążenie oszacowań $f(x_0)$ wynikające z konstrukcji modelu, który może niedostatecznie uwzględniać złożoność problemu; dodawanie nowych danych nie poprawia oszacowania

Ocena modeli

$$bias\left(\hat{f}(x_0)\right)$$

Proste modele są często obciążone błędem systematycznym (**bias**) wynikającym np. z przyjęcia błędnych założeń, pominięcia istotnej cechy. Możliwość zbyt słabego dopasowania do danych (**underfitting**).

$$var\left(\hat{f}(x_0)\right)$$

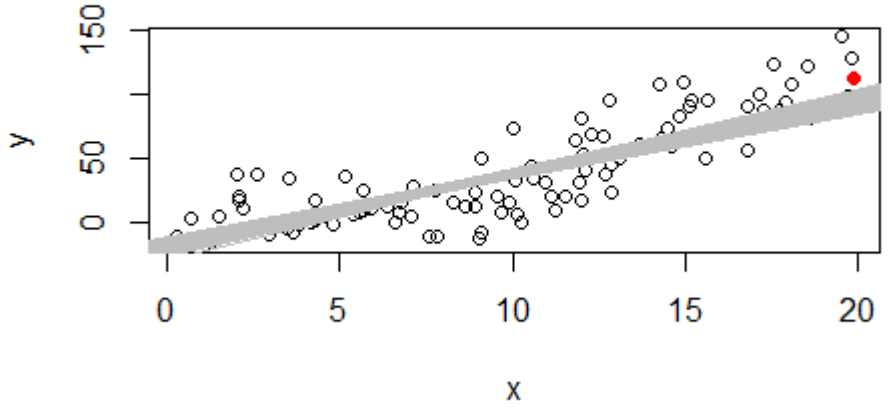
Złożone modele często bardzo dobrze pasują do danych jednak są wrażliwe na drobne zmiany w zbiorze danych. Błędy uzyskanych prognoz mogą być duże (**variance**). Możliwość nadmiernego dopasowania (**overfitting**).

Nie można jednocześnie zminimalizować obciążenia i wariancji.

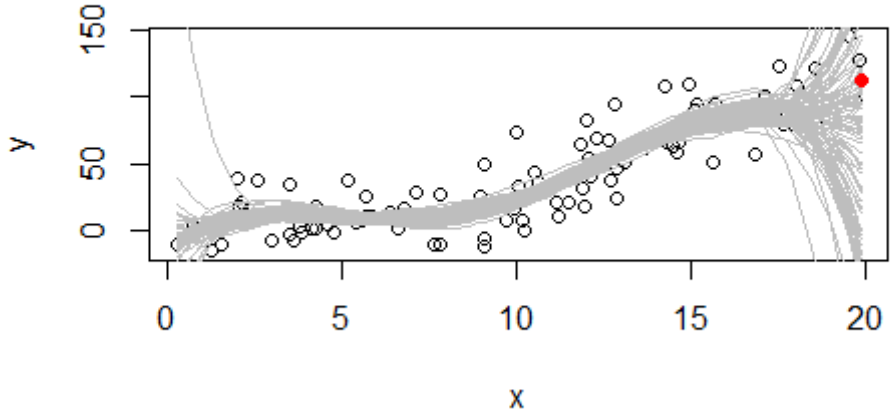
Ocena modeli

Przykład:

Na podstawie 60 spośród pierwszych 90 danych estymujemy model liniowy i wielomian stopnia 6. Interesuje nas prognoza wartości y_{100}, która w tym przypadku wynosi ok. 112.

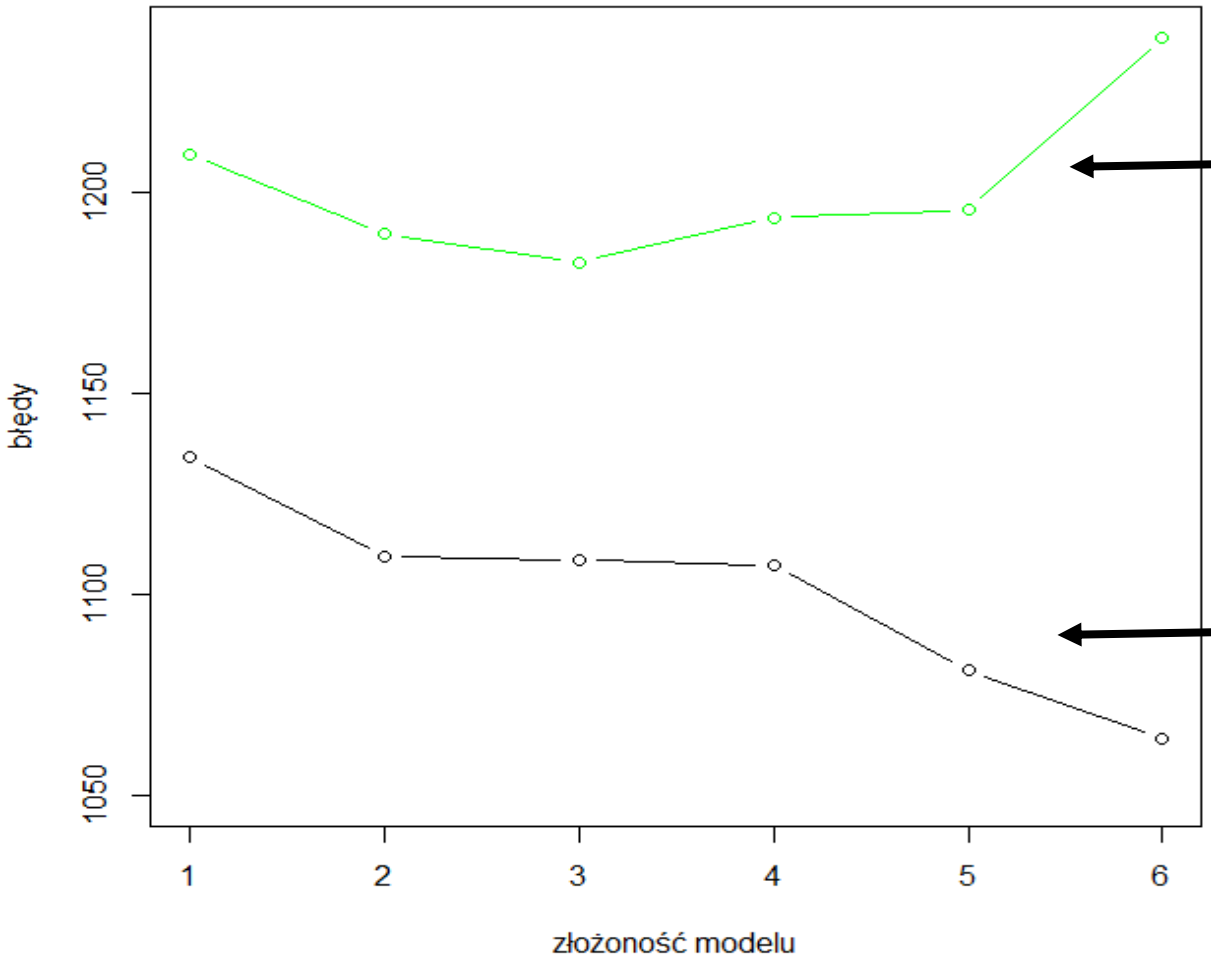


prognozy są obciążone, ale o małym zróżnicowaniu



obciążenie prognoz jest niewielkie, ale ich zróżnicowanie jest duże

Cross validation



błąd generalizacji – błędy prognoz (obliczone na nowych danych)

błąd uczenia – obliczony na danych, na których model był estymowany

Cross validation

Ocena zdolności prognostycznych - sprawdzian krzyżowy (**cross validation**)

- podział zbioru danych na dwa rozłączne zbiory:
  - zbiór uczący (training set) – zwykle jest liczniejszy,
  - zbiór testowy (test set)
- zwykle zbiór uczący jest liczniejszy,
- estymacja i ocena poprawności modelu tylko na podstawie danych ze zbioru uczącego,
- ocena jakości prognoz uzyskanego modelu tylko na podstawie danych ze zbioru testowego.

Cross validation

Uwagi:

- liczność zbioru uczącego i testowego (zwykle ok. 70% i 30%),
- sposób doboru danych do każdego ze zbiorów:
  - błędny dobór może spowodować duże obciążenie uzyskanej oceny jakości prognoz,
- wynik zależy od sposobu podziału danych (ew. błędy systematyczne),
- przed podziałem zbioru warto dokonać permutacji danych (żeby podział był losowy),
- pozwala ocenić błędy prognoz nie tylko na danych, na których model był estymowany, ale również na nowych danych,

Cross validation

Uwagi cd.

- poprawnie wykonany sprawdzian krzyżowy pozwala wybrać odpowiedni model (spośród kilku konkurencyjnych),
- ocena błędu prognozy jest zawyżona (bo model jest estymowany na mniejszej liczbie danych).

Częstą praktyką (zwłaszcza w uczeniu maszynowym) jest podział zbiory treningowego na dwa zbiory:

- zbiór uczący (na którym estymowane są modele),
- zbiór walidacyjny (na którym model jest wstępnie oceniany, dopasowywane są dodatkowe parametry itp).

LOOCV

Leave-One-Out Cross Validation

- metoda oceny własności prognostycznych modelu,

W przypadku n-elementowego zbioru danych (x_i, y_i) dla kolejnych $i = 1, \dots, n$ należy wykonać następujące etapy:

- wybrać daną (x_i, y_i),
- na podstawie pozostałych danych wyestymować parametry modelu,
- na podstawie (x_i, y_i) obliczyć błąd prognozy, np. MSE_i

$$CV_{(n)} = \frac{1}{n}\sum_{i=1}^{n} MSE_i$$

LOOCV

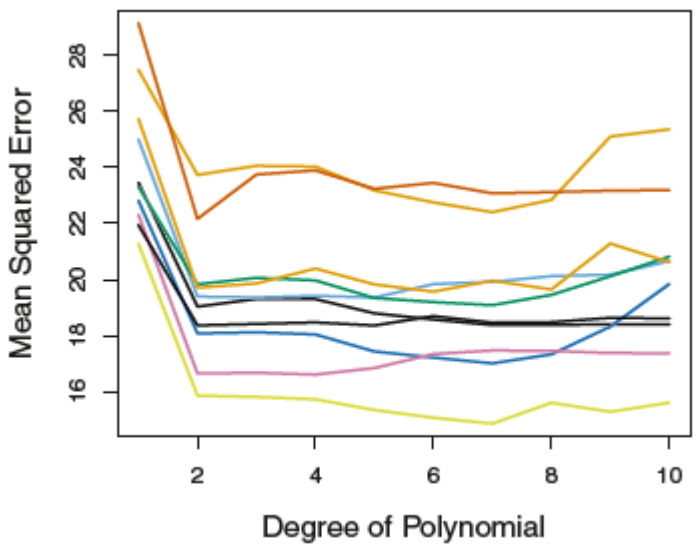
Zalety:

- wynik nie zależy od sposobu podziału danych na zbiór uczący i testowy,
- oszacowanie wariancji błędu prognozy jest mniej obciążone,
- nie zawyża oszacowania MSE tak jak sprawdzian krzyżowy (estymacja modelu prawie na całej próbie),
- ocena nie ma charakteru losowego.

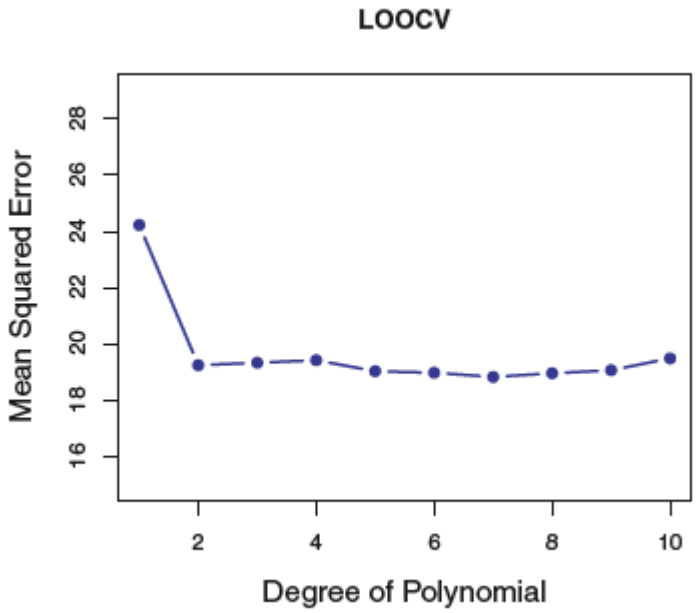
Wady:

- złożoność obliczeniowa.

LOOCV



MSE dla różnych powtórzeń sprawdzianu krzyżowego



MSE dla LOOCV

źródło: *An Introduction to Statistical Learning*

LOOCV

W przypadku regresji liniowej lub wielomianowej można pokazać, że:

$$CV_{(n)} = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{y_i - \hat{y}_i}{1 - h_i}\right),$$

gdzie h_i - wsp. dzwigni dla x_i,. Dla regresji liniowej mamy:

$$h_i = \frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum_{i'=1}^{n}(x_{i'} - \bar{x})^2}$$

Własności:

- $\frac{1}{n} \leq h_i \leq 1$

k-fold CV

k-krotny sprawdzian krzyżowy (**k-fold Cross-Validation**)

- jest alternatywą dla LOOCV,
- łączy zalety CV i LOOCV,
- polega na k-krotnym zastosowaniu procedury sprawdzianu krzyżowego do k różnych rozłącznych zbiorów testowych

	k (w miarę równych) części				
$i = 1$	zbiór uczący				zb. testowy
$i = 2$				zb. testowy	
...	...				
$i = k - 1$		zb. testowy			
$i = k$	zb. testowy				

k-fold CV

k-krotny sprawdzian krzyżowy (**k-fold Cross-Validation**)

	k (w miarę równych) części				
$i = 1$	zbiór uczący				zb. testowy
$i = 2$				zb. testowy	
…	…				
$i = k - 1$		zb. testowy			
$i = k$	zb. testowy				

dla każdego $i = 1, \ldots, k$, na podstawie danych ze zbioru testowego obliczamy

miarę błędu prognozy, np. MSE_i

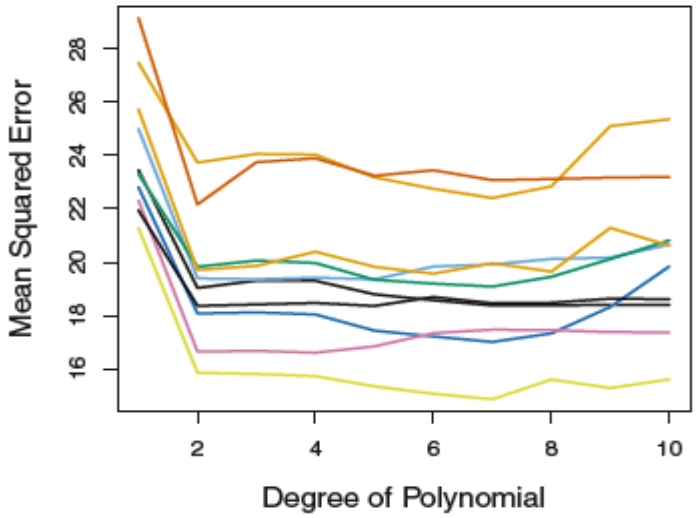
$$CV_{(k)} = \frac{1}{k}\sum_{i=1}^{k} MSE_i$$

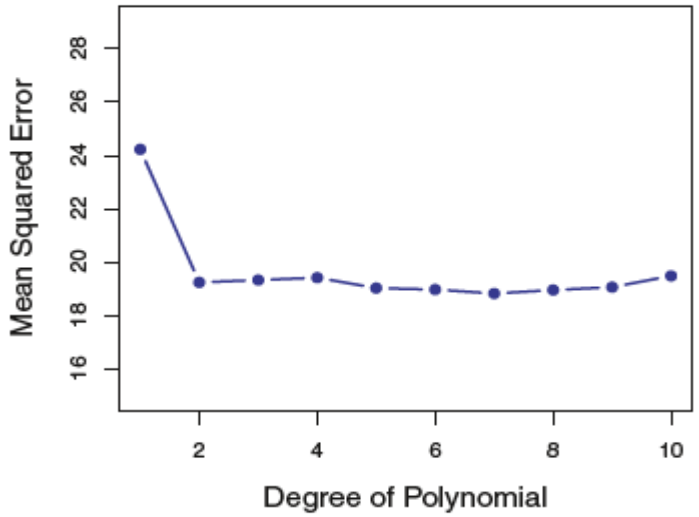
k-fold CV

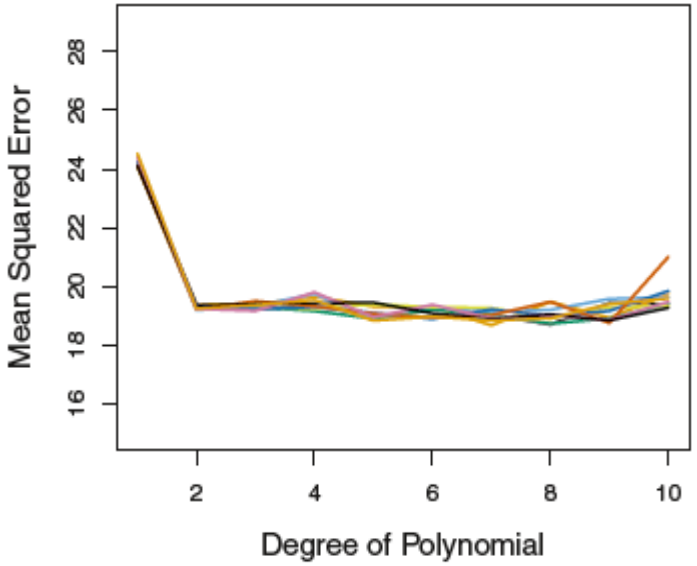
Właściwości k-krotnego sprawdzianu krzyżowego

- LOOCV jest szczególnym przypadkiem k-fold CV (z $k = n$),
- jest mniej wymagający obliczeniowo niż LOOCV (model jest estymowany tylko k razy zamiast n),
- obciążenie oszacowania MSE za pomocą k-fold CV jest większe niż LOOCV, ale mniejsze niż CV,
- wariancja oszacowania MSE za pomocą k-fold CV jest mniejsza niż za pomocą LOOCV, (k-fold CV jest bardziej odporna na zmiany w zbiorze uczącym, LOOCV – średnia z silnie skorelowanych oszacowań)

k-fold CV







MSE dla:

- CV,
- LOOCV,
- *k*-fold CV

źródło: *An Introduction to Statistical Learning*