Sposoby oceniania i wyboru modelu

Stanisław Wilczyński

Uniwersytet Wrocławski

09.11.2017

Wstęp

- Model a rzeczywistość
- Wybór jednego spośród wielu modeli
- Ocenianie modelu

Błąd modelu

X - wektor wejść, Y - zmienna docelowa, $\hat{f}(X)$ - model, $\mathcal{T}$ - zbiór treningowy

$$L(Y, \hat{f}(X)) = (Y - \hat{f}(X))^2$$
$$L(Y, \hat{f}(X)) = |Y - \hat{f}(X)|$$

A dla wielu Xów?

Dla całego modelu błędem jest suma/średnia pojedynczych błedów

Błąd w przypadku dyskretnym

G - zmienna jakościowa (jedna z wartości $1, \ldots, K$),
$p_k(X) = P(G = k|X)$, $\hat{G}(X) = argmax_k \hat{p}_k(X)$

funkcja straty 0-1

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

Funkcja wiarygodności jako funkcja straty

$$L(G, \hat{p}(X)) = -2 \sum_{k=1}^{K} I(G = k) \log \hat{p}_k(X)$$

Rodzaje błędów

Błąd testowy

$$Err_{\mathcal{T}} = E\left[L(Y, \hat{f}(X))|\mathcal{T}\right]$$

Oczekiwany błąd predykcji

$$Err = EErr_{\mathcal{T}}$$

Błąd treningowy

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

Definicja

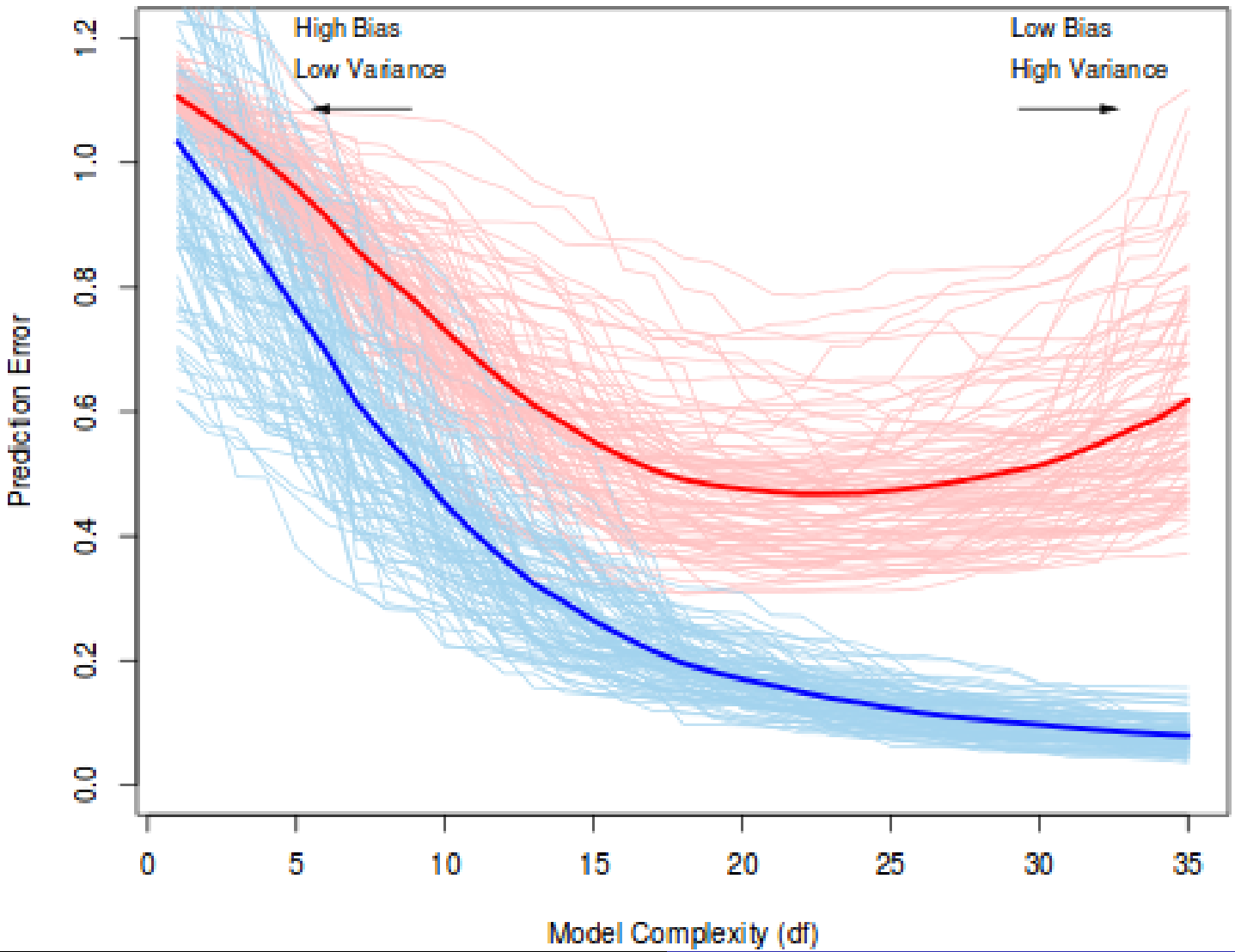
Obciążenie

Błąd wynikający z błędnych założeń modelu (np. za mało parametrów lub złe ich wartości)

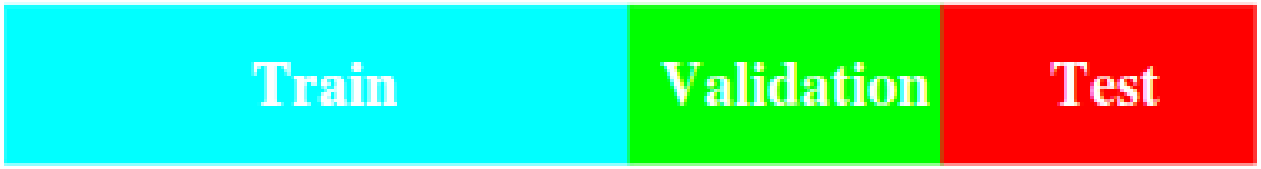
Wariancja

Błąd wynikający z wrażliwości na małe zmiany danych treningowych (np. zbyt dużo parametrów - model "uczy się" szumu w danych)

Ilustracja


High Bias
Low Variance
Low Bias
High Variance
Prediction Error
0.0
0.2
0.4
0.6
0.8
1.0
1.2
0
5
10
15
20
25
30
35
Model Complexity (df)

Dużo danych



- Zbiór treningowy - na tym zbiorze danych uczymy modele
- Zbiór walidacyjny - na tym zbiorze estymujemy błąd predykcji i na jego podstawie wybieramy model
- Zbiór testowy - na tym zbiorze oceniamy wybrany model
- Proporcje?
- Co zrobić kiedy danych jest za mało?

Rozkład błędu

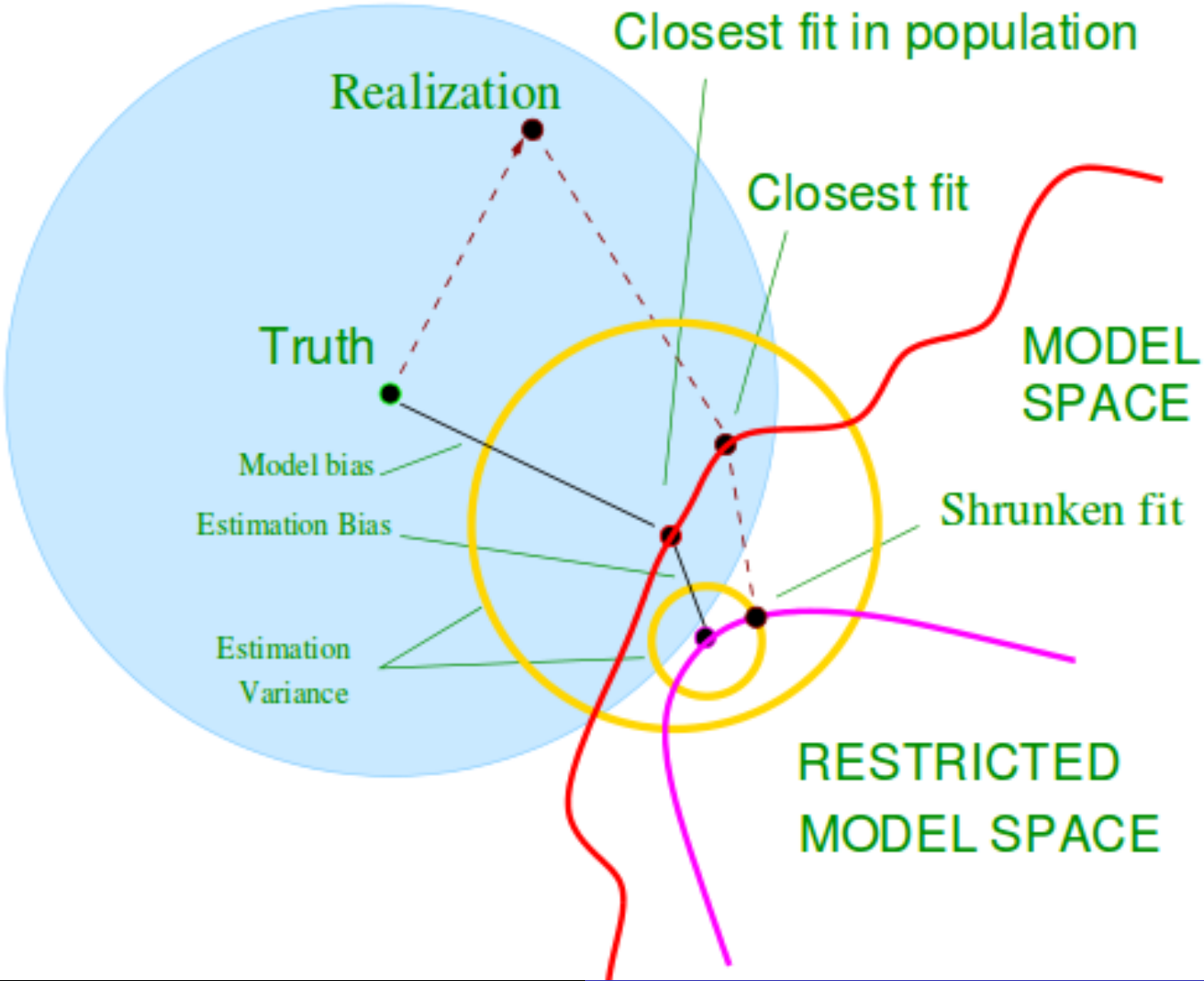
Zakładamy, że $Y = f(X) + \varepsilon$, gdzie $E\varepsilon = 0$ i $Var\varepsilon = \sigma^2$. Wtedy błąd w $X = x_0$ wynosi

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

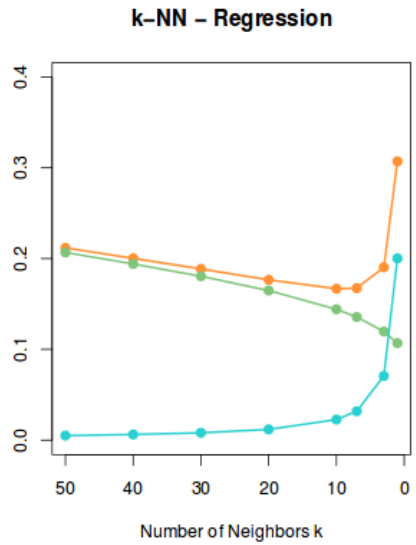
Uwaga

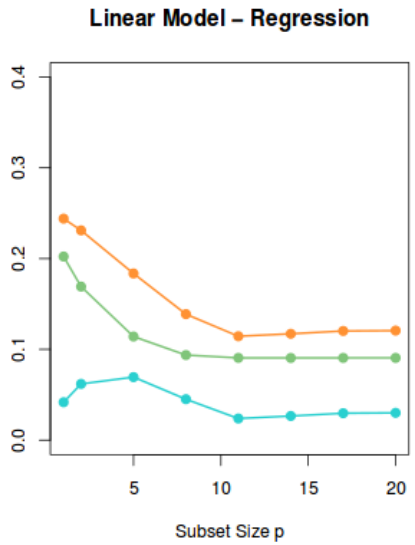
Zwykle przy zwiększaniu stopnia skomplikowania modelu maleje obciążenie, ale rośnie wariancja.

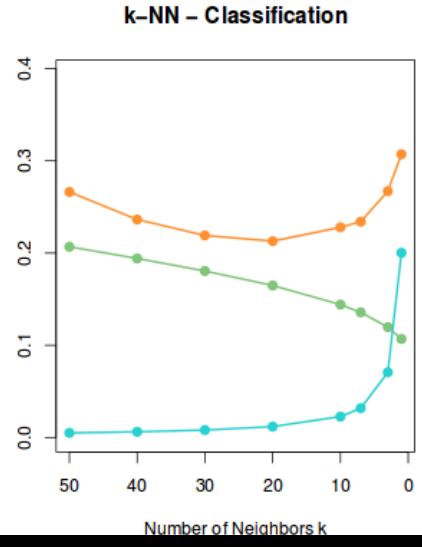
Wariancja i obciążenie w przestrzeni modeli

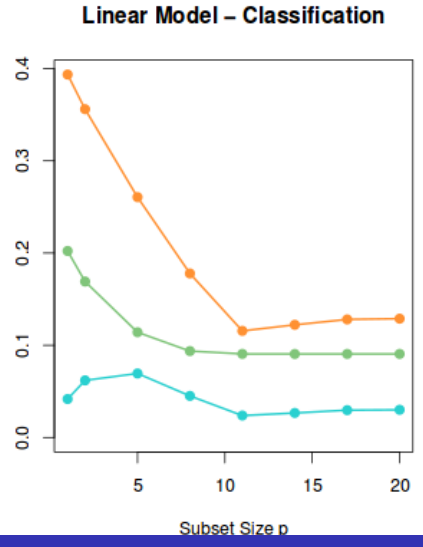


Rozkład w praktyce









Błąd in-sample

Błąd treningowy

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

Błąd in-sample

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

Błąd in-sample lepiej przybliża błąd predykcji, gdyż pozwala zwrócić uwagę na przeuczenie - jeśli ono nastąpi będzie duży w porównaniu do błędu treningowego, który zwykle jest wyraźnie mniejszy niż błąd predykcji.

Optymizm

Optymizm - wzory

$$\begin{aligned} op &= Err_{in} - \overline{err} \\ \omega &= E_y op \end{aligned}$$

Optymizn określa jak bardzo przy tworzeniu naszego modelu polegamy na znajomości klas dla x_i, czyli y_i.

Optymizm cd.

Optymizm dla standardowych L

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

Optymizm dla regresji liniowej

$$\omega = \frac{2d}{N} \sigma^2$$

Szacujemy $E_y(Err_{in})$

$$E_y(Err_{in}) = E_y(\overline{err}) + E_y(op) = \overline{err} + \omega$$

A może by tak użyć błędu in-sample do estymacji błędu predykcji?

Statystyka C_p i AIC

Statystyka C_p - wybieramy model z najmniejszą wartością

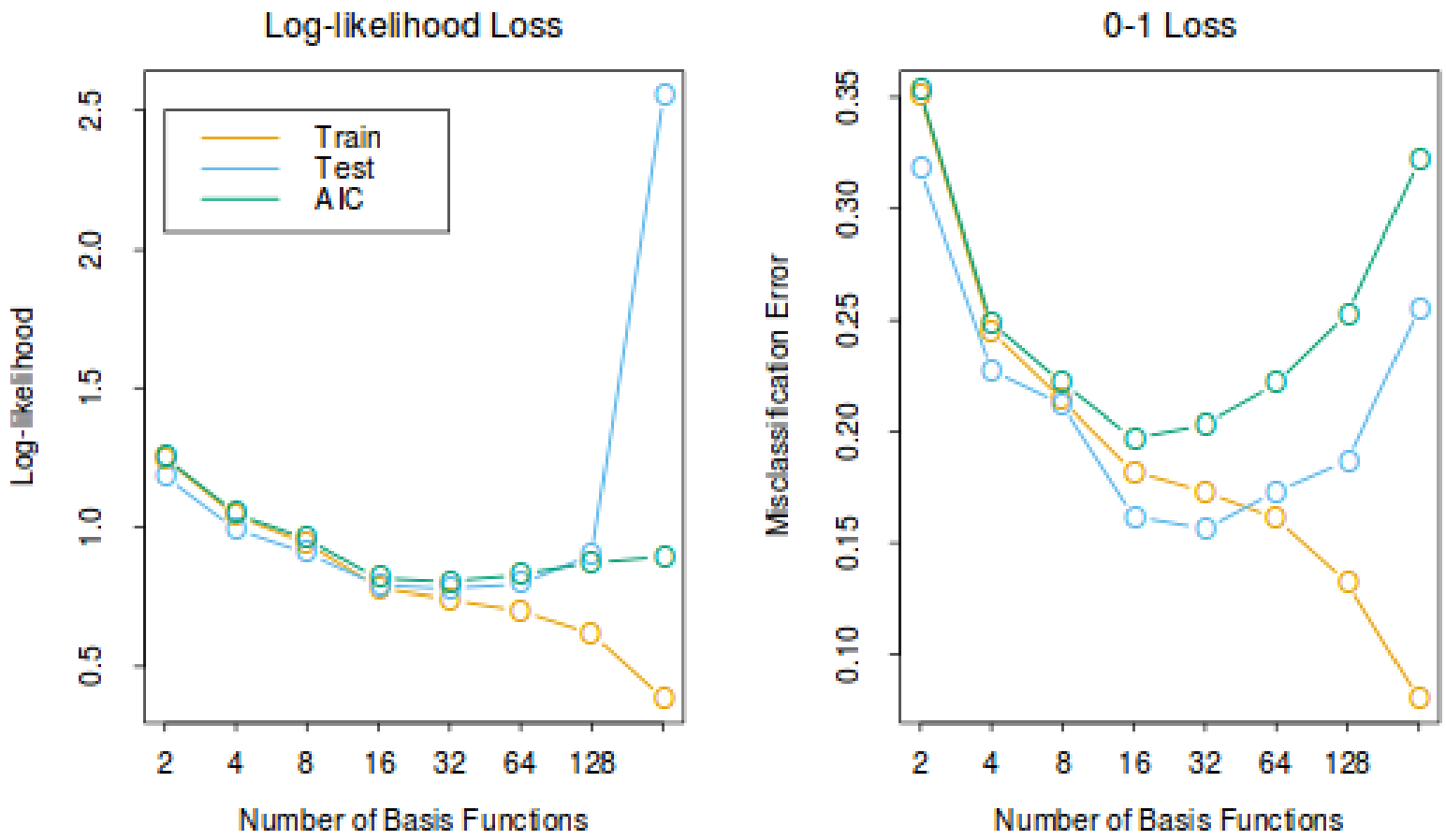
$$C_p = \overline{err} + 2\frac{d}{N}\hat{\sigma}^2$$

Kryterium informacyjne Akaike - wybieramy model z najmniejszą wartością

$$\begin{aligned} loglik &= \sum_{i=1}^{N} \log P_{\hat{\theta}}(y_i) \\ AIC &= -\frac{2}{N} \cdot loglik + 2 \cdot \frac{d}{N} \end{aligned}$$

Dla modelu gaussowskiego (błąd ma rozkład normalny) statystyka C_p i AIC są równoważne.

AIC - ilustracja



Statystyka bayesowska vs statystyka częstościowa

Przykład1

Słyszę dzwoniący telefon. Mam model pozwalający mi identyfikować skąd dochodzi dźwięk.

- Statystyka częstościowa: Na podstawie tego modelu wniokuję gdzie najprawdopodobniej jest telefon.
- Statystyka bayesowska: oprócz modelu wiem, gdzie ostatnio zdarzało mi się kłaść telefon. Wnioskuję, więc nie tylko z modelu, ale i z dodatkowej wiedzy który posiadam. Podsumowując w statystyce bayesowskiej prawdopodobieństwo zdarzeń jest uaktualniane na podstawie wcześniejszych zdarzeń. Wiedzę o tych wcześniejszych zdarzeniach nazywamy wiedzą **apriori**.

Rozkład apriori i aposteriori

W statystyce bayesowskiej prawdopodobieństwo jest subiektywne - sami możemy nadać rozkład apriori na jakąś zmienną.

Przykład2

Załóżmy, że mamy niesymetryczną monetę z parametrem niesymetryczności θ. Nakładamy pewien rozkład apriori na $\theta, \Pi(\theta)$. Niech $f(x|\theta)$ bedzię warunkową funkcją masy prawdopobieństwa wyrzucenia orła lub reszki. Powiedzmy, że rzuciliśmy monetą raz i wypadł x. Wtedy prawdopobieństwo aposteriori naszego parametru niesymetryczności wynosi

$$\Pi(\theta|x) = \frac{f(x|\theta)\Pi(\theta)}{\int_\Theta f(x|\theta_0)\Pi(\theta_0)d\theta_0}$$

Zastosowanie statystyki bayesowskiej jako kryterium wyboru modelu

Załóżmy, że mamy zbiór danych Z, rozważamy modele $M_1, \ldots, M_m$ i odpowiadające im wketory parametrów $\theta_1, \ldots, \theta_n$. Zakładamy rozkład apriori Π na modele, $f(\theta_i|M_i)$ na parametry oraz $g(Z|\theta_i, M_i)$ - gęstość przy danych parametrach. Jako model wybierzemy ten z największą wartością prawdodpodobieństwa aposteriori

$$P(M_i|Z) \propto \Pi(M_i) \cdot P(Z|M_i) = \Pi(M_i) \cdot \int_\Theta g(Z|\theta_i, M_i) f(\theta_i|M_i) d\theta_i$$

Porównanie dwóch modeli - czynnik Bayesa

$$\frac{P(M_i|Z)}{P(M_j|Z)} = \frac{\Pi(M_i)}{\Pi(M_j)} \frac{P(Z|M_i)}{P(Z|M_j)}$$

Zastosowanie statystyki bayesowskiej jako kryterium wyboru modelu cd.

Niestety $P(Z|M_i)$ (całka z poprzedniego slajdu) jest trudna do obliczenia. W związku z tym przybliżamy ją używając aproksymacji Laplace'a. Niech d_i będzie wymiarowością M_i, a N liczbą próbek w Z.

Aproksymacja Laplace'a

$$\log P(Z|M_i) \approx \log g(Z|\hat{\theta}_i, M_i) - \frac{d_i}{2} \log N$$

Bayesowskie kryterium informacyjne Schwarza i AIC

$$\begin{aligned} BIC &= -2 \log g(Z|\hat{\theta}_i, M_i) + d_i \log N = -2loglik + d \log N \\ AIC &= -2loglik + d \end{aligned}$$

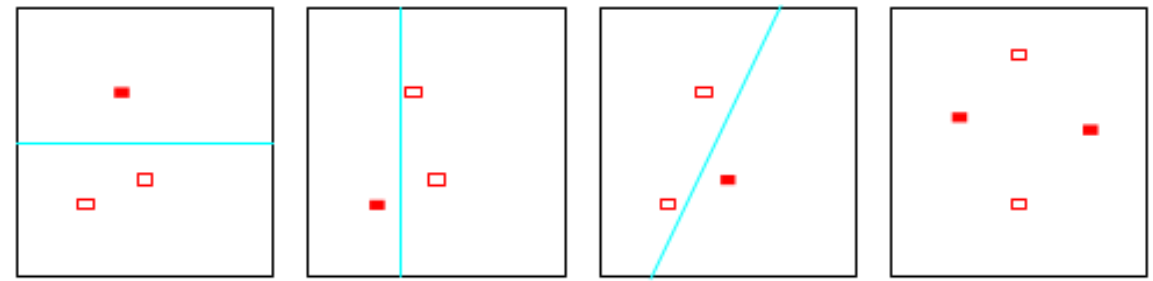
Własności BIC

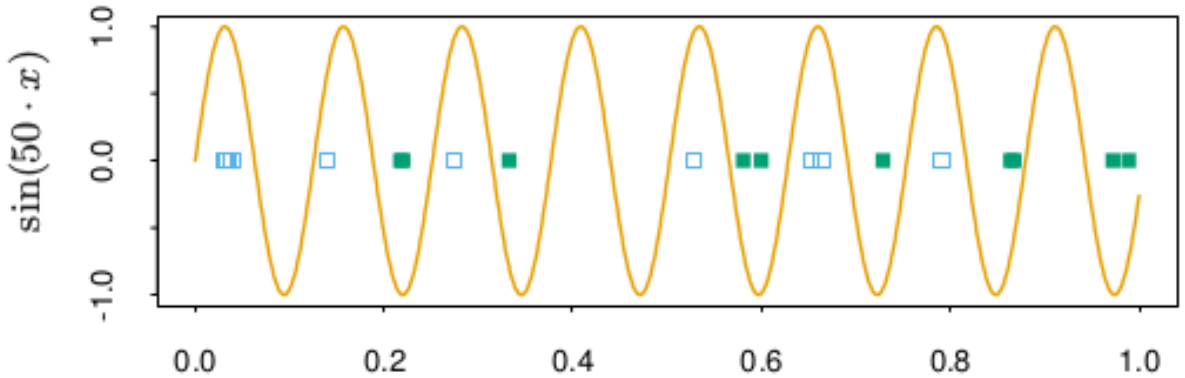
- Podobnie jak w AIC polegamy na estymatorach największej wiarygodności
- Mocne podstawy teoretyczne
- Dla $N > e^2$ BIC penalizuje złożone modele dużo mocniej niż AIC
- Nie tylko dostajemy najbardziej prawdopodobny model ale również o ile jest lepszy od pozostałych modeli
- BIC jest zgodne - jeśli wśród rozważanych modeli jest ten właściwy to przy $N \to \infty$ prawdopodobieństwo wyboru przez BIC tego modelu zbiega do 1.
- BIC działa, jeśli $d << N$
- BIC nie preferuje wybrania "przeuczonego" modelu (w przeciwieństwie do AIC)

Wymiar Wapnika-Czerwonienkisa (VC dimension)

Rozważmy klasy funkcji indykatorowych $\{f(x, \alpha)\}$ indeksowanych parametrem α np. $f_1 = I(\alpha_0 + \alpha_1 x > 0)$, $f_2 = I(\sin(\alpha \cdot x) > 0)$.
Zbiór punktów przestrzeni jest rozdzielany klasą funkcji $\{f(x, \alpha)\}$, gdy niezależnie jak przydzielimy punkty do 2 grup, istnieje reprezentant klasy, który dokładnie je oddziela.
Wymiarem VC klasy $\{f(x, \alpha)\}$ nazywamy najwiekszą liczbę punktów (w jakiejś konfiguracji), która jest rozdzielana przez przedstawiciela klasy.

Przykład VC





Wymiar VC cd.

Zastosowania

Wymiar VC daje możliwość znalezenia górnego ograniczenia na optymizm.

Procedura wybierania modelu przy użyciu VC wygląda następująco:

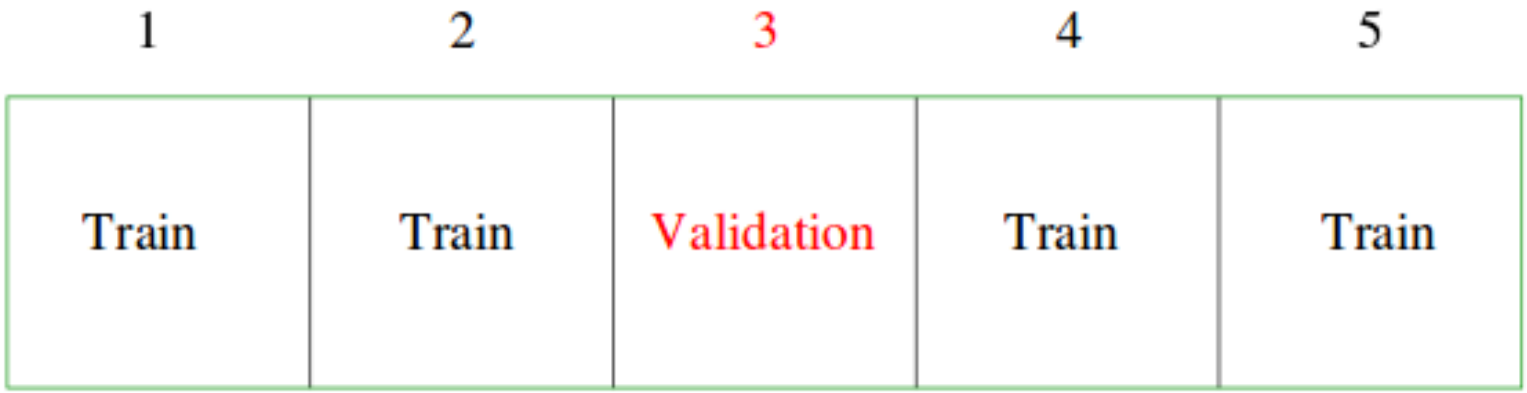
- Rozważamy modele z rosnącymi VC : $h_1 < h_2 < h_3 < \ldots$
- Wybieramy model o najmniejszym górnym ograniczeniu

Walidacja krzyżowa (cross validation)

- Jedna z najprostszych metod oceniania modelu = szacowania błędu predykcji
- Dzielimu zbiór danych na K części
- W każdym z K kroków bierzemy jedną z K części i traktujemy ją jako zbiór walidacyjny, a resztę jako zbiór treningowy
- Uśredniamy wynik po tych K krokach

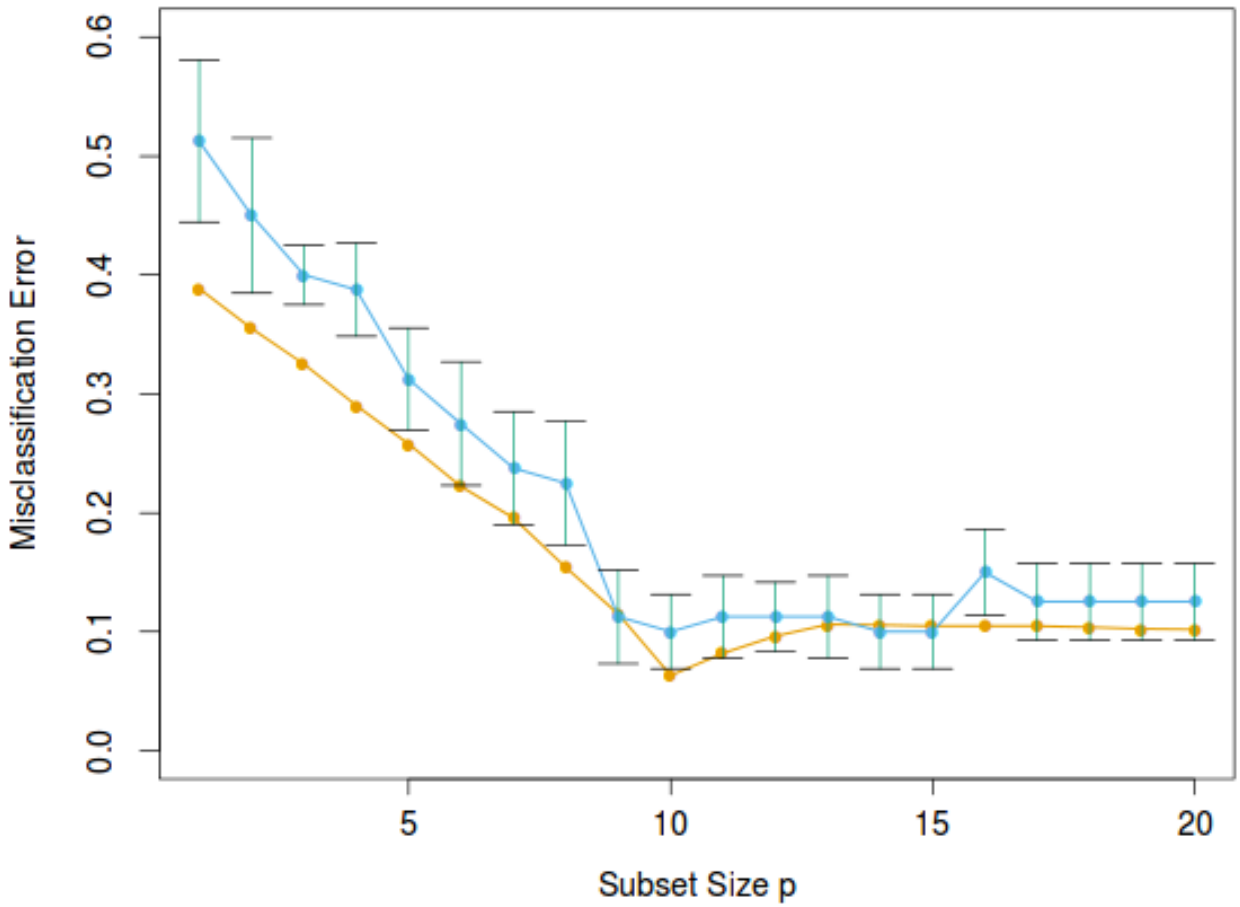
Walidacja krzyżowa cd.

Typowymi wyborami K są $5, 10$ - zwykle zaniżające skuteczność modelu czy N (jednak niezalecane z powodu dużej wariancji).



Walidacja krzyżowa cd.

Poniżej przykład dla 10krotnej walidacji krzyżowej (niebieski) w porównaniu do rzeczywistego błędu predykcji (pomarańczowy).



Jak wykonywać CV - dobre/złe praktyki

- Dane należy rozdzielić do grup losowo (random shuffle)

Czy tak można??

1. Znajdźmy p predyktorów mocno skorelowanych z Y
2. Zbudujmy model używając tylko tych predyktorów
3. Oceńmy jego skuteczność za pomocą CV

Jak wykonywać CV - dobre/złe praktyki

- Dane należy rozdzielić do grup losowo (random shuffle)

Czy tak można??

1. Znajdźmy p predyktorów mocno skorelowanych z Y
2. Zbudujmy model używając tylko tych predyktorów
3. Oceńmy jego skuteczność za pomocą CV

Nie

W zbiór walidacyjny nie jest niezależny od treningowego, predyktory już "widziały" zbiór walidacyjny. Może to prowadzić to ogromnego przesacowania skuteczności modelu.

Jak wykonywać CV - dobre/złe praktyki

- Dane należy rozdzielić do grup losowo (random shuffle)

Czy tak można??

1. Znajdźmy p predyktorów mocno skorelowanych z Y
2. Zbudujmy model używając tylko tych predyktorów
3. Oceńmy jego skuteczność za pomocą CV

Nie

W zbiór walidacyjny nie jest niezależny od treningowego, predyktory już "widziały" zbiór walidacyjny. Może to prowadzić to ogromnego przesacowania skuteczności modelu.

Jak w takim razie zrobić to poprawnie?

Walidacja krzyżowa - kiedy nie stosować?

- Zbiór danych zmienia się w czasie
- Liczba danych jest zbyt mała (żeby podzielić zbiór)
- Jeśli uczenie modelu jest skomplikowane obliczeniowo, CV jest bardzo czasochłonna

Metody bootstrapowe - idea

- Załóżmy, że chcemy obliczyć statystyczną dokładność wartości $S(\mathbf{Z})$ liczonej na zbiorze $\mathbf{Z}$.
- Losujemy ze zwracaniem B zbiorów treningowych $\mathbf{Z}^{*b}$, $b = 1, \ldots, B$ każdy rozmiaru N z oryginalnego zbioru danych.
- Obliczamy interesującą nas wartość $S(\mathbf{Z})$ na każdym zbiorze treningowym i te wartości $S(\mathbf{Z}^{*1}), \ldots, S(\mathbf{Z}^{*B})$ wykorzystujemy do obliczenia statystycznej dokładności $S(\mathbf{Z})$ - obliczamy np. średnią i odchylenie standardowe naszych wartości.

Metody bootstrapowe cd.

Możemy użyć tej metody do estymacji błędu predykcji

Podejście 1

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

Podejście 2

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

Podejście 3

$$\widehat{\mathrm{Err}}^{(.632)} = 0.368 \cdot \overline{err} + 0.632 \cdot \widehat{\mathrm{Err}}^{(1)}$$

Wymiar VC do wybierania modelu

Z prawdopodobieństwe co najmniej $1 - \eta$

$$Err \leqslant \overline{err} + \frac{\epsilon}{2}\left(1 + \sqrt{1 + \frac{4 \cdot \overline{err}}{\epsilon}}\right)$$

$$\text{gdzie} \quad \epsilon = a_1 \frac{h[\log(a_2 N/h) + 1] - \log(\eta/4)}{N}$$

$$\text{oraz } 0 < a_1 \leqslant 4,\ 0 < a_2 \leqslant 2.$$