WYKŁAD IV : Ocena klasyfikatorów. Estymacja prawdopodobieństwa błędnej klasyfikacji.

MiNI PW, semestr letni 2013/2014

Estymacja prawdopodobieństwa błędnej klasyfikacji

$$\mathcal{U} = \{(x_1, y_1), \ldots, (x_n, y_n)\} \quad - \quad \text{próba ucząca}$$

$\hat{d}$ – klasyfikator skonstruowany na podstawie próby uczącej.
Chcemy estymować warunkowy błąd klasyfikacji

$$Err_{\mathcal{U}} = P(\hat{d}(X) \neq Y \mid \mathcal{U}), \qquad \text{gdzie} \ \ (X, Y) \perp \mathcal{U}$$

Błąd klasyfikacji dla przyszłej obserwacji klasyfikatora skonstruowanego na podstawie próby uczącej $\mathcal{U}$.
Błąd prognozy. Zależy od próby uczącej $\mathcal{U}$.
Bezwarunkowy błąd klasyfikacji predykcji

$$Err = E(Err_{\mathcal{U}})$$

Uśredniamy losowość próby uczącej $\mathcal{U}$ na podstawie której skonstruowano klasyfikator.
Uwaga Okazuje się, że z reguły jesteśmy w stanie skonstruować zadowalające estymatory błędu bezwarunkowego Err, ale nie $Err_{\mathcal{U}}$, choć chcielibyśmy szacować ten drugi !!

Estymator przez powtórne podstawienie (resubstitution)

$$\bar{err} = \#\{(x_i, y_i) \in \mathcal{U} : \ \hat{d}(x_i) \neq y_i\}/n.$$

Oszacowanie $\bar{err}$ jest naturalnym estymatorem $Err_{\mathcal{U}}$. Ale elementy próby $\mathcal{U}$ pełnią funkcję niezależnych od $\mathcal{U}$ obserwacji $(\mathbf{X}, Y)$. Jest estymatorem optymistycznym w tym sensie, że daje z reguły wartość $<$ od $Err_{\mathcal{U}}$.
Próba ucząca została użyta do konstrukcji $\hat{d}$ i szacowania błędu.

Uwaga ważna Nie należy porównywać metod klasyfikacji w oparciu o porównanie błędów przez powtórne podstawienie, szczególnie jeśli metody różnią się liczbą parametrów ! QDA z reguły ma mniejszy błąd $\bar{err}$ od LDA, ale niekoniecznie ma mniejsza wartość $Err_{\mathcal{U}}$.
Dla metody 1-NN (najbliższego sąsiada z $k = 1$) $\bar{err} = 0$!!

Estymator metodą próby testowej

Próba testowa $\mathcal{T}$ o liczności m niezależna od próby $\mathcal{U}$,

$$\hat{E}rr_{\mathcal{U}} = \#\{(x_i, y_i) \in \mathcal{T} : \ \hat{d}(x_i) \neq y_i\}/m$$

$$E(\hat{E}rr_{\mathcal{U}} \mid \mathcal{U}) = Err_{\mathcal{U}} \quad - \quad \text{estymator nieobciążony } Err_{\mathcal{U}}$$

W praktyce musimy podzielić próbę z pełną informacją na próbę uczącą i próbę testową, mamy mniej obserwacji do konstrukcji klasyfikatora. Z reguły klasyfikator konstruowany na podstawie wszystkich dostępnych danych.

Kroswalidacja

Kroswalidacja (sprawdzanie krzyżowe)

Dzielimy próbę uczącą na K części ($K = 5$, $K = 10$ lub $K = N$ z reguły)

$$\underbrace{\_\,\_\,\_\,\_\quad \_\,\_\,\_\,\_\,\_\quad \_\,\_\,\_\,\_\,\_\quad \ldots\ldots\quad \_\,\_\,\_\,\_\,\_}_{K-1 \quad \text{części}} \qquad \underbrace{\_\,\_\,\_\,\_\,\_}_{1 \quad \text{część}}$$

Konstruujemy klasyfikator na podstawie $K - 1$ części, testujemy na części pozostałej $\mathcal{U}^{-i}$.
Powtarzamy K razy: K wersji klasyfikatora i K oszacowań prawdopodobieństwa błędnej decyzji

$$\hat{E}rr_{\mathcal{U}^{-1}}, \hat{E}rr_{\mathcal{U}^{-2}}, \ldots, \hat{E}rr_{\mathcal{U}^{-k}}.$$

Oszacowanie ostateczne

$$\hat{Err} = (\hat{E}rr_{\mathcal{U}^{-1}} + \hat{E}rr_{\mathcal{U}^{-2}} + \ldots + \hat{E}rr_{\mathcal{U}^{-K}})/K$$

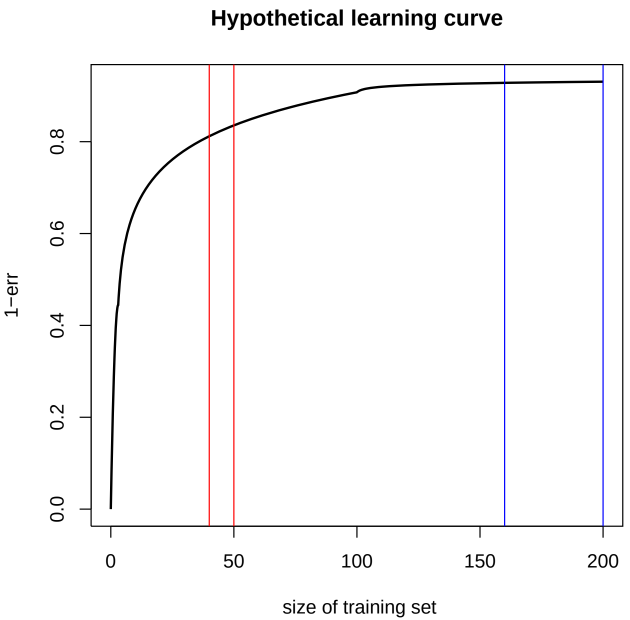
Problem: szacujemy prawdopodobieństwo błędu klasyfikatora opartego na (około) $n(K-1)/K$ elementach, używamy klasyfikatora opartego na n elementach.

Najmniejszy błąd dla $K = n$: $n-1$ elementów użytych do konstrukcji estymatora, sprawdzamy na n–tym, powtarzamy n razy, uśredniamy

$\longrightarrow$ – leave-one-out crossvalidation

najmniejsze obciążenie dla szacowania Err, największa wariancja.

Duża wariancja – wada wszystkich estymatorów kroswalidacyjnych!


Hypothetical learning curve
1−err
0.0
0.2
0.4
0.6
0.8
0
50
100
150
200
size of training set

Przykład. Dane brach dotyczą 5 parametrów geometrycznych 136 muszli znalezionych w jednej z 5 lokalizacji (zmienna LOC, $g = 5$). Estymator prawdopodobieństwa poprawnej klasyfikacji przez powtórne podstawienie metody LDA:

```
brach.lda=lda(LOC ~., data=brach)
 brach.pred=predict(brach.lda, newdata=brach)
 print(tabl <- table(brach$LOC, brach.pred$class))

     1  2  3  4  5
  1 22  0  0  0  3
  2  0 51  0  0  0
  3  0  7 21  0  2
  4  0  0  0 23  2
  5  0  0  0  0 22
 print(procent<-100.0*sum(diag(tabl))/sum(tabl))
 90.84967
```

Analogiczne oszacowanie dla metody QDA wynosi 96.078 (CM(2009), str. 45). Oceńmy prawdopodobieństwa poprawnej klasyfikacji metodą kroswalidacji *n*-krotnej.

```
for ( i in 1:length(brach$LOC)) {
brach.lda=lda(LOC ~ ., data=brach[-c(i), ])
brach.pred=predict(brach.lda, newdata=brach)
if (i==1)  predykcja=brach.pred$class[1]
else
predykcja=c(predykcja,brach.pred$class[i])
}
print(tabl <- table(brach$LOC, predykcja))
print(procent<-100.0*sum(diag(tabl))/sum(tabl))
```

```
 predykcja
     1  2  3  4  5
  1 19  0  0  1  5
  2  0 51  0  0  0
  3  0  8 20  0  2
  4  0  0  0 22  3
  5  0  0  0  0 22
> print(procent<-100.0*sum(diag(tabl))/sum(tabl))
[1] 87.5817
```

Analogiczny wynik dla QDA wynosi 87.58.
Różnica dla dwóch estymatorów (resubstytucji i CV) wynosi dla QDA około 8.5%, dla LDA tylko 3.3%. Różnica wynika z faktu, że QDF lepiej dopasowuje się do próby uczącej niż LDF. Dokładniejsza analiza wskazuje również, że selekcja zmiennych przy użyciu funkcji `step` prowadzi dla obu metod do zmniejszenia oszacowania błędu metodą kroswalidacji, gdy przy ocenie błędu metodą przez powtórne podstawienie błędy rosną !
Estymator błędu przez powtórne podstawienie nie daje realistycznej oceny błędu, również nie powinien być używany do porównania metod klasyfikacyjnych.

CV jako metoda estymacji $Err_{\mathcal{U}}$ i Err

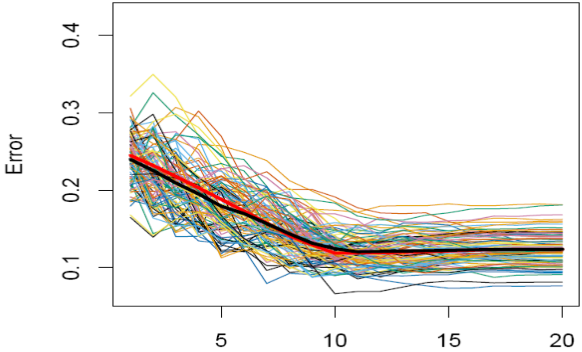
Przykład (ESL, rozdział VII)

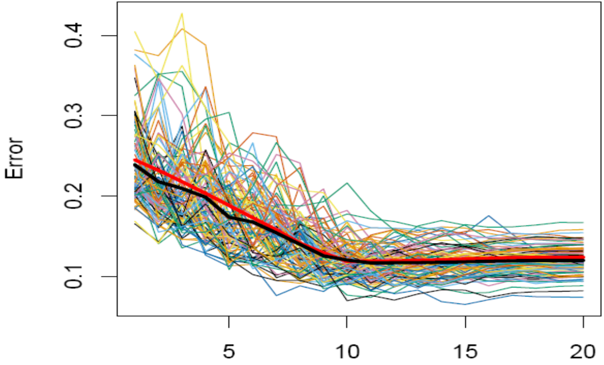
$(Y, \mathbf{X}) \in \{0, 1\} \times R^{20}$: $Y = 1$ jeśli $\sum_{i=1}^{10} X_i > 5$, 0 w przeciwnym przypadku.
CV-10 i CV leave-one-out jak estymatory $Err_{\mathcal{U}}$ i Err liczone dla 100 prób uczących ($n = 80$).
Metoda estymacji: regresja liniowa dla najlepszego podzbioru predyktorów danej liczności ($\mathrm{argmin}_{M:|M|=p} RSS_M$)

Czerwona linia - Err

Czarna linia - ECV_{10} i ECV_n

Pierwszy rysunek - Metoda CV_{10}, drugi Metoda CV_n.


Error
0.1
0.2
0.3
0.4
5
10
15
20


Error
0.1
0.2
0.3
0.4
5
10
15
20

Metoda bootstrap oceny błędu klasyfikacji

Estymator bootstrap 0,632

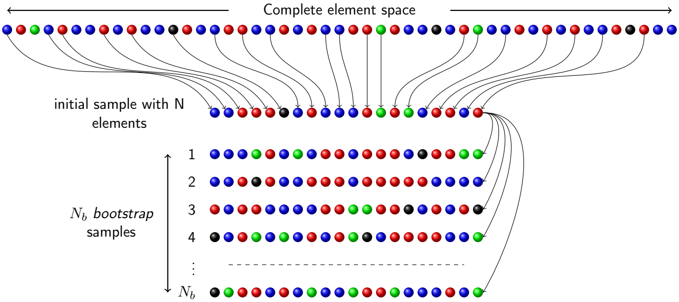
Próba ucząca o liczności n: losujemy m prób bootstrap o liczności n z próby uczącej (tj. prób o liczności n losowanych ze zwracaniem z oryginalnej próby),
Próba bootstrap z reguły nie zawiera wszystkich elementów próby oryginalnej.

$$P(\text{nie wylosowanie ustalonego elementu do pseudopróby}) =$$

$$(\frac{n-1}{n})^n = (1 - \frac{1}{n})^n \approx e^{-1} \approx 0.368$$

Próba bootstrap $\longrightarrow$ *klasyfikator* $\hat{d}$ (na podstawie średnio $0{,}632n$ różnych obserwacji)

Bootstrap



Theorem (B. Efron, Ann. Statist. 1979)

When N tend to infinity, the distribution of average values computed from bootstrap samples is equal to the distribution of average values obtained from ALL samples with N elements which can be constructed from the complete space. Thus the width of the distribution gives an evaluation of the sample quality.

m prób bootstrap daje m klasyfikatorów: $\hat{d}_1, \ldots, \hat{d}_m$
(i) $(x_1, y_1) \longleftarrow$ częstość błędnych zaklasyfikowań przez te spośród klasyfikatorów $\hat{d}_1, \ldots, \hat{d}_m$, które <u>nie wykorzystują</u> (x_1, y_1): $\quad r_1$
(ii) $err_{boot} = (r_1 + r_2 + \ldots + r_n)/n$
estymator pesymistyczny $(E(err_{boot}) > Err)$, gdyż ocena błędu klasyfikacji oparta na mniejszej liczbie różnych obserwacji niż klasyfikator

Modyfikacja tego estymatora – estymator 0,632

$$err_{boot-opt} = (\hat{p}_1 + \hat{p}_2 + \ldots + \hat{p}_m)/m$$

$\hat{p}_i$ – frakcja błędnych sklasyfikowań dla i–tego klasyfikatora bootstrap stosowanego do całej próby
$\hat{p}_i$ – optymistyczny, bo w próbie tylko $\approx 0{,}368$ nowych obserwacji
<u>Estymator 0,632</u>

$$0{,}632 \times err_{boot} + 0{,}368 \times err_{boot-opt}$$

Dotąd: tylko jeden klasyfikator, którego błąd szacowaliśmy. Co gdy mamy l klasyfikatorów $\hat{d}_1, \hat{d}_2, \ldots, \hat{d}_l$? Chcemy wybrać jeden!

Próba z pełną obserwowalnością

$\downarrow$ 50%	$\downarrow$ 25%	$\downarrow$ 25%
<u>próba ucząca</u>	<u>próba walidacyjna</u>	<u>próba testowa</u>
$\downarrow$	$\downarrow$	$\downarrow$
konstrukcja $\hat{d}_1, \ldots, \hat{d}_l$	szacujemy błędy klasyf. wybieramy $\hat{d}_{K^*}$	szacujemy błąd $\hat{d}_{K^*}$

Dla otrzymania <u>nieobciążonej</u> oceny błędu klasyfikacji $\hat{d}_{K^*}$ konieczne jest oszacowanie błędu na niezależnej próbie! To samo tyczy się jednej metody z optymalizowanym parametrem !

Krzywa operacyjno-charakterystyczna ROC

Krzywa operacyjno-charakterystyczna (ROC – receiver operating characteristic curve)
$\hat{d}_t$- reguła klasyfikacyjna dla dwóch populacji zależna od progu t. Na przykład, gdy $\hat{d}_t$ oparta na oszacowaniu $\hat{p}(i|\mathbf{x})$ prawdopodobieństwa aposteriori:

$$\hat{d}_t = 2, \text{ gdy } \log(\hat{p}(2|\mathbf{x})/\hat{p}(1|\mathbf{x})) > t.$$

ROC: wykres punktów postaci $(P(\hat{d}_t = 2|Y = 1), P(\hat{d}_t = 2|Y = 2))$ dla $t \in R$.
$P(\hat{d}_t = 2|Y = 2)$- czułość; $P(\hat{d}_t = 1|Y = 1)$- specyficzność

testu $\hat{d}_t$ dla H_0: obserwacja pochodzi z populacji 1 vs H_1: obserwacja pochodzi z populacji 2. ROC- wykres mocy testu od błędu pierwszego rodzaju (1-specyficzność) dla testów zależnych od t.
Naturalne estymatory czułości i specyficzności: frakcje próbkowe.

Przykład. Zdrowi ($Y = 1$) i chorzy ($Y = 2$) (100 chorych i 200 zdrowych) poddani badaniom diagnostycznym, na podstawie których są sklasyfikowani jako zdrowi ($d = 1$) lub chorzy ($d = 2$).

	d=1	d=2
Zdrowy	TN	FP
Chory	FN	TP

	d=1	d=2
Zdrowy	176	24
Chory	3	97

TN – True Negative

$$\text{Czułość próbkowa} = \hat{P}(d = 2 \mid Chory) = \frac{TP}{TP + FN}$$

$$\text{Specyficzność próbkowa} = \hat{P}(d = 1 \mid Zdrowy) = \frac{TN}{TN + FP}$$

Czyli dla $\quad H_0$: zdrowy, $\quad H_1$: chory
Czułość próbkowa odpowiada oszacowaniu mocy $= 1 - \beta$
Specyficzność próbkowa odpowiada oszacowaniu
$= 1 -$ błąd pierwszego rodzaju

W przykładzie:
Czułość próbkowa $= 97/100 = 0{,}97$, specyficzność próbkowa
$= 176/200 = 0{,}88$.

$$\hat{P}r(\text{błędna decyzja}) = \frac{FP + FN}{TN + FP + FN + TP} = 0{,}09$$

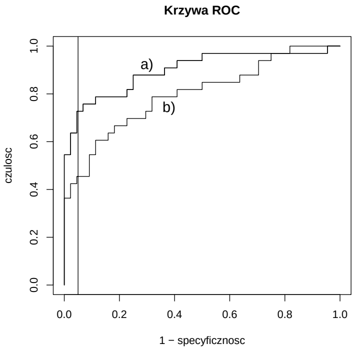
Uwaga. Rozpatrywanie różnych progów związane z różnymi konsekwencjami błędnych decyzji. l_{21} -strata związana decyzją 1, gdy $g = 2$ i l_{12} strata związana decyzją 2, gdy $g = 1$ i chcielibyśmy wiedzieć, jak zmiana strat wpływa na czułość i specyficzność klasyfikatora bayesowskiego

Załóżmy, że l_{12} – stały l_{21} – zmienny

$$d = 2 \iff l_{21} p(2 \mid x) > l_{12} \underbrace{(1 - p(2 \mid x))}_{p(1|x)}, \qquad p(2 \mid x) \geqslant \frac{l_{12}}{l_{12} + l_{21}}$$

zmienny próg przy zmiennym l_{21} (dla każdego progu inna reguła o swojej czułości i specyficzności)

Empiryczne krzywe ROC otrzymujemy przez oszacowanie prawdopodobieństw w definicji krzywych teoretycznych przez odpowiednie frakcje. Na rysunku przedstawiono empiryczne krzywe ROC dla modelu logistycznego dopasowanego do danych `urine` przy zastosowaniu pełnego zbioru atrybutów i dla zbioru `sg`, `mmho` i `urea`. Zaznaczono moc odpowiednich testów na poziomie istotności 0.05.



Krzywa CAP (Cumulative Accuracy Profile)

AUC (Area Under Curve) - pole pod teoretyczną (empiryczną) krzywą ROC.

Krzywa CAP: Zmieniamy współrzędną x-ową w porównaniu z krzywą ROC:

$$(P(\hat{d}_t(X) = 2), P(\hat{d}_t(X) = 2 | Y = 2)), \quad t \in R,$$

$Y = 2$ - 'zły' klient banku, to staramy się maksymalizować procent wykrytych 'złych' klientów przy ustalonym procencie odrzuconych wniosków kredytowych, procent odrzuconych wniosków może się zmieniać.

Nazywana często krzywą LIFT !!!

Krzywa LIFT

Dla metod dających oszacowanie $P(Y = 2|X = x)$.
Niech $\hat{\pi}(x)$ będzie oszacowaniem tego prawdopodobieństwa i q_x kwantylem rzędu x zmiennej losowej $\hat{\pi}(X)$ (dla ustalonej próby uczącej $\mathcal{U}$).

$$LIFT(x) = P(Y = 2|\hat{\pi}(X) \geqslant q_{1-x}).$$

Teoretyczna krzywa LIFT

$$(x, LIFT(x)), \quad x \in R.$$

Punkt (x, y) należy do krzywej LIFT gdy prawdopodobieństwo należenia do klasy 2 wsród procentu x o najwyższej wartości $\hat{\pi}(\cdot)$ wynosi y.
Empiryczna krzywa LIFT. Uciąglona krzywa

$$(\hat{\pi}_{(i)}, \#\{Y_j = 2, \hat{\pi}(x_j) \geqslant \hat{\pi}_{(i)}\}/n) \qquad j = 1, \ldots, n.$$