Jakość procedury klasyfikacyjnej:

poglądowa interpretacja i szacowanie możliwości poprawy na podstawie charakterystyki ROC

Maciej Górkiewicz

mygorkie@cyf-kr.edu.pl

Uniwersytet Jagielloński w Krakowie

Collegium Medicum: Wydział Nauk o Zdrowiu

Instytut Zdrowia Publicznego

Dyrektor: Prof. dr hab. med. Andrzej Pająk

ISCB – polska grupa narodowa, www.iscb.pl Spotkanie naukowe Kraków 04.12.2009

Tematy prezentacji:

1. **Klasyfikacja binarna: Positive vs. Negative**
2. **Założenie o zmiennej ukrytej Z $\in R^1$, Z $\in R^2$**
3. **Błąd pomiarowy: Z $\rightarrow$ cechy X $\rightarrow$ zmienna Y**
4. **Szacowanie możliwości poprawy**
5. **Wnioski i dyskusja**

1. Klasyfikacja binarna: Positive vs. Negative

$TPR = TP / N_{Pos}$; $FPR = FP / N_{Neg}$; ROC: TPR = f(FPR)

		Rozpoznanie Positive „♦”	Rozpoznanie Negative „♣”	
		♣↓ ↓♦ ♣↓ ↓♦	♦↓ ↓♣ ♦↓ ↓♣	
Weryfikacja klasyfikacji	Positive: faktycznie ♦	↓ ↓ ↓ TP ↓ (true Pos.)	....	$\Sigma = N_{Pos}$
	Negative: faktycznie ♣	↓ FP (false Pos.)	....	$\Sigma = N_{Neg}$

1. Klasyfikacja binarna: Positive vs. Negative

Wynik klasyfikacji: informacja o przewidywanej klasie obiektu ...

To jeszcze nie efekt ! To jeszcze nie korzyść !

O efekcie, o korzyści można mówić (warunek *sine qua non*) dopiero wtedy, kiedy:

1: Potrafimy spożytkować tę informację w celu uzyskania określonych korzyści, oraz:

2: Uzyskanie tych korzyści może być trudniejsze / ograniczone bez informacji o przewidywanej klasie obiektu.

Przykładowe korzyści:

Prewencja indywidualna: osoba o klasyfikacji Positive (prognoza? znaczne ryzyko?) podejmuje odpowiednie działania ...

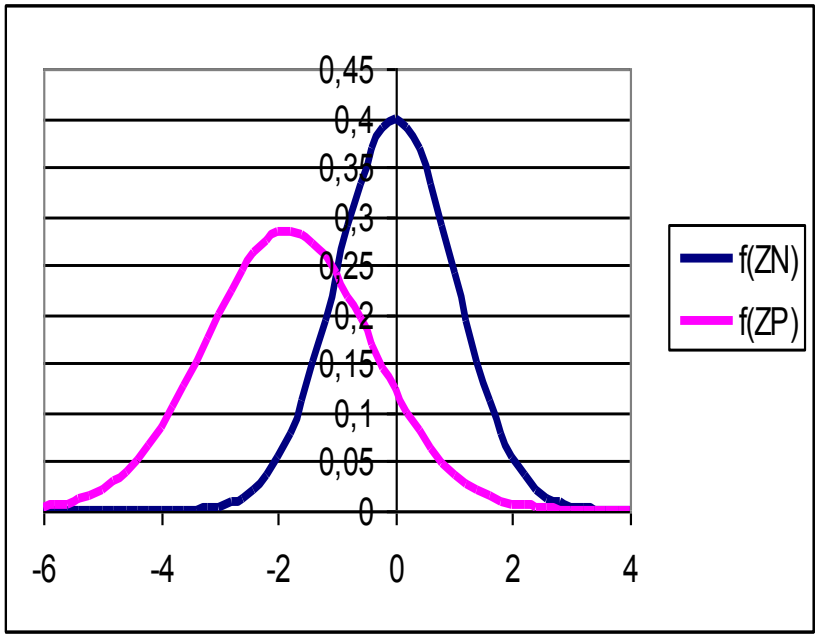
Prewencja społeczna: prognoza liczby osób o klasyfikacji Positive → możność zapobiegania wystąpieniu / skutkom wystąpienia ...

2. Założenie o zmiennej ukrytej Z ...

... Jest taka właściwość Z klasyfikowanych obiektów; bezpośrednio niemierzalna, że istnieje dystrybuanta:

$$\mathbf{Pr(Pos.|\ \zeta \le Z)}$$

W praktyce częściej posługujemy się gęstościami rozkładu zmiennej Z w klasach Pos. i Neg.



Podstawowe pojęcie:

PRÓG DECYZYJNY: $Pr_0(Pos.)$

Reguła klasyfikacji:

Jeśli **$Pr(Pos.|Z) < Pr_0(Pos.) \rightarrow$**

$\rightarrow$ Rozpoznanie.Negative;

przeciwnie:

Jeśli **$Pr(Pos.|Z) \ge Pr_0(Pos.) \rightarrow$**

$\rightarrow$ Rozpoznanie.Positive;

W praktyce częściej próg Z_0; powyżej próg **$Z_0 = 0 \rightarrow$ FPR = 0,5; $Z_0 = -2 \rightarrow$ TPR = 0,5;**

2. Założenie o zmiennej ukrytej Z ...

W wielu zastosowaniach, na przykład w badaniach przesiewowych, wygodnie jest rozgraniczyć to, na co osoby korzystające z wyników klasyfikacji już nie mają wpływu, na przykład stan zdrowia badanej osoby w chwili wykonywania pomiarów, od tego, co jeszcze można modyfikować: wybór terapii, wybór trybu życia.

Z1: Stan obecny i jego możliwe przyczyny	

przeszłość	Stan obecny	przyszłość

	Z2: stan obecny i jego możliwe następstwa

Przy takim podejściu mamy do czynienia ze zmienną ukrytą dwuwymiarową (lub, w innym ujęciu, z dwoma zmiennymi jednowymiarowymi):

$Z \in R^2$; z założenia istnieje mapowanie $[Z_1, Z_2] \rightarrow$ Pr(niepożądane zdarzenie)

Jakość procedury klasyfikacyjnej ... charakterystyki ROC

3. Błąd pomiarowy: Z → cechy X → zmienna Y

			→	$\{X_1\}$	→	{Y1}	→	Pr(Pos|Y1, Y20)
	←	Z1: past				□		
$Pr(Pos\|Z_{10},Z_{20})$			→	$\{X_{12}\}$		{Y12}	→	Pr(Pos|Y1, Y2)
	←	Z2:				↑		
			→	$\{X_2\}$	→	{Y2}	→	Pr(Pos|Y2, Y10)

Legenda:

Kolor niebieski: zmienne nie-obserwowalne: mierzalne cechy {X} klasyfikowanych obiektów

Kolor zielony: zmienne kryterialne {Y} – w przypadku subiektywnych opinii (na przykład klasyfikowania pacjentów przez klinicystów) może być znana tylko ostateczna diagnoza, np. ostateczna klasyfikacja obiektów na Przewidywane.Positive vs. Przewidywane.Negative;

Kolor żółty: zmienne mierzalne, lub dające się estymować na podstawie obserwacji,

Prawdopodobieństwo wystąpienia w badanej populacji obiektu klasy Positive $Pr(Pos|Z_{10},Z_{20})$ – estymowalne na podstawie obserwowanych liczności klas w badanej próbie losowej; prawdopodobieństwa warunkowe Pr(Pos\|Y ...) estymowalne na podstawie próby: na przykład metodą ogólnej regresji logistycznej, Pr(Pos\|Y ...) = f(ΣX); Y* = Σb*X*; lub jako rozkłady empiryczne wartości Y w klasach Positive i Negative.

Jakość procedury klasyfikacyjnej ... charakterystyki ROC

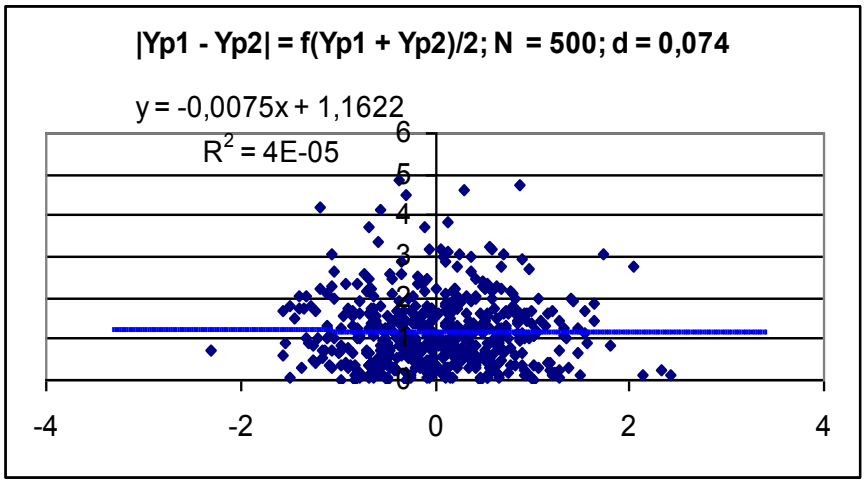
3. Błąd pomiarowy: Z → cechy X → zmienna Y

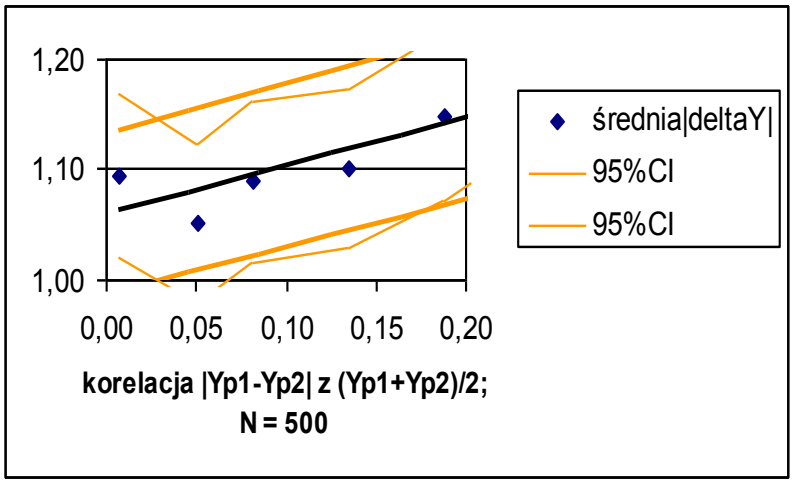
Błąd pomiarowy może być wyrażony w kategoriach rozkładu zmiennej kryterialnej Y* w klasach obiektów Positive i Negativie (faktycznie, a nie według prognozy !):

Mierzone Y*^ = Y* + ε* ; dla rozkładów normalnych wariancja V^ = V_Y + Vε

ε* - błąd pomiarowy zmiennej Y*; gdzie indeks * = 1; 2; 12; pomiar idealny: ε* = 0.

Jeśli ε* ma rozkład normalny N(0; Vε), to SDε może być estymowane na podstawie wyników powtórnych pomiarów; bardzo orientacyjne (Excel) wyniki modelowania:





Dla próby N = 60 szerokość przedziału 95%CI jest 3 razy większa; nadal optymistycznie można szacować **SDε** według dolnej granicy 95%CI dla średniej |Ypomiar1-Ypomiar2|

Jakość procedury klasyfikacyjnej ... charakterystyki ROC

3. Błąd pomiarowy: Z → cechy X → zmienna Y

Błąd pomiarowy może być wyrażony w kategoriach rozkładu zmiennej kryterialnej Y* w klasach obiektów Positive i Negativie (faktycznie, a nie według prognozy !):

Mierzone $Y^{*\wedge} = Y^* + \varepsilon^*$;

ale, mierzone $Y^{*\wedge}$ tylko odzwierciedla zmienną ukrytą Z, czyli: $Y^* = Z^* + \delta^*$; zatem:

Mierzone $Y^{*\wedge} = Z^* + \delta^* + \varepsilon^*$; a z drugiej strony, $Y^* = \Sigma b^* X^*$;

δ^* - błąd wyrażania zmiennej Z* poprzez Y*; gdzie indeks * = 1 (past); 2 (future); 12;

Kiedy porównujemy 2 sposoby pomiaru tak samo zdefiniowanej zmiennej Y, to jest na przykład za pomocą takiego samego równania regresji (te same X'y, te same b'y) to badamy możliwość zmniejszenia składowej ε^* błędu pomiarowego;

Kiedy, trwając przy tak samo pomyślanej (zdefiniowanej nieformalnie) zmiennej Z, porównujemy 2 sposoby wyrażenia jej poprzez różnie zdefiniowane zmienne Y, to badamy możliwość zmniejszenia składowej δ^* błędu pomiarowego; przy różnych ε^*

Kiedy porównujemy dwie różnie pomyślane (zdefiniowane nieformalnie) zmienne Z, na przykład Z1 i Z2, lub Z1 i Z12, wyrażane poprzez różne, odpowiednio zdefiniowane zmienne Y, to badamy skutki zamiany typu składowej δ^* błędu pomiarowego; na przykład *=1 na *=2, lub *=1 na *=12; przy różnych ε^*

Dwie ostatnie sytuacje rozróżniamy raczej według kryteriów heurystycznych ...

Jakość procedury klasyfikacyjnej ... charakterystyki ROC

3. Błąd pomiarowy: Z → cechy X → zmienna Y

Pojęcie <Błąd pomiarowy> jest wygodne dla projektanta procedur pomiarowych; dla projektanta procedury klasyfikacyjnej bardziej dogodne jest pojęcie:

<Jakość zmiennej kryterialnej > , na przykład zmiennej Y, lub zmiennej Z

Jakość ta może być definiowana za pomocą prawdopodobieństwa:

Pr[(Yi | Pos) > [(Yj | Neg); gdzie i-ty obiekt wylosowano z Positive, a j-ty z Negative

Należy zwrócić uwagę, że jak dotąd, pojęcia błędu pomiarowego, jakości zmiennej kryterialnej, były dyskutowane bez nakładania jakichkolwiek warunków na wybór progu decyzyjnego, w szczególności bez zakładania jakiejś jednej określonej wartości progu !

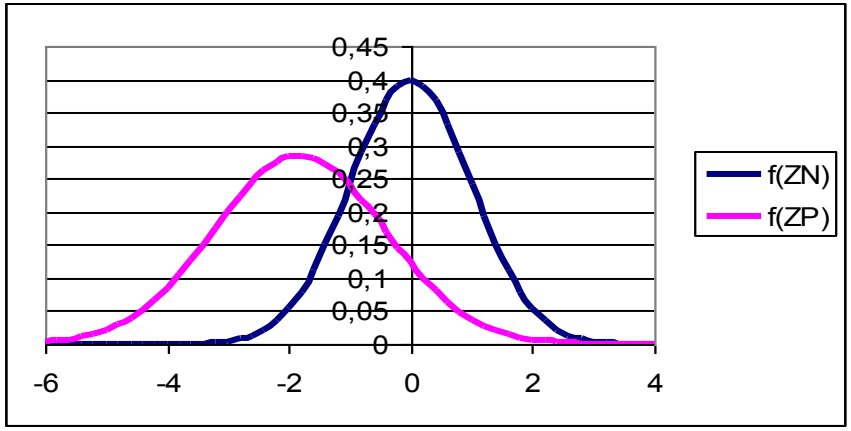
Przy takim podejściu, mając na uwadze późniejsze zdefiniowanie kryteriów wyboru progu decyzyjnego procedury klasyfikacyjnej, wygodnie jest posługiwać się charakterystyką ROC, to jest zależnością między % poprawnie klasyfikowanych obiektów Positive (true positive ratio) TPR a % błędnie klasyfikowanych obiektów klasy Negative (false positive ratio) FPR dla wszystkich możliwych progów Y_0:

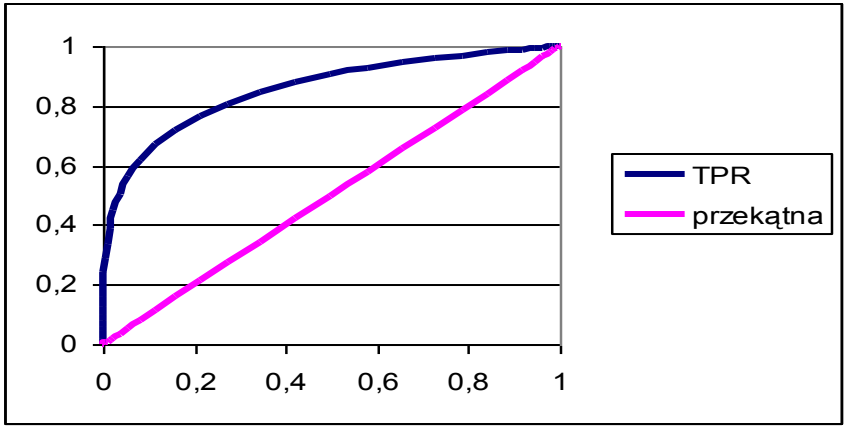
W szczególności, jakość zmiennej Y, definiowana za pomocą prawdopodobieństwa: Pr[(Yi | Pos) > [(Yj | Neg); gdzie i-ty obiekt wylosowano z Positive, a j-ty z Negative, może być estymowana jako pole powierzchni AUC pod krzywą ROC

Jakość procedury klasyfikacyjnej ... charakterystyki ROC

3. Błąd pomiarowy: Z → cechy X → zmienna Y

Konstruowanie charakterystyki ROC jest oczywiste, kiedy znane są rozkłady zmiennej kryterialnej Y w rozpatrywanych klasach; kiedy dostępne są tylko rozkłady empiryczne, zastosowanie mają odpowiednie metody estymacji ...





Podstawowa zaleta charakterystyki ROC polega na tym, że liczne rozsądne kryteria doboru progu decyzyjnego Y_0 zmiennej Y można definiować za pomocą tych samych pojęć TPR i FPR, i przedstawić jako linie ciągłe na płaszczyźnie ROC; ponieważ krzywa ROC jest krzywą parametryczną, z parametrem Y_0 , to punkt przecięcia → optymalne Y_0

... Z drugiej strony, kiedy dla zadanej wartości AUC założymy określoną postać rozkładu zmiennej Y, to możemy badać, jakie wartości parametrów rozkładu odpowiadają temu założeniu : procedury kalkulator1.xls Kalkulator _1: **Niezbędna różnica średnich zmiennej klasyfikacyjnej** kalkulator11.xls **Przykładowe obliczenia wpływu dokładności pomiaru na AUC** przesłane uprzednio ...

Jakość procedury klasyfikacyjnej ... charakterystyki ROC

4. Szacowanie możliwości poprawy ...

... dwa aspekty ulepszania procedury klasyfikacji: korzyści ↑, a koszty ↓
Korzyści ↑ : większy % poprawnych klasyfikacji; w krótszym czasie ...
koszty ↓ : koszty estymowania i korygowania (aktualizowania), potem: co, i jak mierzymy; jak przetwarzamy zgromadzone dane ....

... trzy podstawowe sposoby poprawiania charakterystyki ROC:

1: wykorzystanie kryterium Pareto + wykorzystanie możliwości loterii

2: złożone procedury: ciąg zmiennych Y stosowanych warunkowo w określonej kolejności (drzewa decyzyjne, odsiewanie) ; zazwyczaj prosta postać zależności Y = f(X)

3: złożona zależność Y = f(X); często tylko jedna funkcja kryterialna Y ...

Koszty estymowania: **Kalkulator _2: Niezbędna wielkość próby**

1:Pareto: **Kalkulator_4: Sprawdzenie, czy charakterystyka ROC poprawna (proper ROC)**

2: złożone procedury: **Kalkulator _3: Oszacowanie efektów wstępnego odsiewania obiektów praktycznie na pewno nie-Pozytywne**

W razie potrzeby proszę o uwagi (e-mail: **mygorkie@cyf-kr.edu.pl**)

Jakość procedury klasyfikacyjnej ... charakterystyki ROC

5. Wnioski i dyskusja ...

... Zdaję sobie sprawę, że to tylko wprowadzenie do zastosowań ROC; ale mam nadzieję, że (+ przesłane uprzednio kalkulatory w Excel) nie bez praktycznej przydatności ...

Metoda ROC ma znane, dyskutowane w literaturze ograniczenia:

... że w praktyce często wiele klas: nie ma większego problemy, jeśli klasy uporządkowane hierarchicznie ....

... że w praktyce często mamy do czynienia z sekwencją decyzji, a nie z jednorazowym wyborem np.. Sposoby prowadzenia terapii – przy czym granice między klasami nieostre ...

... że w praktyce często informacja o faktycznych klasach klasyfikowanych obiektów nieosiągalna nawet *post factum*; np.. Rozpoznano.ciąża ... Po 2-3 tygodniach gołym okiem widać: nie ma ciąży; ale co było w chwili klasyfikowania – nigdy się nie dowiemy:

Rozpoznanie OK., ale samoistne poronienie, czy rozpoznanie błędne ...

Analogicznie: wpływ terapii, wpływ profilaktyki: Rozpoznanie.Pos $\rightarrow$ po czasie nie ma Pos...

Jakość procedury klasyfikacyjnej:

poglądowa interpretacja i szacowanie możliwości poprawy na podstawie charakterystyki ROC

Dziękuję za uwagę !

Maciej Górkiewicz

e-mail mygorkie@cyf-kr.edu.pl

Uniwersytet Jagielloński w Krakowie
Collegium Medicum:
Wydział Nauk o Zdrowiu
Instytut Zdrowia Publicznego
ul. Grzegórzecka 20
31-531 Kraków