Zastosowanie krzywych ROC w analizie dyskryminacyjnej

Mirosław Krzyśko, Waldemar Wołyński

Wydział Matematyki i Informatyki UAM Poznań

5 czerwca 2013

Pojęcia wstępne

Załóżmy, że dysponujemy K niezależnymi, prostymi próbami losowymi o liczebnościach, odpowiednio, $n_1, n_2, \ldots, n_K$, pobranymi z K różnych populacji (klas, grup):

$$\begin{aligned}
&\mathbf{X}_{11}, \mathbf{X}_{12}, \ldots, \mathbf{X}_{1n_1} && - && \text{z populacji 1} \\
&\mathbf{X}_{21}, \mathbf{X}_{22}, \ldots, \mathbf{X}_{2n_2} && - && \text{z populacji 2} \\
&\qquad\qquad \ldots \\
&\mathbf{X}_{K1}, \mathbf{X}_{K2}, \ldots, \mathbf{X}_{Kn_K} && - && \text{z populacji } K,
\end{aligned}$$

gdzie $\mathbf{X}_{ij} = (X_{ij1}, X_{ij2}, \ldots, X_{ijp})'$ jest j-tą obserwacją z i-tej populacji zawierającą p obserwowanych cech, $i = 1, 2, \ldots, K$, $j = 1, 2, \ldots, n_i$. Powyższe dane można zapisać w postaci ciągu

$$\mathbf{Z}_1, \mathbf{Z}_2, \ldots, \mathbf{Z}_n,$$

gdzie $\mathbf{Z}_i = (\mathbf{X}_i', Y_i)'$ jest uporządkowaną parą, w której $\mathbf{X}_i = (X_{i1}, X_{i2}, \ldots, X_{ip})' \in \mathcal{X} \subset \mathbb{R}^p$ oznacza i-tą obserwację, natomiast Y_i jest etykietą populacji, do której ta obserwacja należy, przyjmującą wartości w pewnym skończonym zbiorze $\mathcal{Y}$, $i = 1, 2, \ldots, n$, $n = n_1 + n_2 + \cdots + n_K$.

Pojęcia wstępne

Składowe wektora $\mathbf{X}_i = (X_{i1}, X_{i2}, \ldots, X_{ip})'$ nazywać będziemy cechami lub zmiennymi (nazewnictwo przyjęte w statystyce matematycznej). W eksploracji danych używana jest zamiennie nazwa atrybuty.

Próbę $\mathbf{Z}_1, \mathbf{Z}_2, \ldots, \mathbf{Z}_n$ nazywać będziemy próbą uczącą.

Interesuje nas problem predykcji etykiety Y na podstawie wektora cech $\mathbf{X}$. Problem ten nazywany jest klasyfikacją, dyskryminacją, uczeniem się pod nadzorem lub rozpoznawaniem postaci.

Reguła klasyfikacyjna, zwana krótko klasyfikatorem, jest funkcją

$$d : \mathcal{X} \to \mathcal{Y}.$$

Gdy obserwujemy nowy wektor $\mathbf{X}$, to prognozą etykiety Y jest $d(\mathbf{X})$.

Poziom błędu i klasyfikator bayesowski

Naszym celem jest znalezienie takiego klasyfikatora d, który daje dokładną predykcję. Rozpocznijmy od następującej definicji:

Definicja

Rzeczywisty poziom błędu klasyfikatora d jest równy

$$L(d) = P(\{d(\mathbf{X}) \neq Y\}). \tag{1}$$

$L(d)$ jest prawdopodobieństwem zdarzenia, że klasyfikator d błędnie zaklasyfikuje nową obserwację $(\mathbf{X}', Y)'$, niezależną od próby uczącej, pod warunkiem, że próba ucząca jest ustalona.

Poziom błędu i klasyfikator bayesowski

Weźmy pod uwagę przypadek dwóch klas, tj. gdy $\mathcal{Y} = \{1, 0\}$.
Niech

$$\begin{aligned} r(\mathbf{x}) &= \mathrm{E}(Y|\mathbf{X} = \mathbf{x}) = 1 \cdot \mathrm{P}(Y = 1|\mathbf{X} = \mathbf{x}) + 0 \cdot \mathrm{P}(Y = 0|\mathbf{X} = \mathbf{x}) \\ &= \mathrm{P}(Y = 1|\mathbf{X} = \mathbf{x}) \end{aligned}$$

oznacza funkcję regresji zmiennej Y względem wektora $\mathbf{X}$.
Z twierdzenia Bayes'a mamy

$$\begin{aligned} r(\mathbf{x}) &= \mathrm{P}(Y = 1|\mathbf{X} = \mathbf{x}) \\ &= \frac{\pi_1 f_1(\mathbf{x})}{\pi_1 f_1(\mathbf{x}) + \pi_0 f_0(\mathbf{x})}, \end{aligned}$$

gdzie

$$f_1(\mathbf{x}) = f(\mathbf{x}|Y = 1), \quad f_0(\mathbf{x}) = f(\mathbf{x}|Y = 0),$$

$$\pi_1 = \mathrm{P}(Y = 1), \quad \pi_0 = \mathrm{P}(Y = 0), \quad \pi_1 + \pi_0 = 1.$$

Definicja

Klasyfikator postaci

$$d_B(\mathbf{x}) = \begin{cases} 1, & jeżeli\ r(\mathbf{x}) > \frac{1}{2}, \\ 0, & poza\ tym. \end{cases}$$

nazywać będziemy klasyfikatorem bayesowskim.

Klasyfikator bayesowski zapisać można w innych równoważnych postaciach:

$$d_B(\mathbf{x}) = \begin{cases} 1, & \text{jeżeli } \mathrm{P}(Y = 1|\mathbf{X} = \mathbf{x}) > \mathrm{P}(Y = 0|\mathbf{X} = \mathbf{x}), \\ 0, & \text{poza tym.} \end{cases}$$

lub

$$d_B(\mathbf{x}) = \begin{cases} 1, & \text{jeżeli } \pi_1 f_1(\mathbf{x}) > \pi_0 f_0(\mathbf{x}), \\ 0, & \text{poza tym.} \end{cases}$$

Poziom błędu i klasyfikator bayesowski

Definicja

Zbiór postaci

$$\{\mathbf{x}: \ \mathrm{P}(Y=1|\mathbf{X}=\mathbf{x}) = \mathrm{P}(Y=0|\mathbf{X}=\mathbf{x})\}$$

nazywać będziemy powierzchnią rozdzielającą grupy 1 i 0.

Twierdzenie

Klasyfikator bayesowski jest optymalny, tj. jeżeli d jest jakimkolwiek innym klasyfikatorem, to $L(d_B) \leq L(d)$, gdzie $L(d)$ jest rzeczywistym poziomem błędu klasyfikatora d danym wzorem (1).

Poziom błędu i klasyfikator bayesowski

Niestety, klasyfikator bayesowski zależy od wielkości nam nieznanych i stąd w praktyce jesteśmy zmuszeni posługiwać się jego oceną skonstruowaną z próby uczącej. Znane są trzy główne podejścia.

1. Wybieramy pewną klasę klasyfikatorów $\mathcal{D}$ i znajdujemy $\hat{d} \in \mathcal{D}$, który minimalizuje ocenę $L(d)$.
2. Znajdujemy ocenę $\hat{r}$ funkcji regresji r i definiujemy

$$\hat{d}(\mathbf{x}) = \begin{cases} 1, & \text{jeżeli } \hat{r}(\mathbf{x}) > \frac{1}{2}, \\ 0, & \text{poza tym.} \end{cases}$$

3. Estymujemy f_1 z tych $\mathbf{X}_i$ dla których $Y_i = 1$, estymujemy f_0 z tych $\mathbf{X}_i$ dla których $Y_i = 0$. Niech $\hat{\pi}_1 = \frac{1}{n}\sum_{i=1}^{n} Y_i$. Definiujemy

$$\hat{r}(\mathbf{x}) = \hat{\mathrm{P}}(Y = 1|\mathbf{X} = \mathbf{x}) = \frac{\hat{\pi}_1 \hat{f}_1(\mathbf{x})}{\hat{\pi}_1 \hat{f}_1(\mathbf{x}) + (1 - \hat{\pi}_1)\hat{f}_0(\mathbf{x})}$$

i

$$\hat{d}(\mathbf{x}) = \begin{cases} 1, & \text{jeżeli } \hat{r}(\mathbf{x}) > \frac{1}{2}, \\ 0, & \text{poza tym.} \end{cases}$$

Krzywe ROC

Niech etykieta $Y \in \mathcal{Y} = \{1, 0\}$ i niech X będzie obserwowaną zmienną absolutnie ciągłą.
Reguła klasyfikacyjna ma postać

$$d(x) = \begin{cases} 1, & \text{jeżeli } x > c, \\ 0, & \text{jeżeli } x \leq c, \end{cases}$$

gdzie c jest wartością progową.
Dla wybranego progu c $(-\infty \leq c \leq +\infty)$ niech

$$\mathrm{P}(d(X) = 1 | Y = 1) = \mathrm{P}(X > c | Y = 1) = \bar{G}(c) = 1 - G(c),$$

gdzie $\bar{G}$ jest funkcją przeżycia, a G dystrybuantą zmiennej X w grupie o etykiecie 1 oraz niech

$$\mathrm{P}(d(X) = 1 | Y = 0) = \mathrm{P}(X > c | Y = 0) = \bar{F}(c) = 1 - F(c),$$

gdzie $\bar{F}$ jest funkcją przeżycia, a F dystrybuantą zmiennej X w grupie o etykiecie 0.

Krzywe ROC

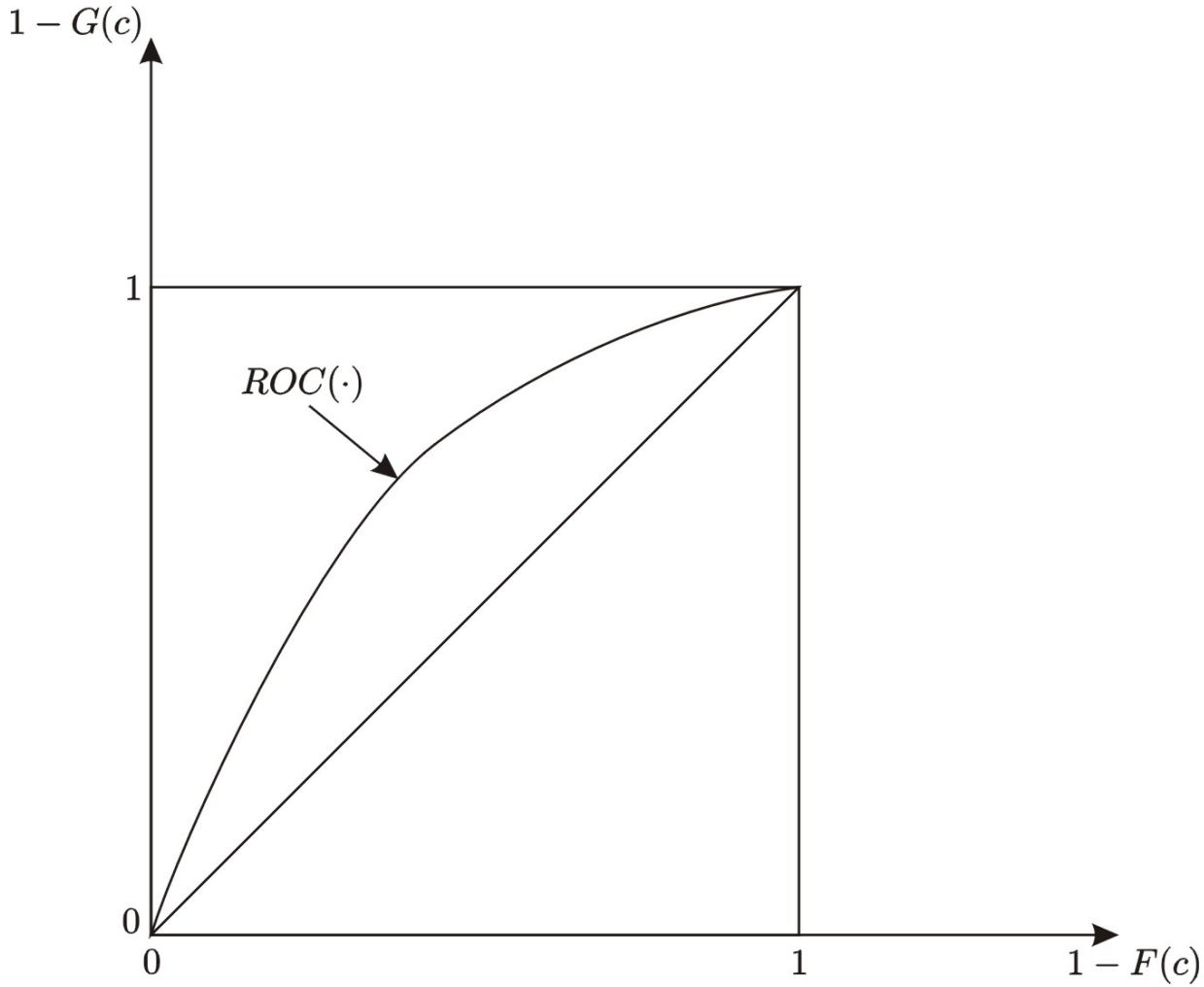
Definicja

Krzywa ROC ma postać

$$\mathrm{ROC}(\cdot) = \{(1 - F(c), 1 - G(c)) : \ -\infty \leq c \leq \infty\}.$$

W terminologii statystyki matematycznej krzywa ROC jest wykresem zależności mocy od rozmiaru testu statystycznego z obszarem odrzucenia $\{x : \ x > c\}$, przy zmieniającym się progu c.
Ponieważ $F(+\infty) = G(+\infty) = 1$ oraz $F(-\infty) = G(-\infty) = 0$, to krzywa $\mathrm{ROC}(\cdot)$ łączy wierzchołki $(0, 0)$ i $(1, 1)$ jednostkowego kwadratu.

Krzywe ROC



Przykład krzywej ROC.

Krzywe ROC

Krzywe ROC znajdują liczne zastosowania w diagnostyce technicznej i medycznej. Np. w medycynie grupa o etykiecie $Y = 0$ może oznaczać grupę ludzi zdrowych (grupa kontrolna), a grupa o etykiecie $Y = 1$ grupę ludzi chorujących na określoną chorobę. Zmienną X możemy interpretować jako wartość (wynik) pewnego testu diagnostycznego, a c jako wartość progową w tym teście.
W diagnostyce medycznej prawdopodobieństwo

$$\mathrm{P}(d(X) = 1|Y = 1) = \mathrm{P}(X > c|Y = 1)$$

nosi nazwę czułości reguły diagnostycznej i jest oznaczane przez $\mathrm{SE}(c)$. Jego oceną z próby uczącej jest frakcja osób chorych, u których prawidłowo rozpoznano chorobę. Frakcja ta jest oznaczana przez $\mathrm{TPF}(c)$. Wyrażenie

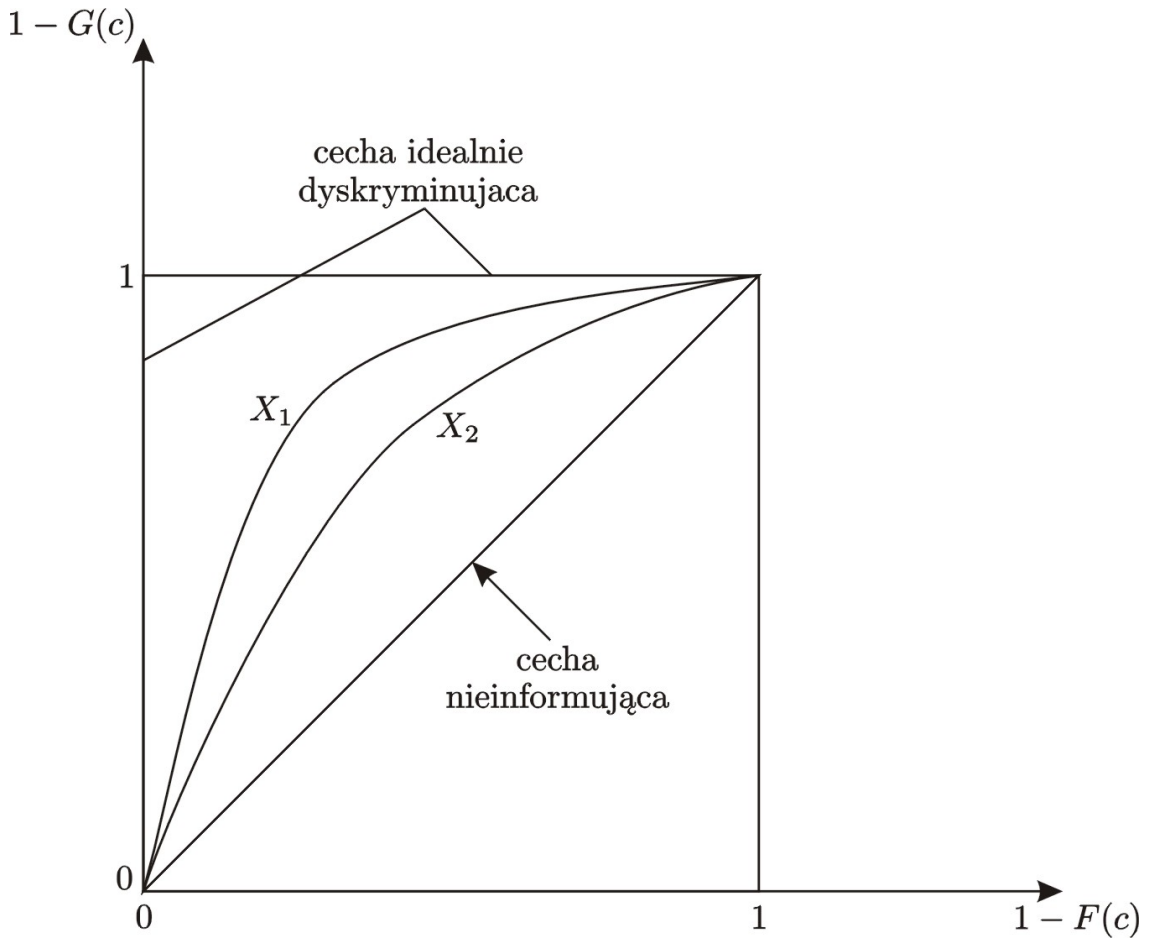
$$\mathrm{P}(d(X) = 1|Y = 0) = 1 - \mathrm{P}(X > c|Y = 0)$$

nazywa się swoistością reguły diagnostycznej i jest oznaczane przez $\mathrm{SP}(c)$.

Krzywe ROC

Oceną prawdopodobieństwa $\mathrm{P}(d(X) = 1|Y = 0) = 1 - \mathrm{P}(X > c|Y = 0)$ z próby uczącej jest frakcja osób zdrowych, u których błędnie stwierdzono występowanie choroby. Frakcja ta jest oznaczana przez $\mathrm{FPF}(c)$.
Wróćmy teraz do ogólnych własności krzywych ROC. Jeżeli cecha X jest nieinformująca, tj. rozkłady prawdopodobieństwa cechy X w dwóch grupach o etykietach 0 i 1 są identyczne ($F(c) = G(c)$), to krzywa ROC ma postać $\mathrm{ROC}(t) = t$, dla $t \in [0, 1]$. Jest to przekątna kwadratu jednostkowego łącząca punkty $(0, 0)$ i $(1, 1)$.
Z drugiej strony dla zmiennej idealnie oddzielającej grupy oraz wybranego progu c mamy $\mathrm{P}(X > c|Y = 1) = 1$ oraz $\mathrm{P}(X > c|Y = 0) = 0$ lub $1 - G(c) = 1$ oraz $1 - F(c) = 0$, tj. krzywa ROC pokrywa się z bokami kwadratu łączącymi punkty $(0, 0)$ i $(0, 1)$ oraz $(0, 1)$ i $(1, 1)$. Dla większości cech krzywa ROC leży między krzywymi odpowiadającymi cechom nieinformującym i idealnie rozdzielającym grupy. Cechy silnie dyskryminujące grupy mają krzywe ROC leżące blisko wierzchołka $(0, 1)$ kwadratu.

Krzywe ROC



Krzywe ROC dla dwóch cech, gdzie cecha X_1 jest cechą jednoznacznie silniej dyskryminującą od cechy X_2. Dla porównania pokazano również krzywe ROC dla cech: nieinformującej oraz idealnie dyskryminującej.

Krzywe ROC

Przytoczymy pewne własności charakteryzujące krzywe ROC.

Własność 1. *Krzywe ROC są niezmiennicze względem ściśle rosnących przekształceń cechy X.*

Własność 2. *Krzywa ROC ma następującą reprezentację:*

$$\mathrm{ROC}(t) = \bar{G}(\bar{F}^{-1}(t)),\ \ t \in [0,1]$$

lub

$$\mathrm{ROC}(t) = 1 - G(F^{-1}(1-t)),\ \ t \in [0,1].$$

Im rozkłady zmiennych X_1 i X_0 będą bardziej zróżnicowane, tym klasyfikator będzie lepszy. Miarą zróżnicowania rozkładów w dwóch grupach jest pole pod krzywą ROC.

Własność 3. *Niech* $\mathrm{AUC} = \int_0^1 \mathrm{ROC}(t)dt$ *będzie polem pod krzywą ROC. Wówczas*

$$\mathrm{AUC} = \mathrm{P}(X_1 > X_0),$$

gdzie X_1 *i* X_0 *są zmiennymi losowymi niezależnymi oraz*

$$X_1 = (X|Y = 1) \sim G, \quad X_0 = (X|Y = 0) \sim F.$$

Własność 4. (Estymacja krzywej ROC)
Niech $X_{i1}, X_{i2}, \ldots, X_{iN_i}$ będzie próbą uczącą z grupy o etykiecie i, gdzie $i = 0, 1$. Zakładamy, że próby te są niezależne. Estymatorem z próby krzywej ROC jest

$$\widehat{\mathrm{ROC}}(t) = 1 - \hat{G}_{N_1}(\hat{F}_{N_0}^{-1}(1-t)), \ \ t \in [0,1],$$

gdzie $\hat{G}_{N_1}$ i $\hat{F}_{N_0}$ są dystrybuantami z próby postaci:

$$\hat{G}_{N_1}(t) = \frac{1}{N_1}\sum_{j=1}^{N_1} \mathrm{I}_{(-\infty,t]}(X_{1j}), \ \ \hat{F}_{N_0}(t) = \frac{1}{N_0}\sum_{j=1}^{N_0} \mathrm{I}_{(-\infty,t]}(X_{0j}),$$

gdzie

$$\mathrm{I}_{(-\infty,t]}(X_{ij}) = \begin{cases} 1, & \text{gdy } X_{ij} \leq t, \\ 0, & \text{gdy } X_{ij} > t, \end{cases}$$

dla $j = 1, 2, \ldots, N_i$, $i = 0, 1$.

Krzywe ROC

Estymator $\widehat{\mathrm{ROC}}(t)$ jest niemalejącą funkcją schodkową. Z Własności 3 wiemy, że jeżeli X_1 i X_0 są niezależnymi zmiennymi losowymi takimi, że $X_1 = (X|Y=1) \sim G$ i $X_0 = (X|Y=0) \sim F$, to $\mathrm{AUC} = \mathrm{P}(X_1 > X_0)$. Stąd

$$\widehat{\mathrm{AUC}} = \frac{1}{N_0 N_1} \sum_{i=1}^{N_1} \sum_{j=1}^{N_0} \mathrm{I}(X_{1i} > X_{0j}).$$

Wyrażenie to jest znaną statystyką U Manna-Whitneya.

Własność 5. *Jeżeli $X_1 \sim N(\mu_1, \sigma_1^2)$ i $X_0 \sim N(\mu_0, \sigma_0^2)$, to*

$$\mathrm{ROC}(t) = \Phi(a + b\Phi^{-1}(t)), \tag{2}$$

gdzie

$$a = \frac{\mu_1 - \mu_0}{\sigma_1}, \quad b = \frac{\sigma_0}{\sigma_1} \tag{3}$$

oraz Φ oznacza dystrybuantę rozkładu $N(0,1)$.

Własność 6. *Pole pod krzywą ROC postaci (2) jest równe*

$$\mathrm{AUC} = \Phi\left(\frac{a}{\sqrt{1+b^2}}\right), \tag{4}$$

gdzie a i b dane są wzorem (3).

Uogólnimy teraz własności 5 i 6 na przypadek wielowymiarowy. Zamiast obserwowanej zmiennej losowej X weźmiemy pod uwagę wektor losowy $\mathbf{X} = (X_1, X_2, \ldots, X_p)'$ i założymy, że wektor ten w grupie o etykiecie 1 ma p-wymiarowy rozkład normalny $N_p(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1)$ oraz w grupie o etykiecie 0 ma p-wymiarowy rozkład normalny $N_p(\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0)$, gdzie $\boldsymbol{\Sigma}_1$ i $\boldsymbol{\Sigma}_0$ są macierzami kowariancji określonymi dodatnio ($\boldsymbol{\Sigma}_1 > 0,\ \boldsymbol{\Sigma}_0 > 0$).

Krzywe ROC

Weźmiemy pod uwagę regułę klasyfikacyjną opartą na kombinacji liniowej

$$U(\mathbf{a}) = \mathbf{a}'\mathbf{X} = a_1X_1 + a_2X_2 + \cdots + a_pX_p$$

postaci

$$d(\mathbf{x}) = \begin{cases} 1, & \text{jeżeli } u(\mathbf{a}) = \mathbf{a}'\mathbf{x} > c, \\ 0, & \text{jeżeli } u(\mathbf{a}) = \mathbf{a}'\mathbf{x} \leq c, \end{cases} \tag{5}$$

gdzie $-\infty \leq c \leq +\infty$ jest wartością progową.
Niech U_1 i U_0 będą niezależnymi zmiennymi losowymi reprezentującymi kombinację liniową $U(\mathbf{a}) = \mathbf{a}'\mathbf{X}$ w grupie o etykietach 1 i 0.
Wówczas

$$U_1(\mathbf{a}) \sim N(\mathbf{a}'\boldsymbol{\mu}_1, \mathbf{a}'\boldsymbol{\Sigma}_1\mathbf{a}) \text{ i } U_0(\mathbf{a}) \sim N(\mathbf{a}'\boldsymbol{\mu}_0, \mathbf{a}'\boldsymbol{\Sigma}_0\mathbf{a}).$$

Wówczas zgodnie z Własnościami 5 i 6, krzywa ROC dla zmiennych U_1 i U_0 ma postać (2), a pole pod nią jest równe (4), gdzie

$$a = \frac{\mathbf{a}'\boldsymbol{\mu}_1 - \mathbf{a}'\boldsymbol{\mu}_0}{\mathbf{a}'\boldsymbol{\Sigma}_1\mathbf{a}}, \quad b = \frac{\mathbf{a}'\boldsymbol{\Sigma}_0\mathbf{a}}{\mathbf{a}'\boldsymbol{\Sigma}_1\mathbf{a}}.$$

Własność 7. (Krzyśko, 1999)
Wektor $\mathbf{a}$, *dla którego pole* $\mathrm{AUC}(\mathbf{a})$ *pod krzywą ROC jest maksymalne ma postać:*

$$\mathbf{a} = (\mathbf{\Sigma}_1 + \mathbf{\Sigma}_0)^{-1}\boldsymbol{\delta}, \tag{6}$$

gdzie

$$\boldsymbol{\delta} = \boldsymbol{\mu}_1 - \boldsymbol{\mu}_0.$$

Zajmiemy się teraz dopuszczalnością naszej reguły klasyfikacyjnej. Prawdopodobieństwa błędnych zaklasyfikowań oparte na regule (5) są równe:

$$\mathrm{P}(d(\mathbf{X}) = 1|Y = 0) = \mathrm{P}(U > c|Y = 0) = 1 - F(c) = 1 - \Phi\left(\frac{c - \mathbf{a}'\boldsymbol{\mu}_0}{(\mathbf{a}'\mathbf{\Sigma}_0\mathbf{a})^{\frac{1}{2}}}\right)$$

oraz

$$\mathrm{P}(d(\mathbf{X}) = 0|Y = 1) = \mathrm{P}(U \leq c|Y = 1) = G(c) = \Phi\left(\frac{c - \mathbf{a}'\boldsymbol{\mu}_1}{(\mathbf{a}'\mathbf{\Sigma}_1\mathbf{a})^{\frac{1}{2}}}\right).$$

Krzywe ROC

Przypomnijmy, że reguła klasyfikacyjna z progiem c_1 jest lepsza od reguły z progiem c_2, jeżeli prawdopodobieństwa błędnych zaklasyfikowań spełniają nierówności:

$$1 - F(c_1) \leq 1 - F(c_2) \text{ i } G(c_1) \leq G(c_2)$$

oraz co najmniej jedna z nich jest ostra.
Drugą nierówność możemy równoważnie zapisać w postaci

$$1 - G(c_1) \geq 1 - G(c_2).$$

Zauważmy, że punkty $(1 - F(c), 1 - G(c))$ przy zmieniającym się c definiują krzywą ROC. Przypomnijmy, że reguła klasyfikacyjna z progiem c jest dopuszczalna, jeżeli nie istnieje reguła od niej lepsza.

Własność 8. *Reguła klasyfikacyjna (5), gdzie wektor* $\mathbf{a}$ *dany jest wzorem (6) oraz*

$$c = \mathbf{a}'\boldsymbol{\mu}_0 + \mathbf{a}'\boldsymbol{\Sigma}_0\mathbf{a} = \mathbf{a}'\boldsymbol{\mu}_1 - \mathbf{a}'\boldsymbol{\Sigma}_1\mathbf{a}$$

jest regułą dopuszczalną w klasie reguł liniowych.

Załóżmy, teraz że $Y \in \{1, 2, \ldots, K\}$, gdzie $K \geq 3$.
Niech

$$p_i(\mathbf{x}) = P(Y = i|\mathbf{X} = \mathbf{x})$$

będzie prawdopodobieństwem a posteriori przynależności testowego punktu $\mathbf{x}$ do i-tej grupy i niech

$$\mathbf{p}(\mathbf{x}) = (p_1(\mathbf{x}), p_2(\mathbf{x}), \ldots, p_K(\mathbf{x}))',$$

gdzie $p_i(\mathbf{x}) > 0$, $\sum_{i=1}^{K} p_i(\mathbf{x}) = 1$, $i = 1, 2, \ldots, K$.

Ponieważ wszystkie obliczenia wykonywane są dla każdego $\mathbf{x}$ oddzielnie, to w dalszym ciągu będziemy $\mathbf{x}$ pomijać pisząc p_i zamiast $p_i(\mathbf{x})$.

Weźmy pod uwagę tylko parę i-tą oraz j-tą, $i, j = 1, 2, \ldots, K$, $j \neq i$. Par takich możemy utworzyć

$$\binom{K}{2} = \frac{1}{2}K(K-1).$$

Niech

$$\mu_{ij} = P(Y = i | Y = i \text{ lub } Y = j) = \frac{p_i}{p_i + p_j},$$

Ponadto, mamy

$$\mu_{ji} = 1 - \mu_{ij}, \quad i, j = 1, 2, \ldots, K, \quad j \neq i.$$

Niech r_{ij} będzie nieobciążoną oceną warunkowego prawdopodobieństwa μ_{ij}.
Nasz model ma postać:

$$\mu_{ij} = E(r_{ij}) = \frac{p_i}{p_i + p_j}, \quad i, j = 1, 2, \ldots, K, \quad j \neq i.$$

Ponieważ $\sum_{i=1}^{K} p_i = 1$, to ocenie podlega $K - 1$ niezależnych parametrów, natomiast liczba równań w naszym układzie jest równa $\frac{1}{2}K(K-1)$. Stad nie jest możliwe znalezienie ocen $\hat{p}_i$ w taki sposób by

$$\hat{\mu}_{ij} = \frac{\hat{p}_i}{\hat{p}_i + \hat{p}_j} = r_{ij},$$

dla wszystkich i, j (układ $K(K-1)/2$ równań z $K-1$ niewidomymi jest zazwyczaj sprzeczny).
Zatem możemy tylko żądać, by $\hat{\mu}_{ij}$ były bliskie wartościom r_{ij}.

Hastie i Tibshirani (1998) jako kryterium bliskości wybrali ważoną odległość Kullbacka-Leiblera między r_{ij} oraz μ_{ij}:

$$l(\mathbf{p}) = \sum_{i<j} n_{ij} \left[r_{ij} \ln \frac{r_{ij}}{\mu_{ij}} + (1 - r_{ij}) \ln \frac{1 - r_{ij}}{1 - \mu_{ij}} \right],$$

gdzie $\mathbf{p} = (p_1, p_2, \ldots, p_K)'$, $p_i > 0$, $\sum_{i=1}^{K} p_i = 1$, $i = 1, 2, \ldots, K$, n_{ij} jest liczbą elementów próby uczącej należących do połączonych klas i-tej oraz j-tej.
Włączenie do odległości Kullbacka-Leiblera liczebności n_{ij} jako wag jest związane z tym, że dla poszczególnych par grup oceny r_{ij} sa uzyskiwane z różną precyzją.
Chcemy znaleźć wektor $\hat{\mathbf{p}}$ minimalizujący funkcję $l(\mathbf{p})$, pod warunkiem, że $p_i > 0$, $i = 1, \ldots, K$, $\sum_{i=1}^{K} p_i = 1$.

Z warunku koniecznego istnienia ekstremum funkcji $l(\mathbf{p})$ otrzymujemy następujący układ równań:

$$\sum_{j \neq i} n_{ij} \mu_{ij} = \sum_{j \neq i} n_{ij} r_{ij}, \quad i = 1, 2, \ldots, K, \tag{7}$$

gdzie $p_i > 0$, $\sum_{i=1}^{K} p_i = 1$.
Ponieważ nie można uzyskać jawnego rozwiązania układu (7), to Hastie i Tibshirani zastosowali następujący algorytm iteracyjny i pokazali jego zbieżność.

ALGORYTM

1. Startujemy z pewnymi wartościami początkowymi $\tilde{p}_1, \tilde{p}_2, \ldots \tilde{p}_K$, z nich wyliczamy

$$\tilde{\mu}_{ij} = \frac{\tilde{p}_i}{\tilde{p}_i + \tilde{p}_j}, \quad i, j = 1, \ldots, K, \quad j \neq i.$$

2. Wartości $\tilde{p}_i$ mnożymy przez iloraz

$$(\sum_{j \neq i} n_{ij} r_{ij}) / (\sum_{j \neq i} n_{ij} \tilde{\mu}_{ij})$$

a następnie uzyskane iloczyny normalizujemy dzieląc każdy z iloczynów przez ich sumę (normalizacja jest konieczna, ponieważ $p_1 + p_2 + \cdots + p_K = 1$). Uzyskujemy w ten sposób przybliżone oceny prawdopodobieństw $p_1, p_2, \ldots, p_K$.

3. Powtarzamy czynności z punktu 1 i 2 aż do uzyskania zbieżności procedury.

W wyniku tego postępowania uzyskujemy następujący klasyfikator

$$\hat{d}(\mathbf{x}) = \arg \max_{1 \leq i \leq K} \hat{p}_i(\mathbf{x}).$$

Weźmy pod uwagę tożsamość postaci

$$p_i = \sum_{j \neq i} \left(\frac{p_i + p_j}{K-1} \right) \left(\frac{p_i}{p_i + p_j} \right), \quad i = 1, 2, \ldots, K.$$

Oceniając $\frac{p_i}{p_i+p_j}$ przez r_{ij} oraz oceniając $p_i + p_j$ przez $\frac{2}{K}$ otrzymujemy prostą nieiteracyjną ocenę prawdopodobieństwa p_i:

$$\tilde{p}_i = \frac{2}{K(K-1)} \sum_{j \neq i} r_{ij}, \quad i = 1, 2, \ldots, K.$$

Oceny te możemy przyjąć jako punkty startowe w naszym algorytmie.

Wagi n_{ij} w odległości Kullbacka-Leiblera mogą trochę poprawić efektywność ocen, ale nie dają dużego efektu nawet wtedy, gdy liczebności klas są bardzo różne. Stąd w dalszym ciągu założymy równość wag.

Twierdzenie (Hastie i Tibshirani)

$$\tilde{p}_i > \tilde{p}_j \quad \textit{wtedy i tylko wtedy, gdy} \quad \hat{p}_i > \hat{p}_j.$$

Ponieważ $\tilde{p}_i$ zachowują ten sam porządek co $\hat{p}_i$, to nasz klasyfikator może być oparty na ocenach $\tilde{p}_i$.
Pomijając stały czynnik $2/K(K-1)$ w $\tilde{p}_i$ otrzymujemy

$$\hat{d}(\mathbf{x}) = \arg \max_{1 \le i \le K} \sum_{j \neq i} r_{ij}(\mathbf{x}).$$

Inną nieiteracyjną ocenę $\tilde{p}_i$ podali Price i inni (1995).
Prawdziwa jest tożsamość:

$$\sum_{j \neq i}(p_i + p_j) - (K-2)p_i = 1, \quad i = 1, 2, \ldots, K.$$

Ale $\frac{p_i}{p_i+p_j} = \mu_{ij}$ lub $p_i + p_j = \frac{p_i}{\mu_{ij}}$. Stąd

$$\sum_{j \neq i} \frac{p_i}{\mu_{ij}} - (K-2)p_i = 1 \quad \text{lub} \quad p_i = \left(\sum_{j \neq i} \frac{1}{\mu_{ij}} - (K-2) \right)^{-1}.$$

Oceniając μ_{ij} przez r_{ij} otrzymujemy:

$$\tilde{p}_i = \left(\sum_{j \neq i} \frac{1}{r_{ij}} - (K-2) \right)^{-1}.$$

Klasyfikatory wieloklasowe

Omówione procedury mają łatwą do zauważenia wadę. Klasyfikując obserwację $\mathbf{x}$ musimy, za pomocą klasyfikatorów binarnych, oszacować $K(K-1)/2$ prawdopodobieństw μ_{ij}. Zauważmy jednak, że jeżeli klasyfikowana obserwacja w rzeczywistości pochodzi z i-tej klasy, to w oszacowaniu tych prawdopodobieństw obserwacje uczące z i-tej klasy biorą udział tylko $K-1$ razy. Do pozostałych $(K-1)(K-2)/2$ oszacowań, obserwacje uczące z i-tej klasy nie są w ogóle wykorzystywane, co oznacza, że ich wynik może być w zasadzie zupełnie dowolny.

Przykładowo, niech $K = 3$ oraz klasyfikowana obserwacja pochodzi z pierwszej klasy.
Na podstawie próby uczącej obliczono, że $r_{12} = 0.58$, $r_{13} = 0.69$.
Ponieważ do oszacowania μ_{23} nie wykorzystujemy obserwacji uczących z pierwszej klasy, zatem oznaczmy dowolny wynik tego oszacowania przez x, $0 \leq x \leq 1$.
Pomimo, że oba oszacowania r_{12} i r_{13} wskazują na klasę pierwszą, aż w 19% źle dobrana wartość x powoduje błędną decyzję.

x	$\hat{d}(\mathbf{x})$
0.00 - 0.04	3
0.04 - 0.85	1
0.85 - 1.00	2

Nieiteracyjna ocena prawdopodobieństwa p_i zaproponowana przez Jelonka i Stefanowskiego (1998).
Wprowadzili oni skojarzony z każdym klasyfikatorem binarnym współczynnik wiarygodności t_{ij} postaci

$$t_{ij} = \frac{v_i}{v_i + e_j},$$

gdzie v_i jest liczbą poprawnie zaklasyfikowanych obserwacji z klasy i-tej, natomiast e_j jest liczbą błędnie zaklasyfikowanych obiektów z klasy j-tej.
Obliczanie współczynników wiarygodności przeprowadzane jest w fazie uczenia klasyfikatorów, tj z próby uczącej. Klasyfikator ten jest postaci

$$\hat{d}(\mathbf{x}) = \arg \max_{1 \leq i \leq K} \sum_{j \neq i} t_{ij} r_{ij}(\mathbf{x}).$$

Nieiteracyjna ocena prawdopodobieństwa p_i pochodząca od Moreiry i Mayoraza (1998).
Weźmy ponownie pod uwagę tożsamość

$$p_i = \sum_{j \neq i} \left(\frac{p_i + p_j}{K-1} \right) \left(\frac{p_i}{p_i + p_j} \right), \quad i = 1, 2, \ldots, K.$$

Niech $q_{ij} = P(Y = i \text{ lub } Y = j | Y \in \{1, 2, \ldots, K\}) = p_i + p_j$ i niech s_{ij} będzie oceną z próby uczącej prawdopodobieństwa q_{ij}.
Oceniając $p_i + p_j$ przez s_{ij}, oceniając $\frac{p_i}{p_i + p_j}$ przez r_{ij} oraz pomijając stały czynnik $(K-1)^{-1}$ otrzymujemy klasyfikator postaci

$$\hat{d}(\mathbf{x}) = \arg \max_{1 \leq i \leq K} \sum_{j \neq i} s_{ij}(\mathbf{x}) r_{ij}(\mathbf{x}).$$

Krzyśko i Wołyński (2008) podali cztery inne nieiteracyjne oceny prawdopodobieństwa p_i.

Weźmy pod uwagę tożsamość

$$\prod_{j\neq i}(p_i+p_j)\frac{p_i}{p_i+p_j}=p_i^{K-1},\quad i=1,2,\ldots,K. \tag{8}$$

Oceniając p_i+p_j przez $2/K$ oraz $\frac{p_i}{p_i+p_j}$ przez r_{ij} otrzymujemy

$$\tilde{p}_i=\sqrt[K-1]{\frac{2}{K}\prod_{j\neq i}r_{ij}}$$

oraz

$$\hat{d}(\mathbf{x})=\arg\max_{1\leq i\leq K}\prod_{j\neq i}r_{ij}(\mathbf{x}).$$

W tożsamości (8) oceniając $p_i + p_j$ przez s_{ij} oraz oceniając $\frac{p_i}{p_i+p_j}$ przez r_{ij} otrzymujemy

$$\tilde{p}_i = \sqrt[K-1]{\prod_{j \neq i} s_{ij} r_{ij}}$$

oraz

$$\hat{d}(\mathbf{x}) = \arg \max_{1 \leq i \leq K} \prod_{j \neq i} s_{ij}(\mathbf{x}) r_{ij}(\mathbf{x}).$$

Weźmy pod uwagę tożsamość

$$\sum_{j\neq i} \frac{p_i}{p_i+p_j}(p_i+p_j)^2 = p_i + (K-2)p_i^2, \quad i = 1,2,\ldots,K.$$

Oceniając $\frac{p_i}{p_i+p_j}$ przez r_{ij} oraz $p_i + p_j$ przez s_{ij} otrzymujemy równanie

$$(K-2)\tilde{p}_i^2 + \tilde{p}_i - \sum_{j\neq i} s_{ij}^2 r_{ij} = 0.$$

Dodatni pierwiastek tego równania jest równy

$$\tilde{p}_i = \frac{1}{2(K-2)}\left(-1 + \sqrt{1 + 4(K-2)\sum_{j\neq i} s_{ij}^2 r_{ij}}\right).$$

Stąd

$$\hat{d}(\mathbf{x}) = \arg\max_{1\leq i\leq K} \prod_{j\neq i} s_{ij}^2(\mathbf{x}) r_{ij}(\mathbf{x}).$$

Weźmy pod uwagę tożsamość

$$\prod_{j\neq i}(p_i+p_j)^2\frac{p_i}{p_i+p_j}=p_i\prod_{j\neq i}(p_i+p_j),\quad i=1,2,\ldots,K.$$

lub równoważnie

$$p_i=\frac{\prod_{j\neq i}(p_i+p_j)^2\frac{p_i}{p_i+p_j}}{\prod_{j\neq i}(p_i+p_j)}.$$

Zastępując po stronie prawej p_i+p_j przez s_{ij} oraz $\frac{p_i}{p_i+p_j}$ przez r_{ij} otrzymujemy

$$\tilde{p}_i=\frac{\prod_{j\neq i}s_{ij}^2 r_{ij}}{\prod_{j\neq i}s_{ij}}.$$

Stąd

$$\hat{d}(\mathbf{x})=\arg\max_{1\leq i\leq K}\tilde{p}_i(\mathbf{x}).$$

Literatura

1. T. Hastie, R. Tibshirani (1998), Classification by pairwise coupling, *The Annals of Statistics* **26**, 451-471.
2. M. Krzyśko (1999), Linear discriminant functions which maximize the area under the ROC curve. *Discussiones Mathematicae: Algebra and Stochastic Methods* **19**, 335-344.
3. M. Krzyśko, W. Wołyński (2008), New variants of pairwise classification, *European Journal of Operational Research* **199**, 512-519.
4. J. Jelonek, J. Stefanowski (1998), Experiments on solving multiclass learning problems by n^2–classifier. In *Proceedings of the 10th European Conference on Machine Learning (ECML-98)*, 172-177.
5. M. Moreira, E. Mayoraz (1998), Improved pairwise coupling classification with correcting classifiers. In *Proceedings of the 10th European Conference on Machine Learning (ECML-98)*, 160-171.
6. D. Price, S. Knerr, L. Personnaz, G. Dreyfus (1995), Pairwise neural network classifiers with probabilistic outputs. In *Advances in Neural Information Processing Systems 7 (NIPS-94)*, 1109-1116.