Wstęp do uczenia maszynowego

Klasyfikacja

Ewa Szczurek + BW (modyfikacje)

bartek@mimuw.edu.pl
Instytut Informatyki
Uniwersytet Warszawski

8 kwietnia 2024

Klasyfikacja, metody podstawowe

Zmienna objaśniana jest **jakościowa** (a nie ilościowa, jak w przypadku regresji liniowej).

Przykładowe metody klasyfikacji:

- K najbliższych sąsiadów (KNN)
- Regresja logistyczna
- Liniowa analiza dyskryminacyjna
- Kwadratowa analiza dyskryminacyjna

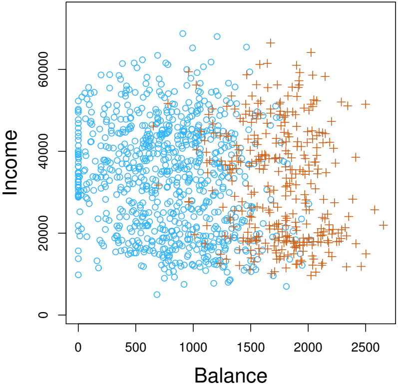
Inne metody będą omówione później.

Przykład problemu klasyfikacji: 'Brak spłaty' ('Default')

Opis:

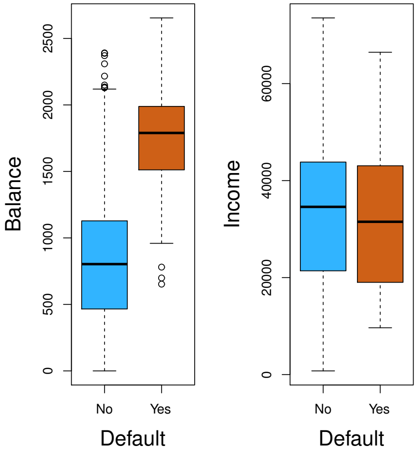
- System kart kredytowych.
- Predyktory:
  - 'Dochód' ('Income')
  - 'Zadłużenie' ('Balance'; miesięczne zadłużenie na karcie)
  - 'Student' (czy osoba jest studentem; wartości: 'TAK lub 'NIE')
- Zmienna objaśniana Y='Default' (czy osoba zaniechała spłacania karty kredytowej; wartości: 'TAK' lub 'NIE')

Ilustracja danych 'Default' dla 10000 klientów



- niebieski: osoby spłacające
- pomarańczowy: osoby, które zaniechały spłacania
- wygląda na to, że zaniechanie spłaty jest częstsze u osób z większym zadłużeniem

Dane 'Default': związek zaniechania spłat z zadłużeniem i dochodami



- Rzeczywiste dane z obserwacji rzadko ukazują tak czyste zależności

Regresja logistyczna

Model logistyczny: przypadek z jedną zmienną predyktorową

Chcemy wyestymować prawdopodobieństwo tego, że zmienna objaśniana Y daje odpowiedź 1, pod warunkiem zmiennej X

$$p_1(X) := Pr(Y = 1 \mid X).$$

- Przybliżanie prostą regresji liniowej dla binarnego kodowania Y

$$p_1(X) = \beta_0 + \beta_1 X.$$

  - Jeśli $\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X > 0.5$, przewidujemy klasę 1, w przeciwnym wypadku 0
  - Dla niektórych wartości X uzyskamy wartości poza przedziałem $[0, 1]$

Model logistyczny: przypadek z jedną zmienną predyktorową

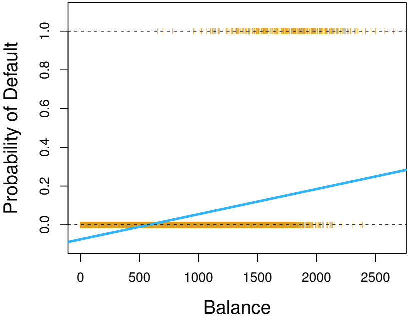
Chcemy wyestymować prawdopodobieństwo tego, że zmienna objaśniana Y daje odpowiedź 1, pod warunkiem zmiennej X

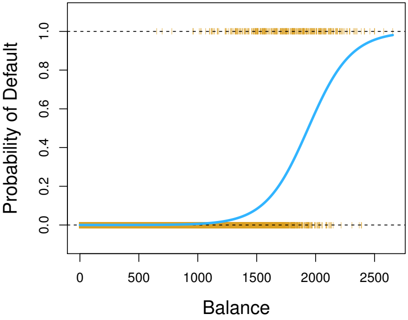
$$p_1(X) := Pr(Y = 1 \mid X).$$

- Przybliżanie **funkcją logistyczną** w $[0, 1]$

$$p_1(X) = \frac{e^{\beta_0 + \beta_1 X}}{1 + e^{\beta_0 + \beta_1 X}}.$$

Przybliżenie prawdopodobieństwa zdarzenia zaniechania spłacania karty w zależności od predyktora 'zadłużenie'





- Prosta regresji liniowej dla binarnego kodowania Y
- Funkcja logistyczna

Regresja logistyczna

Regresja logistyczna przybliża $p_1(X)$ funkcją logistyczną

$$p_1(X) = \frac{e^{\beta_0 + \beta_1 X}}{1 + e^{\beta_0 + \beta_1 X}}$$

Wówczas **iloraz szans** (*odds ratio*) w $[0, \infty)$:

$$\frac{p_1(X)}{p_0(X)} = \frac{p_1(X)}{1 - p_1(X)} = e^{\beta_0 + \beta_1 X}$$

Liniowa zależność pojawia się po zastosowaniu logarytmu:

$$\log\left(\frac{p_1(X)}{1 - p_1(X)}\right) = \beta_0 + \beta_1 X.$$

(ang. *log odds, logit - logistic unit*)

Regresja logistyczna: interpretacja wartości parametrów

Dla regresji liniowej

- β_0 to wartość bazowa Y: średnia Y przy wszystkich predyktorach równych 0
- β_1 to liczba jednostek o ile zmienia się Y gdy zwiększymy X o jedną jednostkę

Dla regresji logistycznej

- β_0 to wartość bazowa logarytmu ilorazu szans: średnia logit przy wszystkich predyktorach równych 0
- β_1 to zmiana logarytmu ilorazu szans gdy zwiększymy X o jedną jednostkę
- Zmiana $p_1(X)$ przy zmianie X o jedną jednostkę zależy od wartości X
- Ale, wiadomo, że
  - Gdy $\beta_1 > 0$, wzrost X zwiększy $p_1(X)$
  - Gdy $\beta_1 < 0$, wzrost X zmniejszy $p_1(X)$

Estymacja parametrów w modelu logistycznej regresji metodą maksymalizacji wiarygodności

Szukamy wartości parametrów $\hat{\beta}_0, \hat{\beta}_1$ tak aby prawdopodobieństwo $p(X)$ dawało wartość bliską jedności dla wszystkich obserwacji, gdzie zaniechano spłaty karty; oraz bliską zera dla tych, gdzie tak się nie stało.

Osiąga się to przez maksymalizację **funkcji wiarogodności**:

$$L(\beta, D) = \prod_{i:\ y_i=1} p_1(x_i) \prod_{j:\ y_j=0} (1 - p_1(x_j)).$$

gdzie D to dane (pary (x_i, y_i) dla $i = 1 \ldots n$).

Wyestymowane parametry regresji logistycznej dla danych 'Default' z predyktorem 'balance'

Parametr	Estymacja	Std. błąd	z-statystyka	p-wartość
$\hat{\beta}_0$	-10.6513	0.3612	-29.5	< 0.0001
balance	0.0055	0.0002	24.9	< 0.0001

- z-statystyka pełni podobną rolę jak t-statystyka w modelu regresji liniowej
- Na przykład, z-statystyka przy założeniu $H_0 : \beta_1 = 0$ to $\hat{\beta}_1 / SE[\hat{\beta}_1]$
- Ta hipoteza sugeruje, że prawdopodobieństwo 'Default' nie zależy od 'balance'.

Jakie jest prawdopodobieństwo zaniechania spłat w zależności od zadłużenia na karcie?

Przy zadłużenieu $X = 1000$ mamy

$$\hat{p}_1(X) = \frac{e^{\hat{\beta}_0+\hat{\beta}_1 X}}{1+e^{\hat{\beta}_0+\hat{\beta}_1 X}} = \frac{e^{-10.6513+0.0055\times 1000}}{1+e^{-10.6513+0.0055\times 1000}} = 0.00576$$

Natomiast dla zadłużenia 2000 mamy

$$\hat{p}_1(X) = \frac{e^{\hat{\beta}_0+\hat{\beta}_1 X}}{1+e^{\hat{\beta}_0+\hat{\beta}_1 X}} = \frac{e^{-10.6513+0.0055\times 2000}}{1+e^{-10.6513+0.0055\times 2000}} = 0.586$$

Estymacja parametrów dla danych 'Default' oraz jakościowego predyktora 'student'

Parametr	Estymacja	Std. błąd	z-stat.	p-wartość
$\hat{\beta}_0$	-3.5041	0.0707	-49.55	< 0.0001
`student (='TAK')`	0.4049	0.1150	3.52	0.0004

Prawdopodobieństwo zaniechania spłat karty osoby, która jest studentem wyliczamy na podstawie parametrów:

$$\hat{Pr}(\texttt{default} = Yes|\texttt{student} = Yes) = \frac{e^{-3.5041+0.4049\times 1}}{1+e^{-3.5041+0.4049\times 1}} = 0.0431$$

Natomiast dla osoby nie będącej studentem:

$$\hat{Pr}(\texttt{default} = Yes|\texttt{student} = No) = \frac{e^{-3.5041+0.4049\times 0}}{1+e^{-3.5041+0.4049\times 0}} = 0.0292$$

Regresja logistyczna dla modelu z wieloma predyktorami

Mamy p predyktorów: $X_1, \ldots, X_p$. Funkcja logistyczna w tym przypadku to:

$$p_1(X) = \frac{e^{\beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p}}{1 + e^{\beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p}}.$$

Natomiast logarytm ilorazu szans jest równy

$$\log\left(\frac{p_1(X)}{1 - p_1(X)}\right) = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p.$$

Regresja logistyczna dla danych 'default'

Parametr	Estymacja	Std. błąd	z-stat.	p-wartość
$\hat{\beta}_0$	-10.8690	0.4923	-22.08	< 0.0001
`balance`	0.0057	0.0002	24.74	< 0.0001
`income`	0.0030	0.0082	0.37	0.7115
`student (='TAK')`	-0.6468	0.2362	-2.74	0.0062

Czy mamy do czynienia z paradoksem?

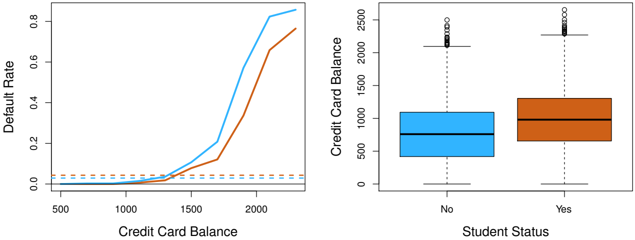
W modelu opartym tylko na tej zmiennej

- parametr związany ze zmienną '`student`', jest dodatni
- zatem prawdopodobieństwo zaprzestania spłat przez studenta jest **wyższe**, niż dla nie-studenta.

W modelu opartym na trzech predyktorach

- parametr związany ze zmienną '`student`' jest ujemny
- zatem, przy ustalonych wartościach zmiennych '`balance`' oraz '`income`', prawdopodobieństwo zaprzestania spłat przez studenta jest **mniejsze** niż dla nie-studenta.

Wyjaśnienie paradoksu



- czerwona linia: '`Student`', niebieska linia: '`nie-student`'
- przerywane: p-stwo zaniechania, uśrednione po '`balance`' oraz '`income`'
- dla banku, student jest mniej ryzykowny niż nie-student **z tym samym '`balance`'**
- '`student`' oraz '`balance`' są zależne (studenci zwykle mają większe zadłużenie).
- Zjawisko zwane **zakłócaniem** (ang. confounding).

Przewidywanie prawdopodobieństwa zaniechania spłat zadłużenia na karcie

Używając wyestymowanych parametrów obliczmy prawdopodobieństwo zaniechania spłat karty dla studenta, mającego dochód 40000 oraz zadłużenie na karcie 1500.

$$\hat{p}_1(X) = \frac{e^{-10.869+0.00574\times 1500+0.003\times 40-0.6468\times 1}}{1+e^{-10.869+0.00574\times 1500+0.003\times 40-0.6468\times 1}} = 0.058.$$

Dla osoby nie będącej studentem, ale mającej ten sam dochód i zadłużenie mamy:

$$\hat{p}_1(X) = \frac{e^{-10.869+0.00574\times 1500+0.003\times 40-0.6468\times 0}}{1+e^{-10.869+0.00574\times 1500+0.003\times 40-0.6468\times 0}} = 0.105.$$

Liniowa analiza dyskryminacyjna (LDA)

LDA: podstawowy pomysł

- Poprzednio modelowaliśmy bezpośrednio prawdopodobieństwo $Pr(Y = k \mid X = x)$.
- Teraz będziemy modelować rozkład wartości predyktorów X, dla każdej zadanej wartości Y z osobna. Korzystając z twierdzenia Bayesa będziemy mogli odzyskać interesujące nas prawdopodobieństwo $Pr(Y = k \mid X = x)$.
- **Korzyści w porównaniu z regresją logistyczną:**
  - Estymacje są bardziej stabilne przy modelu LDA, gdy klasy są dobrze rozdzielone,
  - lub gdy n jest małe, ale rozkład dla każdego predyktora X jest w przybliżeniu normalny.

Zastosowanie twierdzenia Bayesa do klasyfikacji

- Przypuśćmy, że mamy do czynienia z problemem klasyfikacji dla $K \geq 2$ klas.
- Dla $1 \leq k \leq K$, niech π_k przedstawia prawdopodobieństwo tego, że losowo wybrana obserwacja pochodzi z klasy o numerze k (czyli $\pi_k = Pr(Y = k)$). Jest to tzw. prawdopodobieństwo *a priori* (Ang. *prior*).
- Niech $f_k(X)$ będzie funkcją gęstości dla zmiennej losowej X, przy założeniu że Y należy do k-tej klasy.
- Wówczas twierdzenie Bayesa mówi:

$$Pr(Y = k \mid X = x) = \frac{\pi_k f_k(x)}{\sum_{i=1}^{K} \pi_i f_i(x)}.$$

- Przyjmujemy oznaczenie $p_k(x) = Pr(Y = k \mid X = x)$. Jest to tzw. prawdopodobieństwo *a posteriori* (Ang. *posterior*)

LDA dla $p = 1$

- Prior π_k estymujemy z danych obserwowanych jako proporcję liczby przypadków $Y = k$ do liczby wszystkich przypadków.
- Dla estymacji funkcji f_k przyjmujemy założenie, że dane pochodzą z rozkładu normalnego o średniej μ_k oraz wariancji σ_k^2. Wówczas

$$f_k(x) = \frac{1}{\sqrt{2\pi}\sigma_k} \exp\left(\frac{-(x-\mu_k)^2}{2\sigma_k^2}\right).$$

- Przyjmujemy dalsze założenie: $\sigma_1^2 = \ldots \sigma_K^2 = \sigma^2$. Wówczas, na mocy twierdzenia Bayesa, dostajemy

$$\begin{aligned} p_k(x) &= \frac{\pi_k \frac{1}{\sqrt{2\pi}\sigma} \exp\left(\frac{-(x-\mu_k)^2}{2\sigma^2}\right)}{\sum_{i=1}^K \pi_i \frac{1}{\sqrt{2\pi}\sigma} \exp\left(\frac{-(x-\mu_i)^2}{2\sigma^2}\right)} = \frac{\pi_k \exp\left(\frac{-(x-\mu_k)^2}{2\sigma^2}\right)}{\sum_{i=1}^K \pi_i \exp\left(\frac{-(x-\mu_i)^2}{2\sigma^2}\right)} \\ &= \frac{\pi_k \exp\left(\frac{2\mu_k x - \mu_k^2}{2\sigma^2}\right)}{\sum_{i=1}^K \pi_i \exp\left(\frac{2\mu_i x - \mu_i^2}{2\sigma^2}\right)} \end{aligned}$$

LDA dla $p = 1$

- Bayesowski klasyfikator przypisuje obserwację $X = x$ do tej klasy k, dla której $p_k(x)$ przyjmuje wartość największą.
- Ponieważ mianownik jest tu stały, to k o największej wartości $p_k(x)$ jest wyznaczone przez licznik.
- Stosując logarytm przypisujemy obserwację do klasy k, dla której osiągana jest największa wartość **funkcji dyskryminującej**:

$$\delta_k(x) = x \cdot \frac{\mu_k}{\sigma^2} - \frac{\mu_k^2}{2\sigma^2} + \log(\pi_k).$$

- Jest to funkcja liniowa od x, stąd L w LDA.

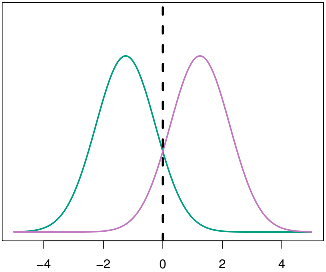
LDA dla $K = 2$ oraz $p = 1$

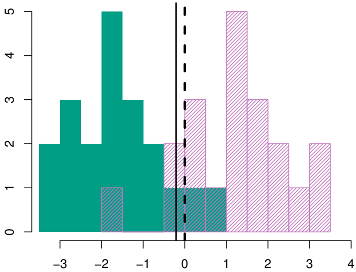
Jeśli $\pi_1 = \pi_2$ to klasyfikator bayesowski przypisuje obserwację do klasy 1 jeśli $2x(\mu_1 - \mu_2) > \mu_1^2 - \mu_2^2$, oraz do klasy 2 w przeciwnym przypadku.

Zatem bayesowska granica decyzyjna w tym przypadku to punkt

$$x = \frac{\mu_1^2 - \mu_2^2}{2(\mu_1 - \mu_2)} = \frac{\mu_1 + \mu_2}{2}.$$

LDA dla $K = 2$ oraz $p = 1$, ilustracja





- dwie funcje gęstości o rozkładzie normalnym.
- linia przerywana - bayesowska linia decyzyjna.

- Histogram losowych obserwacji (po 20 z każdej z tych klas)
- Ciągła czarna linia - linia decyzyjna otrzymana w modelu LDA z danych treningowych.

Estymowanie parametrów dla LDA o K klasach i jednej zmiennej objaśniającej

- Mamy n_k obserwacji w klasie k-tej. $n = n_1 + \ldots + n_K$.
- Wówczas estymujemy parametry następująco:
  - $$\hat{\mu}_k = \frac{1}{n_k} \sum_{i: y_i = k} x_i$$
  - $$\hat{\sigma}^2 = \frac{1}{n - K} \sum_{k=1}^{K} \sum_{i: y_i = k} (x_i - \hat{\mu}_k)^2$$
  - $$\hat{\pi}_k = n_k / n$$
- Parametry te wstawiamy do wzoru na funkcję dyskryminacyjną.

LDA przy więcej niż jednym predyktorze (1)

Mamy $p > 1$ predyktorów $X_1, \ldots, X_p$. Podstawowe założenie to, że $X = [X_1, \ldots, X_p]^T$ jest wylosowane z **wielowymiarowego rozkładu normalnego**: $X \sim N(\mu, \Sigma)$, gdzie $\mu = \mathbb{E}(X) \in \mathbb{R}^p$ jest **wektorem wartości oczekiwanych**, a

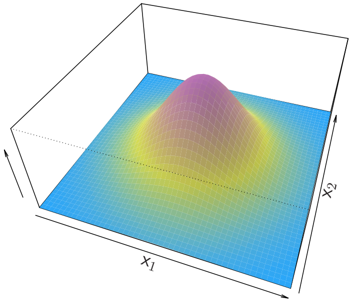
$$\Sigma = Cov(X) = \mathbb{E}((X - \mu)(X - \mu)^T)$$

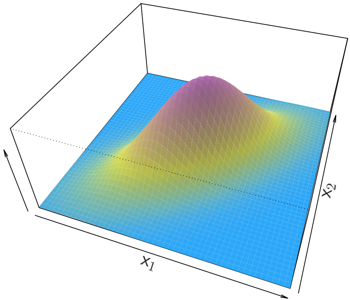
jest $p \times p$ **macierzą kowariancji**.

Funkcja gęstości wielowymiarowego rozkładu normalnego:

$$f(x) = \frac{1}{(2\pi)^{p/2}\sqrt{det \Sigma}} \exp\left(-\frac{1}{2}(x - \mu)^T \Sigma^{-1}(x - \mu)\right)$$

Przykład rozkładów normalnych dwuwymiarowych





- Rozkład normalny przy predyktorach X_1, X_2 nieskorelowanych.
- Rozkład normalny przy predyktorach X_1, X_2 skorelowanych (współczynnik korelacji 0.7).

LDA przy więcej niż jednym predyktorze (2)

Przyjmujemy, że dane z k-tej klasy są losowane z rozkładu normalnego (p-wymiarowego) o wartości oczekiwanej $\mu_k \in \mathbb{R}^p$ oraz wspólnej macierzy kowariancji Σ.

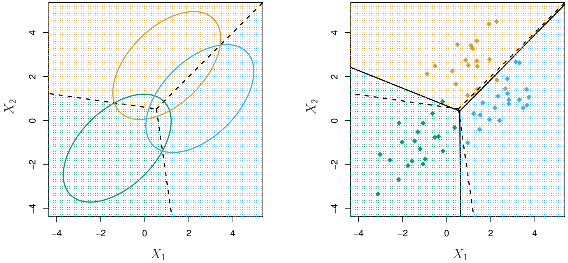
Funkcja dyskryminująca:

$$\delta_k(x) = x^T \Sigma^{-1} \mu_k - \frac{1}{2} \mu_k^T \Sigma^{-1} \mu_k + \log \pi_k$$

Jest to **funkcja liniowa** od x (stąd **L**DA). Dla danych $x \in \mathbb{R}^p$ wybieramy tę klasę k, dla której wartość $\delta_k(x)$ jest największa.

Przykład z trzema klasami

Obserwacje losowane z trzech rozkładów normalnych dwuwymiarowych, o różnych średnich i wspólnej macierzy kowariancji



- elipsy - obszary zawierające 95% prawdopodobieństwa
- przerywane linie - bayesowskie granice decyzyjne dla tego modelu.

- linie ciągłe- granice decyzyjne LDA na podstawie 20 obserwacji wygenerowanych z każdej klasy
- Testowy błąd bayesowski to 0.0746, a dla LDA to 0.0770.

Model LDA dla 'Default'

- trenowany na 10000 danych
- w oparciu o dwa predyktory 'balance' i 'student'
- otrzymany błąd treningowy wynosi 2.75%, ale nie koniecznie oznacza to, że model jest dobry!
- **macierz błędu** (ang. *confusion matrix*) wyjaśnia, dlaczego

		True default status		
		No	Yes	Total
Predicted	No	9,644	252	9,896
default status	Yes	23	81	104
	Total	9,667	333	10,000

- tylko 3.33% wszystkich osoób zaniechało spłacania karty
- trywialny klasyfikator, który każdego klasyfikuje jako 'non-default' popełnia błąd tylko trochę gorszy niż ten wytrenowany.

Miary jakości klasyfikacji

- **Czułość** (*sensitivity, true-positive rate*):

$$TPR = \frac{TP}{TP + FN} = \frac{TP}{P}$$

Tutaj: TPR=81/333=24.3%

- **Swoistość** (*specificity*):

$$SP = \frac{TN}{TN + FP} = \frac{TN}{N}$$

Tutaj SP=9644/9667= 99.8%.

- Co zrobić żeby poprawić czułość? Obniżyć próg dla klasy '`default`'. Obecnie klasyfikujemy do '`default`' jeśli

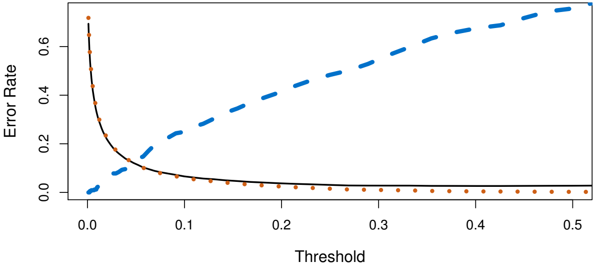
$$Pr(\texttt{default} = Yes \mid X = x) > 0.5$$

Macierz błędu dla progu dla 'default' równym 20%

		True default status		
		No	Yes	Total
Predicted	No	9,432	138	9,570
default status	Yes	235	195	430
	Total	9,667	333	10,000

- Teraz TPR=195/333=58.6 (dużo lepiej) oraz
- SP=9432/9667=97.6% (tylko trochę gorzej)

Zależność poziomu błędów od progu



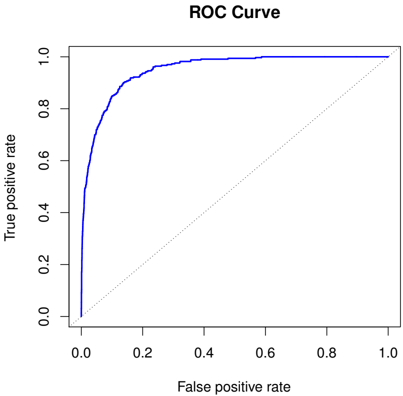
- **Niebieska** przerywana linia - osoby, które zaniechały spłat ale zostały źle sklasyfikowane (1-TPR)
- **Pomarańczowa** kropkowana linia - proporcja błędów wśród osób spłacających kartę (FPR=1-SP)
- **Czarna** ciągła linia - łączny błąd metody (1- *accuracy*)

Miary związane z macierzą błędów

		Predicted class		
		− or Null	+ or Non-null	Total
True class	− or Null	True Neg. (TN)	False Pos. (FP)	N
	+ or Non-null	False Neg. (FN)	True Pos. (TP)	P
	Total	N*	P*	

Name	Definition	Synonyms
False Pos. rate	FP/N	Type I error, 1−Specificity
True Pos. rate	TP/P	1−Type II error, power, sensitivity, recall
Pos. Pred. value	TP/P*	Precision, 1−false discovery proportion
Neg. Pred. value	TN/N*	

Krzywa ROC (*Receiver Operating Characteristics*)



Wykres przy zmieniającym się progu dla zaklasyfikowania jako 'default'.

Ogólna miara klasyfikatora: **AUC** (pole powierzchni pod krzywą ROC, *area under the curve*)

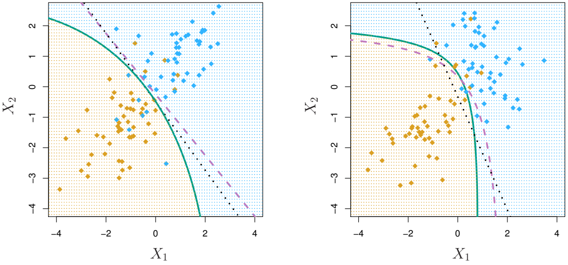
Kwadratowa analiza dyskryminacyjna, QDA

Przyjmujemy, że dane pochodzą z (wielowymiarowego) rozkładu normalnego, ale dane z k-tej klasy są generowane ze specyficzną średnią i **specyficzną macierzą kowariancji** $X \sim N(\mu_k, \Sigma_k)$.
Wówczas funkcja dyskryminująca dla klasy k wygląda następująco:

$$\delta_k(x) = -\frac{1}{2}(x - \mu_k)^T \Sigma_k^{-1}(x - \mu_k) - \frac{1}{2}\log(det(\Sigma_k)) + \log \pi_k.$$

Jest to funkcja **kwadratowa** od x.
Niebezpieczeństwo przeuczenia (overfitting) przy QDA: dla każdej klasy musimy wyestymować $p(p+1)/2$ parametrów macierzy kowariancji więc przy K klasach łączna liczba parametrów dla estymowania macierzy kowariancji wynosi $Kp(p+1)/2$. Łącznie $Kp(p+1)/2 + Kp$ parametrów. Dla LDA liczba parametrów to $p(p+1)/2 + Kp$.

LDA a QDA



Granice decyzjne: *bayesowska* (purpurowa przerywana); *LDA* (czarna kropkowana); *QDA* (zielona ciągła). Dwie klasy i dwa predyktory.
Lewy panel: $\Sigma_1 = \Sigma_2$. Korelacja pomiędzy X_1 i X_2 w obu klasach jest 0.7
Prawy panel: $\Sigma_1 \neq \Sigma_2$. Korelacja pomiędzy X_1 a X_2 w pierwszej klasie (pomarańczowej) jest 0.7, a w drugiej klasie (niebieskiej) -0.7.

Żadna z metd klasyfikacji nie jest lepsza od pozostałych we wszystkich sytuacjach

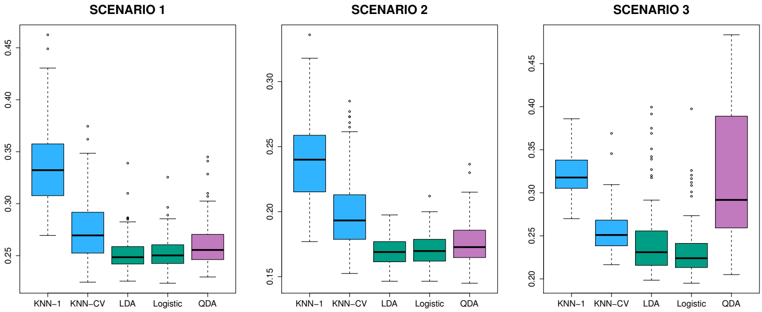
- Logistyczna regresja
- Liniowa analiza dyskryminacyjna (LDA)
- Kwadratowa analiza dyskryminacyjna (QDA)
- KNN
- Logistyczna regresja oraz LDA są podobnymi metodami – obie estymują liniową granicę decyzyjną, ale estymacja jest wykonana różnymi metodami (dla logistycznej regresji to maksymalna wiarygodność, a dla LDA estymacja parametrów rozkładów normalnych).
- KNN jest metodą całkowicie nieparametryczną. Wymaga wybrania K. Dalej porównujemy KNN dla $K = 1$ (KNN-1) i K wybranego automatycznie (KNN-CV)
- O QDA można myśleć jak o czymś pomiędzy LDA i KNN.

Błędy dla trzech liniowych scenariuszy danych ($p = 2$)

(S1) Po 20 danych treningowych losowanych z rozkładu normalnego. Obserwacje z każdej klasy są nieskorelowane, o różnych średnich dla obu klas.

(S2) Jak wyżej, ale predyktory X_1 i X_2 są skorelowane (-0.5).

(S3) X_1 i X_2 są losowane z t-rozkładu.

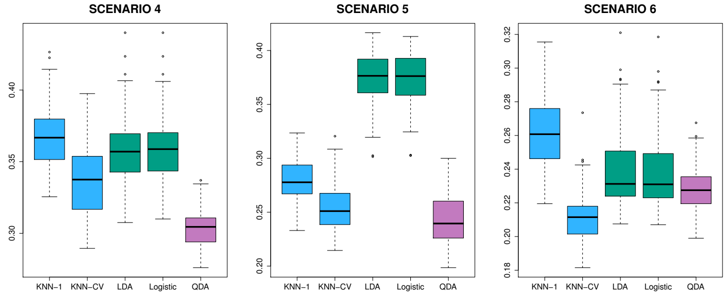


Błędy dla trzech nieliniowych scenariuszy danych ($p = 2$)

(S4) Dane generowane z rozkładu normalnego. Predyktory X_1 i X_2 mają korelację 0.5 w pierwszej klasie i -0.5 w drugiej.

(S5) Dane generowane z rozkładu normalnego o nieskorelowanych predyktorach, ale odpowiedzi były losowane z użyciem logistycznej funkcji od X_1^2, X_2^2 oraz $X_1 X_2$.

(S6) Jak wyżej, ale odpowiedzi losowane z innnej mocno nieliniowej funkcji.



Podsumowanie

Poznaliśmy nowe metody klasyfikacji

- regresja logistyczna
- LDA
- QDA

Materiał dodatkowy

Szacowanie istotności parametrów w regresji logistycznej

Test Walda dla jednego parametru

- $H_0 : \theta = \theta_0$.
- Dany estymator MLE $\hat{\theta}$ z danych D.
- Test korzysta z faktu, że

$$z = \frac{\hat{\theta} - \theta_0}{SE(\hat{\theta})} \sim N(0, 1)$$

gdzie $SE(\hat{\theta})$ to pierwiastek wariancji estymatora $\hat{\theta}$.

Wyprowadzenie dla wektora k parametrów

- W ogólności, dla wektora parametrów θ wymiaru k zachodzi

$$\sqrt{n}(\hat{\theta} - \theta) \sim N(0, I(\theta)^{-1}),$$

gdzie $I(\theta)$ (macierz $k \times k$) to **informacja Fishera** postaci

$$I(\theta) = \left[\mathbb{E}\left[-\frac{\delta^2}{\delta\theta_i\delta\theta_j}\log(f(D;\theta))\right]\right]$$

a $\log(f(D;\theta)) = \ell(\theta, D)$ to log wiarogodność.

- Zatem $\hat{\theta} \sim N(\theta, \frac{1}{n}I(\theta)^{-1})$ i wariancja estymatora $V = \mathrm{Var}(\hat{\theta}) = \frac{1}{n}I(\theta)^{-1}$.

Szacowanie wariancji estymatora dla wektora k parametrów

- Ponieważ

$$\left[-\frac{\delta^2}{\delta\theta_i\delta\theta_j}\ell(\theta, D)\right] = \left[-\frac{\delta^2}{\delta\theta_i\delta\theta_j}\sum_{i=1}^{n}\log(f(D_i;\theta))\right]$$

to $I(\theta)$ można przybliżyć przez

$$J(\hat{\theta}) = \left[\frac{1}{n}\sum_{i=1}^{n} -\frac{\delta^2}{\delta\theta_i\delta\theta_j}\log(f(D_i;\theta))\right]_{\theta=\hat{\theta}}.$$

- Stąd, mamy estymator wariancji estymatora

$$\hat{V} = \frac{1}{n}J(\hat{\theta})^{-1} = \left[-\frac{\delta^2}{\delta\theta_i\delta\theta_j}\ell(\theta, D)\right]^{-1}_{\theta=\hat{\theta}}.$$

- Zauważmy, że nasz estymator jest odwrotnością macierzy Hessego

$$H = \left[-\frac{\delta^2}{\delta\theta_i\delta\theta_j}\ell(\theta, D)\right]$$

w punkcie $\hat{\theta}$.

Test Walda dla k parametrów

- $H_0 : \theta = \theta_0$.
- Przybliżamy błąd standardowy dla $\hat{\theta}_j$ przez j-ty element na diagonali $\hat{V}$, $\hat{SE}(\hat{\theta}_j) = \hat{V}_{jj}$.
- Wówczas

$$z = \frac{\hat{\theta} - \theta_0}{\hat{SE}(\hat{\theta})} \sim N(0, 1)$$

i możemy korzystać z wyznaczania zbioru krytycznego lub p-wartości z rozkładu standardowego normalnego.