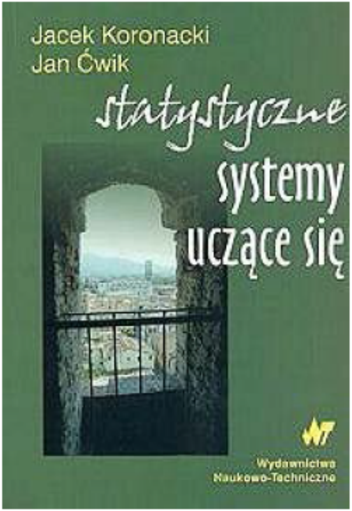
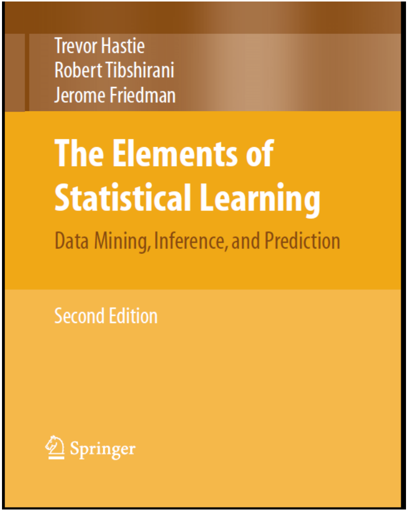
WYKŁAD I: PROBLEM KLASYFIKACJI POD NADZOREM, LINIOWA ANALIZA DYSKRYMINACYJNA

Wydział Matematyki i Nauk Informacyjnych PW, semestr letni 2013/14


Jacek Koronacki
Jan Ćwik
statystyczne
systemy
uczące się
Wydawnictwa
Naukowo-Techniczne


Trevor Hastie
Robert Tibshirani
Jerome Friedman
The Elements of
Statistical Learning
Data Mining, Inference, and Prediction
Second Edition
Springer

Problem klasyfikacji (pod nadzorem) – LDA

Model sytuacji praktycznej: n par losowych postaci

$$\mathcal{U} = (\mathbf{X}_1, Y_1), (\mathbf{X}_2, Y_2), \ldots, (\mathbf{X}_n, Y_n) \sim P_{\mathbf{X},Y}$$

$\mathbf{X}_i$ – wektor zmiennych objaśniających (atrybutów) dla i–tego osobnika,
$\mathbf{X}_i \in \mathcal{X}$ (przestrzeń p-wymiarowa)
Y_i – etykieta przynależności do klasy dla i–tego osobnika,
$Y_i \in \mathcal{G} = \{1, 2, \ldots, g\}$.
$\#\{\mathbf{X}_i : \quad Y_i = j\} = n_j, \quad n_1 + n_2 + \cdots + n_g = n$.

$\mathbf{X}_i$ – wektory losowe, których współrzędne mogą mieć dowolny charakter (zmienne ciągłe, dyskretne, porządkowe, nominalne), rozkład cechy w różnych klasach $(\mathbf{X}|Y = i)$ może być różny.

Cel: na podstawie próby uczącej (treningowej) $\mathcal{U}$ skonstruować klasyfikator, czyli funkcję określającą na podstawie wektora atrybutów $\mathbf{x}$ przynależność do jednej z g klas.

Obserwujemy konkretne wartości zmiennych
$(\mathbf{X}_i, Y_i) : \ \{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)\}$.

Klasyfikacja pod nadzorem (z nauczycielem), analiza dyskryminacyjna

Konstrukcja $\quad d: \ \mathcal{X} \longrightarrow \mathcal{G}$

d – reguła klasyfikacyjna (dyskryminacyjna)

d jest funkcją próby $\{(\mathbf{X}_i, Y_i), \ i = 1, 2, \ldots, n\}$.

d ma „dobrze" przewidywać indeks klasy, z której pochodzi **nowa** obserwacja.

Określenie „pod nadzorem" odpowiada sytuacji, gdy dysponujemy próbą uczącą, która zawiera pełną informację o obserwacjach (tzn. wektor atrybutów i indeks klasy).

Inna sytuacja – mamy tylko wartości atrybutów i szukamy w danych naturalnych skupień: klasyfikacja bez nauczyciela (analiza skupień), np. podział klientów ze względu na zachowania konsumenckie.

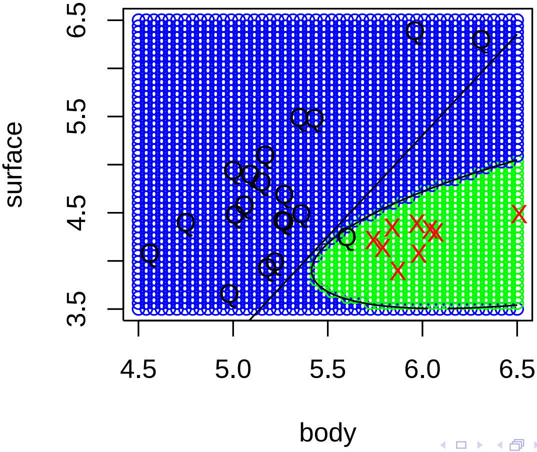
Przykłady.
1) Klasy: „chory”, „zdrowy”
Klasyfikacja pozwala lepiej zrozumieć zależność między faktem zachorowania a atrybutami.

Wariant problemu : przynależność do grupy ryzyka (zachorowania) lub nie.
Grupa ucząca: oparta o historie pacjentów cierpiących na pewną chorobę (sięgającą czasów, gdy na nią nie cierpieli) oraz analogiczna historia badań ludzi zdrowych.
2) Klasy: „dobrzy” i „źli” klienci
– z punktu widzenia ich wypłacalności bankowej – tzw. scoring bankowy.
– z punktu widzenia wierności świadczycielowi usług (np. sieci telefonii komórkowej), churning w CRM

3) Poczta śmieciowa (spam),
dotąd $\#(\mathcal{G}) = 2$, ale może być $\#(\mathcal{G}) > 2$
4) Automatyczne rozpoznawanie cyfr kodów pocztowych:
atrybuty – zaczernienie lub nie odpowiedniego piksela pola, w które wpisuje się cyfrę.
5) Automatyczne rozpoznawanie zapachów (projekt: sztuczny nos, klasyfikacja 'chory', 'zdrowy' na podstawie oddechu)

Uwaga. Podstawowa trudność analizy klasyfikacyjnej polega na **nierozłączności klas**. Możemy mieć obiekty o takich samych lub podobnych wartościach atrybutów należące do różnych klas. Dlatego stosuje się tu modelowanie probabilistyczne.

Dane `earthquake` dotyczące klasyfikacji wybuchów nuklearnych (X) i trzęsień ziemi (Q) na podstawie zmiennyh sejsmologicznych (`body` i `surface`). Klasy zawierają 20 i 9 obserwacji odpowiednio.



Liniowa analiza dyskryminacyjna (LDA – Linear Discriminant Analysis)

Sir R. Fisher, 1936 – podejście bezmodelowe EDA. Przypadek $g = 2$, $x_i \in R^p$, dwie podgrupy uczące odpowiadające $y = 1$ i $y = 2$.
Idea: znajdź kierunek **a**, który najlepiej rozdziela podgrupy uczące po zrzutowaniu na ten kierunek, przy uwzględnieniu zmienności wewnątrzgrupowej rzutów.

$x_{11}, x_{12}, \ldots, x_{1n_1}$ – obserwacje z klasy 1
$x_{21}, x_{22}, \ldots, x_{2n_2}$ – obserwacje z klasy 2

$$\bar{x}_k = \frac{1}{n_k} \sum_{i=1}^{n_i} x_{ki}, \qquad k = 1, 2$$

Jak oceniać zmienność wewnątrzgrupową? (podobny problem jak w ANOVA)

Założenie: klasy charakteryzują się taką samą macierzą kowariancji.
S_1, S_2 – próbkowe macierze kowariancji w klasach.
Macierz kowariancji wewnątrzgrupowej (wspólnej dla obu klas) – uogólnienie połączonego estymatora wariancji)

$$\mathbf{W} = \frac{1}{n-2}\sum_{k=1}^{2}(n_k - 1)S_k = \frac{1}{n-2}\sum_{k=1}^{2}\Big\{\sum_{i=1}^{n_k}(x_{ki} - \bar{x}_k)(x_{ki} - \bar{x}_k)'\Big\}$$

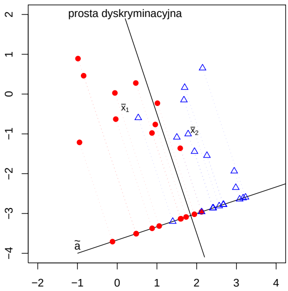
(w grupach centrujemy przez średnią w grupie, a nie przez średnią globalną!)
Użyteczny fakt omówiony przy okazji PCA: $\mathbf{a}$ dowolny wektor o długości 1. $\mathbf{x}_1, \ldots, \mathbf{x}_m$ wektory o próbkowej macierzy kowariancji $\mathbf{S}$, to wariancja rzutów $\mathbf{a}'\mathbf{x}_1, \ldots, \mathbf{a}'\mathbf{x}_m$ wynosi $\mathbf{a}'\mathbf{S}\mathbf{a}$.

Empiryczna wariancja rzutów obu prób na kierunek **a** oceniana przez $\mathbf{a}'\mathbf{W}\mathbf{a}$.

Odległość zrzutowanych środków wynosi $\mathbf{a}'\bar{\mathbf{x}}_2 - \mathbf{a}'\bar{\mathbf{x}}_1$ i jej wariancja może być estymowana przez $(1/n_1 + 1/n_2)\mathbf{a}'\mathbf{W}\mathbf{a}$. Studentyzowana wartość zrzutowanych środków wynosi (z dokładnością do stałej)

$$\frac{(\mathbf{a}'\bar{\mathbf{x}}_2 - \mathbf{a}'\bar{\mathbf{x}}_1)^2}{\mathbf{a}'\mathbf{W}\mathbf{a}}$$

Rys. 1.3.



Metoda Fishera: Rozpatrz rzuty $\bar{\mathbf{x}}_1$ i $\bar{\mathbf{x}}_2$ na kierunek a i maksymalizuj względem $\mathbf{a}$

$$\operatorname*{argmax}_{\mathbf{a}} CRIT(\mathbf{a}) =: \frac{(\mathbf{a}'\bar{\mathbf{x}}_2 - \mathbf{a}'\bar{\mathbf{x}}_1)^2}{\mathbf{a}'\mathbf{W}\mathbf{a}}$$

Szukamy $\tilde{\mathbf{a}}$ jako punktu stacjonarnego

$$\frac{d\,CRIT(\mathbf{a})}{d\mathbf{a}} = \frac{2(\mathbf{a}'\bar{\mathbf{x}}_2 - \mathbf{a}'\bar{\mathbf{x}}_1)(\bar{\mathbf{x}}_2 - \bar{\mathbf{x}}_1)\mathbf{a}'\mathbf{W}\mathbf{a} - 2(\mathbf{a}'\bar{\mathbf{x}}_2 - \mathbf{a}'\bar{\mathbf{x}}_1)^2\mathbf{W}\mathbf{a}}{(\mathbf{a}'\mathbf{W}\mathbf{a})^2} = 0$$

Wektory $\bar{\mathbf{x}}_2 - \bar{\mathbf{x}}_1$ i $\mathbf{W}\tilde{\mathbf{a}}$ muszą być współliniowe. Zatem kierunek $\tilde{\mathbf{a}}$ spełnia

$$\mathbf{W}\tilde{\mathbf{a}} = \bar{\mathbf{x}}_2 - \bar{\mathbf{x}}_1$$

i

$$\tilde{\mathbf{a}} = \mathbf{W}^{-1}(\bar{\mathbf{x}}_2 - \bar{\mathbf{x}}_1).$$

Reguła klasyfikacyjna:

Jeśli $\tilde{\mathbf{a}} = \mathrm{argmax}...$, to zaklasyfikuj wektor $\mathbf{x}$ do klasy j, jeśli

$$|\tilde{\mathbf{a}}'\mathbf{x} - \tilde{\mathbf{a}}'\bar{\mathbf{x}}_j| < |\tilde{\mathbf{a}}'\mathbf{x} - \tilde{\mathbf{a}}'\bar{\mathbf{x}}_k|,$$

dla $k \neq j, k, j \in \{1, 2\}$.

Postać reguły klasyfikacyjnej

Równanie hiperpłaszczyzny prostopadłej do $\tilde{\mathbf{a}}$ i przechodzącej przez $(\bar{\mathbf{x}}_1 + \bar{\mathbf{x}}_2)/2$

$$\tilde{\mathbf{a}}'(\mathbf{x} - (\bar{\mathbf{x}}_1 + \bar{\mathbf{x}}_2)/2) = 0$$

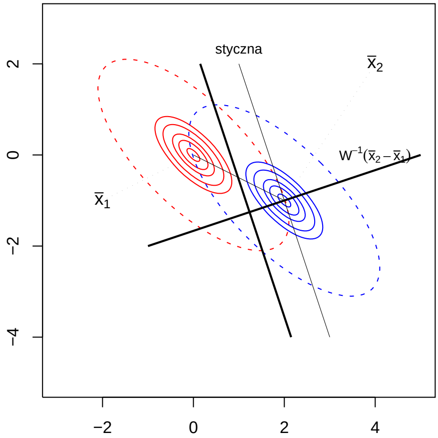
gdzie $\tilde{\mathbf{a}} = \mathbf{W}^{-1}(\bar{\mathbf{x}}_2 - \bar{\mathbf{x}}_1)$.
Reguła Fishera:
Przypisz punkt do klasy 2, jeśli

$$(\bar{\mathbf{x}}_2 - \bar{\mathbf{x}}_1)'\mathbf{W}^{-1}(\mathbf{x} - \frac{1}{2}(\bar{\mathbf{x}}_2 + \bar{\mathbf{x}}_1)) > 0$$

oraz do klasy 1 w przeciwnym przypadku.
$\tilde{\mathbf{a}}$ – (pierwszy) wektor kanoniczny.
$\tilde{\mathbf{a}}'\mathbf{x}$ – (pierwsza) zmienna kanoniczna.

Uwaga Wektor **a** nie jest z reguły wektorem o kierunku $\bar{\mathbf{x}}_2 - \bar{\mathbf{x}}_1$- łączącym środki grup !

Rys. 1.4.



Wektory kanoniczne

Dla $g = 2$ kierunek $\mathbf{a}_1 = \mathbf{W}^{-1}(\bar{\mathbf{x}}_2 - \bar{\mathbf{x}}_1)$ maksymalizuje

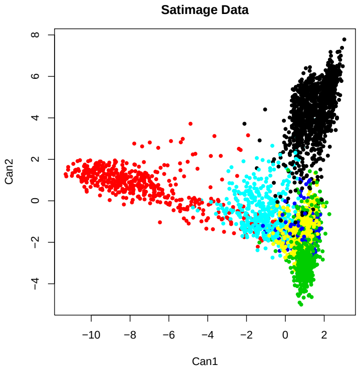
$$\frac{\text{wariancja międzygrupowa rzutów}}{\text{wariancja wewnątrzgrupowa rzutów}} = \frac{\mathbf{a}'\mathbf{B}\mathbf{a}}{\mathbf{a}'\mathbf{W}\mathbf{a}}, \qquad (*)$$

gdzie wariancja międzygrupowa rzutów jest wariancją zrzutowanych średnich (liczonych z krotnością n_k) i $\mathbf{B}$ jest macierzą kowariancji międzygrupowej

$$\mathbf{B} = \frac{1}{(g-1)} \sum_{i=1}^{g} n_k(\bar{\mathbf{x}}_i - \bar{\mathbf{x}})(\bar{\mathbf{x}}_i - \bar{\mathbf{x}})'.$$

Dla dowolnej liczby klas g wektor maksymalizujący $(*)$: pierwszy wektor kanoniczny $\mathbf{a}_1$, wektor $\mathbf{a}_2$ taki, że $\mathbf{a}_2'\mathbf{W}\mathbf{a}_1 = 0$ (nieskorelowanie rzutów w tych kierunkach) i maksymalizujący $(*)$ - drugi wektor kanoniczny itd. Są to wektory własne odpowiadające kolejnym wartościom własnym macierzy $\mathbf{W}^{-1}\mathbf{B}$.

$\mathbf{a}_i'\mathbf{x}_j$: wartość i-tej zmiennej kanonicznej dla j-tej obserwacji. Wykres dwóch pierwszych zmiennych kanonicznych często używany w wizualizacji danych.



Reguła Bayesa, LDA,QDA

Załóżmy chwilowo, że znamy rozkład $P_{\mathbf{X},Y}$ pary $(\mathbf{X}, Y)$. Rozkład ten opisany jest przez:
$p(\mathbf{x}|k)$, $k = 1, 2, \ldots, g$ – dyskretny rozkład prawdopodobieństwa lub gęstość w $k-$tej populacji;
prawdopodobieństwa apriori $\pi_k = P(Y = k)$.

Reguła Bayesa (przy znanych $p(\mathbf{x}|k)$!):

zaobserwowany wektor $\mathbf{x}$ zaklasyfikuj do populacji k, dla której wartość prawdopodobieństwa aposteriori $p(k|\mathbf{x})$ jest największa

$$k = \underset{l=1,\ldots,g}{\operatorname{argmax}}\ p(l|\mathbf{x}).$$

Klasyfikator oparty na regule Bayesa – klasyfikator bayesowski.

Tw. Bayesa $\Longrightarrow$

$$p(k|\mathbf{x}) = \frac{\pi_k p(\mathbf{x}|k)}{\sum_{l=1}^{g} \pi_l p(\mathbf{x}|l)}$$

Reguła Bayesa równoważna maksymalizacji licznika

$$k = \underset{l=1,\dots,g}{\operatorname{argmax}}\ \pi_l p(\mathbf{x}|l) \qquad (*)$$

Cały czas zakładamy, że znamy $p(\mathbf{x}|k)$.

Niech $g = 2$ i rozkłady cechy $\mathbf{X}$ w klasach są normalne z taką samą macierzą kowariancji:

$$p(\mathbf{x}|k) \ \sim \ N(\mathbf{m_k}, \Sigma),$$

czyli

$$p(\mathbf{x}|k) = \frac{1}{(2\pi)^{p/2}|\Sigma|^{1/2}} \exp\big(-\frac{1}{2}(\mathbf{x} - \mathbf{m}_k)'\Sigma^{-1}(\mathbf{x} - \mathbf{m}_k)\big)$$

Σ jest taka sama dla wszystkich klas (założenie).

Dla wartości k spełniającej $(*)$ maksymalne jest

$$\log \pi_k p(\mathbf{x}|k) = \log \pi_k + \log p(\mathbf{x}|k)$$

Maksymalizujemy

$$-\frac{1}{2}(\mathbf{x}-\mathbf{m}_k)'\Sigma^{-1}(\mathbf{x}-\mathbf{m}_k)+\log\pi_k+C$$

$$\delta_k(\mathbf{x}):=-\frac{1}{2}(\mathbf{x}-\mathbf{m}_k)'\Sigma^{-1}(\mathbf{x}-\mathbf{m}_k)+\log\pi_k$$

$\delta_k(\cdot)$ – funkcja dyskryminacyjna dla klasy k

$$\delta_{12}=\log\frac{p(1|\mathbf{x})}{p(2|\mathbf{x})}=\log\frac{p(\mathbf{x}|1)}{p(\mathbf{x}|2)}+\log\frac{\pi_1}{\pi_2}=$$

$$=(\mathbf{m}_1-\mathbf{m}_2)'\Sigma^{-1}\mathbf{x}-\frac{1}{2}(\mathbf{m}_1-\mathbf{m}_2)'\Sigma^{-1}(\mathbf{m}_1+\mathbf{m}_2)+\log\frac{\pi_1}{\pi_2}$$

Jeśli $\delta_{12}>0$, to klasyfikujemy do klasy 1,
jeśli $\delta_{12}<0$, to klasyfikujemy do klasy 2,
jeśli $\delta_{12}=0$, to klasyfikujemy gdziekolwiek (lub zawieszamy decyzję).

$\delta_{12}(\cdot)$ jest liniowa! Równanie $\delta_{12}(\mathbf{x})=0$ jest równaniem hiperpłaszczyzny dyskryminacyjnej, za pomocą której rozdzielamy obie klasy.

Linear Discriminant Analysis (LDA)

Uwaga. Jeśli $\pi_1 = \pi_2$ i $\bar{\mathbf{x}}_i \longrightarrow \mathbf{m}_i$, $\mathbf{W} \longrightarrow \Sigma$, to otrzymujemy metodę LDA Fishera.

Dla $g > 2$,

$$\delta_{kl} = \log \frac{p(k|\mathbf{x})}{p(l|\mathbf{x})} = \log \frac{p(\mathbf{x}|k)}{p(\mathbf{x}|l)} + \log \frac{\pi_k}{\pi_l} =$$

$$= (\mathbf{m}_k - \mathbf{m}_l)'\Sigma^{-1}\mathbf{x} - \frac{1}{2}(\mathbf{m}_k - \mathbf{m}_l)'\Sigma^{-1}(\mathbf{m}_k + \mathbf{m}_l) + \log \frac{\pi_k}{\pi_l}$$

Dla $g = 3$,
$\delta_{12}(\mathbf{x}) > 0$ i $\delta_{13}(\mathbf{x}) > 0 \Longrightarrow$ klasa 1,
$\delta_{12}(\mathbf{x}) < 0$ i $\delta_{23}(\mathbf{x}) > 0 \Longrightarrow$ klasa 2,
$\delta_{13}(\mathbf{x}) < 0$ i $\delta_{23}(\mathbf{x}) < 0 \Longrightarrow$ klasa 3.

$\delta_{kl}(\cdot)$ – funkcja dyskryminacyjna między klasami k i l.

Przy podstawieniu $\mathbf{W}$ w miejsce Σ i $\bar{\mathbf{x}}_i$ w miejsce $\mathbf{m}_i$ otrzymujemy metodę LDA (Linear Discriminant Analysis)

Uwagi.
(1) dla równych prawdopodobieństw apriori $\pi_1 = \pi_2 = \cdots = \pi_g$,
Reguła Bayesa $\equiv$ maksymalizacji $p(\mathbf{x}|k)$
(dyskryminacja metodą największej wiarogodności),
(2) w tej sytuacji maksymalizacja $p(\mathbf{x}|k) \equiv$ minimalizacji odległości Mahalanobisa

$$\mathrm{argmin}(\mathbf{x} - \mathbf{m}_k)'\Sigma^{-1}(\mathbf{x} - \mathbf{m}_k)$$

Sytuacja nierównych macierzy kowariancji Σ_1, Σ_2 (odpowiednio w klasie 1 i 2)

$\delta_k(\cdot)$ – funkcja dyskryminacyjna dla klasy k

$$\delta_k(\mathbf{x}) := -\frac{1}{2}(\mathbf{x} - \mathbf{m}_k)'\Sigma_k^{-1}(\mathbf{x} - \mathbf{m}_k) + \log \pi_k$$

$\delta_{kl}(\mathbf{x}) = \delta_k(\mathbf{x}) - \delta_l(\mathbf{x})$ forma kwadratowa $\mathbf{x}$, a nie funkcja liniowa.

Quadratic Discrminant Analysis (QDA)

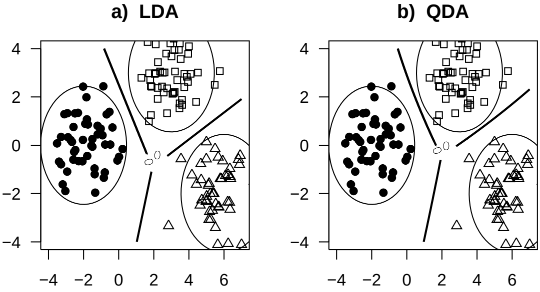
Powierzchnia rozdzielająca klasy k i l opisywana równaniem kwadratowym

$$\{\mathbf{x}:\ \delta_k(\mathbf{x}) \ = \ \delta_l(\mathbf{x})\}$$

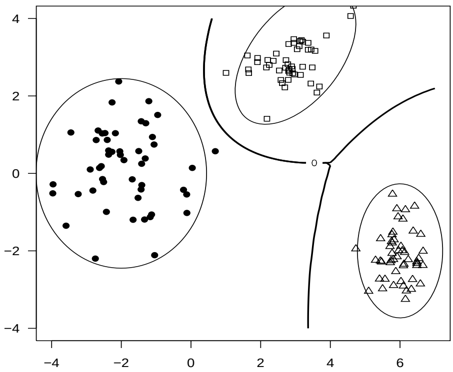
Dla $g = 2$: klasyfikujemy do klasy 2, gdy

$$\delta_{21}(\mathbf{x}) = \frac{1}{2}\log\frac{|\Sigma_1|}{|\Sigma_2|} + \mathbf{x}'(\Sigma_2^{-1}\mathbf{m}_2 - \Sigma_1^{-1}\mathbf{m}_1) -$$

$$\frac{1}{2}\mathbf{x}'(\Sigma_2^{-1} - \Sigma_1^{-1})\mathbf{x} - \frac{1}{2}\mathbf{m_2}'\Sigma_2^{-1}\mathbf{m}_2 + \frac{1}{2}\mathbf{m_1}'\Sigma_1^{-1}\mathbf{m}_1 + \log\frac{\pi_2}{\pi_1} \quad > \quad 0$$



Klasyfikacja 3 prób z rozkładów normalnych o takich samych macierzach kowariancji: metody LDA i QDA



Klasyfikacja 3 prób z rozkładów normalnych o różnych macierzach kowariancji: metoda QDA

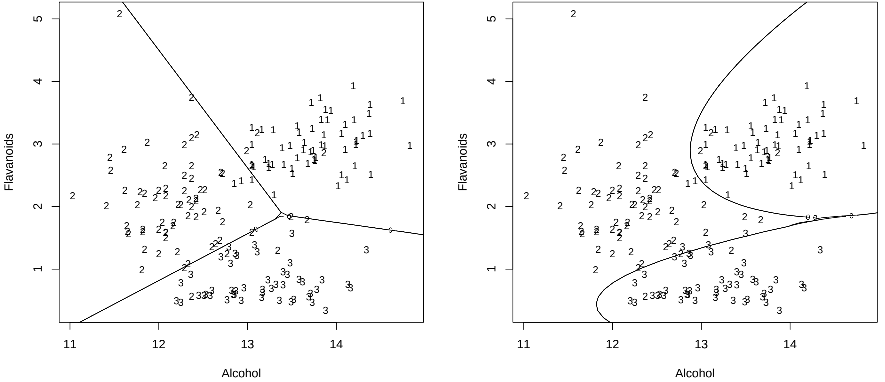
Zbiór `wine.data` zawiera dane dotyczące wyników chemicznej analizy win pochodzących z tego samego regionu Włoch, ale od trzech różnych plantatorów (g=3).
Liczba obserwacji: 177 (klasa 1-58, klasa 2 - 71, klasa 3-48 obserwacji).
Przeprowadźmy analizę LDA i QDA w oparciu o `V2` (zawartość flawonoidów) i `V8` (zawartość alkoholu).

```
wina.lda=lda(V1~V2+V8, data=wina)
wina.pred=predict(wina.lda)
print(table(wina$V1,wina.pred$class))
# tabela klasyfikacji dla metody LDA

     1  2  3
  1 56  3  2
  2  4 60  7
  3  0  0 48

[1] Procent poprawnej klasyfikacji dla proby treningowej:
[1] 0.92135
```

Analogicznie analiza QDA

```
wina.qda=qda(V1~V2+V8, data=wina)
wina.pred=predict(wina.lda)
print(table(wina$V1,wina.pred$class))
     1  2  3
  1 57  2  0
  2  4 65  2
  3  0  3 45
[1] Procent poprawnej klasyfikacji dla proby treningowej:
[1] 0.93820
```

Próba ucząca została podobnie sklasyfikowana przez obie metody, ale obszary przynależności do klas dla obu metod są specyfikowane bardzo różnie (por. różnice dla V8=12.5 i dużych wartości V2).



Flavanoids
Alcohol
Flavanoids
Alcohol

Uwagi

- Można pokazać, że pierwszy wektor kanoniczny $\mathbf{W}^{-1}(\bar{\mathbf{x}}_2 - \bar{\mathbf{x}}_1)$ otrzymuje się również jako estymator współczynników kierunkowych w metodzie MNK, jeśli zakodujemy klasy jako $Y = \pm 1$. To tlumaczy, dlaczego LDA jest odporna na odstępstwa od normalności w klasach.
- Czasami zamiast stosować QDA stosuje się LDA do rozszerzonego zestawu predyktorów $(X_1, \ldots, X_p, X_1X_2, \ldots, X_{p-1}X_p, X_1^2, \ldots, X_p^2)$. Z reguły działa podobnie jak QDA.
- Regularyzowana forma QDA: $\hat{\Sigma}_k(\alpha) = \alpha\hat{\Sigma}_k + (1-\alpha)\hat{\Sigma}$. α wybierana na podstawie działania na zbiorze testowym albo metodą walidacji krzyżowej.
- LDA i QDA mimo swojej prostoty działają często lepiej niż wiele znacznie bardziej wyrafinowanych metod klasyfikacyjnych.