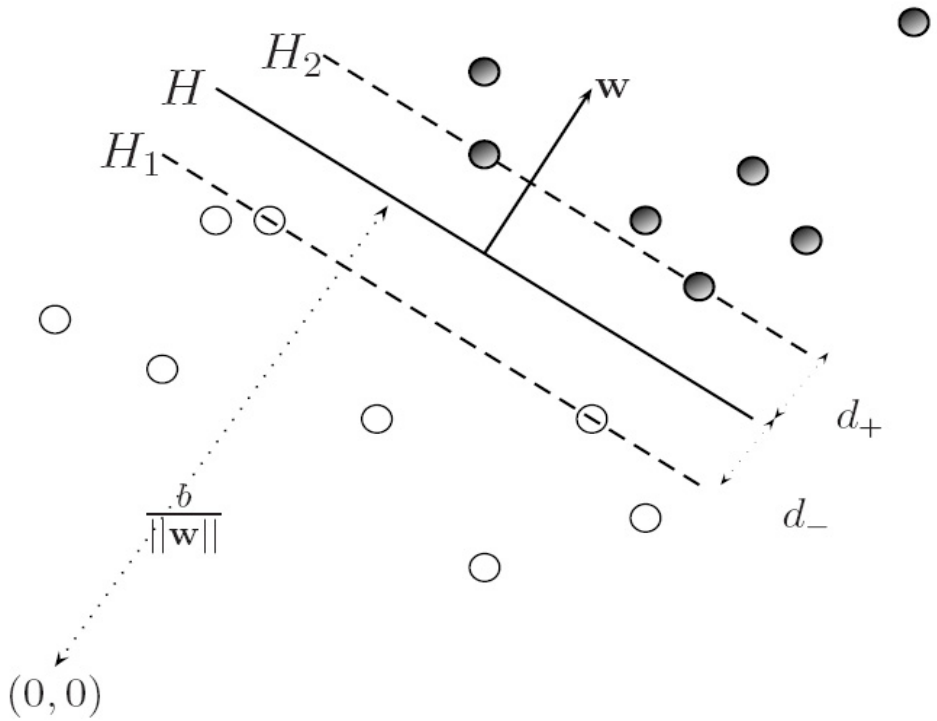
WYKŁAD V: Maszyny wektorów podpierających (SVM). Empiryczne reguły bayesowskie

MiNI PW, semestr letni 2013/2014

SVM – Support Vector Machines

Rozpatrzmy sytuację dwóch klas $g = 2$, $Y = \pm 1$.
Idea dla przypadku liniowo separowalnego (istnieje hiperpłaszczyzna oddzielająca zbiory uczące z różnych klas).

Konstruujemy dwie równoległe hiperpłaszczyzny, we wnętrzu których nie leży ani jeden element próby uczącej, oddalone maksymalnie od siebie.
Wektory leżące na hiperpłaszczyznach – wektory podpierające.



Wyznaczanie hiperpłaszczyzn podpierających (przypadek liniowo separowalny: istnieje hiperpłaszczyzna postaci $\mathbf{x}'\mathbf{w} + b = 0$ rozdzielająca dwie podpróby)

$$(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n), \quad y_i \in \{-1, 1\}, \quad \mathbf{x}_i \in R^p$$

Szukamy wektora $\mathbf{w} \in R^p$ i $b \in R$ takich, że

$$\mathbf{x}_i'\mathbf{w} + b \geqslant +1, \quad \text{gdy } y_i = +1$$

$$\mathbf{x}_i'\mathbf{w} + b \leqslant -1, \quad \text{gdy } y_i = -1$$

Równoważnie

$$y_i(\mathbf{x}_i'\mathbf{w} + b) - 1 \geqslant 0 \quad \text{dla wszystkich } i$$

Przypomnienie: odległość $\mathbf{x}_0$ od hiperpłaszczyzny $f(\mathbf{x}) = \mathbf{x}'\mathbf{w} - \mathbf{a} = 0$ wynosi $|f(\mathbf{x}_0)|/||\mathbf{w}||$.

Hiperpłaszczyzna H_1

$$\mathbf{x}_i'\mathbf{w} + b = 1$$

jest odległa od początku układu współrzędnych (0,0) o $|1 - b|/\|\mathbf{w}\|$

Hiperpłaszczyzna H_2

$$\mathbf{x}_i'\mathbf{w} + b = -1$$

jest odległa od początku układu współrzędnych (0,0) o $|-1 - b|/\|\mathbf{w}\|$

Odległość między hiperpłaszczyznami H_1 i H_2 wynosi

$$\frac{2}{\|\mathbf{w}\|}$$

Problem minimalizacji

$$\frac{\|\mathbf{w}\|^2}{2}$$

Przy ograniczeniach

$$y_i(\mathbf{x}_i'\mathbf{w} + b) - 1 \geqslant 0, \quad i = 1, \ldots, n$$

Optymalna hiperpłaszczyzna umieszczona w środku między hiperpłaszczyznami, tak aby odległości $d_+ + d_-$ od niej do H_1 i H_2 były równe

$$d_+ = d_-$$

Problem minimalizacji funkcji kwadratowej na R^p przy ograniczeniach liniowych

$\alpha_1, \alpha_2, \ldots, \alpha_n \geqslant 0$ – mnożniki Lagrange'a
Szukamy punktu siodłowego funkcji

$$L(\mathbf{w}, b, \boldsymbol{\alpha}) = \frac{1}{2}(\mathbf{w}'\mathbf{w}) - \sum_{i=1}^{n} \alpha_i \Big\{ [(\mathbf{x}_i'\mathbf{w}) + b] y_i - 1 \Big\}$$

Warunki Karusha–Kuhna–Tuckera

$$\frac{\partial L(\mathbf{w}, b, \boldsymbol{\alpha})}{\partial b} = -\sum_{i=1}^{n} \alpha_i y_i = 0$$

$$\frac{\partial L(\mathbf{w}, b, \boldsymbol{\alpha})}{\partial \mathbf{w}} = \mathbf{w} - \sum_{i=1}^{n} \alpha_i y_i \mathbf{x}_i = 0$$

$$\alpha_i \Big\{ y_i(b + \mathbf{x}_i'\mathbf{w}^0) - 1 \Big\} = 0, \quad i = 1, \ldots, n$$

Powyższe równania dają

$$\sum_{i=1}^{n} \alpha_i y_i = 0$$

$$\mathbf{w}^0 = \sum_{i=1}^{n} y_i \ \alpha_i \ \mathbf{x}_i$$

Podstawienie w $L(\mathbf{w}, b, \boldsymbol{\alpha})$ daje

$$L(\boldsymbol{\alpha}) = \frac{1}{2}\|\mathbf{w}^0\|^2 - \sum_{i=1}^{n} \alpha_i \Big\{ y_i(b^0 + \mathbf{x}_i' \mathbf{w}^0) - 1 \Big\} = \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{n} \alpha_i \alpha_j y_i y_j (\mathbf{x}_i' \mathbf{x}_j)$$

Minimalizacja $L(\boldsymbol{\alpha})$ przy warunkach

$$\alpha_i \geqslant 0, \quad i = 1, \ldots, n, \qquad \sum_{i=1}^{n} \alpha_i y_i = 0$$

$\boldsymbol{\alpha}^0 = (\alpha_1^0, \ldots, \alpha_n^0)$ – rozwiązanie, to

$$\mathbf{w}^0 = \sum_{i=1}^{n} \alpha_i^0 \ y_i \mathbf{x}_i$$

Ale $\alpha_i^0 = 0$ dla i takiego, że $\mathbf{x}_i$ nie jest wektorem podpierającym, to jest gdy

$$(\mathbf{x}_i'\mathbf{w}^0 + b^0)y_i \neq 1$$

Stąd

$$\mathbf{w}^0 = \sum_{\text{wektory podp.}}^{n} \alpha_i^0 \; y_i \; \mathbf{x}_i$$

Optymalna hiperpłaszczyzna dyskryminacyjna

$$\sum_{\text{wektory podp.}} y_i \; \alpha_i^0 \; \mathbf{x}_i'\mathbf{x} + b^0 = 0,$$

gdzie

$$b^0 = \frac{1}{2}\left[(\mathbf{w}^{0\prime}\mathbf{x}^*(1)) + (\mathbf{w}^{0\prime}\mathbf{x}^*(-1))\right],$$

gdzie $\mathbf{x}^*(1)$ jest dowolnym wektorem podpierającym z klasy 1, a $\mathbf{x}^*(-1)$ jest dowolnym wektorem podpierającym z klasy -1.

Sytuacja klas nieseparowalnych

Stałe $\xi_i \geqslant 0$ osłabiające warunek liniowej separowalności (kary za nieidealne rozdzielenie podprób przez hiperpłaszczyznę dyskryminacyjną)

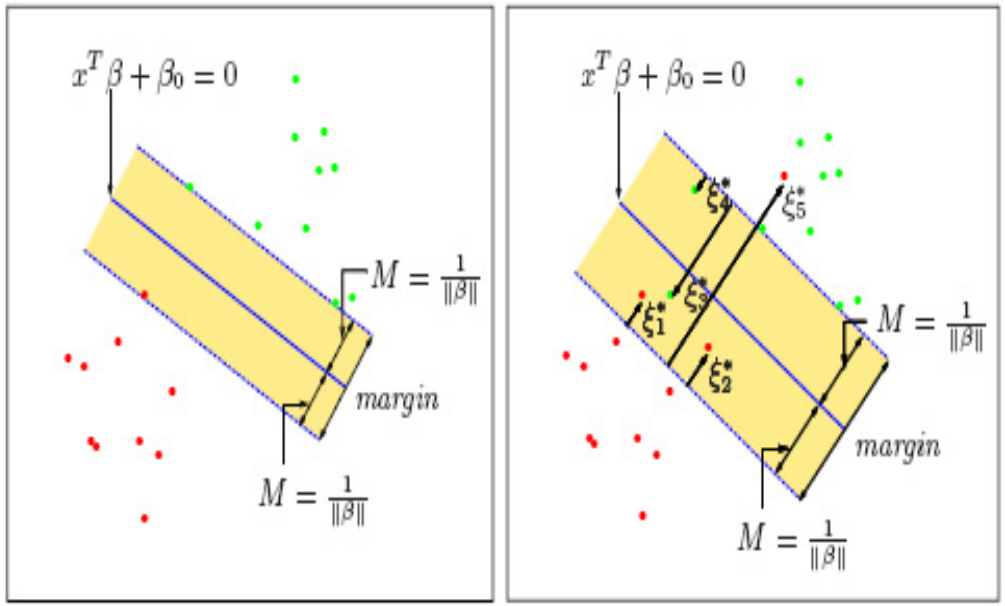
$$\mathbf{x}_i'\mathbf{w} + b \geqslant 1 - \xi_i, \quad \text{gdy} \quad y_i = +1$$

$$\mathbf{x}_i'\mathbf{w} + b \leqslant -1 + \xi_i, \quad \text{gdy} \quad y_i = -1$$

W takiej sytuacji rozwiązujemy problem optymalizacyjny

$$\frac{\|\mathbf{w}\|^2}{2} + C\sum_{i=1}^{n}\xi_i$$

C -parametr kosztu (cost parameter)



Rozwiązanie problemu optymalizacji: minimalizacja funkcji $L(\alpha)$ przy ograniczeniach $0 \leqslant \alpha_i \leqslant C$ i $\sum_{i=1}^{n} \alpha_i y_i = 0$.

Metoda jądrowa Z reguły łączy się przedstawioną wyżej metodę SVM z przekształceniem oryginalnych zmiennych.

Oparte to jest na spostrzeżeniu, że wyznaczanie optymalnej hiperpłaszczyzny rozdzielającej w metodzie SVM wymaga wyznaczenia iloczynów skalarnych $\mathbf{x}_i'\mathbf{x}_j$ i $\mathbf{x}_i'\mathbf{x}$

$$\Phi: \quad R^p \longrightarrow R^N$$

$$(\mathbf{x}_i, y_i) \longrightarrow (\Phi(\mathbf{x}_i), y_i), \quad i = 1, \ldots, n$$

Metoda SVM w przestrzeni nowych cech $\Phi(\mathbf{x})$.
Możemy dokonać nieliniowych przekształceń atrybutów.
Wielomiany stopnia drugiego:
przekształcenia liniowe w przestrzeni o $2p + p(p-1)/2 = p(p+3)/2$ współrzędnych

$$z^{(1)} = x^{(1)}, \ldots, z^{(p)} = x^{(p)}$$

$$z^{(p+1)} = \left(x^{(1)}\right)^2, \ldots, z^{(2p)} = \left(x^{(p)}\right)^2$$

$$z^{(2p+1)} = x^{(1)}x^{(2)}, \ldots, z^{(p(p+3)/2)} = x^{(p)}x^{(p-1)}$$

Aby stosować SVM musimy tylko umieć liczyć iloczyny skalarne w przestrzeni R^N

$$K(\mathbf{x}, \mathbf{y}) = \Phi(\mathbf{x})' \ \Phi(\mathbf{y})$$

okazuje się, że można scharakteryzować funkcje $K(\mathbf{x}, \mathbf{y})$, które wyznaczają iloczyn skalarny w R^N i zamiast wybierać Φ można wybierać $K(\mathbf{x}, \mathbf{y})$!

Jądro wielomianowe stopnia d

$$K(\mathbf{x}, \mathbf{y}) = (\mathbf{x}'\mathbf{y} + 1)^d,$$

Jądro gaussowskie radialne

$$K(\mathbf{x}, \mathbf{y}) = \exp(-\gamma \|\mathbf{x} - \mathbf{y}\|^2 / 2\sigma^2)$$

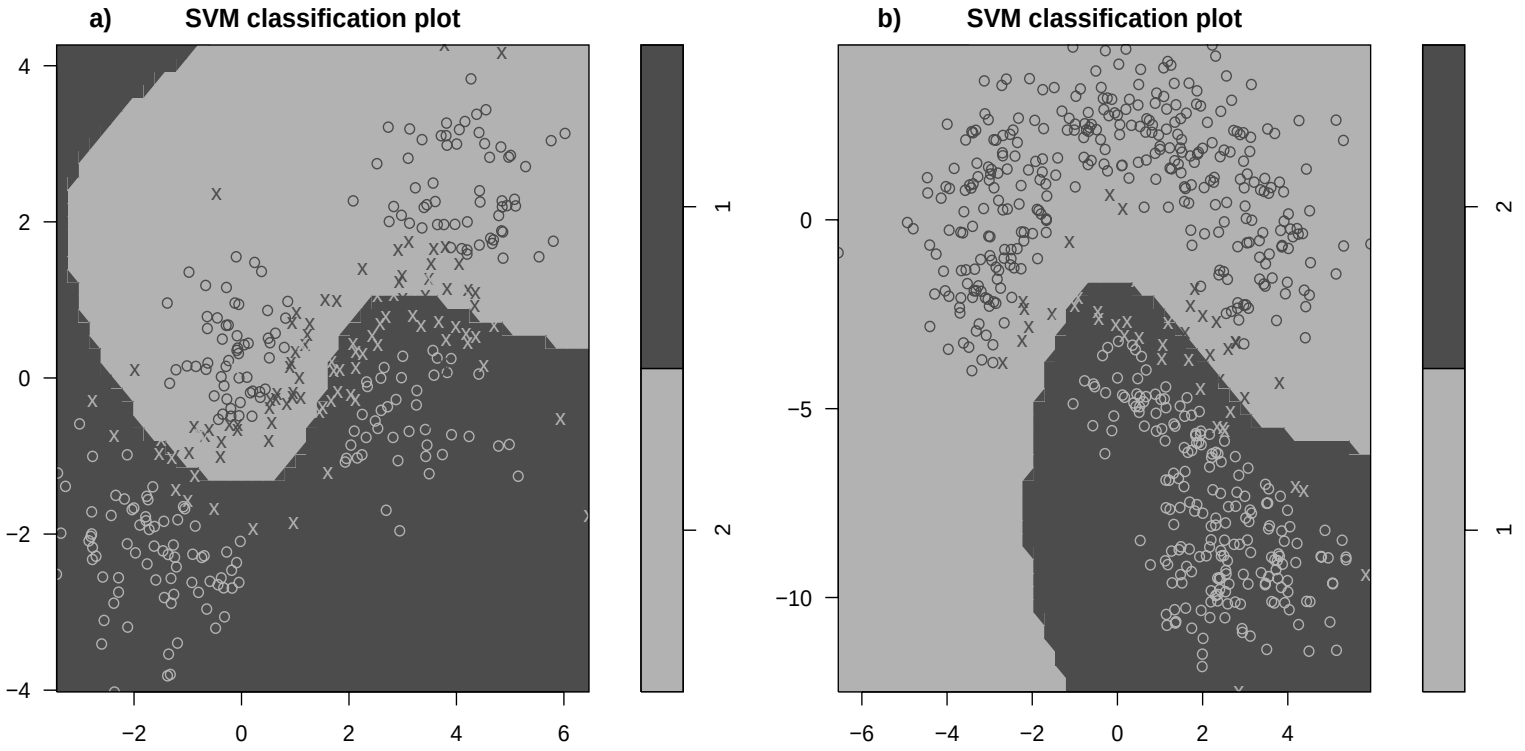
γ-parametr jądra radialnego Jądro Laplace'a

$$K(\mathbf{x}, \mathbf{y}) = \exp(-\|\mathbf{x} - \mathbf{y}\|)$$

Jądro wielomianowe stopnia d $K(\mathbf{x}, \mathbf{y}) = (\mathbf{x}'\mathbf{y} + 1)^d$, opisuje iloczyn skalarny w przestrzeni, której współrzędne zawierają wszystkie iloczyny oryginalnych zmiennych $x^{(j)}$ stopnia d i stopni niższych. Reguła decyzyjna ma postać

$$\mathrm{sgn}\Big(\sum_{\text{wektory podp.}} y_i \alpha_i{}^0 K(\mathbf{x}_i, \mathbf{x}) + b^0\Big)$$

Podejście jądrowe połączone z SVM umożliwia budowę klasyfikatorów o nieliniowych, elastycznych kształtach w oryginalnej przestrzeni obserwacji.



pakiet `e1071`, funkcja `svm`, opcja kernel= "radial".
Pakiet zawiera funkcje `tune.svm` pozwalającą wybrać optymalne parametry kosztu C i γ.

Empiryczne reguły bayesowskie

Naiwną regułę bayesowską (por. wykład VIII) można wykorzystywać również w przypadku, gdy atrybuty są ciągłe, wymaga to jednak estymacji gęstości $p(x^{(i)}|k)$ (prawdopodobieństwa a priori estymowane są przez frakcje elementów z odpowiednich klas).
Rozpatrzmy ogólnie problem estymacji gęstości prawdopodobieństwa $p(x)$ na podstawie prostej próby losowej $X_1, \ldots, X_n$ losowanej z rozkładu o tej gęstosci. Najprostszy estymator gęstości p- histogram (otrzymywany w R instrukcją `hist(..,prob=T)`). Estymator jądrowy: dla ustalonej gęstości prawdopodobieństwa K (jądra) i parametru wygładzającego h_n (zależy od liczności próby n i często od danych)

$$\hat{p}_n(x) = \frac{1}{nh_n} \sum_{i=1}^{n} K\left(\frac{x - X_i}{h_n}\right)$$

W przypadku szczególnym jądra jednostajnego na przedziale $[-1/2, 1/2]$ estymator jądrowy ma postać

$$\hat{p}_n(x) = \frac{\text{liczba punktów w } [x - h_n/2, x + h_n/2]}{nh_n}$$

Analiza teoretyczna wskazuje na duży wpływ parametru wygładzającego na zachowanie się $\hat{p}_n(x)$. Ogólna zależność: $h_n \uparrow \Rightarrow$ wariancja $\hat{p}_n(x) \downarrow$, natomiast gdy $h_n \downarrow \Rightarrow$ obciążenie $\hat{p}_n(x) \downarrow$. Wybór h_n powinien równoważyć te dwie tendencje. Propozycje:

- Parametr wygładzający Silvermana:

$$h_n = (4/3)^{1/5} \tilde{\sigma} n^{-1/5} \qquad \text{gdzie } \tilde{\sigma} = min(S, IQR/1.34)$$

- Parametr wygładzający Sheathera-Jonesa;
- Metoda oparta na $k(n)$ najbliższych sąsiadach, gdzie $k(n) \in N$ - parametr metody:
  $h_n = R_n$: odległość od x do $k(n)$-tego najbliższego sąsiada spośród obserwacji $X_1, \ldots, X_n$.

Metoda najbliższego sąsiada

W problemie klasyfikacji przy modyfikacji powyższej definicji R_n do $\tilde{R}_n$: promień najmniejszej kuli zawierającej $k(n)$ obserwacji w próbie połączonej predyktorów z obu klas,
empiryczna reguła bayesowska ma postać:

wybierz klasę, do której należy najwięcej obserwacji w tak wybranej kuli

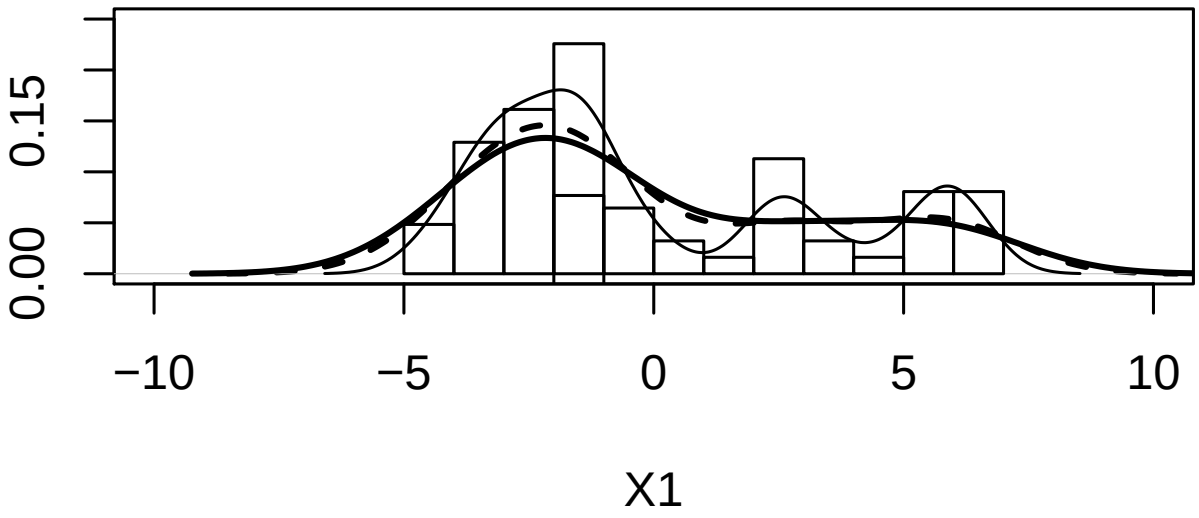
Reguła kNN ($k(n)$ najbliższych sąsiadów).
Taka sama definicja dla wielowymiarowego wektora atrybutów !

Przykład. Przeprowadźmy estymację gęstości dla zmiennej $X1$ danych geny3PC (pierwsza składowa główna dla danych Alizadeha, por. CM (2009), str. 79). Dane zawierają obserwacje z 3 populacji, można się spodziewać, że gęstość X1 może być wielomodalna. bw=bw.nrd oznacza parametr wygładzający Silvermana, bw.SJ Sheathera-Jonesa, parametr wygładzający domyślny bw.nrd0 : w definicji parametru Silvermana $(4/3)^{1/5} = 1.06$ zastąpione jest przez 0.9.

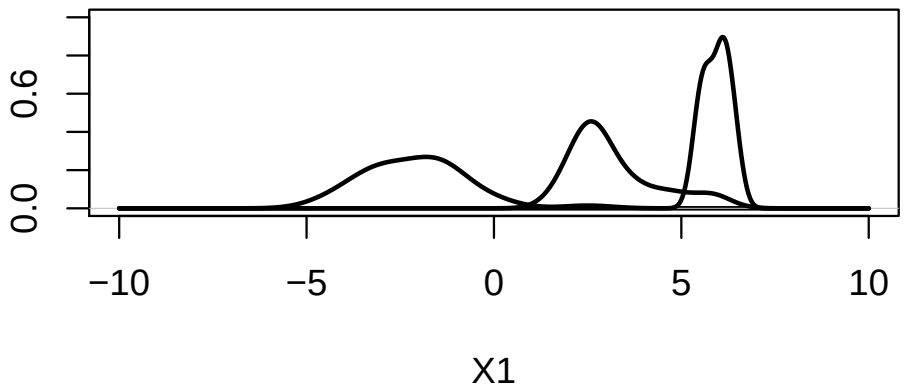
```
dens1<-density(geny3.PC$X1, kernel="gaussian")
dens2<-density(geny3.PC$X1, bw=bw.nrd(geny3.PC$X1),
kernel="gaussian")
dens3<-density(geny3.PC$X1, bw=bw.SJ(geny3.PC$X1),
kernel="gaussian")
```

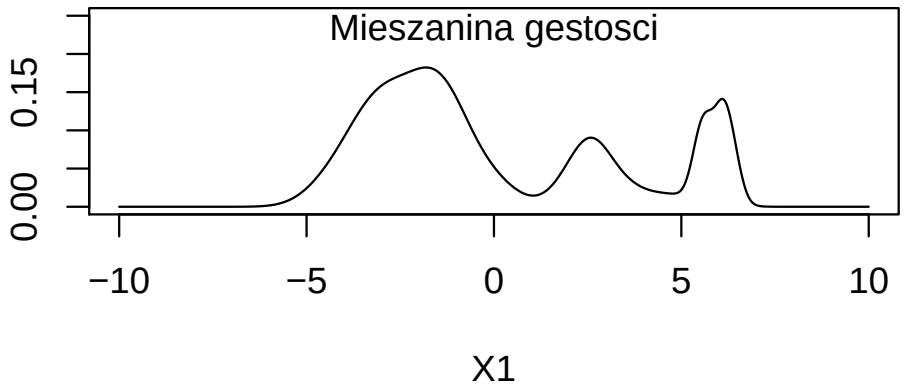
linia ciągła zwykła - rozstęp Sheathera-Jonesa
linia ciągła pogrubiona - rozstęp Silvermana
linia przerywana - rozstęp bw.nrd0.

Estymator jądrowy z rozstępem Sheathera-Jonesa najbardziej adekwatnie opisuje mody histogramów.

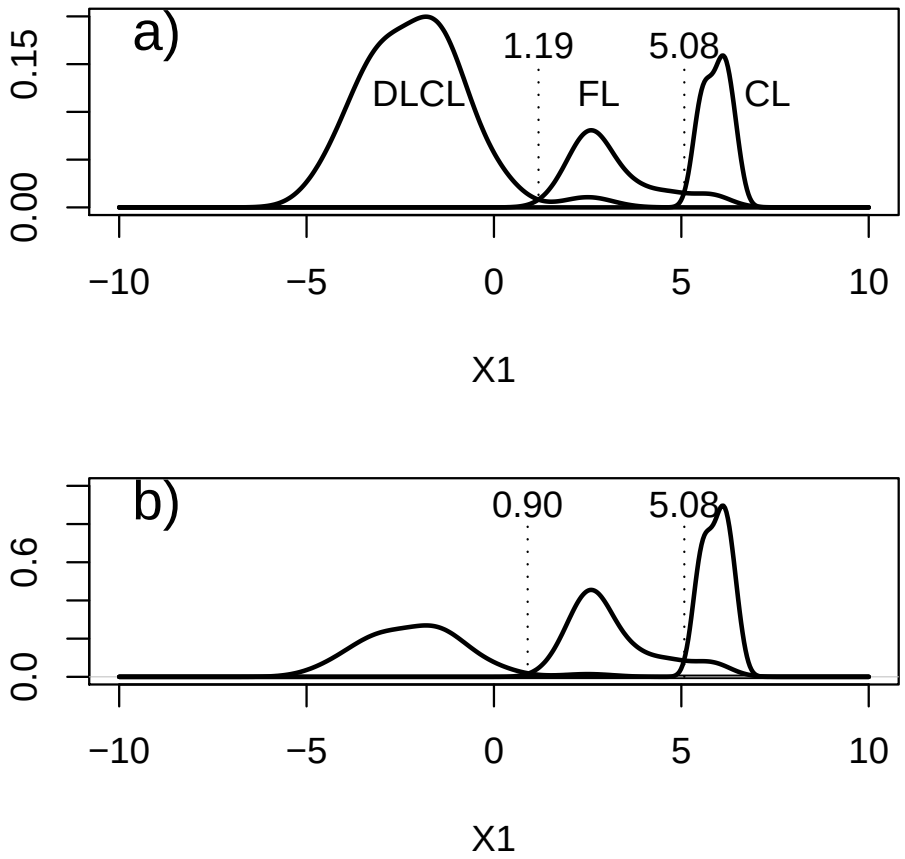


Wyznaczmy estymatory gęstości w poszczególnych klasach. Gęstości w grupach są jednomodalne i omawiane parametry wygładzające działaja bardzo podobnie. Stosujemy `bw=bw.nrd` (rozstęp Silvermana).
Wyznaczmy jeszcze inny estymator gęstości $X1$ z wzoru $\sum_{i=1}^{3} \hat{\pi}_i \hat{f}_i(x)$.

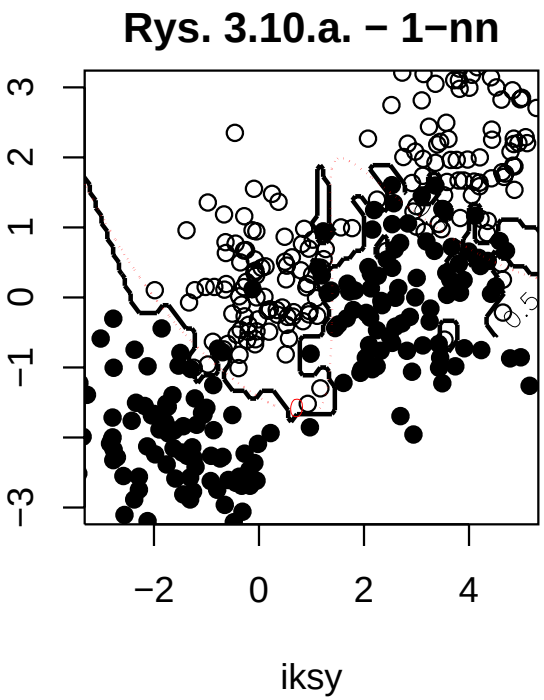


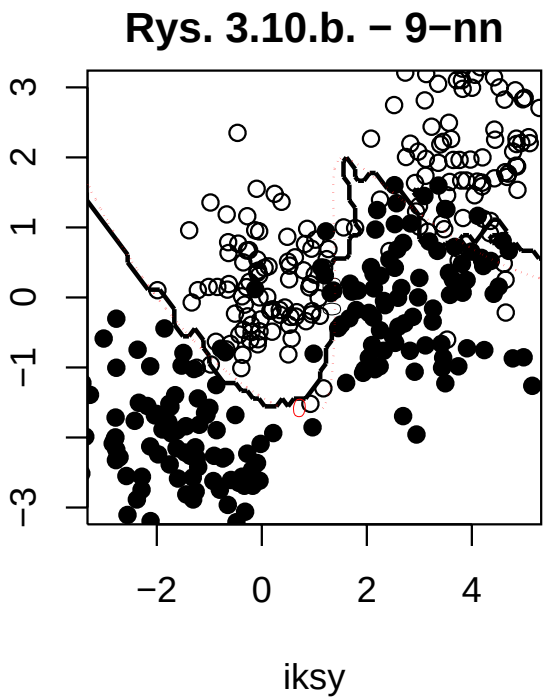


Wykorzystajmy jeszcze otrzymane estymatory do konstrukcji empirycznej reguły bayesowskiej ($\hat{\pi}_i \hat{f}_i(x) > \hat{\pi}_j \hat{f}_j(x)$ dla $j \neq i$ to klasyfikuj do klasy i) przy $\hat{\pi}_1 = 46/62$, $\hat{\pi}_2 = 11/62$ i $\hat{\pi}_3 = 9/62$ oraz przy $\hat{\pi}_i = 1/3$



Działanie metody kNN dla $k = 1$ i $k = 9$ dla symulowanego zbioru danych. Znacznie większa regularność granicy obszarów decyzyjnych dla drugiego przypadku.





Metody prototypowe

Metoda kNN jest metodą prototypową: za prototyp uważamy obserwację leżącą najbliżej klasyfikowanej obserwacji. W wielu metodach prototypy nie muszą być przykładami z próby treningowej.

Metoda K-średnich (g-średnich)

Metoda analizy skupień może być wykorzystna w klasyfikacji:
Mając grupę punktów wybieramy w jakiś sposób R prototypów. Dla każdego punktu znajdujemy najbliższy prototyp - wyznaczamy w ten sposób punkty $\mathcal{C}_i$ najbliższe prototypowi i. Wyliczamy środki ciężkości $\mathcal{C}_i$. Stają się one nowymi prototypami .. itd
Wykorzystanie w klasyfikacji: w każdej klasie umieszczamy R prototypów i stosujemy algorytm k-średnich. Później metoda najbliższego sąsiada: szukamy najbliższego spośród $g \times K$ prototypów do $\mathbf{x}$ i klasyfikujemy $\mathbf{x}$ do klasy najbliższego prototypu.

Metody prototypowe cd. Learning Vector Quantization

- 1. Wybierz po R prototypów w każdej klasie $m_1(k), \ldots m_R(k), k = 1, \ldots, g$ (może być losowy wybór R punktów z każdej klasy).
- 2. Losowo wybierz (ze zwracaniem) punkt z próby uczącej i znajdź najbliższy prototyp $m_j(k)$.
- (i) Jesli $y_i = k$ (elementy są w tej samej klasie) przesuń prototyp w kierunku punktu x_i.

$$m_j(k) = m_j(k) + \eta(x_i - m_j(k)),$$

$\eta > 0$

- (ii) Jesli $y_i \neq k$ (elementy są w różnych klasach) oddal prototyp od punktu x_i.

$$m_j(k) = m_j(k) - \eta(x_i - m_j(k))$$

- Powtarzaj krok 2 zmniejszając w każdej iteracji η.

Wada: sam algorytm, nie odpowiada optymalizacji żadnego kryterium.

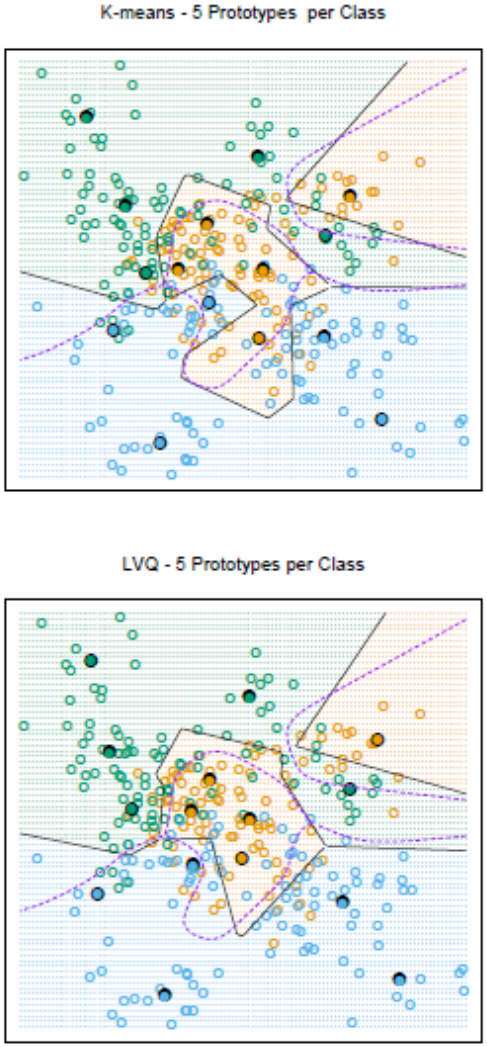


FIGURE 13.1. *Simulated example with three classes and five prototypes per class. The data in each class are generated from a mixture of Gaussians. In the upper panel, the prototypes were found by applying the K-means clustering algorithm separately in each class. In the lower panel, the LVQ algorithm (starting from the K-means solution) moves the prototypes away from the decision boundary. The broken purple curve in the background is the Bayes decision boundary.*