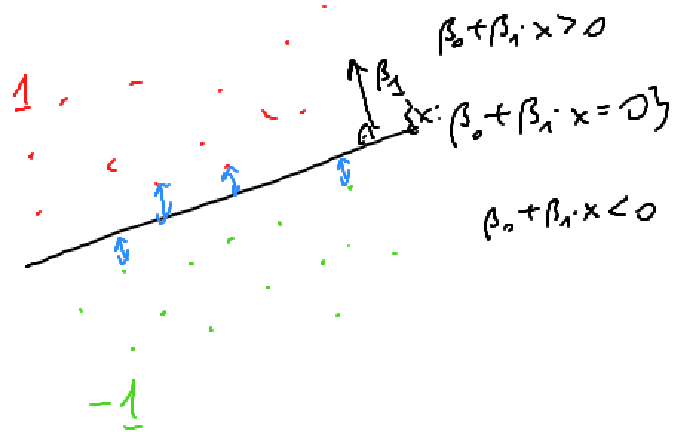
"Maximal margin classifier"

$$(x_i, y_i) \in \mathbb{R}^p \times \{-1,1\}$$

Zakładamy, że istnieją $\beta_0 \in \mathbb{R}$, $\beta_1 \in \mathbb{R}^p$:

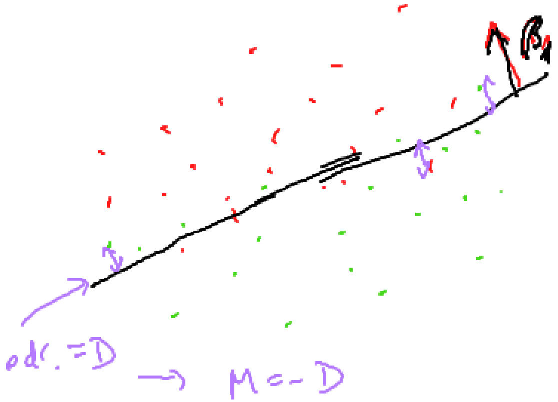
$$(\beta_0 + \beta_1 \cdot x_i)\, y_i > 0 \qquad , i=1,\dots,n$$



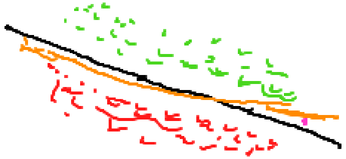
Wybieramy β_0, β_1 tak, aby zmaksymalizować M takie, że

$$(\beta_0 + \beta_1 \cdot x_i)\, y_i \geq M \qquad , i=1,\dots,n$$

przy założeniu, że $\|\beta_1\|^2 = 1$.



Klasyfikacja: $\operatorname{sgn}(\beta_0 + \beta_1 \cdot x)$



"Support vector classifier" („klasyfikator wektorów nośnych" ?)

$C \geq 0$ – parametr ustalony z góry

· Wybieramy β_0, β_1 tak, aby zmaksymalizować M takie, że

zachodzi $y_i \cdot (\beta_0 + \beta_1 \cdot x_i) \geq M(1 - \varepsilon_i)$, $i = 1, \dots, n$, $\|\beta_1\|^2 = 1$

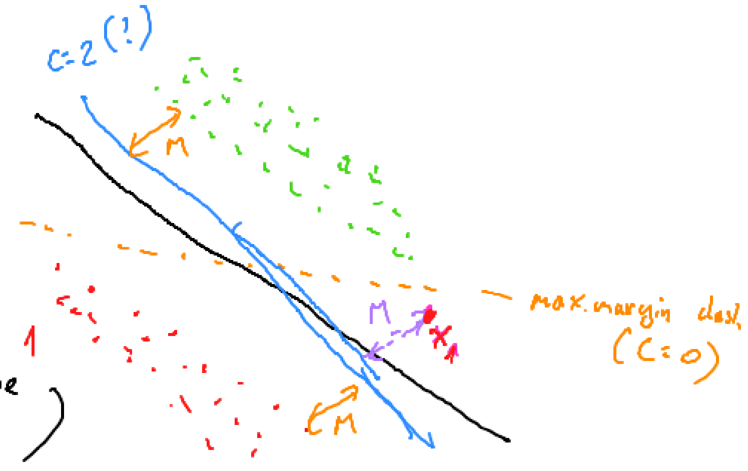
dla pewnych $\varepsilon_i \geq 0$, $\sum_{i=1}^{n} \varepsilon_i \leq C$

(tw. Karusha–Kuhna–Tuckera)

Klasyfikacja: $\operatorname{sgn}(\beta_0 + \beta_1 \cdot x)$ $\langle x_i, x_j \rangle$

Znajdujemy $f(x) = \beta_0 + \beta_1 \cdot x = \sum_{i=1}^{n} \alpha_i y_i \langle x, x_i \rangle + \beta_0$

dowodzi się, że f jest takiej postaci (dlaczego tak jest nie będzie wyjaśnione)



$\varepsilon_1 = C$, $\varepsilon_2 = \dots = \varepsilon_n = 0$

tj. $C = 2$

$$\begin{bmatrix} \langle x_1, x_1 \rangle & \langle x_1, x_2 \rangle & \dots & \langle x_1, x_n \rangle \\ \vdots & & & \\ \langle x_n, x_1 \rangle & \dots & & \langle x_n, x_n \rangle \end{bmatrix}$$

← ~~wyznacznik~~ macierz Grama T te iloczyny wyznaczają położenie $(x_1, \dots, x_n)$ względem 0

$\langle Ux_i, Ux_j \rangle = \langle x_i, x_j \rangle$

„Support vector machine" (maszyna wektorów nośnych)

W metodzie SVC zastępujemy $\langle \cdot , \cdot \rangle$ pewną funkcją $K(\cdot , \cdot)$ symetryczną nieujemnie określoną

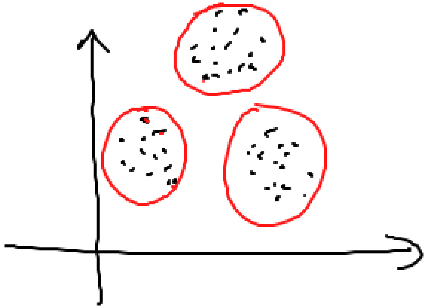
np. $$K(x, x') = \exp(-\gamma \|x - x'\|^2)$$

$$K(x, x') = (1 + \|x - x'\|^2)^p$$

jest w scikit

Klasteryzacja

uczenie nienadzorowane



Cel: podzielić dane na pewną liczbę grup tak, aby dane wewnątrz każdej grupy były podobne do siebie, a dane z różnych grup niepodobne do siebie

„K-średnich"   (K-means)

Zadajemy z góry liczbę klas K.

dane: $x_1, \dots, x_n \in \mathbb{R}^p$

Notacja: $C_1, C_2, \dots, C_K \subset \{1, 2, \dots, n\}$ podzbiory zbioru indeksów

$$C_1 \cup \dots \cup C_K = \{1, \dots, n\}$$

$$C_i \cap C_j = \emptyset \text{ dla } i \neq j$$

$\rightsquigarrow$ C_i – klaster i-ty $\{x_l : l \in C_i\}$ – dane z klastra i-tego

Pomysł: podział na klastry jest dobry, jeśli łączna wariancja wewnątrzklastrowa jest mała:

czyli minimalizujemy $\sum_{j=1}^{K} W(C_j)$

częsty wybór: $W(C_k) = \frac{1}{|C_k|} \sum_{i,i' \in C_k} \| x_i - x_{i'} \|^2$

<u>Algorytm K-średnich</u> pozwala znaleźć pewne rozbicie C_k, dla którego $\sum_{j=1}^{K} W(C_j)$ jest w miarę mała.

1) Losowo przydzielamy obserwacje do jednej z klas $1, \dots, K$
w ten sposób dostajemy pewien początkowy podział na klastry $C_1, \dots, C_K \neq \emptyset$
← o to trzeba zadbać.

2) Iterujemy:

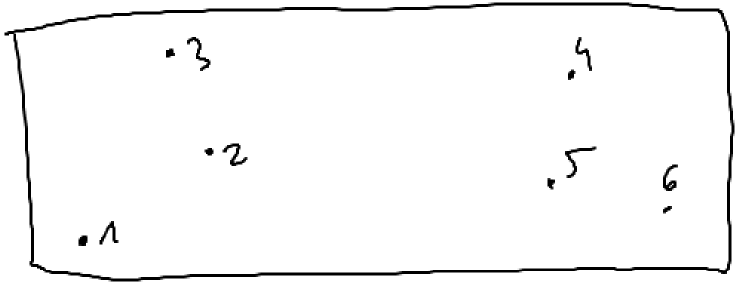
(a) dla każdego klastra znajdujemy jego centroid: $\bar{x}_k = \frac{1}{|C_k|} \sum_{i \in C_k} x_i \in \mathbb{R}^p$, $k = 1, \dots, K$

(b) przyporządkowujemy (na nowo) każde ~~x_i~~ $i \in \{1, \dots, n\}$ do takiego C_j, aby $\| x_i - \bar{x}_j \|$ było minimalne

tak długo, aż (b) nie zmieni $\sum_{j=1}^{K} W(C_j)$.

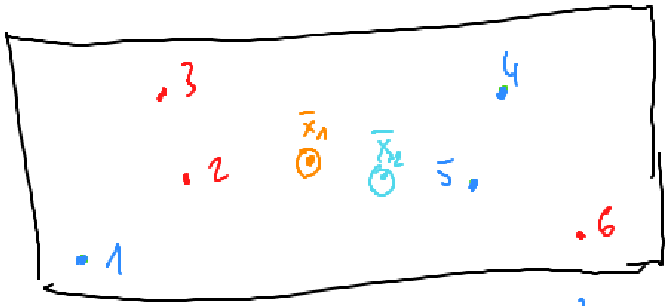
Np. 6 punktów

$K=2$

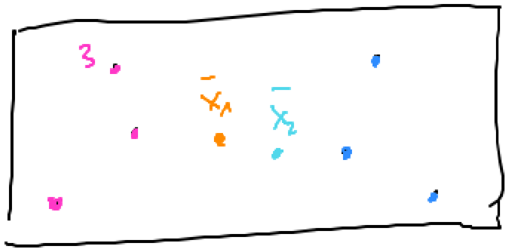


b) Jeśli x_i jest bliżej $\bar{x}_1$ niż $\bar{x}_2$, to $i \in \tilde{C}_1$, a w przeciwnym razie $i \in \tilde{C}_2$

1)



$C_1 = \{2, 3, 6\}$ $C_2 = \{1, 4, 5\}$

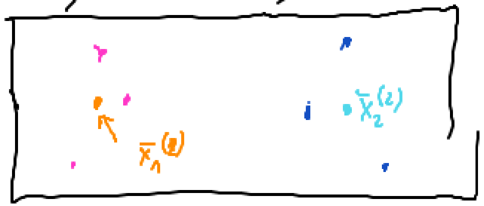


$\tilde{C}_1 = \{1, 2, 3\}$ $\tilde{C}_2 = \{4, 5, 6\}$

a) znajdujemy centroidy:

$$\bar{x}_1 = \frac{1}{3}(x_2 + x_3 + x_6) \qquad \bar{x}_2 = \frac{1}{3}(x_1 + x_4 + x_5)$$

a_2) znajdujemy centroidy



b_2) $\tilde{\tilde{C}}_1 = \tilde{C}_1$, $\tilde{\tilde{C}}_2 = \tilde{C}_2$ nic się nie zmieniło

Ostateczna postać C_j nie zależy od (losowego) wyboru w 1)

Kroki 2a–2b nie zwiększają wyrażenia $\sum_{j=1}^{K} W(C_j)$, gdzie $W(C_j)=\frac{1}{|C_j|}\sum_{i,i'\in C_j}\|x_i-x_{i'}\|^2$,

dlatego po pełnej liczbie kroków $\sum_{j=1}^{k} W(C_j)$ nie będzie się zmieniać. $\bar{x}_j=\frac{1}{|C_j|}\sum_{i\in C_j} x_i$

Równość:
$$W(C_j)=\frac{1}{|C_j|}\sum_{i,k\in C_j}\|(x_i-\bar{x}_j)+(\bar{x}_j+x_k)\|^2=$$

$$=\frac{1}{|C_j|}\sum_{i,k}\langle(x_i-\bar{x}_j)+(\bar{x}_j-x_k),(x_i-\bar{x}_j)+(\bar{x}_j-x_k)\rangle=\frac{1}{|C_j|}\Big(\underbrace{\sum_{i,k\in C_j}\|x_i-\bar{x}_j\|^2}_{=\sum_{k\in C_j}\left(\sum_i\|x_i-\bar{x}_j\|^2\right)=|C_j|\cdot\sum_i\|x_i-\bar{x}_j\|^2}+2\sum_{i,k}\langle x_i-\bar{x}_j,\bar{x}_j-x_k\rangle+\sum_{i,k}\|\bar{x}_j-x_k\|^2\Big)$$

$$=2\sum_i\|x_i-\bar{x}_j\|^2+\frac{1}{|C_j|}2\Big\langle\underbrace{\sum_i(x_i-\bar{x}_j)}_{=0},\underbrace{\sum_k(\bar{x}_j-x_k)}_{=0}\Big\rangle$$

$$W(C_j)=2\sum_i\|x_i-\bar{x}_j\|^2$$

$$\sum_i(x_i-\bar{x}_j)=\sum_i x_i-\sum_i\bar{x}_j=\sum x_i-|C_j|\cdot\frac{1}{|C_j|}\sum_i x_i=0\in\mathbb{R}^p$$

Druga obserwacja: $$\sum_{i\in C_j}\|x_i-\bar{x}_j\|^2 = \min_{a\in\mathbb{R}^p}\sum_{i\in C_j}\|x_i-a\|^2$$

<u>Uzasadnienie:</u>

$$\sum_{i\in C_j}\|x_i-\bar{x}_j+\bar{x}_j-a\|^2 = \sum_{i\in C_j}\langle \underline{(x_i-\bar{x}_j)}+(\bar{x}_j-a),\ \underline{(x_i-\bar{x}_j)}+(\bar{x}_j-a)\rangle =$$

$$= \sum_{i\in C_j}\|x_i-\bar{x}_j\|^2 + 2\underbrace{\sum_{i\in C_j}\langle x_i-\bar{x}_j,\ \bar{x}_j-a\rangle}_{=\ \langle \sum_{i\in C_j}(x_i-\bar{x}_j),\ \bar{x}_j-a\rangle\ =\ 0} + \sum_{i\in C_j}\|\bar{x}_j-a\|^2 =$$

$$= \sum_{i\in C_j}\|x_i-\bar{x}_j\|^2 + |C_j|\cdot\|\bar{x}_j-a\|^2$$

$$\sum_{j=1}^{K} W(C_j) = 2 \underbrace{\sum_{j=1}^{K} \sum_{i \in C_j}}_{} \|x_i - \bar{x}_j\|^2 = 2 \sum_{j=1}^{K} \min_{a \in \mathbb{R}^p} \sum_{i \in C_j} \|x_i - a\|^2$$

$C_1, \dots, C_K$

a) $\bar{x}_1, \dots, \bar{x}_K$

b) zmieniamy C_j na $\tilde{C}_j$ $\longrightarrow$ ~~zmienia się~~

$$2 \sum_{j=1}^{K} \sum_{i \in \tilde{C}_j} \|x_i - \bar{x}_j\|^2 \geq$$

$$\geq 2 \sum_{j=1}^{K} \sum_{i \in \tilde{C}_j} \|x_i - \tilde{\bar{x}}_j\|^2 = \sum_{j=1}^{K} W(\tilde{C}_j)$$

centroidy dla $\tilde{C}_j$

Zadanie 4

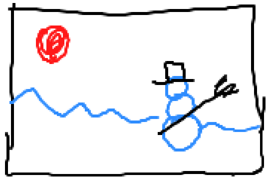
Zapoznać się z SVM lub k-means (do wyboru) do 23

1,2,3

k-means: redukcja liczby kolorów

często 2^{24} kolorów $\approx 16.7 \cdot 10^6$

każdy piksel — jedna obserwacja $\in \mathbb{R}^3$
(r,g,b)



(dobrze wybrać dane wymagające pewnej obróbki, w tym lub innym zadaniu)