Bayesowski:

$$\arg\max_k P(Y=k \mid X=x_0) \to \text{klasa odpowiadająca obserwacji } x_0$$

$\boxed{6}_{x_0}$ $\boxed{4}_{x_1}$ $\boxed{8}_{x_2}$

$P(Y=6 \mid X=x_0) = 0.99$ $\quad P(Y=4 \mid X=x_1) = 1$ $\quad P(Y=8 \mid X=x_2) = 0.9$

$P(Y=0 \mid X=x_0) = 0.01$ $\quad P(Y=6 \mid X=x_2) = 0.04$

$P(Y=0 \mid X=x_2) = 0.06$

LDA $p=1$

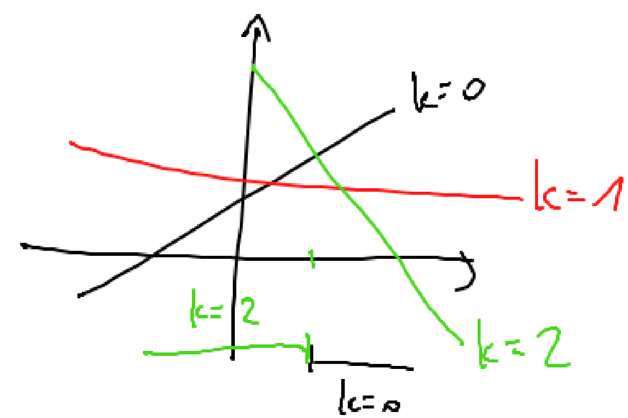
$$P(Y=k \mid X=x) = \frac{\pi_k f_k(x)}{\sum_{l=1}^{K} \pi_l f_l(x)}$$

$$\pi_k = P(Y=k) \quad \hat{\pi}_k = \frac{\#\{i : y_i = k\}}{N} =: \frac{n_k}{N}$$

($i=1,\dots,N$ ↑ obserwacje)

$$P(X \in A \mid Y=k) = \int_A f_k(x)\,dx$$

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



LDA: $\sigma_1 = \sigma_2 = \dots = \sigma_K =: \sigma$

klasyfikacja: $\arg\max_k \left(\log \pi_k + \frac{\mu_k}{\sigma^2} x - \frac{\mu_k^2}{2\sigma^2}\right)$

$$\hat{\sigma}^2 = \frac{1}{N-K} \sum_{k=1}^{K} \sum_{i : y_i = k} (x_i - \hat{\mu}_k)^2 \qquad \hat{\mu}_k = \frac{1}{n_k} \sum_{i : y_i = k} \overbrace{x_i}^{\in \mathbb{R}}$$

$\underline{p>1}$:

$$f_k(x) = \frac{1}{(2\pi)^{p/2} |\det \Sigma_k|^{1/2}} \exp\left(-\frac{1}{2}(x-\mu_k)^T \Sigma_k^{-1} (x-\mu_k)\right)$$

$$\Sigma = \operatorname{Cov} X = \left[\operatorname{Cov}(X_i, X_j)\right]_{i,j}$$
$$\operatorname{Cov}(X_i, X_j) = E\left[(X_i - EX_i)(X_j - EX_j)\right]$$

LDA: $\Sigma_1 = \Sigma_2 = \ldots = \Sigma_k =: \Sigma$

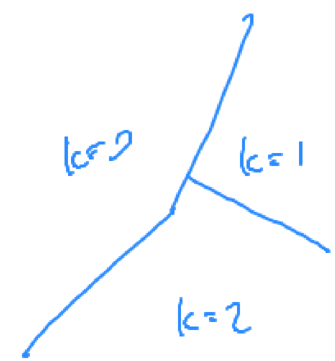
$$\hat{\mu}_k = \frac{1}{n_k} \sum_{i: y_i = k} \overbrace{x_i}^{\in \mathbb{R}^p}$$

$$\hat{\Sigma} = \frac{1}{N-K} \sum_{k=1}^{K} \sum_{i: y_i = k} (x_i - \hat{\mu}_k)(x_i - \hat{\mu}_k)^T$$

iloczyn:

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

$\operatorname{argmax}_k \delta_k(x)$ - klasa, do której przyporządkowujemy x



quadratic discriminant analysis

QDA: jak LDA, ale bez założenia, że $\Sigma_1 = \Sigma_2 = \dots = \Sigma_K$

liniowa analiza dyskryminacyjna kwadratowa

$$\hat{\Sigma}_k = \frac{1}{n_k - K} \sum_{i: y_i = k} (x_i - \hat{\mu}_k)(x_i - \hat{\mu}_k)^T$$

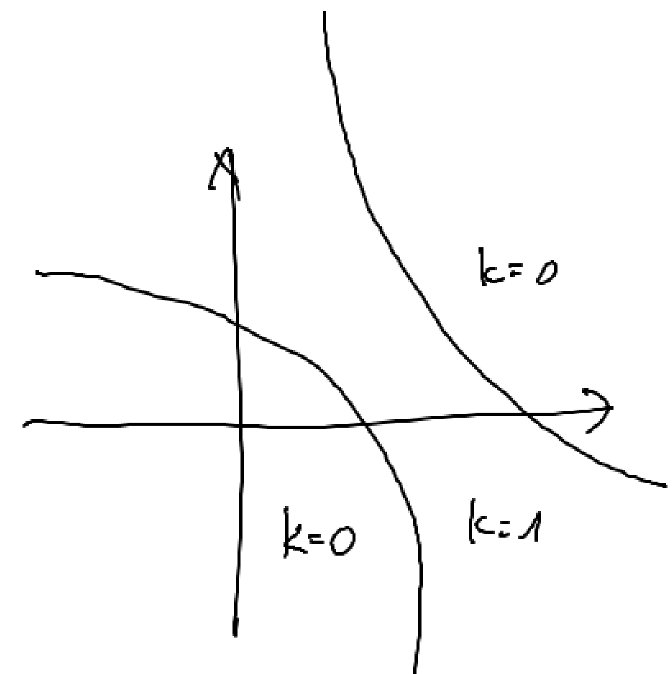
$$\boxed{\delta_k(x) = -\frac{1}{2}(x-\mu_k)^T \Sigma_k^{-1} (x-\mu_k) - \frac{1}{2}\log|\Sigma_k| + \log \pi_k}$$

$$x^T \Sigma_k^{-1} x - \mu_k^T \Sigma_k^{-1} x - x^T \Sigma_k^{-1} \mu_k + \mu_k^T \Sigma_k^{-1} \mu_k$$

$$(A+B)\cdot C = A\cdot C + B\cdot C$$

klasyfikacja: $\arg\max_k \delta_k(x)$

dla $p=2$: granice pomiędzy obszarami decyzyjnymi są (kawałkami) krzywymi stożkowymi



przykładowe obszary decyzyjne

Maximal margin classifier

Załóżmy, że $(x_i, y_i) \in \mathbb{R}^p \times \{-1, 1\}$ są takie, że istnieją $\beta_0 \in \mathbb{R}$, $\beta_1 \in \mathbb{R}^p$

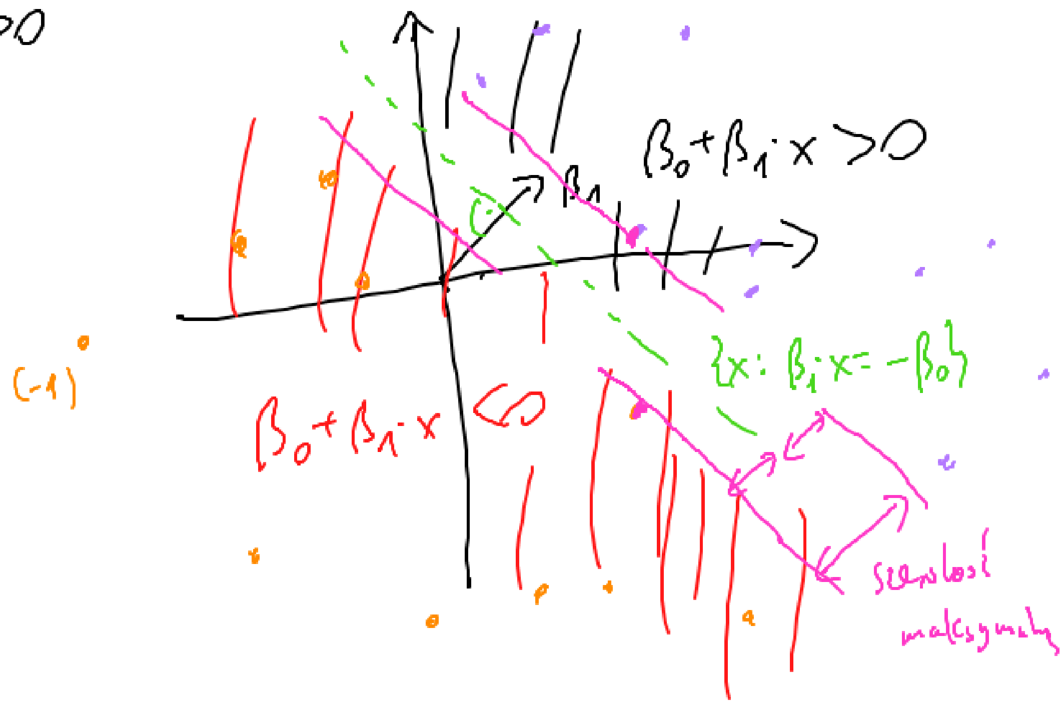
takie, że

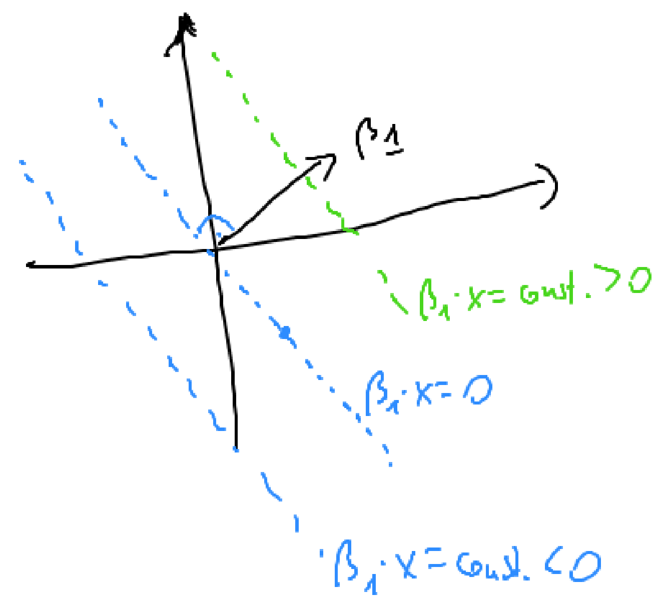
$$\Leftrightarrow \begin{cases} \beta_0 + \beta_1 x_i > 0 & \text{gdy } y_i = 1 \\ \beta_0 + \beta_1 \cdot x_i < 0 & \text{gdy } y_i = -1 \end{cases}$$

$$\beta_0 + \beta_1 x = 0$$

$$x \cdot \beta_1 = -\beta_0$$

$$(\beta_0 + \beta_1 x_i) y_i > 0$$

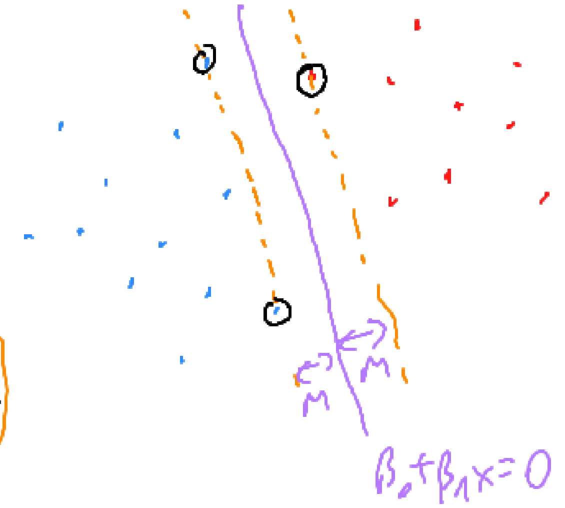




Wybieramy $\beta_0 \in \mathbb{R}, \beta_1 \in \mathbb{R}^p$ tak, aby

$$M = \min_i \underbrace{(\beta_0 + \beta_1 \cdot x_i)\, y_i}_{|\cdot|} = \text{odległość } x_i \text{ od } \{x : \beta_0 + \beta_1 \cdot x\}$$

było możliwie duże, przy ograniczeniu, że $\boxed{\|\beta_1\|_2^2 = 1.}$

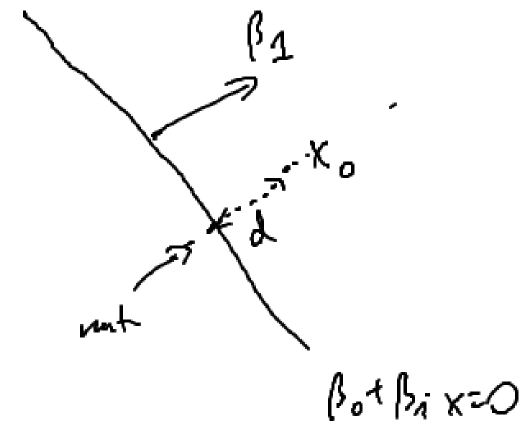


Takie optymalne M będzie odległością między hiperpłaszczyzną $\{x : \beta_0 + \beta_1 x = 0\}$

oraz zbiorami:

$\{x_i : i \text{ takie, że } y_i = 1\}$

$\{x_i : i \text{ takie, że } y_i = -1\}$



Otrzymane dla tego optymalnego M bety β_0, β_1 dają nam ostateczny klasyfikator:

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

$d = ?$

rzut: jest postaci $x_0 + \beta_1 \cdot t = x_0 - \beta_0 \beta_1 - (\beta_1 x_0) \cdot \beta_1$

oraz $\beta_0 + \beta_1 (x_0 + \beta_1 t) = 0$

$$\beta_0 + \beta_1 x_0 + \underbrace{|\beta_1|^2}_{1} \cdot t = 0$$

$$t = -\beta_0 - \beta_1 x_0$$

$$\|x_0 - \text{rzut}\| = \|\beta_0 \beta_1 + (\beta_1 x_0)\beta_1\| = |\beta_0 + \beta_1 x_0| \cdot \underbrace{\|\beta_1\|}_{1}$$