# Jądrowe klasyfikatory liniowe

Waldemar Wołyński

Wydział Matematyki i Informatyki UAM Poznań

Wisła, 9 grudnia 2009

# Zagadnienie klasyfikacyjne

Binarnym zagadnieniem klasyfikacyjnym nazywamy problem przyporządkowania obiektu opisanego przez $p$-wymiarowy wektor obserwowanych cech $\mathbf{X} = (X_1, X_2, \ldots, X_p)'$ do jednej z dwóch populacji (grup,klas) $G_0, G_1$.

O populacjach zakładamy, że mają $p$-wymiarowe rozkłady prawdopodobieństwa z nieznanymi wektorami wartości oczekiwanych $\boldsymbol{\mu}_0$ i $\boldsymbol{\mu}_1$ oraz nieznanymi macierzami kowariancji $\boldsymbol{\Sigma}_0$ i $\boldsymbol{\Sigma}_1$.

Rozwiązanie zagadnienia klasyfikacyjnego polega na podaniu reguły klasyfikacyjnej (klasyfikatora) pozwalającego na przyporządkowanie obiektu do jednej z klas:

$$d : \mathcal{X} \to \{0, 1\}.$$

# Zagadnienie klasyfikacyjne

Poszukujemy klasyfikatora liniowego postaci

$$d(\mathbf{x}) = I(\mathbf{a}'\mathbf{x} > m) = I(< \mathbf{a}, \mathbf{x} >> m),$$

spełniającego warunek

$$\mathbf{a} = \arg \max_{\mathbf{w} \in \mathbf{R}^p} J(\mathbf{w}).$$

Klasyfikator wyznaczamy przy użyciu $n$-elementowej próby uczącej $\mathcal{L}_n$, przy czym

$$\mathcal{L}_n = \mathcal{L}_0 \cup \mathcal{L}_1 = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n\},$$

gdzie

$$\mathcal{L}_0 = \{\mathbf{x}_1^0, \ldots, \mathbf{x}_{n_0}^0\}, \quad \mathcal{L}_1 = \{\mathbf{x}_1^1, \ldots, \mathbf{x}_{n_1}^1\}.$$

# Zagadnienie klasyfikacyjne

Jako estymatory nieznanych parametrów przyjmujemy

$$\hat{\boldsymbol{\mu}}_i = \bar{\mathbf{x}}_i = \frac{1}{n_i} \sum_{\mathbf{x} \in \mathcal{L}_i} \mathbf{x}, \quad i = 0, 1,$$

$$\hat{\boldsymbol{\Sigma}}_i = \mathbf{S}_i = \frac{1}{n_i - 1} \sum_{\mathbf{x} \in \mathcal{L}_i} (\mathbf{x} - \bar{\mathbf{x}}_i)(\mathbf{x} - \bar{\mathbf{x}}_i)', \quad i = 0, 1.$$

Ponadto, niech

$$\mathbf{S}_B = (\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_0)(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_0)',$$

$$\mathbf{S}_W = \frac{1}{n - 2}[(n_0 - 1)\mathbf{S}_0 + (n_1 - 1)\mathbf{S}_1].$$

# Przypadek $\mathbf{\Sigma}_0 = \mathbf{\Sigma}_1$ (Fisher,1936)

Jako miarę odległości pomiędzy grupami $G_0$ i $G_1$ przyjmujemy

$$J(\mathbf{w}) = \frac{\mathbf{w}'\mathbf{S}_B\mathbf{w}}{\mathbf{w}'\mathbf{S}_W\mathbf{w}}.$$

W wyniku maksymalizacji miary $J(\mathbf{w})$ otrzymujemy klasyfikator liniowy, dla którego

$$\mathbf{a} = \mathbf{S}_W^{-1}(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_0), \quad m = \frac{1}{2}(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_0)'\mathbf{S}_W^{-1}(\bar{\mathbf{x}}_1 + \bar{\mathbf{x}}_0).$$

# Przypadek $\mathbf{\Sigma}_0 \neq \mathbf{\Sigma}_1$ (Anderson i Bahadur,1962)

## Twierdzenie

*Przy dodatkowym założeniu normalności rozkładów populacji, dla dowolnego $\theta$ takiego, że macierz $\mathbf{\Sigma}_1 + \theta\mathbf{\Sigma}_0$ jest dodatnio określona, klasyfikator liniowy, dla którego*

$$\mathbf{a} = (\mathbf{\Sigma}_1 + \theta\mathbf{\Sigma}_0)^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0), \quad m = \mathbf{a}'\boldsymbol{\mu}_0 + \theta\mathbf{a}'\mathbf{\Sigma}_0\mathbf{a}$$

*jest dopuszczalny*

# Przypadek $\mathbf{\Sigma}_0 \neq \mathbf{\Sigma}_1$

Kryteria maksymalizujące odległości probabilistyczne (Schumway i Unger, 1974 oraz Krzyśko i Wołyński, 1997)

- odległość Chernoffa:

$$J_1(\mathbf{w}) = \frac{1}{2}s(1-s)\frac{\mathbf{w}'\mathbf{S}_B\mathbf{w}}{(1-s)\mathbf{w}'\mathbf{S}_1\mathbf{w} + s\mathbf{w}'\mathbf{S}_0\mathbf{w}} + \frac{1}{2}\ln[(1-s)\mathbf{w}'\mathbf{S}_1\mathbf{w} + s\mathbf{w}'\mathbf{S}_0\mathbf{w}] - \frac{1}{2}(1-s)\ln(\mathbf{w}'\mathbf{S}_1\mathbf{w}) - \frac{1}{2}s\ln(\mathbf{w}'\mathbf{S}_0\mathbf{w}), \quad s \in [0,1].$$

- odległość Morisity:

$$J_2(\mathbf{w}) = \frac{1}{2}\frac{\mathbf{w}'\mathbf{S}_B\mathbf{w}}{\mathbf{w}'(\mathbf{S}_1+\mathbf{S}_0)\mathbf{w}} + \frac{1}{2}\ln[\mathbf{w}'(\mathbf{S}_1+\mathbf{S}_0)\mathbf{w}] + \ln[(\mathbf{w}'\mathbf{S}_1\mathbf{w})^{\frac{1}{2}} + (\mathbf{w}'\mathbf{S}_0\mathbf{w})^{\frac{1}{2}}] - \frac{1}{2}\ln[(\mathbf{w}'\mathbf{S}_1\mathbf{w})(\mathbf{w}'\mathbf{S}_0\mathbf{w})] - \ln(2\sqrt{2}).$$

# Przypadek $\mathbf{\Sigma}_0 \neq \mathbf{\Sigma}_1$

Wszystkie rozważane miary odległości pomiędzy grupami $J(\mathbf{w})$ są funkcjami wektora $\mathbf{w}$ jedynie poprzez wyrażenia: $\mathbf{w}'\mathbf{S}_B\mathbf{w}$, $\mathbf{w}'\mathbf{S}_i\mathbf{w}$, $i = 0, 1$.

Zauważmy, że

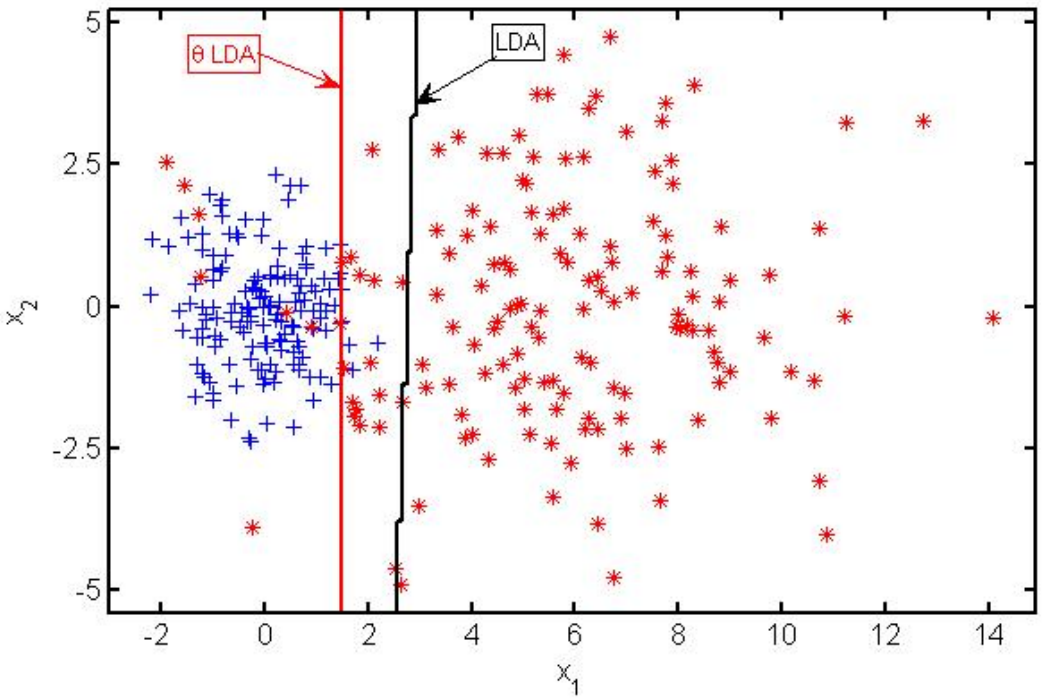
$$\mathbf{w}'\mathbf{S}_B\mathbf{w} = \left( \frac{1}{n_1} \sum_{k=1}^{n_1} < \mathbf{w}, \mathbf{x}_k^1 > - \frac{1}{n_0} \sum_{k=1}^{n_0} < \mathbf{w}, \mathbf{x}_k^0 > \right)^2$$

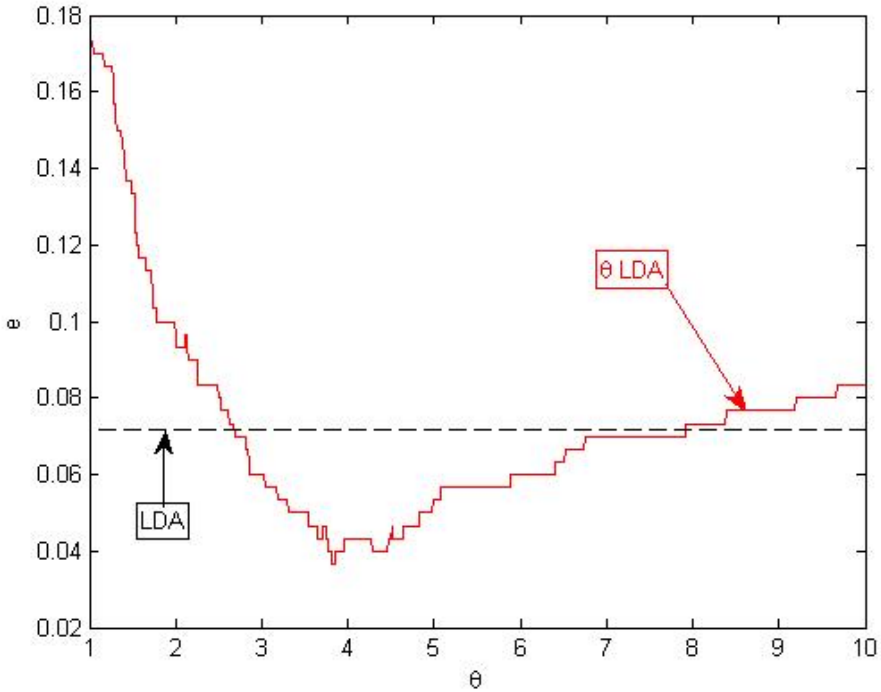
oraz

$$\mathbf{w}'\mathbf{S}_i\mathbf{w} = \frac{1}{n_i - 1} \sum_{k=1}^{n_i} \left( < \mathbf{w}, \mathbf{x}_k^i > - \frac{1}{n_i} \sum_{k=1}^{n_i} < \mathbf{w}, \mathbf{x}_k^i > \right)^2, \quad i = 0, 1.$$

# Przykład

$$\boldsymbol{\mu}_0 = \begin{bmatrix} 0 \\ 0 \end{bmatrix}, \quad \boldsymbol{\Sigma}_0 = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}, \quad \boldsymbol{\mu}_1 = \begin{bmatrix} 6 \\ 0 \end{bmatrix}, \quad \boldsymbol{\Sigma}_1 = \begin{bmatrix} 9 & 0 \\ 0 & 4 \end{bmatrix}.$$




LDA: $e_R = 0.074$, $e_{CV} = 0.086$.

$\theta$ LDA: $e_R = 0.037$ dla $\theta = 3.83$, $e_{CV} = 0.041$.

# Klasyfikatory jądrowe

Niech

$$\mathbf{x} \mapsto \Phi(\mathbf{x}) = x \in \mathcal{H}.$$

Zatem

$$\mathcal{L}_n^{\mathcal{H}} = \mathcal{L}_0^{\mathcal{H}} \cup \mathcal{L}_1^{\mathcal{H}} = \{\Phi(\mathbf{x}_1), \Phi(\mathbf{x}_2), \ldots, \Phi(\mathbf{x}_n)\}$$

oraz

$$d(\mathbf{x}) = I(< a, \Phi(\mathbf{x}) >> m),$$

gdzie

$$a = \arg \max_{w \in Lin(\mathcal{L}_n^{\mathcal{H}})} J(w).$$

# "Kernel trick"

Mamy

$$a = \sum_{j=1}^{n} \alpha_j \Phi(\mathbf{x}_j).$$

Zatem

$$< a, \Phi(\mathbf{x}) > = < \sum_{j=1}^{n} \alpha_j \Phi(\mathbf{x}_j), \Phi(\mathbf{x}) > = \sum_{j=1}^{n} \alpha_j < \Phi(\mathbf{x}_j), \Phi(\mathbf{x}) >$$

$$= \sum_{j=1}^{n} \alpha_j K(\mathbf{x}_j, \mathbf{x}),$$

gdzie $K$ jest jądrem.

Typy jąder:

- wielomianowe – $K(\mathbf{x}, \mathbf{y}) = (1 + \mathbf{x}'\mathbf{y})^c, \quad c > 0,$
- normalne – $K(\mathbf{x}, \mathbf{y}) = \exp(-\|\mathbf{x} - \mathbf{y}\|^2/c), \quad c > 0.$

# Własności

Niech

$$\bar{\mathbf{x}}_i^{\mathcal{H}} = \frac{1}{n_i} \sum_{\mathbf{x} \in \mathcal{L}_i} \Phi(\mathbf{x}), \quad \mathbf{S}_i^{\mathcal{H}} = \frac{1}{n_i - 1} \sum_{\mathbf{x} \in \mathcal{L}_i} (\Phi(\mathbf{x}) - \bar{\mathbf{x}}_i^{\mathcal{H}})(\Phi(\mathbf{x}) - \bar{\mathbf{x}}_i^{\mathcal{H}})', \quad i = 0, 1.$$

Wtedy

$$w' \mathbf{S}_B^{\mathcal{H}} w = \boldsymbol{\alpha}' \mathbf{M} \boldsymbol{\alpha},$$

gdzie

$$\mathbf{M} = (\mathbf{M}_1 - \mathbf{M}_0)(\mathbf{M}_1 - \mathbf{M}_0)', \quad (M_i)_j = \frac{1}{n_i} \sum_{k=1}^{n_i} K(\mathbf{x}_j, \mathbf{x}_k^i)$$

oraz

$$w' \mathbf{S}_i^{\mathcal{H}} w = \boldsymbol{\alpha}' \mathbf{N}_i \boldsymbol{\alpha},$$

gdzie

$$\mathbf{N}_i = \frac{1}{n_i - 1} \mathbf{K}_i (\mathbf{I} - \frac{1}{n_i} \mathbf{1}\mathbf{1}') \mathbf{K}_i', \quad (K_i)_{kl} = K(\mathbf{x}_k, \mathbf{x}_l^i), \quad i = 0, 1.$$

# Przypadek $\mathbf{\Sigma}_0 = \mathbf{\Sigma}_1$ (Mika, Rätsch, Weston, Schölkopf, Müller,1999)

Niech

$$J(\mathbf{w}) = \frac{w'\mathbf{M}w}{w'\mathbf{N}w}, \quad \mathbf{N} = \frac{1}{n-2}[(n_0 - 1)\mathbf{N}_0 + (n_1 - 1)\mathbf{N}_1].$$

Wtedy

$$a = \mathbf{N}^{-1}(\mathbf{M}_1 - \mathbf{M}_0), \quad m = \frac{1}{2}(\mathbf{M}_1 - \mathbf{M}_0)'\mathbf{N}^{-1}(\mathbf{M}_1 + \mathbf{M}_0).$$

# Przypadek $\mathbf{\Sigma}_0 \neq \mathbf{\Sigma}_1$

Dla odległości Chernoffa:

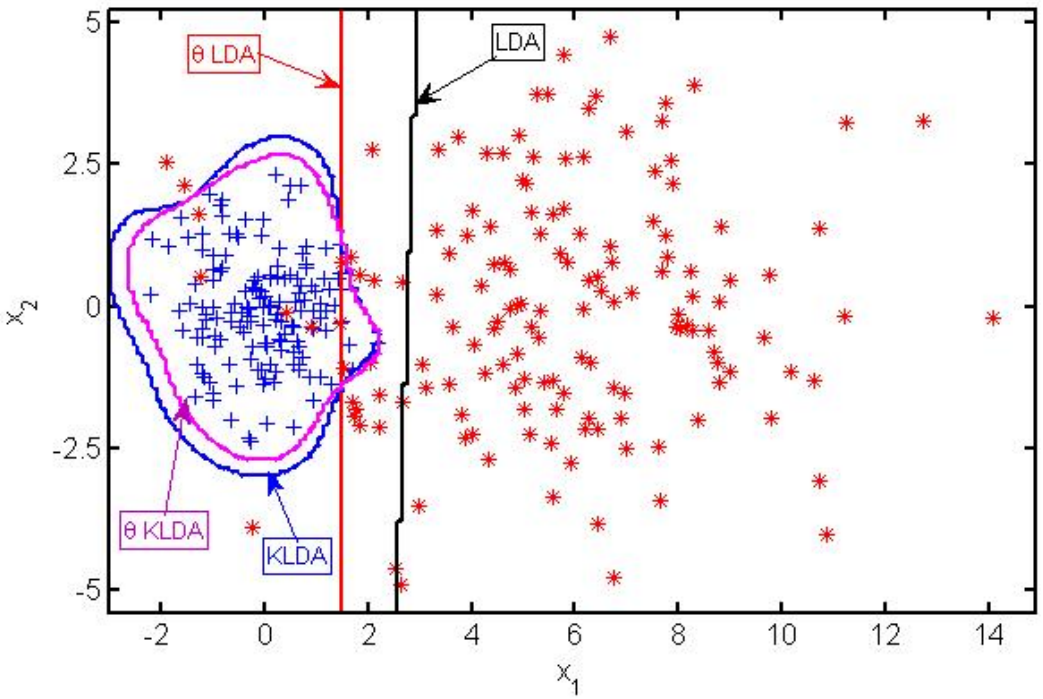
$$J_1(w) = \frac{1}{2}s(1-s)\frac{w'\mathbf{M}w}{(1-s)w'\mathbf{N}_1 w + sw'\mathbf{N}_0 w} + \frac{1}{2}\ln[(1-s)w'\mathbf{N}_1 w + sw'\mathbf{N}_0 w] - \frac{1}{2}(1-s)\ln(w'\mathbf{N}_1 w) - \frac{1}{2}s\ln(w'\mathbf{N}_0 w), \quad s \in [0,1].$$

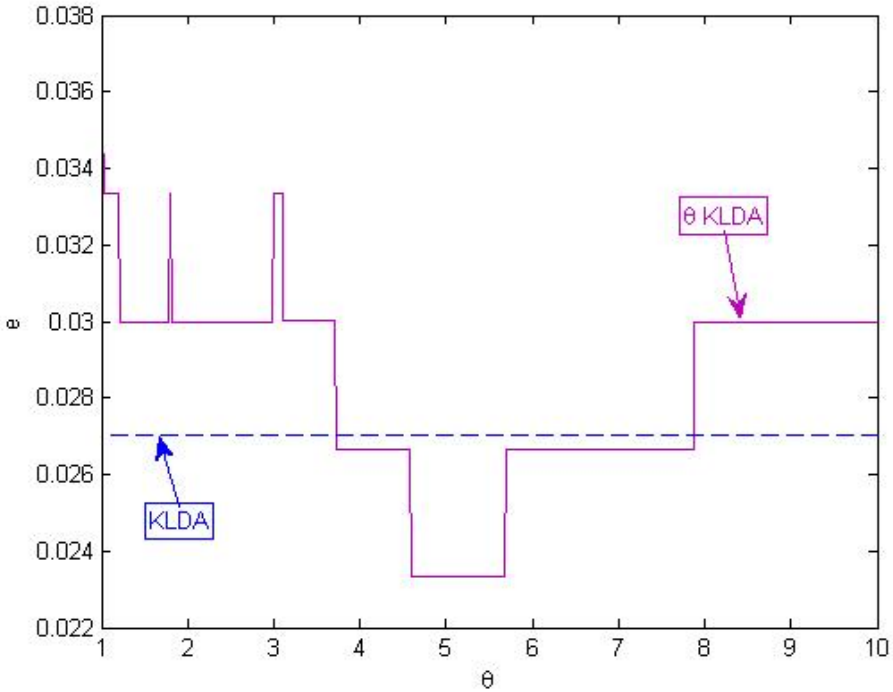
Otrzymujemy

$$a = (\mathbf{N}_1 + \theta\mathbf{N}_0)^{-1}(\mathbf{M}_1 - \mathbf{M}_0), \quad m = a'\mathbf{M}_0 + \theta a'\mathbf{N}_0 a.$$

# Przykład c.d.

$$\boldsymbol{\mu}_0 = \begin{bmatrix} 0 \\ 0 \end{bmatrix}, \quad \boldsymbol{\Sigma}_0 = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}, \quad \boldsymbol{\mu}_1 = \begin{bmatrix} 6 \\ 0 \end{bmatrix}, \quad \boldsymbol{\Sigma}_1 = \begin{bmatrix} 9 & 0 \\ 0 & 4 \end{bmatrix}.$$




KLDA: $e_R = 0.027$, $e_{CV} = 0.058$.

$\theta$ KLDA: $e_R = 0.023$ dla $\theta = 5.64$, $e_{CV} = 0.043$.

# Zagadnienie wieloklasowe

Niech liczba klas $K > 2$.

Oznaczmy

$$p_{ij}(\mathbf{x}) = \mathrm{P}(\mathbf{X} \in G_i | \ \mathbf{X} \in G_i \cup G_j, \ \mathbf{x}), \ i, j = 1, \ldots, K, \ i \neq j.$$

Przyjmujemy

$$p_{ij}(\mathbf{x}) = 1/(1 + \exp(\mathbf{a}'_{ij}\mathbf{x} - m_{ij})), \ i, j = 1, \ldots, K, \ i \neq j.$$

Ponadto

$$p_i(\mathbf{x}) = \mathrm{P}(\mathbf{X} \in G_i | \mathbf{x}), \ i = 1, \ldots, K.$$

Wtedy

$$p_{ij}(\mathbf{x}) = \frac{p_i(\mathbf{x})}{p_i(\mathbf{x}) + p_j(\mathbf{x})}, \ i < j, \ (p_{ji}(\mathbf{x}) = 1 - p_{ij}(\mathbf{x})).$$

Układ $K(K-1)/2$ równań z $K$ niewiadomymi (zazwyczaj sprzeczny).

# Metoda PWC (Hastie i Tibshirani, 1998)

Oszacowanie prawdopodobieństw a posteriori $p_i(\mathbf{x})$ poprzez minimalizację odległości Kullbacka-Leiblera postaci:

$$\rho_{KL}(\hat{p}_1(\mathbf{x}), \ldots, \hat{p}_K(\mathbf{x})) = \sum_{j \neq i} n_{ij} \left[ \hat{p}_{ij}(\mathbf{x}) \log \frac{\hat{p}_{ij}(\mathbf{x})}{p_{ij}(\mathbf{x})} \right],$$

gdzie liczebności $n_{ij} = n_i + n_j$ pełnią rolę wag.

Do wyznaczenia tych oszacowań wykorzystujemy algorytm Bradley'a-Terry'ego.

# Ważona metoda PWC

Niech $q_{ij}(\mathbf{x}) = p_i(\mathbf{x}) + p_j(\mathbf{x})$. Do oszacowania prawdopodobieństwa $q_{ij}(\mathbf{x})$ wykorzystujemy klasyfikatory binarne uczone na próbie zawierającej tylko dwie klasy. Do pierwszej z nich zaliczamy obserwacje z klas $i$-tej i $j$-tej, a do drugiej pozostałe obserwacje.

Ważona procedura sumacyjna:

$$\hat{p}_i(\mathbf{x}) = \frac{1}{K-1} \sum_{j \neq i} \hat{q}_{ij}(\mathbf{x}) \hat{p}_{ij}(\mathbf{x}), \;\; i = 1, 2, \ldots, K.$$

$$\hat{d}(\mathbf{x}) = \arg \max_{1 \leqslant i \leqslant K} \sum_{j \neq i} \hat{q}_{ij}(\mathbf{x}) \hat{p}_{ij}(\mathbf{x}).$$

Ważona procedura iloczynowa:

$$\hat{p}_i(\mathbf{x}) = \sqrt[K-1]{\prod_{j \neq i} \hat{q}_{ij}(\mathbf{x}) \hat{p}_{ij}(\mathbf{x})}, \;\; i = 1, 2, \ldots, K.$$

$$\hat{d}(\mathbf{x}) = \arg \max_{1 \leqslant i \leqslant K} \prod_{j \neq i} \hat{q}_{ij}(\mathbf{x}) \hat{p}_{ij}(\mathbf{x}).$$

# Literatura


1. Anderson T.W., Bahadur R.R. (1962), Classification into two multivariate normal distributions with different covariance matrices, *Ann. Math. Statist.* **33**, 420–431.
2. Krzyśko M., Wołyński W. (1997), Linear discriminant functions for stationary time series, *Biometrical Journal* **39**, 955-973.
3. Mika S., Rätsch G., Weston J., Schölkopf B., Müller K.R. (1999), Fisher discriminant analysis with kernels, In Y.-H. Hu, J. Larsen, E. Wilson, and S. Douglas, editors, *Neural Networks for Signal Processing IX*, 41-48.
4. T. Hastie, R. Tibshirani (1998), Classification by pairwise coupling, *The Annals of Statistics* **26**, 451-471.
5. M. Krzyśko, W. Wołyński (2009), New variants of pairwise classification, *European Journal of Operational Research* **199**, 512–519.