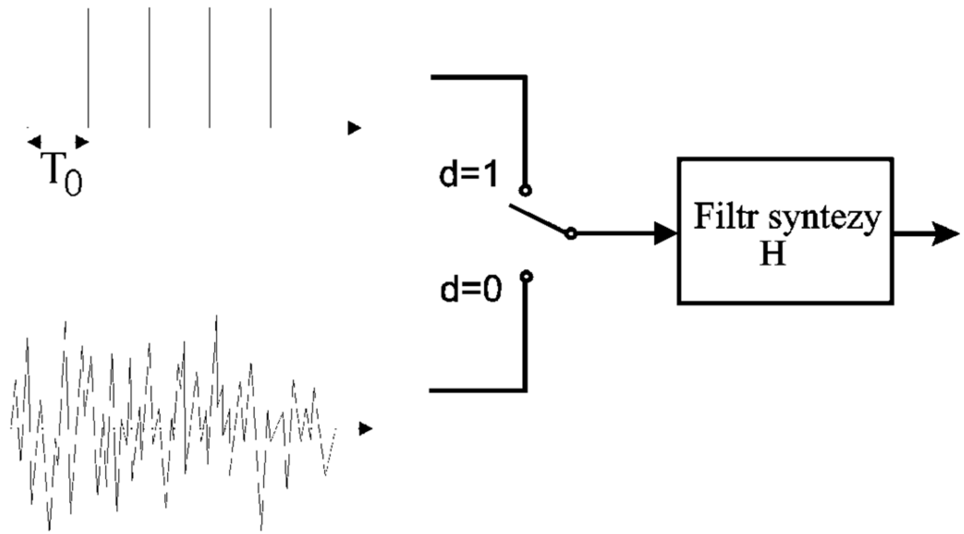
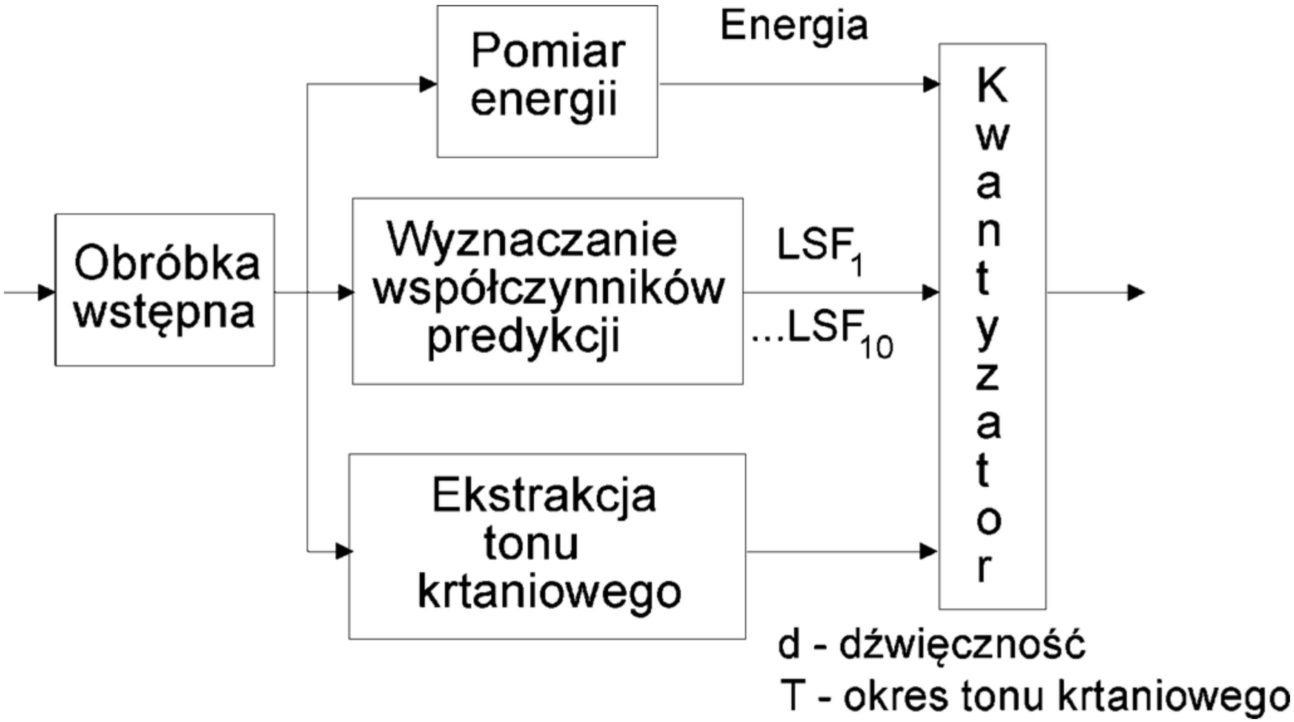
Przykład: rozpoznawanie mowy dźwięcznej i bezdźwięcznej w wokoderze



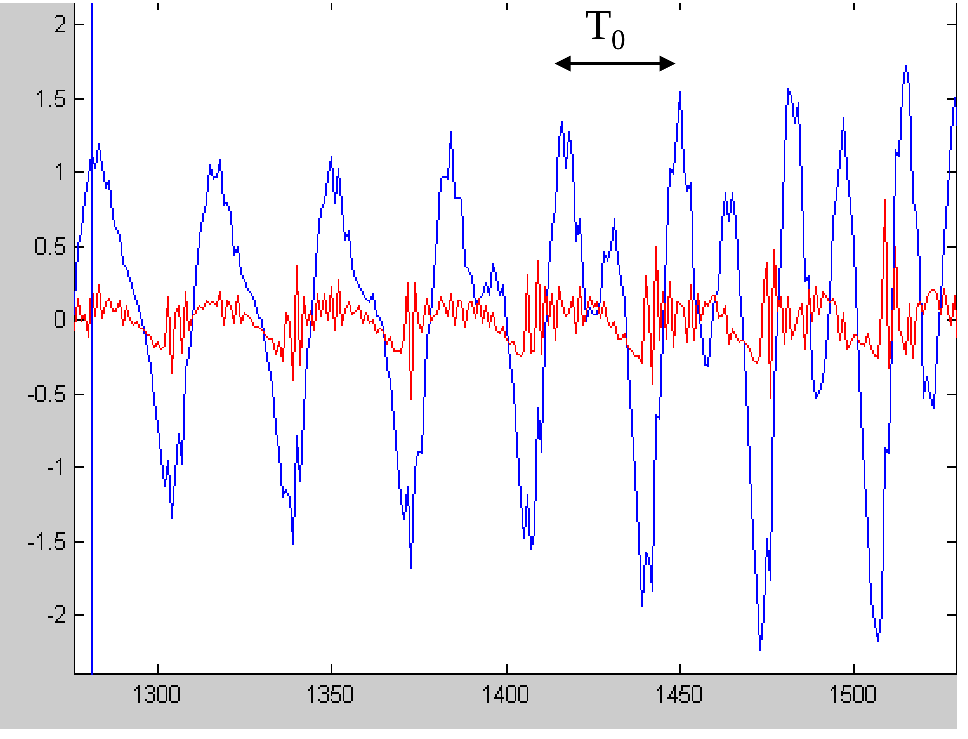
Synteza mowy w odbiorniku wokodera:
d=1 - mowa dźwięczna (T_0 = okres tonu krtaniowego)
d=0 - mowa bezdźwięczna

Wokoder – nadajnik



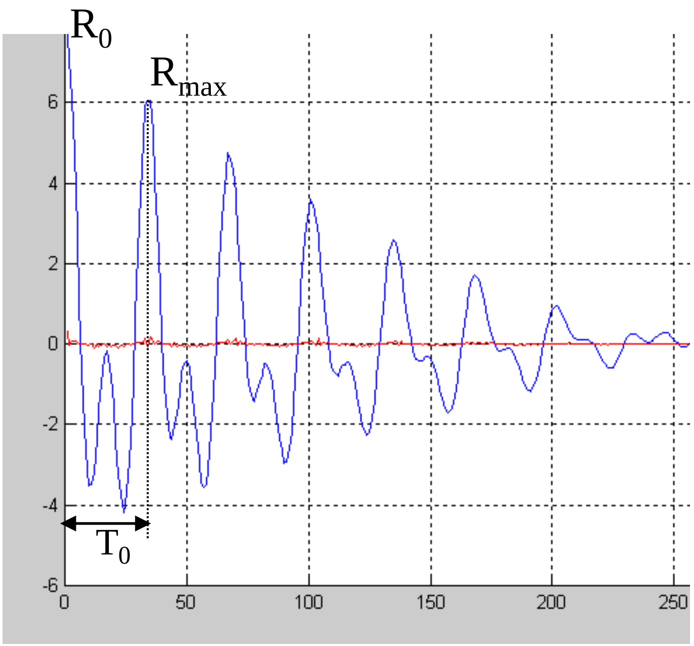
Ekstrakcja tonu krtaniowego

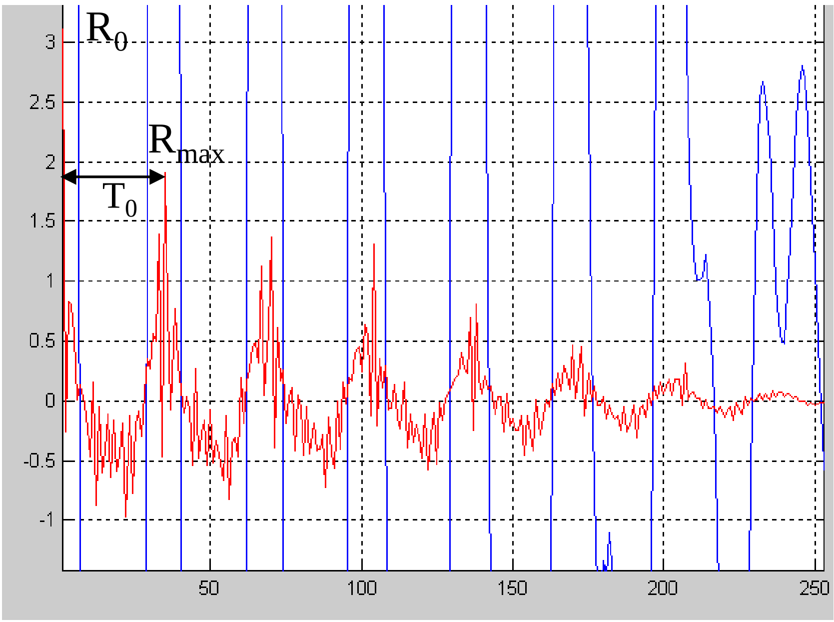
1. Przebieg czasowy sygnału i błędu predykcji



Mowa dźwięczna: T_0 = okres tonu krtaniowego

2. Autokorelacja fragmentu sygnału i błędu predykcji



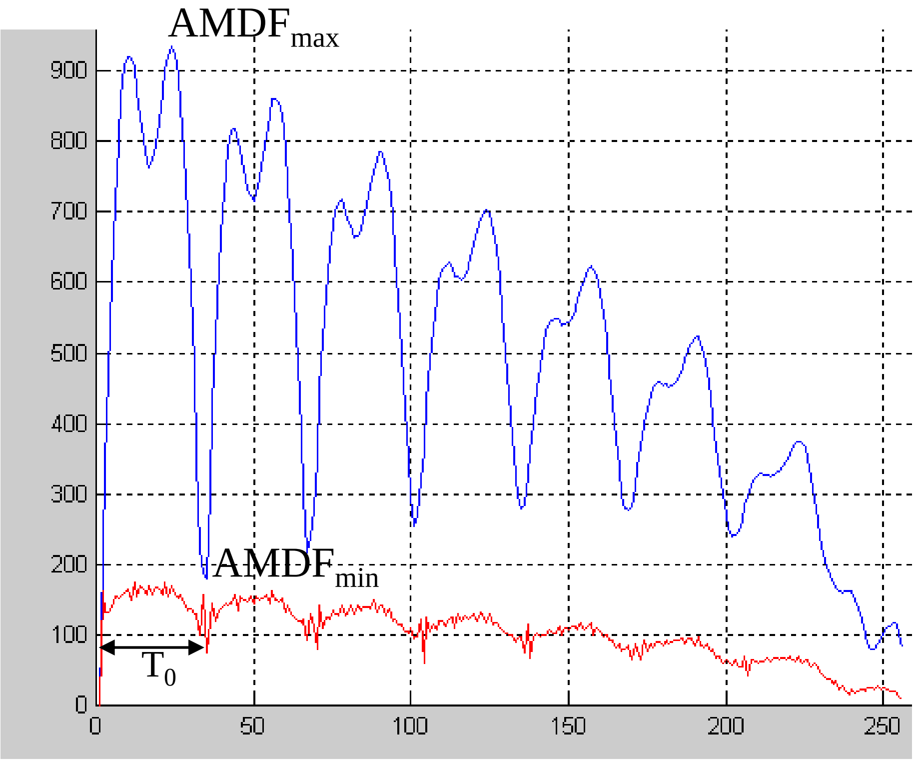


R_{max} – maksimum autokorelacji w zakresie 2 – 20 ms

gdy R_{max} / R_0 > progu -> mowa dźwięczna

3. AMDF (average magnitude difference function)

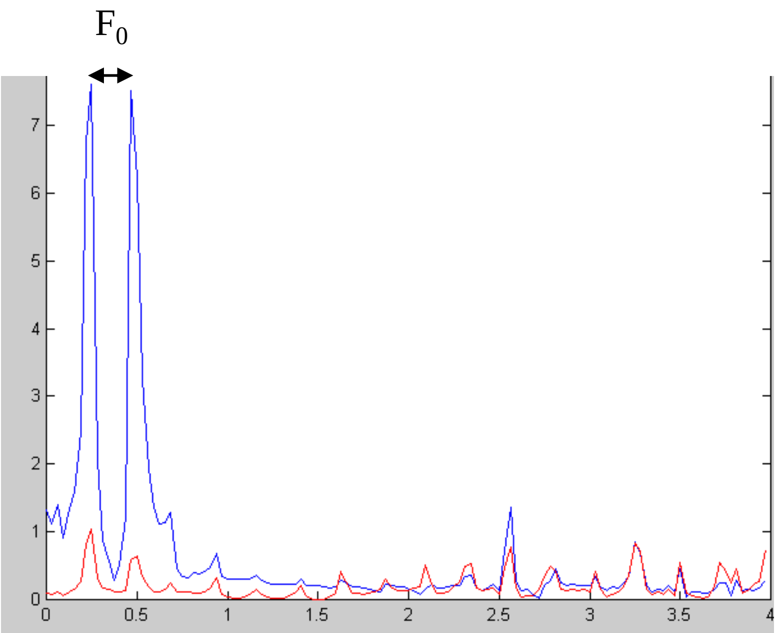
$$AMDF_k = \Sigma_i \, |x_i - x_{i-k}|$$

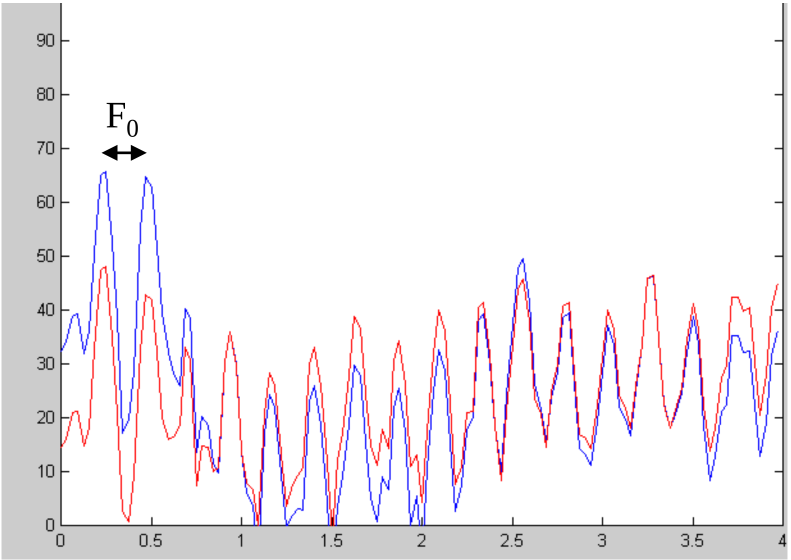


$AMDF_{min}$ – w zakresie 2 – 20 ms

gdy $AMDF_{max}$ / $AMDF_{min}$ > progu -> mowa dźwięczna

4. Widmo sygnału i błędu predykcji





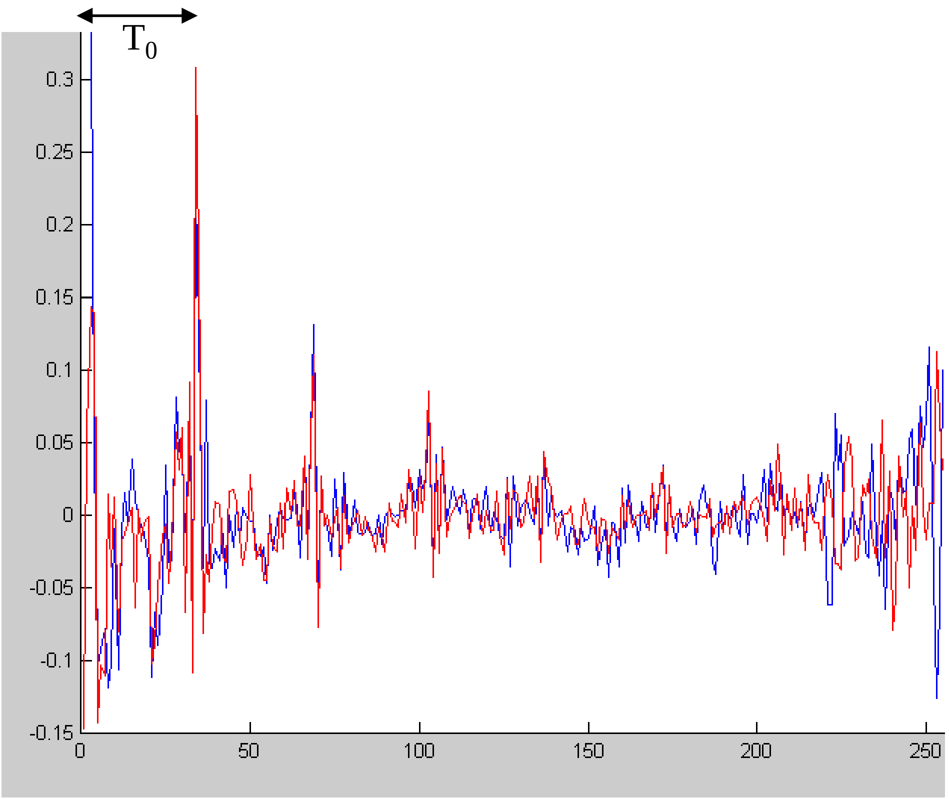
skala liniowa

skala logarytmiczna (dB)

$F_0=1/T_0$ – częstotliwość tonu krtaniowego

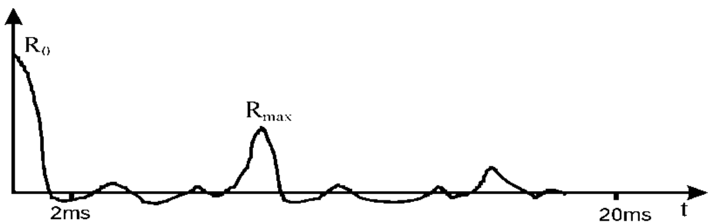
5. Cepstrum sygnału i błędu predykcji

cepstrum = IDFT (log (abs(DFT(sygnał)))

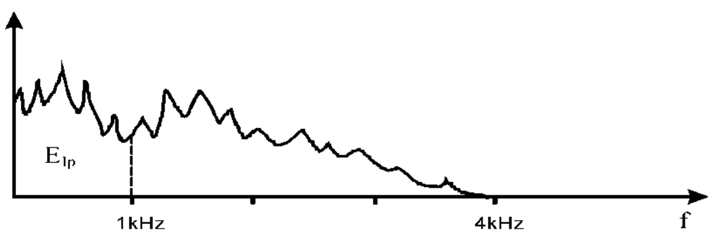


Widmo amplitudy jest funkcją parzystą (na poprzednim slajdzie pokazano jedynie jego prawostronną część), więc IDFT (a także DFT) zlogarytmowanego widma amplitudy można interpretować jako rozkład w szereg cosinusów. Z tego względu zamiast IDFT lub DFT można wykorzystać DCT.

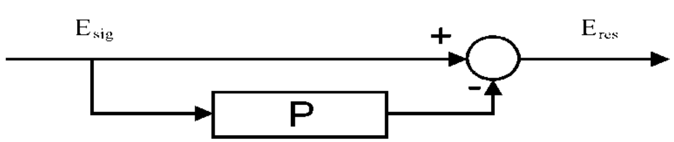
Klasyfikacja mowa dźwięczna / bezdźwięczna
Parametry



znormalizowane max. autokorelacji: $r_{max} = R_{max} / R_0$

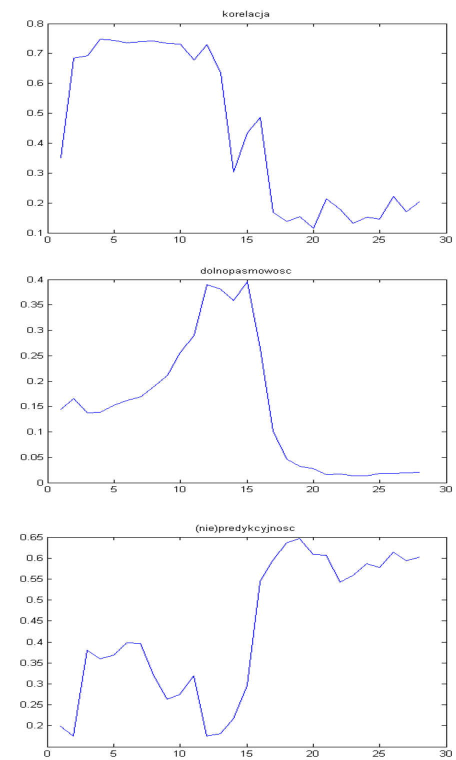


dolnopasmowość: $L_{lp} = E_{lp} / E_{sig}$

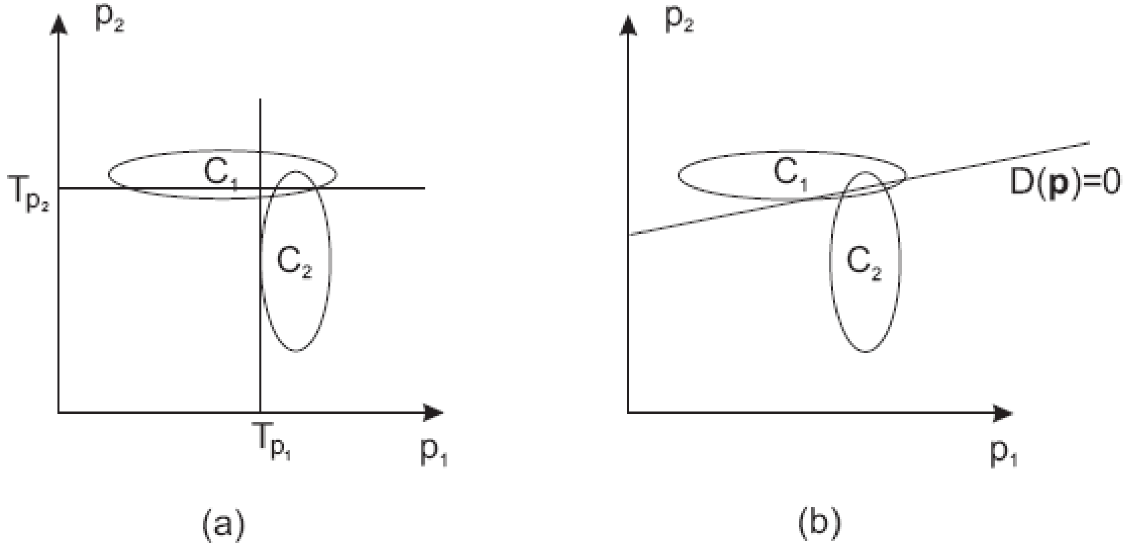


predykcyjność: $L_{res} = E_{res} / E_{sig}$

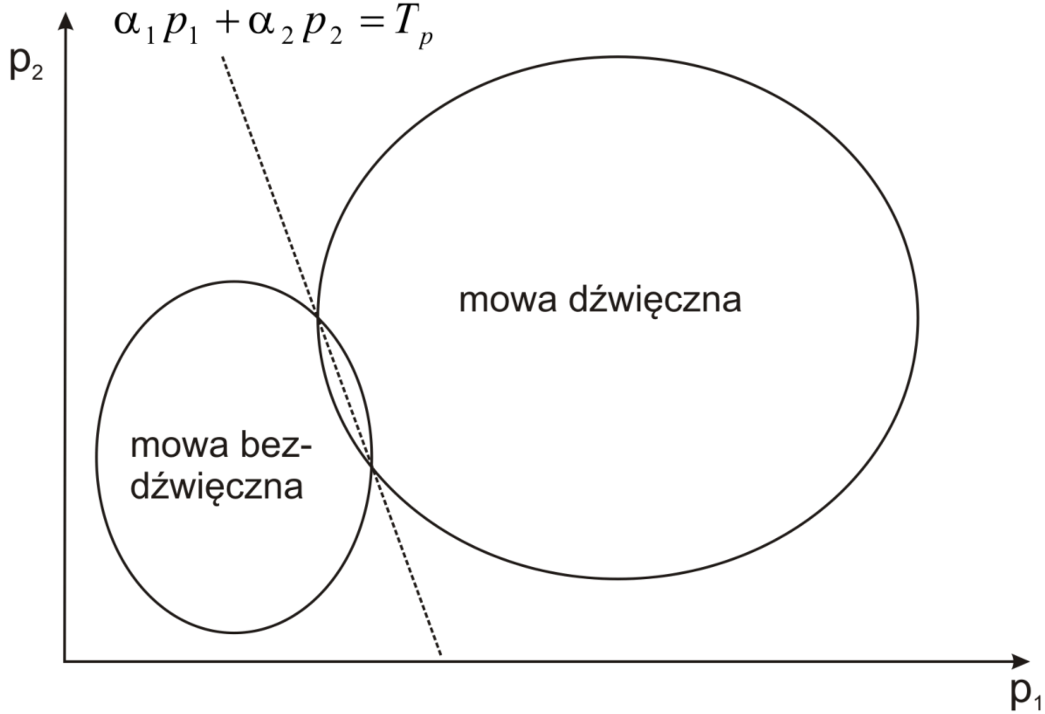
O Ś



DYSKRYMINATOR PROGOWY
i DYSKRYMINATOR LINIOWY



Dyskryminator mowa dźwięczna/bezdźwięczna wykorzystujący kilka parametrów

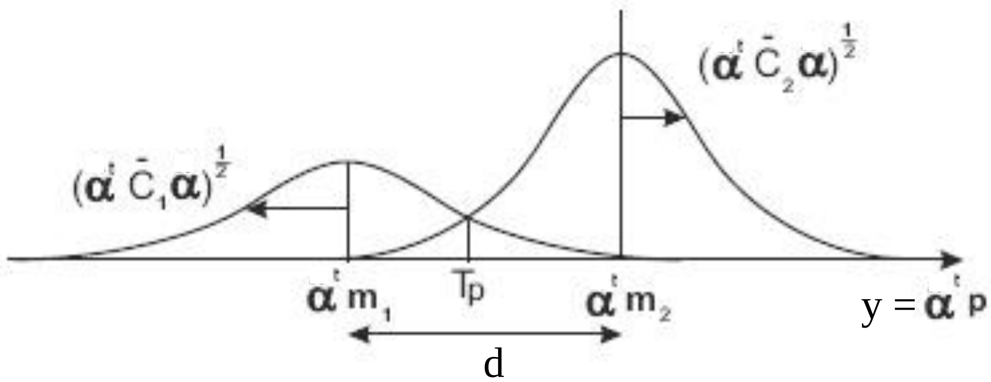


p_1 to np. $AMDF_{max}/AMDF_{min}$
p_2 to np. R_{max}/R_0

Dyskryminator liniowy:

Gdy $\alpha_1 p_1 + \alpha_2 p_2 > T_p \longrightarrow$ mowa dźwięczna

DYSKRYMINATOR LINIOWY



$$y = \boldsymbol{\alpha}^t \mathbf{p} = \sum_{i=1}^{N} \alpha_i \, p_i$$ - funkcja dyskryminująca, T_p - próg

$$\mathbf{p} = [\, p_1, p_2, \ldots, p_N \,]^t$$ - wektor parametrów

MET. FISHERA OBLICZANIA DYSKRYMINATORA LINIOWEGO

Dla sekwencji treningowej oblicza się dla obu klas:

wartości średnie wektora parametrów

$$\mathbf{m}_1 = \frac{1}{n_1} \sum_{\mathbf{p}^j \in C_1} \mathbf{p}^j$$

$$\mathbf{m}_2 = \frac{1}{n_2} \sum_{\mathbf{p}^j \in C_2} \mathbf{p}^j$$

macierze kowariancji

$$\bar{\mathbf{C}}_1 = \frac{1}{n_1} \sum_{\mathbf{p}^j \in C_1} \left(\mathbf{p}^j - \mathbf{m}_1\right) \left(\mathbf{p}^j - \mathbf{m}_1\right)^t$$

$$\bar{\mathbf{C}}_2 = \frac{1}{n_2} \sum_{\mathbf{p}^j \in C_2} \left(\mathbf{p}^j - \mathbf{m}_2\right) \left(\mathbf{p}^j - \mathbf{m}_2\right)^t$$

Dyskryminator Fishera, c.d.

$$\overline{C} = \tfrac{1}{2}[\overline{C}_1 + \overline{C}_2]$$

$$\alpha = \overline{C}^{-1}[m_2 - m_1]$$

Próg T_p dla funkcji dyskryminującej wybiera się eksperymentalnie między $\alpha^t m_1$ i $\alpha^t m_2$.

Dyskryminator Fishera, wyprowadzenie

Wartość funkcji dyskryminującej $y = \alpha^t p = \sum_{i=1}^{N} \alpha_i p_i$ jest zmienną losową:

$$E(y) = E[\alpha^t p] = \alpha^t E(p) = \alpha^t m$$

$$\sigma^2 = E(y - \alpha^t m)^2 = E[(\alpha^t p - \alpha^t m)^2] = E[\alpha^t (p - m)]^2 =$$

$$= E[\alpha^t (p-m)(p-m)^t \alpha] = \alpha^t E[(p-m)(p-m)^t]\alpha = \alpha^t C \alpha$$

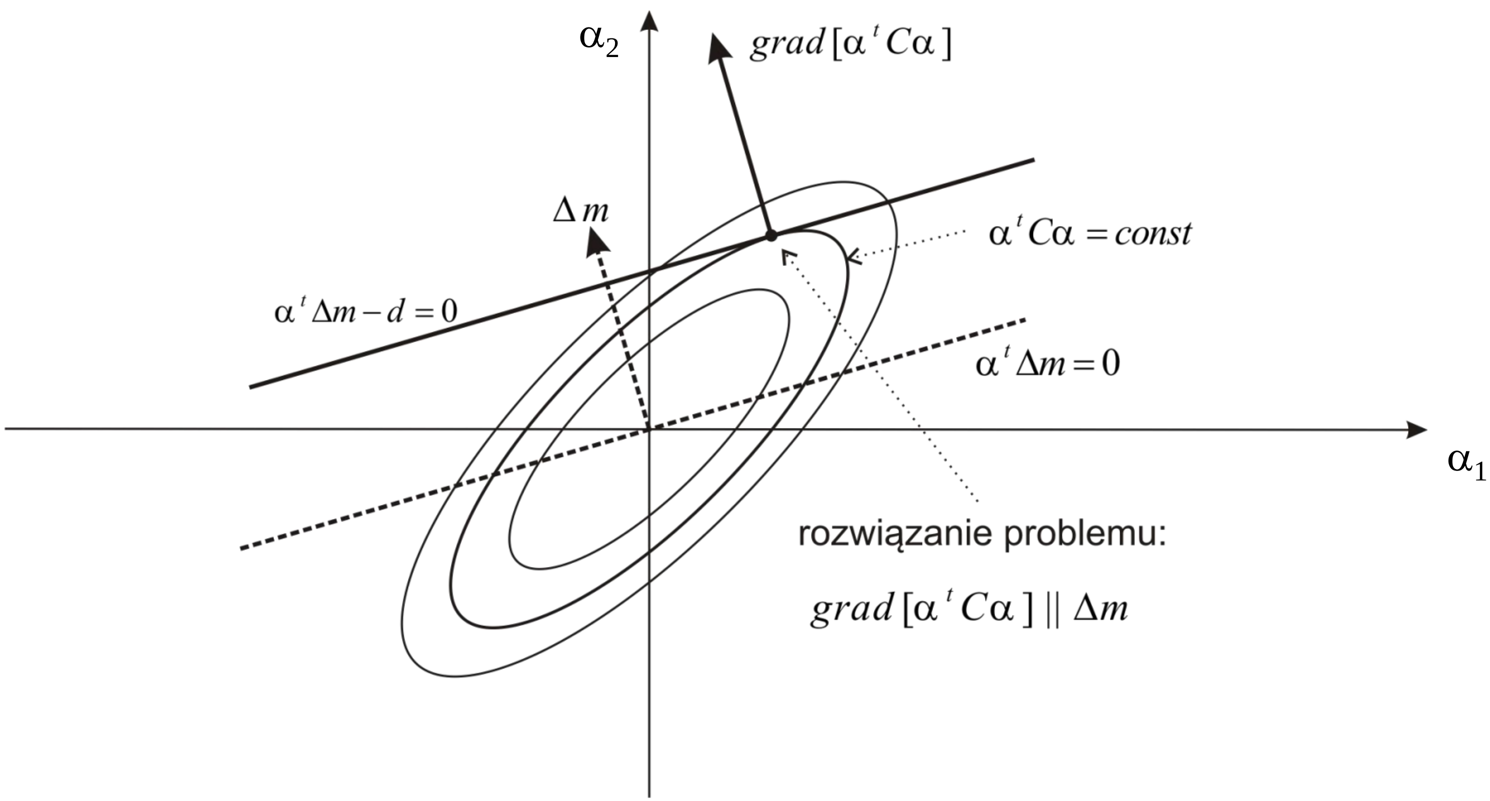
Zadanie separacji 2 klas: min wariancja y przy stałym odstępie wartości średnich

- zakładamy równość wariancji w obu klasach lub podstawiamy $C = \frac{1}{2}[C_1 + C_2]$

$$\boxed{\min \sigma^2 = \min \alpha^t C \alpha, \qquad \alpha^t m_2 - \alpha^t m_1 = \alpha^t \Delta m = d = const}$$

Dyskryminator Fishera, wyprowadzenie

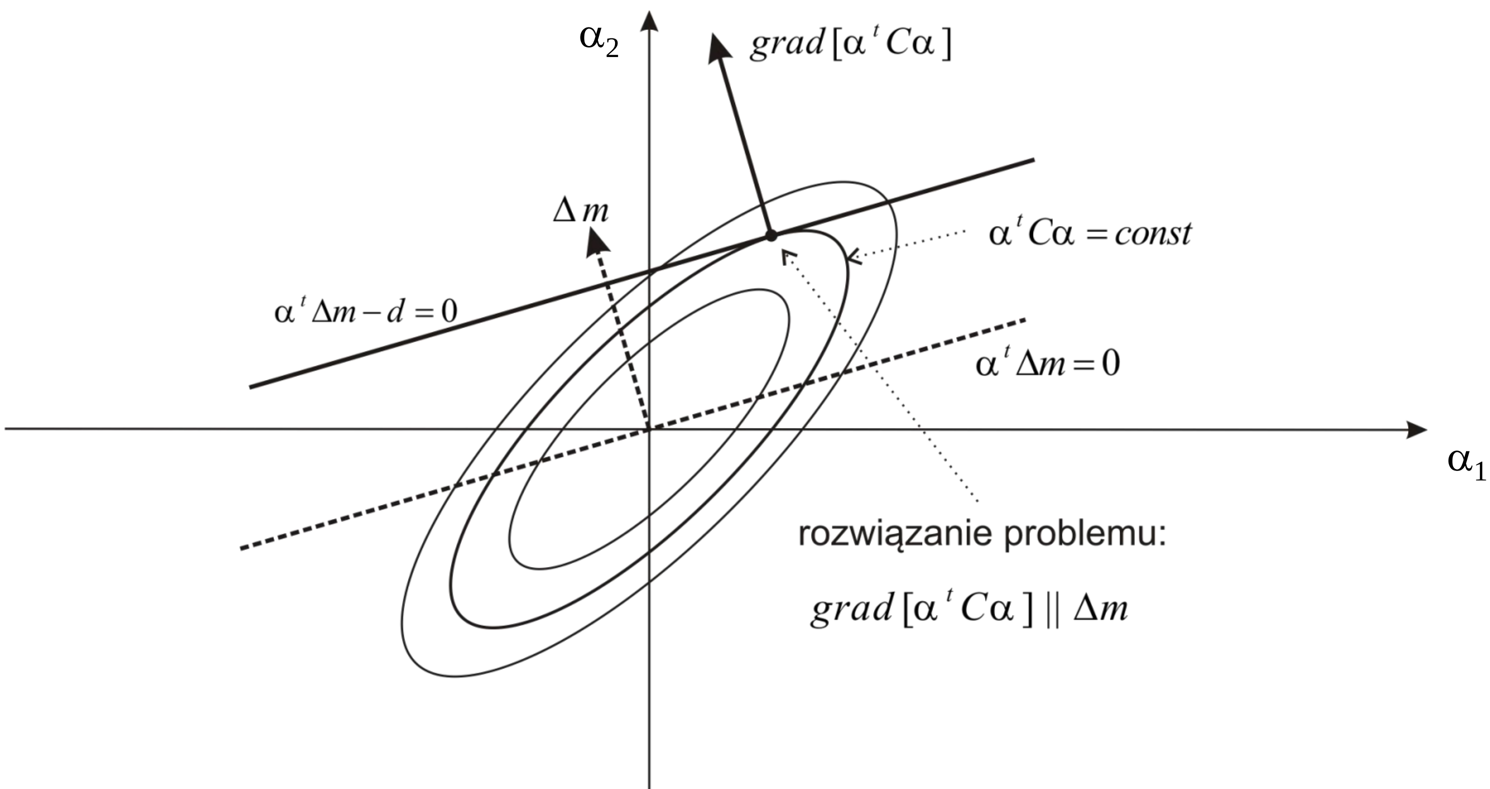
$$\min \sigma^2 = \min \alpha^t C\alpha, \qquad \alpha^t m_2 - \alpha^t m_1 = \alpha^t \Delta m = d = const$$



$$grad\, \alpha^t C\alpha = \frac{\partial}{\partial \alpha}\alpha^t C\alpha = 2C\alpha \quad \Longrightarrow \quad 2C\alpha = \lambda \Delta m$$

Dyskryminator Fishera, wyprowadzenie

$$\boxed{\min \sigma^2 = \min \alpha^t C\alpha, \qquad \alpha^t m_2 - \alpha^t m_1 = \alpha^t \Delta m = d = const}$$



$$2C\alpha = \lambda \Delta m \Longrightarrow \alpha = C^{-1}\tfrac{\lambda}{2}\Delta m, \qquad \Delta m^t \alpha = \Delta m^t C^{-1}\tfrac{\lambda}{2}\Delta m = d$$

d możemy wybrać dowolnie, np. tak, aby λ=2. Wówczas α obliczamy z $\boxed{C\alpha = \Delta m}$

Metoda mnożników Lagrange'a

Wykorzystaliśmy szczególny przypadek zastosowania metody mnożników Lagrange'a:

Szukamy minimum $F(x), \quad x = [x_1, ..., x_N]$

przy ograniczeniach $g_1(x) = 0, \quad g_2(x) = 0, ..., \ g_M(x) = 0$

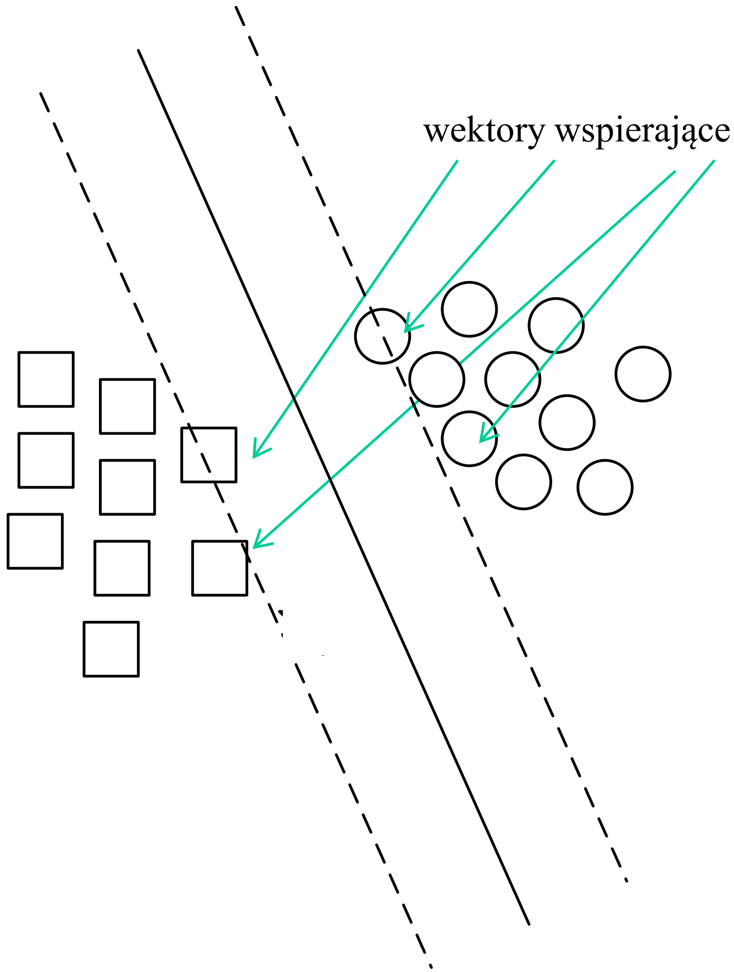
Tworzymy funkcję Lagrange'a $L(x,\lambda) = F(x) + \lambda^t g(x) = F(x) + \lambda_1 g_1(x) + \ldots + \lambda_M g_M(x),$

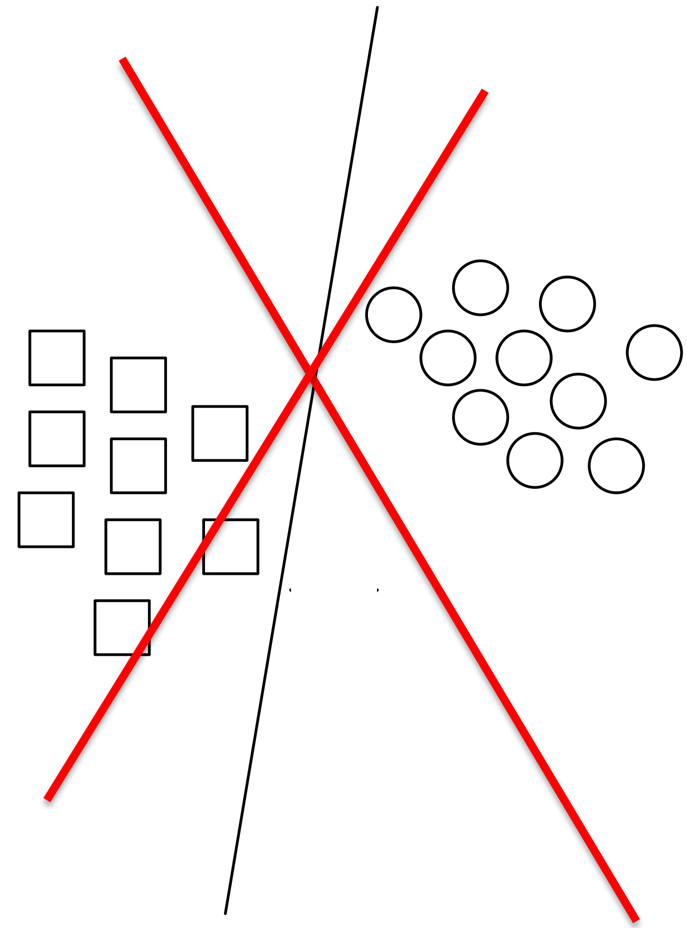
Rozwiązanie:

$$\frac{\partial}{\partial x} L(x,\lambda) = 0$$

$$\frac{\partial}{\partial \lambda} L(x,\lambda) = 0$$

SVM* – wektory wspierające

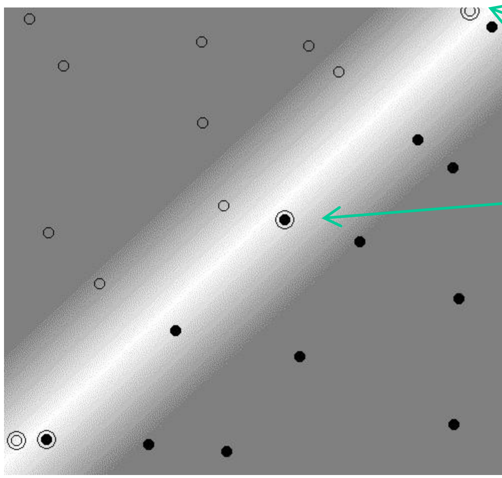




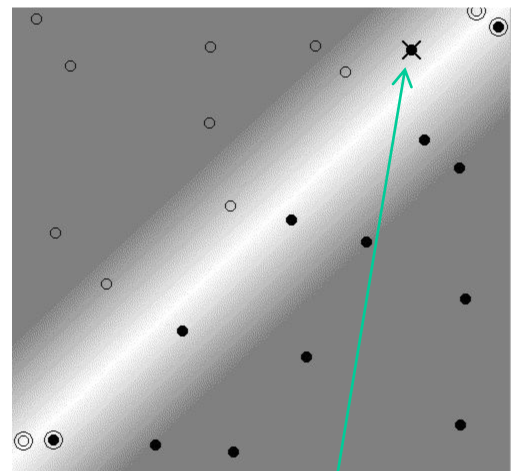
Cel: powiększenie odstępu (margin)

*Support Vector Machines

SVM – wektory wspierające

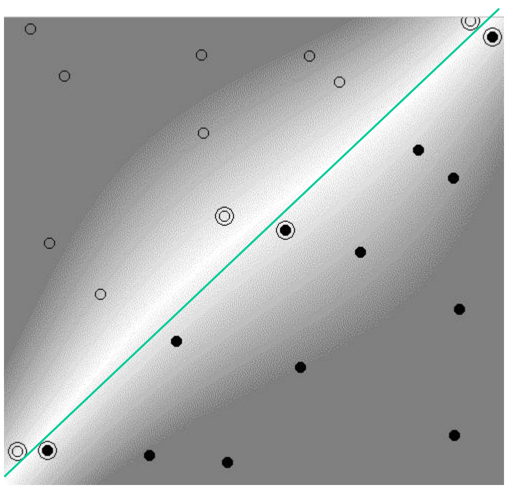


wektory wspierające

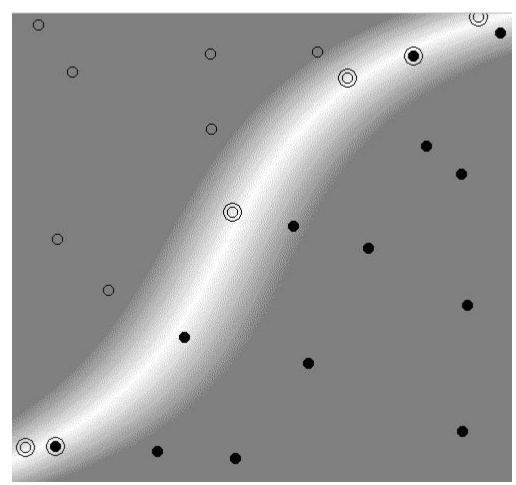


Cel: powiększenie odstępu (margin)

Klasy liniowo nieseparowalne



nieliniowe
przekształcenie



Obrazki: CHRISTOPHER J.C. BURGES

O KLASYFIKATORACH NIECO OGÓLNIEJ

Załóżmy że znane są prawdopodobieństwa (lub gęstości prawdopodobieństwa) cech (wektorów parametrów **x**, na poprzednich slajdach oznaczany jako **p**) dla każdej z klas:

$$f_k(\mathbf{x}), \quad k = 1, \ldots, K$$

oraz prawdopodobieństwa a priori pojawienia się każdej z klas: $P_1, P_2, \ldots, P_K$.

Błędnemu uznaniu *k*-tej klasy za *j*-tą przypisujemy ryzyko $r(j|k)$,
np. $r(j|k) = 0$ gdy $j=k$, oraz $r(j|k)=1$ w pozostałych przypadkach.

Dla określonego klasyfikatora można obliczyć prawdopodobieństwa błędnych decyzji: $p(j|k)$ i oszacować średnie ryzyko:

$$v = \sum_k \sum_j P_k \; p(j\,|\,k)\, r(j\,|\,k)$$

Klasyfikator projektuje się na minimum v.

KLASYFIKATOR BAYESA

Jeśli każdej błędnej decyzji jest przypisane jednakowe ryzyko
(np. $r(j|k) = 0$ gdy $j=k$, oraz $r(j|k)=1$ w pozostałych przypadkach),
wówczas klasyfikator, na podstawie obserwacji wektora parametrów **x**, winien wskazać najbardziej prawdopodobną klasę:

$$k = \arg\max_{i} P(i \,|\, \mathbf{x})$$

Ze wzoru Bayesa:

$$P(i \,|\, \mathbf{x}) = \frac{P_i \ f_i(\mathbf{x})}{\sum_{j=1}^{K} P_j \ f_j(\mathbf{x})}$$

otrzymuje się następujący klasyfikator:

$$k = \arg\max_{i} P_i \ f_i(\mathbf{x})$$

Kiedy dyskryminator liniowy jest klasyfikatorem Bayesa?

Założenia:

N-wymiarowe rozkłady gaussowskie $f_i(\mathbf{x}) = \frac{1}{(2\pi)^{N/2} |\mathbf{C}_i|^{1/2}} \exp[-\frac{1}{2}(\mathbf{x}-\mathbf{m}_i)^t \mathbf{C}_i^{-1}(\mathbf{x}-\mathbf{m}_i)]$

Te same macierze kowariancji dla wszystkich klas $\mathbf{C}_i = \mathbf{C}, \quad \forall i$

Optymalny klasyfikator jest (przy w/w założeniach) liniowy:

$$k = \arg\max_i P_i f_i(\mathbf{x}) = \arg\max_i \ln(P_i f_i(\mathbf{x})) =$$

$$= \arg\max_i \left(-\ln((2\pi)^{N/2} |\mathbf{C}|^{1/2}) - \frac{1}{2}(\mathbf{x}-\mathbf{m}_i)^t \mathbf{C}^{-1}(\mathbf{x}-\mathbf{m}_i) + \ln(P_i) \right) =$$

$$= \arg\max_i \left(-\frac{1}{2}(\mathbf{x}-\mathbf{m}_i)^t \mathbf{C}^{-1}(\mathbf{x}-\mathbf{m}_i) + \ln(P_i) \right)$$

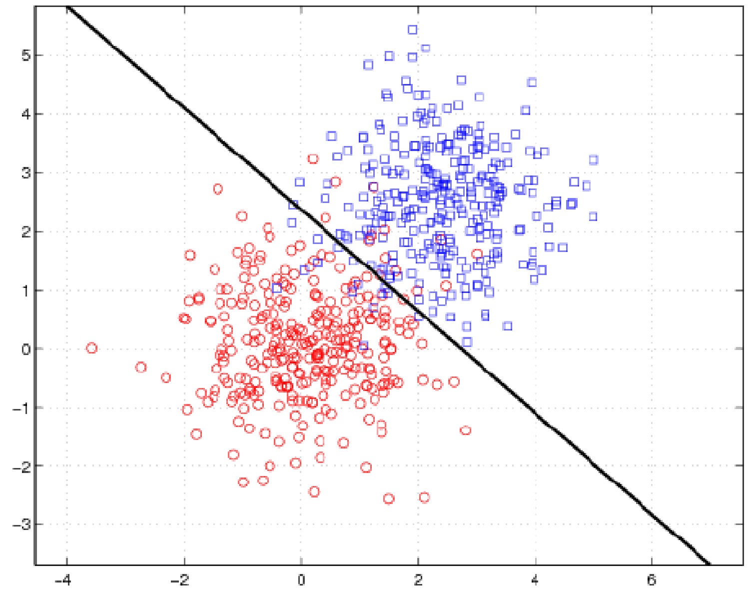
Kiedy dyskryminator liniowy jest klasyfikatorem Bayesa?

Funkcja dyskryminująca

$$\delta_i(\mathbf{x}) = -\frac{1}{2}(\mathbf{x}-\mathbf{m}_i)^t\,\mathbf{C}^{-1}(\mathbf{x}-\mathbf{m}_i) + \ln(P_i) =$$

$$= \underbrace{\mathbf{x}^t\,\mathbf{C}^{-1}\mathbf{m}_i - \frac{1}{2}\mathbf{m}_i^t\,\mathbf{C}^{-1}\mathbf{m}_i}_{\text{funkcja liniowa}} - \frac{1}{2}\underbrace{\mathbf{x}^t\,\mathbf{C}^{-1}\,\mathbf{x}}_{\substack{\text{można pominąć,}\\ \text{bo nie zależy id klasy i}}} + \ln(P_i)$$

Granica między klasą k i j:



$$\mathbf{x}: \quad \delta_k(\mathbf{x}) = \delta_j(\mathbf{x})$$

Dyskryminator kwadratowy

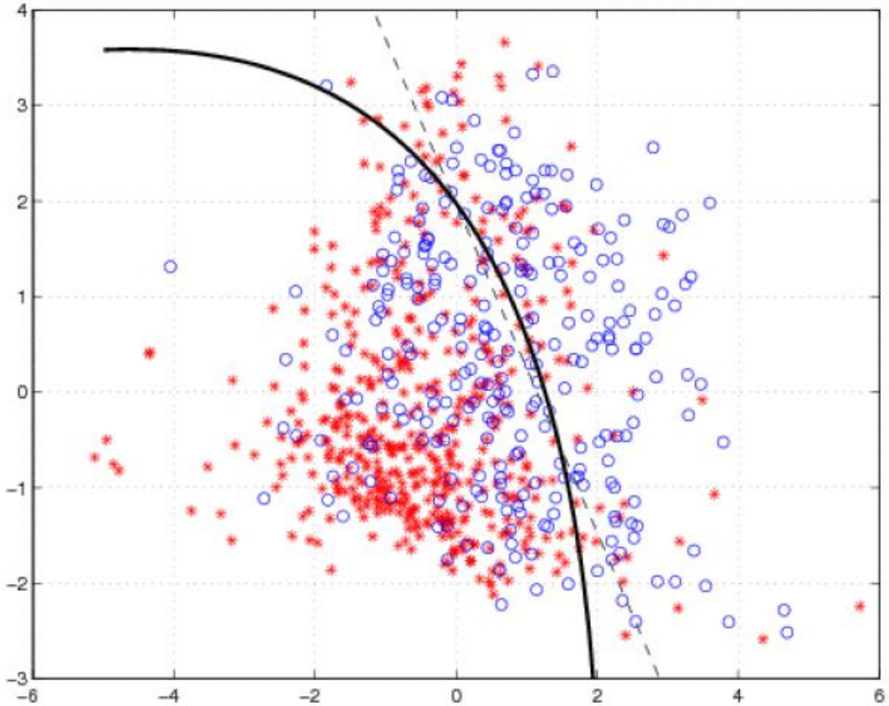
W każdej klasie inna funkcja kowariancji $\mathbf{C}_i$

Funkcja dyskryminująca nie jest liniowa (jest kwadratowa)

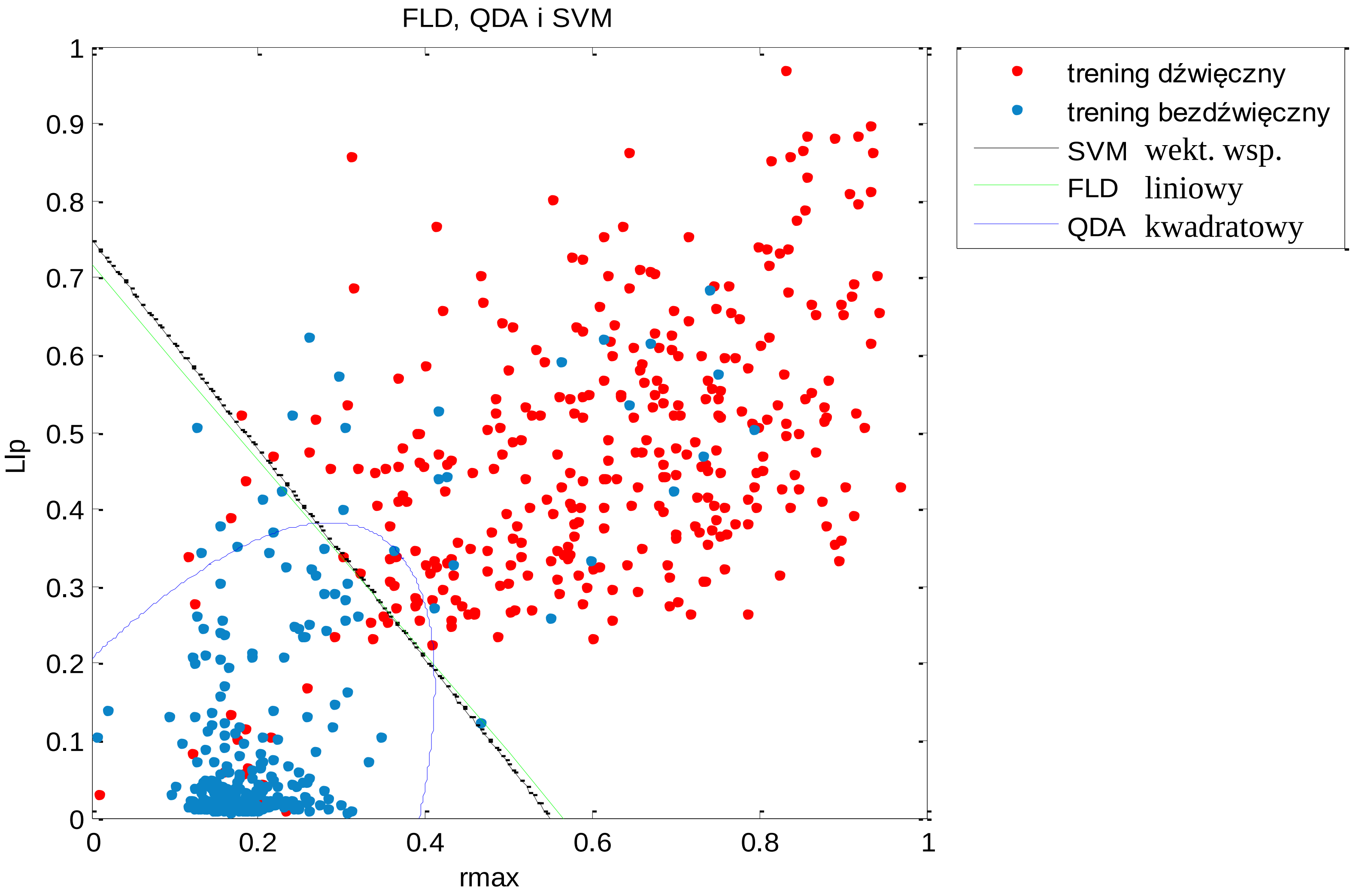
$$\delta_i(\mathbf{x}) = -\ln(|\mathbf{C}_i|^{1/2}) - \frac{1}{2}(\mathbf{x}-\mathbf{m}_i)^t \mathbf{C}_i^{-1}(\mathbf{x}-\mathbf{m}_i) + \ln(P_i)$$

Klasyfikacja:

$$k = \arg\max_i \delta_i(\mathbf{x})$$



Dyskryminatory - porównanie



Błąd w bazie i w teście

$$R(\alpha) \leq R_{emp}(\alpha) + \sqrt{\left(\frac{h(\log(2l/h)+1) - \log(\eta/4)}{l}\right)}$$

Błąd w teście

Błąd w bazie

parametry obliczone w bazie

wielkość bazy

1-η: prawdopodobieństwo, że nierówność jest spełniona

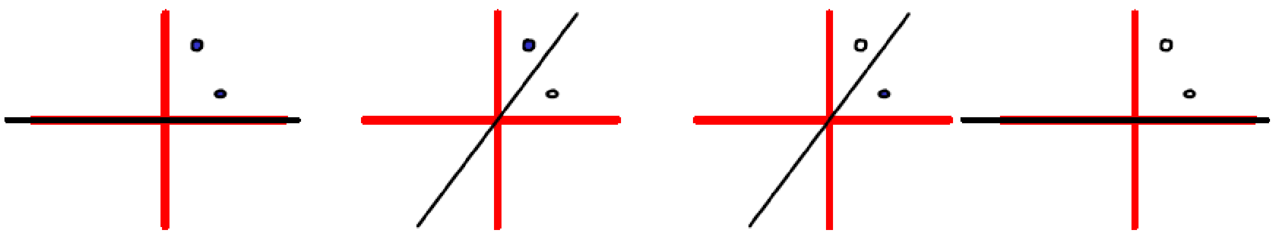
h – **VC dimension** (Vapnik, Czerwonenkis)

VC dim jest miarą złożoności dyskryminatora, ma związek z liczbą parametrów

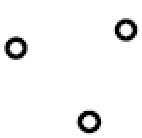
Jak obliczyć VC – dimension?

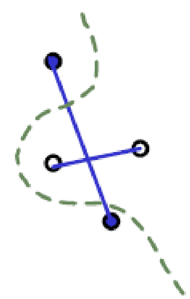
Jest to największa liczba obserwacji, którą klasyfikator jest zdolny bezbłędnie porozdzielać

Np. klasyfikator liniowy na płaszczyźnie (2 parametry) może rozdzielić 2 obserwacje,



może też rozdzielić 3 obserwacje, ale 4 już nie





Wniosek: VCdim=3

Rysunki z prezentacji Andrew W. Moore

Jak obliczyć VC – dimension?

Klasyfikator liniowy w przestrzeni N- wymiarowej może rozdzielić N+1 obserwacji

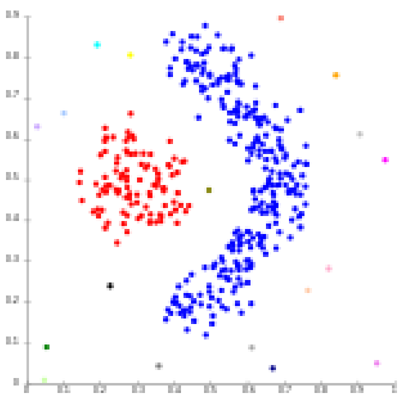
Wniosek: VCdim=N+1

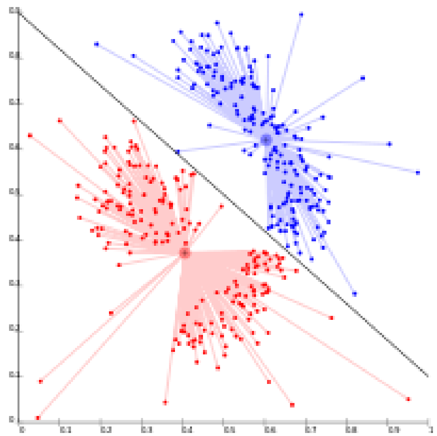
Uwaga: wzór (poniżej) najczęściej szacuje błąd z dużym nadmiarem

$$R(\alpha) \leq R_{emp}(\alpha) + \sqrt{\left(\frac{h(\log(2l/h)+1) - \log(\eta/4)}{l}\right)}$$

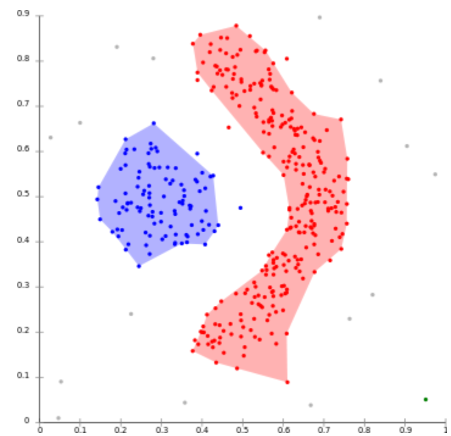
Uczenie „bez nauczyciela*”

Brak bazy uczącej:
Metody: K-średnich (K-means), identyfikacja rozkładów prawdopodobieństwa, etc.





K- średnich



DBSCAN
density-based spatial clustering of applications with noise

*unsupervised learning

Rysunki z Wikipedii