KAPITAŁ LUDZKI
NARODOWA STRATEGIA SPÓJNOŚCI

UNIA EUROPEJSKA
EUROPEJSKI
FUNDUSZ SPOŁECZNY

Prezentacja multimedialna
współfinansowana przez Unię Europejską
w ramach Europejskiego Funduszu Społecznego
w projekcie

*„Innowacyjna dydaktyka bez ograniczeń
– zintegrowany rozwój Politechniki Łódzkiej –
zarządzanie Uczelnią,
nowoczesna oferta edukacyjna
i wzmacniania zdolności do zatrudniania
osób niepełnosprawnych"*

dr inż. Artur Klepaczko

Eksploracja danych
Klasyfikatory liniowe

Zadanie nr 13 – Studia podyplomowe „Przetwarzanie i analiza obrazów biomedycznych"

Politechnika Łódzka
Instytut Elektroniki

90-924 Łódź, ul. Żeromskiego 116,
tel. 042 631 28 83
www.kapitalludzki.p.lodz.pl

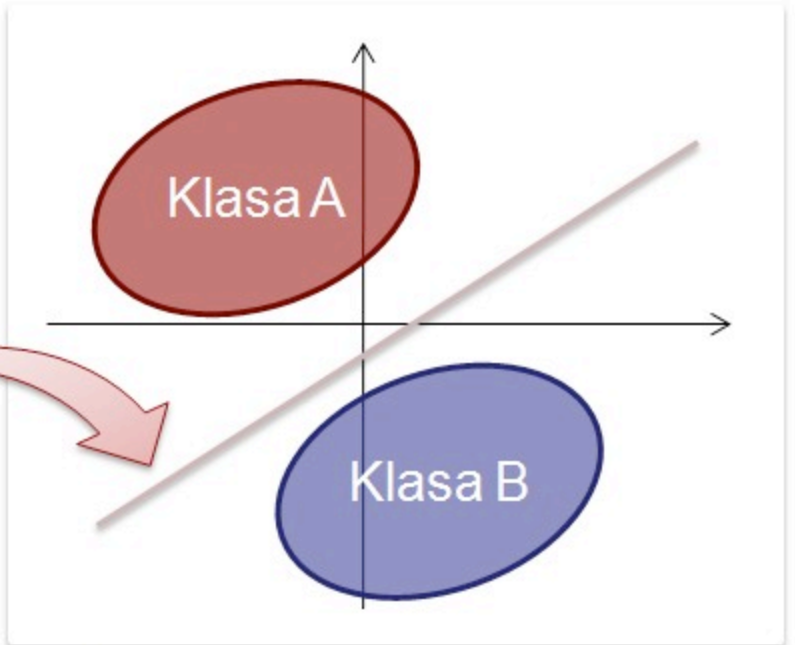
Prezentacja multimedialna współfinansowana przez Unię Europejską w ramach Europejskiego Funduszu Społecznego

Dyskryminacja liniowa

Klasy separowalne liniowo

Funkcja dyskryminacyjna (postać ogólna)

$$y(\mathbf{x}) = \mathbf{w}^T \mathbf{x} + w_0$$

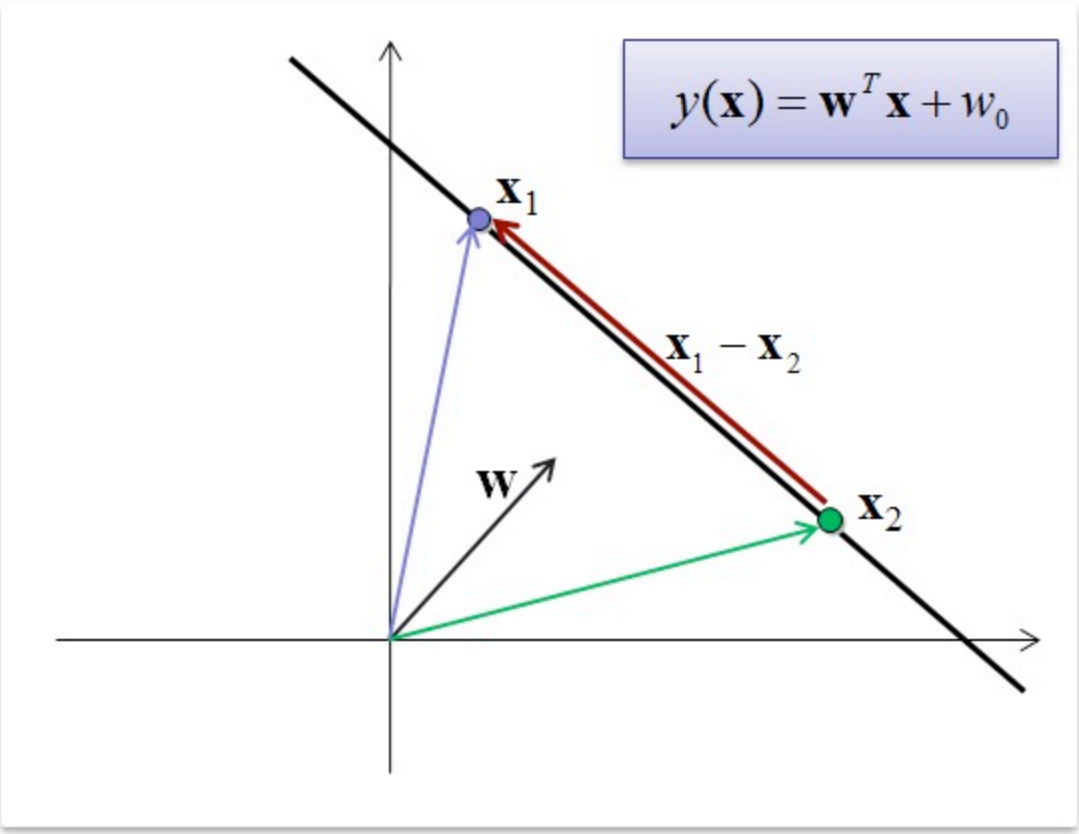


Przestrzeń dwuwymiarowa	– linia
Przestrzeń trójwymiarowa	– płaszczyzna
Przestrzeń n-wymiarowa	– hiperpłaszczyzna

$$y(\mathbf{x}) \geq 0 \quad \Rightarrow \quad \mathbf{x} \in A$$
$$y(\mathbf{x}) < 0 \quad \Rightarrow \quad \mathbf{x} \in B$$

Prezentacja multimedialna współfinansowana przez Unię Europejską w ramach Europejskiego Funduszu Społecznego

Funkcja liniowa — wektor wag



$$y(\mathbf{x}_1) = y(\mathbf{x}_2) = 0$$

$$\mathbf{w}^T\mathbf{x}_1 + w_0 = \mathbf{w}^T\mathbf{x}_2 + w_0 \Rightarrow$$
$$\mathbf{w}^T(\mathbf{x}_1 - \mathbf{x}_2) = 0$$

Iloczyn skalarny wektorów o niezerowej długości jest równy 0 tylko wtedy, gdy wektory te są ortogonalne (prostopadłe).

Wektor wag określa orientację hiperpłaszczyzny decyzyjnej.

Prezentacja multimedialna współfinansowana przez Unię Europejską w ramach Europejskiego Funduszu Społecznego

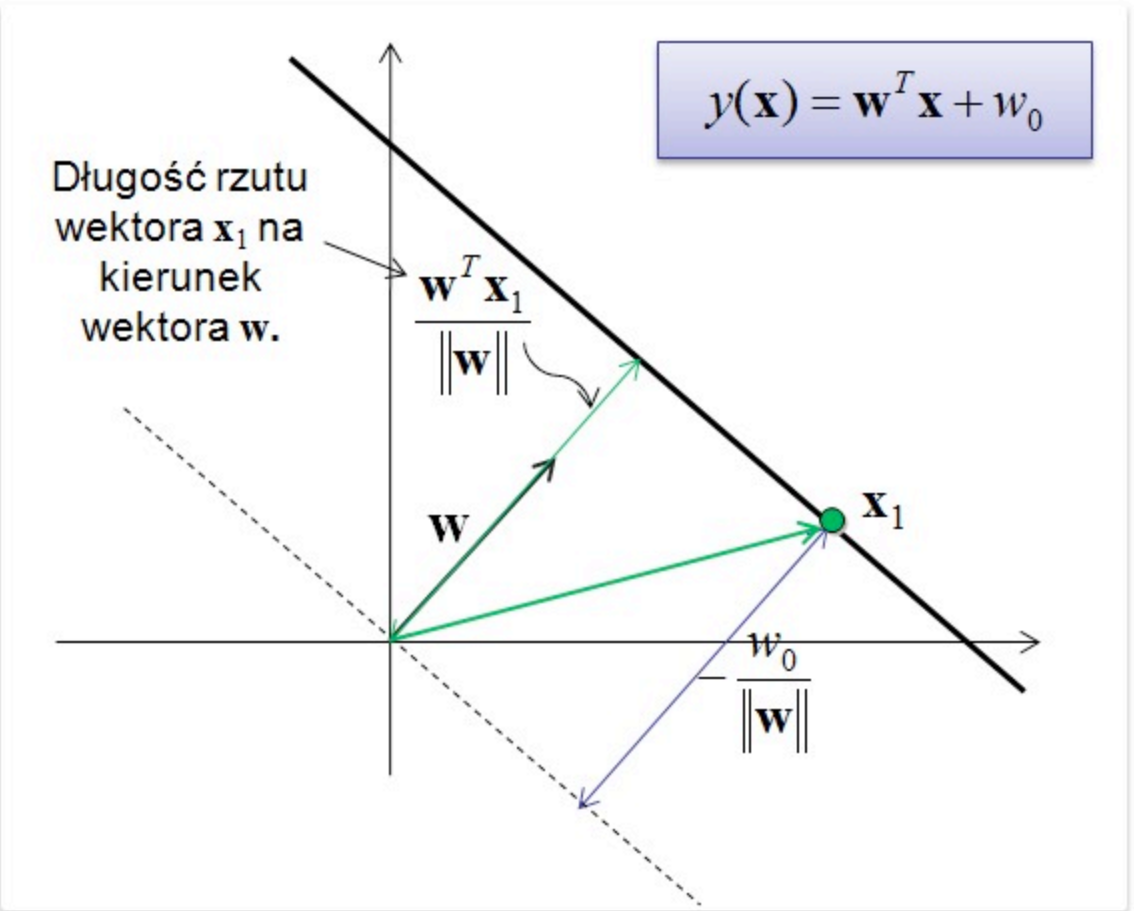
Funkcja liniowa — odchylenie

$$y(\mathbf{x}_1) = 0$$

$$\mathbf{w}^T\mathbf{x}_1 + w_0 = 0 \Rightarrow$$

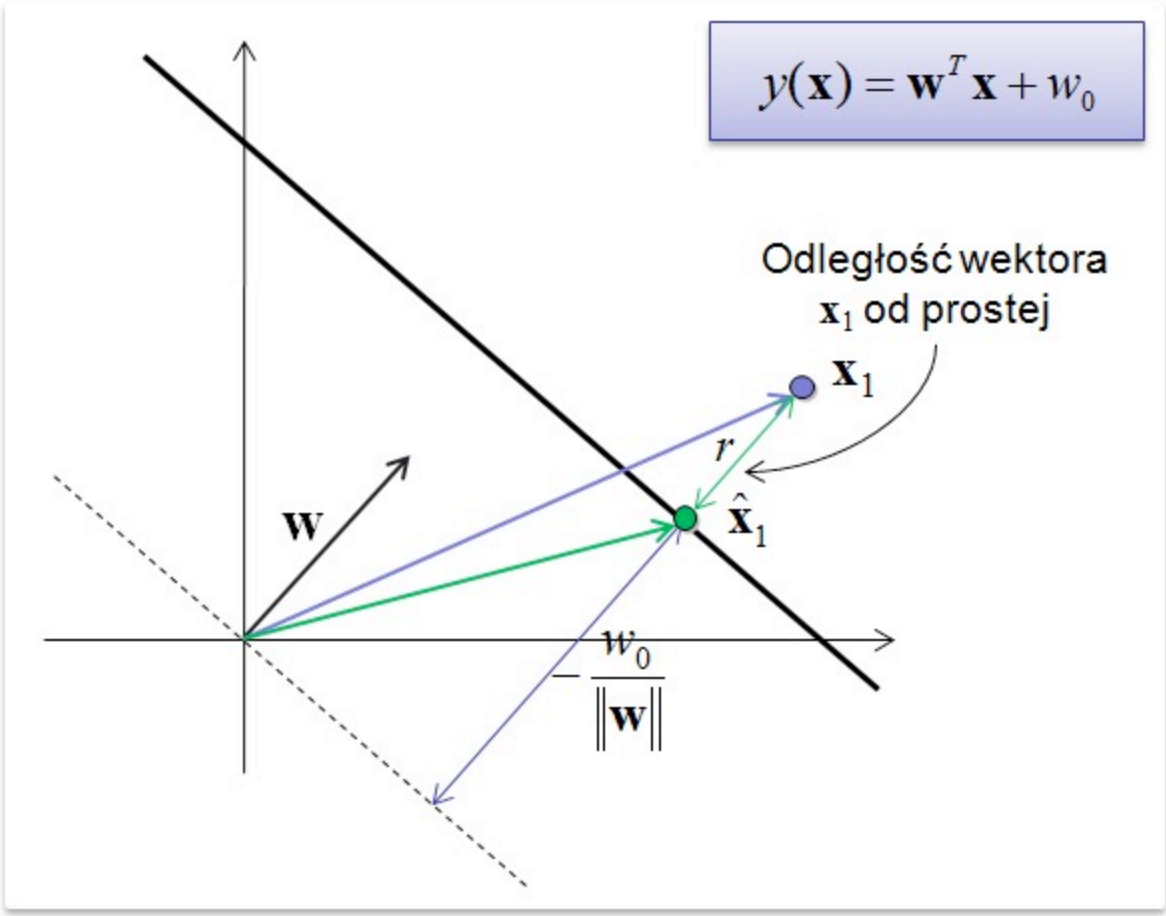
$$\frac{\mathbf{w}^T\mathbf{x}_1}{\|\mathbf{w}\|} = -\frac{w_0}{\|\mathbf{w}\|}$$

Parametr odchylenia określa położenie hiperpłaszczyzny decyzyjnej.



Prezentacja multimedialna współfinansowana przez Unię Europejską w ramach Europejskiego Funduszu Społecznego

Odległość punktu od linii prostej



$$\mathbf{x}_1 = \hat{\mathbf{x}}_1 + r\frac{\mathbf{w}}{\|\mathbf{w}\|}$$

$$\mathbf{w}^T\mathbf{x}_1 = \mathbf{w}^T\hat{\mathbf{x}}_1 + r\frac{\mathbf{w}^T\mathbf{w}}{\|\mathbf{w}\|}$$

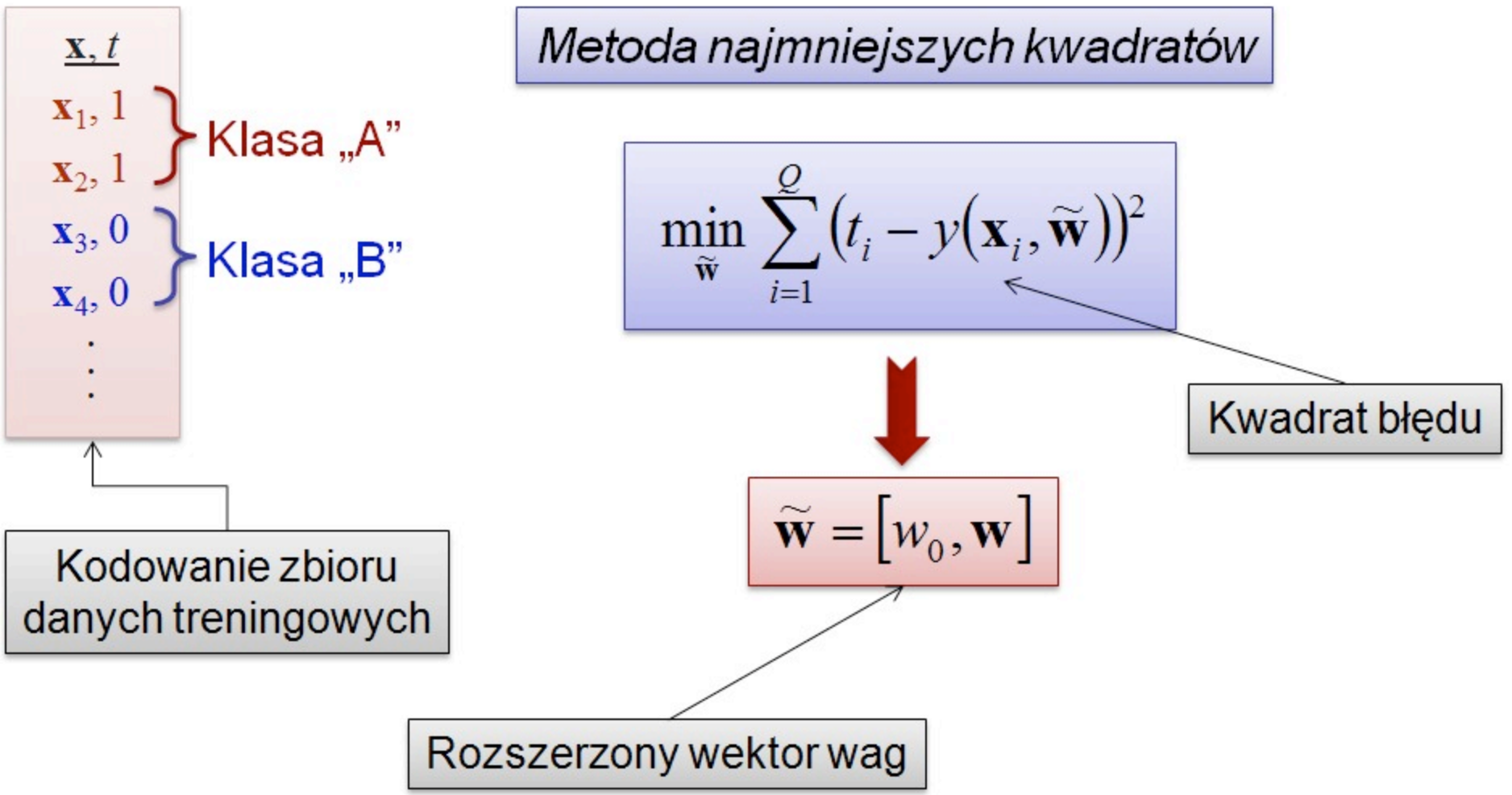
$$\mathbf{w}^T\mathbf{x}_1 + w_0 = \mathbf{w}^T\hat{\mathbf{x}}_1 + w_0 + r\frac{\mathbf{w}^T\mathbf{w}}{\|\mathbf{w}\|}$$

$$y(\mathbf{x}_1) = r\frac{\|\mathbf{w}\|^2}{\|\mathbf{w}\|}$$

$$r = \frac{y(\mathbf{x}_1)}{\|\mathbf{w}\|}$$

Prezentacja multimedialna współfinansowana przez Unię Europejską w ramach Europejskiego Funduszu Społecznego

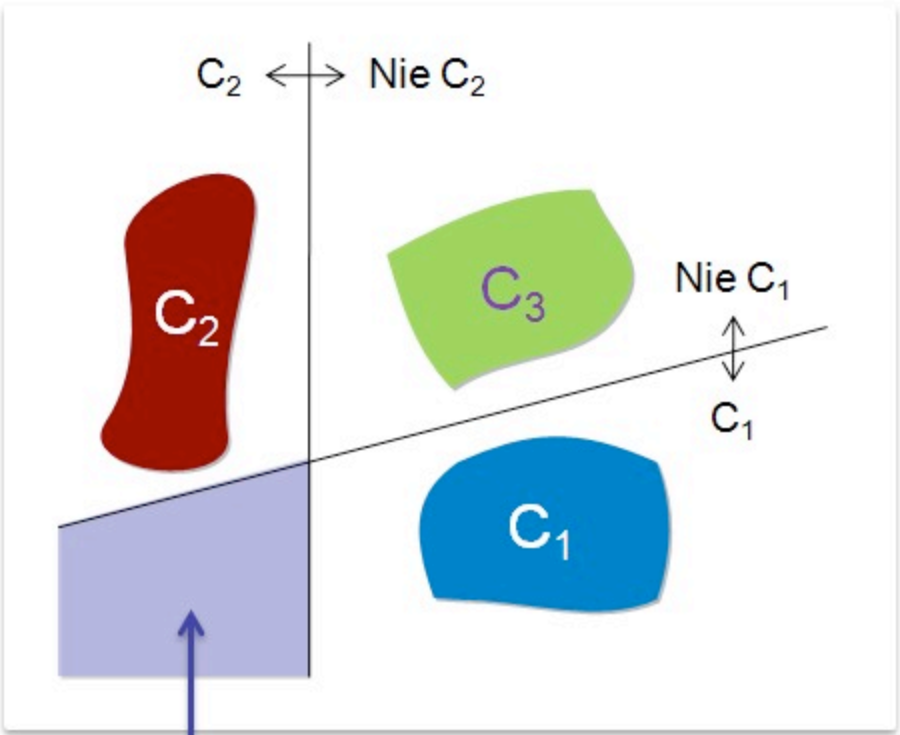
Uczenie klasyfikatora — regresja liniowa



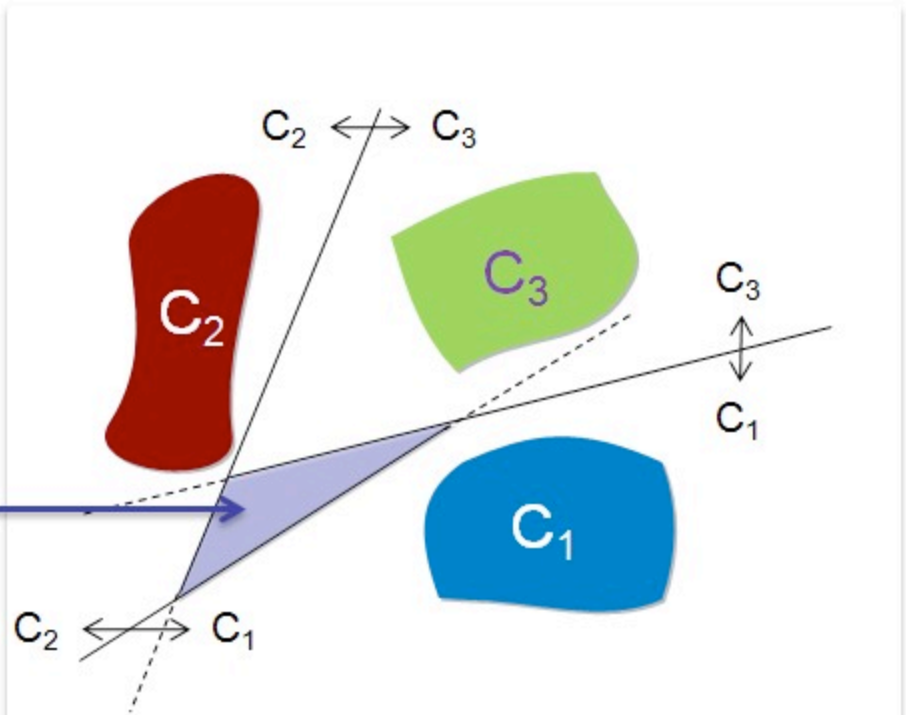
Prezentacja multimedialna współfinansowana przez Unię Europejską w ramach Europejskiego Funduszu Społecznego

Rozwinięcie do przypadku wielu klas

1. Zestaw K–1 linii decyzyjnych



2. Zestaw $K(K–1)/2$ linii decyzyjnych

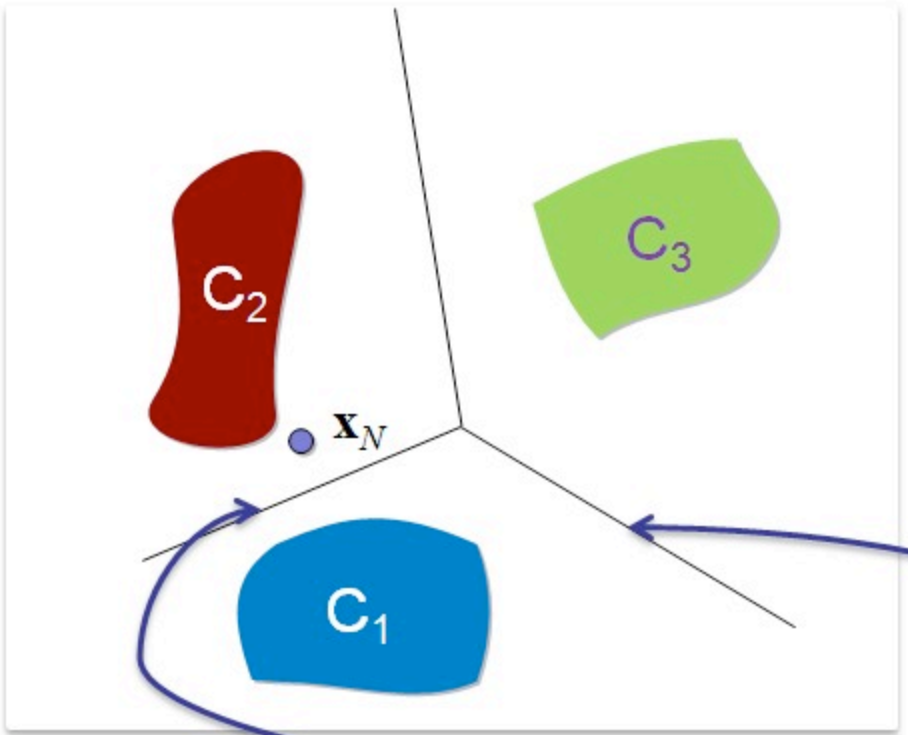


Obszary niejednoznaczne

Prezentacja multimedialna współfinansowana przez Unię Europejską w ramach Europejskiego Funduszu Społecznego

Rozwiązanie dla problemu wielu klas

3. Zestaw *K* linii decyzyjnych



$y(\mathbf{x})$ — funkcja przynależności wektora do danej klasy

Np.

$y_1(\mathbf{x}_N) = 0.3$

$y_2(\mathbf{x}_N) = 0.6$

$y_3(\mathbf{x}_N) = 0.1$

→ Wektor $\mathbf{x}_N$ klasyfikujemy do C_2

$y_1(\mathbf{x}) - y_3(\mathbf{x})$

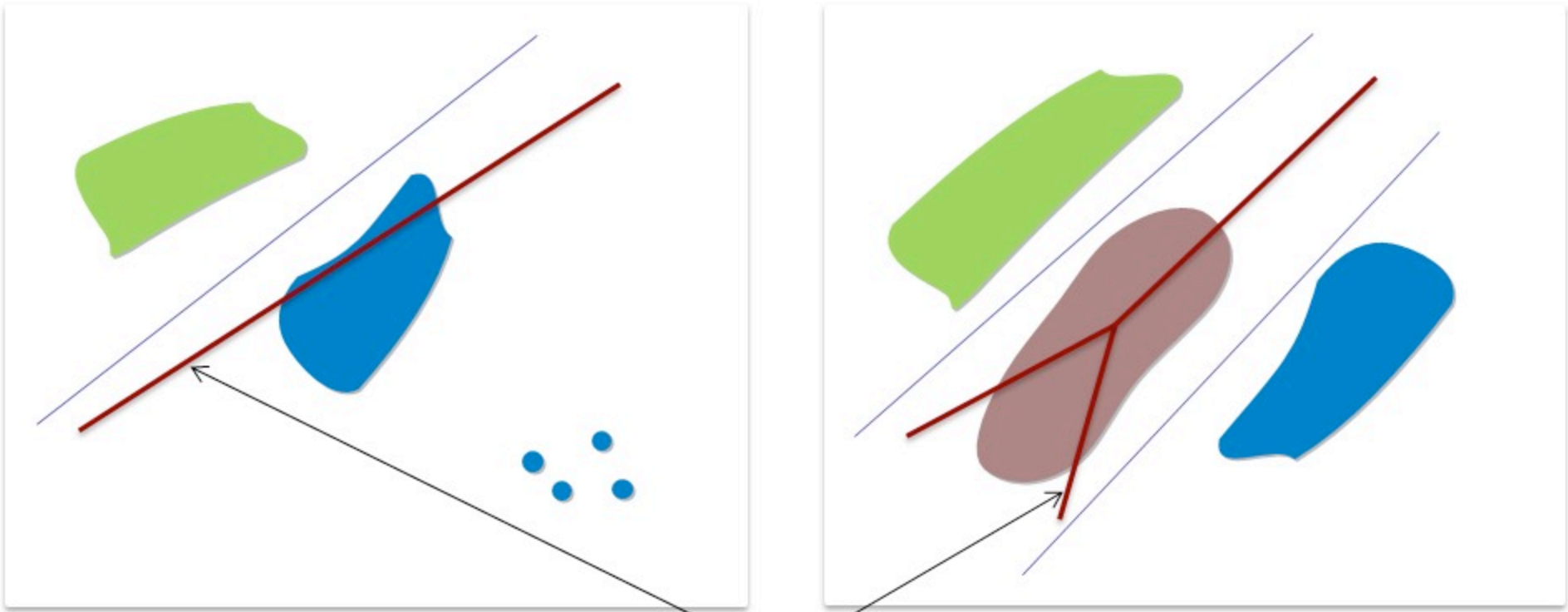
$y_1(\mathbf{x}) - y_2(\mathbf{x})$

Prezentacja multimedialna współfinansowana przez Unię Europejską w ramach Europejskiego Funduszu Społecznego

Regresja liniowa — ograniczenia

Wrażliwość na wartości oddalone (ang. *outliers)*

Słaba separacja w przypadku wielu klas



Linie decyzyjne wyznaczone metodą najmniejszych kwadratów.

Prezentacja multimedialna współfinansowana przez Unię Europejską w ramach Europejskiego Funduszu Społecznego

Regresja a szacowanie prawdopodobieństwa

Regresja liniowa

$$\Pr[y=1|\mathbf{x}] = w_0 + \mathbf{w}^T\mathbf{x}$$

Prawdopodobieństwo przynależności do klasy zakodowanej jako y=1.

Jednakże, wartość funkcji decyzyjnej wykracza poza zakres (0,1).

Transformacja zmiennej docelowej

$$\Pr[y=1|\mathbf{x}] \rightarrow \ln\frac{\Pr[y=1|\mathbf{x}]}{1-\Pr[y=1|\mathbf{x}]}$$

Funkcja *logitowa*

Regresja logistyczna

Prezentacja multimedialna współfinansowana przez Unię Europejską w ramach Europejskiego Funduszu Społecznego

Regresja logistyczna

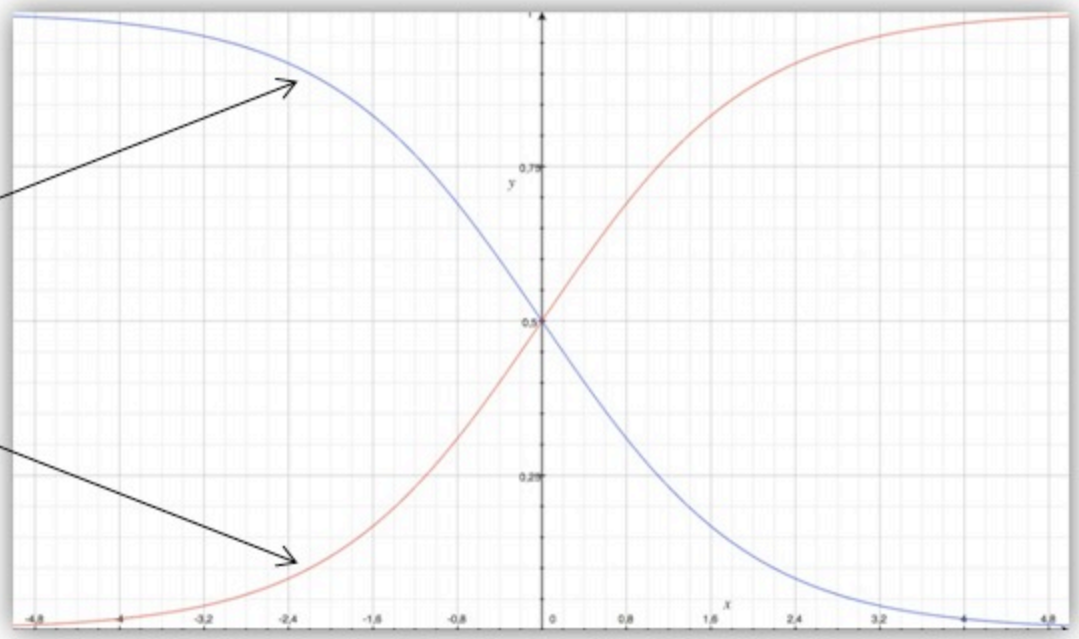
$$\ln \frac{\Pr[y=0 \mid \mathbf{x}]}{1-\Pr[y=0 \mid \mathbf{x}]} = w_0 + \mathbf{w}^T \mathbf{x}$$

Aproksymacja zmiennej docelowej określonej funkcją logitową przy użyciu funkcji liniowej.

Rozwikłanie modelu

$$\Pr[y=0 \mid \mathbf{x}] = \frac{\exp(w_0 + \mathbf{w}^T \mathbf{x})}{1+\exp(w_0 + \mathbf{w}^T \mathbf{x})}$$

$$\Pr[y=1 \mid \mathbf{x}] = \frac{1}{1+\exp(w_0 + \mathbf{w}^T \mathbf{x})}$$



Wartości obu oszacowań mieszczą się zawsze w zakresie (0,1)

Prezentacja multimedialna współfinansowana przez Unię Europejską w ramach Europejskiego Funduszu Społecznego

Metoda największej wiarygodności

Funkcja wiarygodności

$$LL = \sum_{i=1}^{Q} (1 - t_i) \log(1 - \Pr[1 \mid \mathbf{x}_i]) + t_i \log(\Pr[1 \mid \mathbf{x}_i])$$

Maksymalizacja

Parametry modelu logistycznego → $\mathbf{w}$, w_0

Reguła klasyfikacyjna

Jeśli

$$\Pr[1 \mid \mathbf{x}_N] > \Pr[0 \mid \mathbf{x}_N]$$

to wektor $\mathbf{x}_N$ zaliczamy do klasy zakodowanej jako 1.

Regresja logistyczna w zastosowaniu do zbioru *Iris*

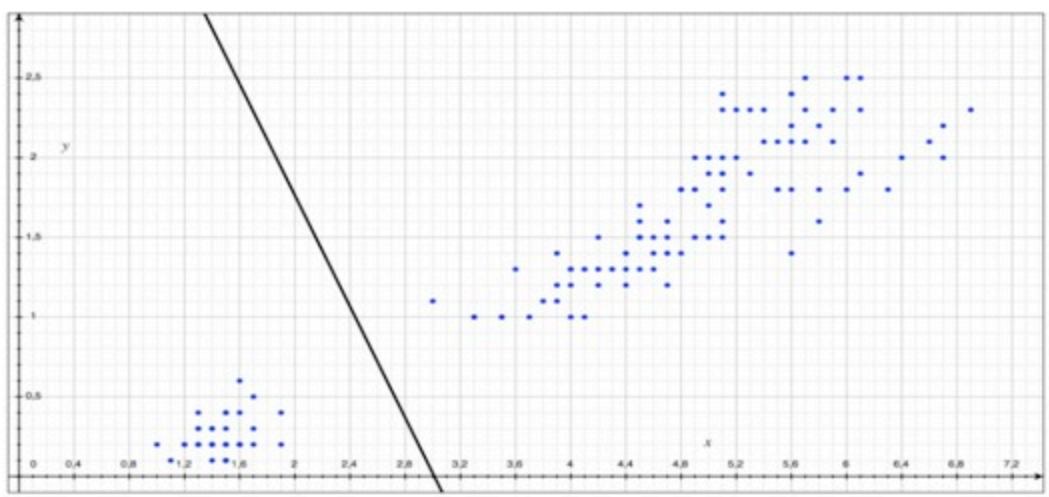
SimpleLogistic:

Class 0 :
29.99 +
[petallength] * -9.96 +
[petalwidth] * -5.71

Class 1 :
-6.15 +
[sepallength] * 1.67 +
[sepalwidth] * 0.82 +
[petallength] * -0.74 +
[petalwidth] * -1.28

Class 2 :
-34.94 +
[sepallength] * -0.4 +
[sepalwidth] * -3.76 +
[petallength] * 6.27 +
[petalwidth] * 10.89

Niekiedy wystarcza podprzestrzeń atrybutów do dobrej separacji wektorów danych z określonej klasy.



=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances	141	94	%
Incorrectly Classified Instances	9	6	%

Prezentacja multimedialna współfinansowana przez Unię Europejską w ramach Europejskiego Funduszu Społecznego

Porównywanie modeli

Obecność w modelu zmiennych nieistotnych zwiększa błędy oszacowania współczynników istotnych zmiennych i w efekcie pogarsza dopasowanie modelu.

Koronacki J., Ćwik J., *Statystyczne systemy uczące się*, WNT, Warszawa 2005

Kryterium informacyjne Akaike (AIC)

$$AIC = -LL + p$$

Funkcja wiarygodności

Liczba parametrów

Minimalizacja
(selekcja cech istotnych)

Lepsza jakość modelu

Regresja logistyczna z użyciem kryterium AIC — zbiór *Iris*

Bez diagnozy modelu

Class 0 :
29.99 +
[petallength] * -9.96 +
[petalwidth] * -5.71

Class 1 :
-6.15 +
[sepallength] * 1.67 +
[sepalwidth] * 0.82 +
[petallength] * -0.74 +
[petalwidth] * -1.28

Class 2 :
-34.94 +
[sepallength] * -0.4 +
[sepalwidth] * -3.76 +
[petallength] * 6.27 +
[petalwidth] * 10.89

Correctly Classified Instances	141	94 %
Incorrectly Classified Instances	9	6 %

Class 0 :
12.27 +
[petallength] * -2.77 +
[petalwidth] * -5.71

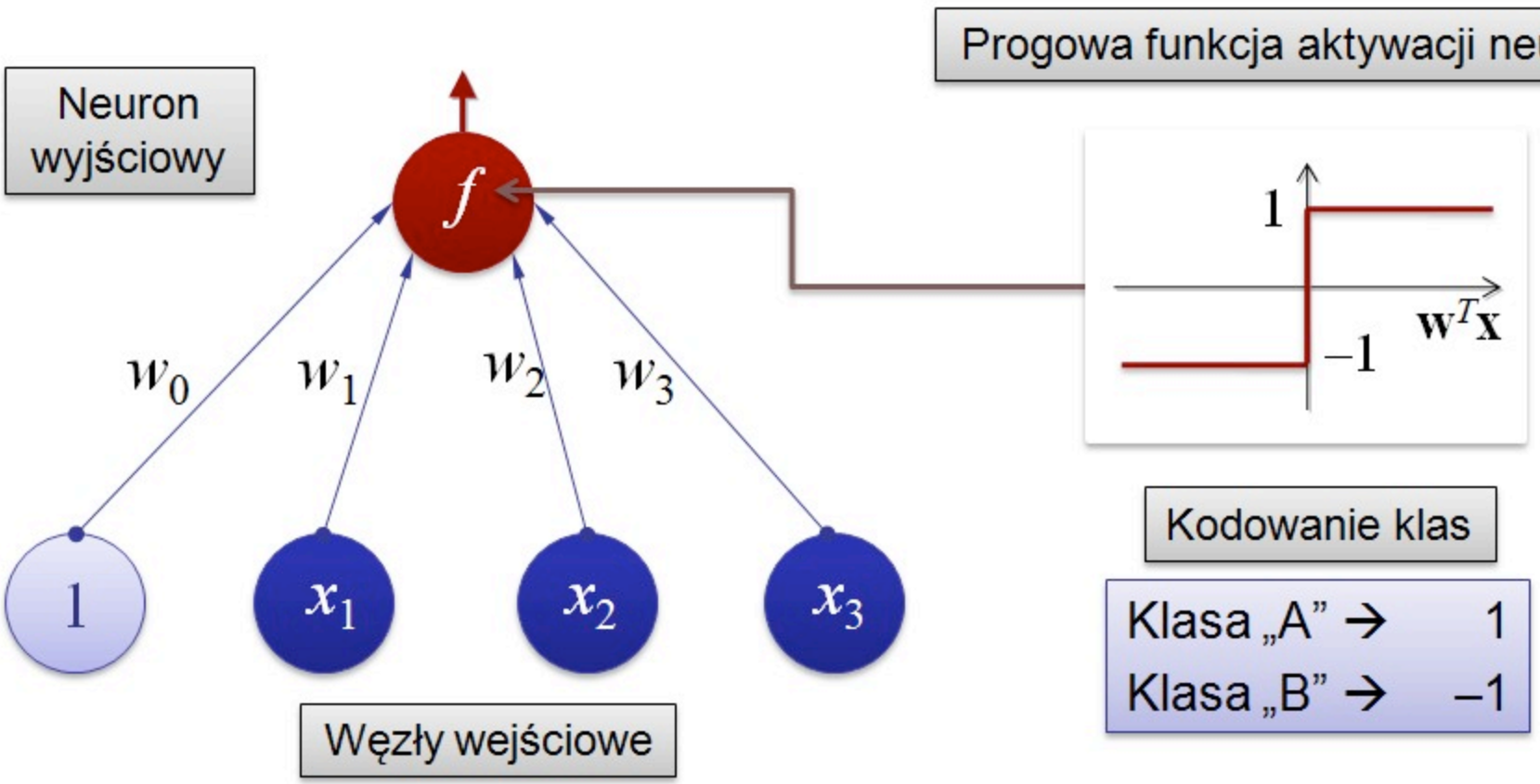
Class 1 :
0.95 +
[sepallength] * 0.51 +
[sepalwidth] * -0.61 +
[petallength] * -0.12 +
[petalwidth] * -0.65

Class 2 :
-22.85 +
[sepalwidth] * -2.24 +
[petallength] * 3.87 +
[petalwidth] * 6.69

Uproszczenie modelu może przyczynić się do poprawy jakości klasyfikacji.

Correctly Classified Instances	145	96.6667 %
Incorrectly Classified Instances	5	3.3333 %

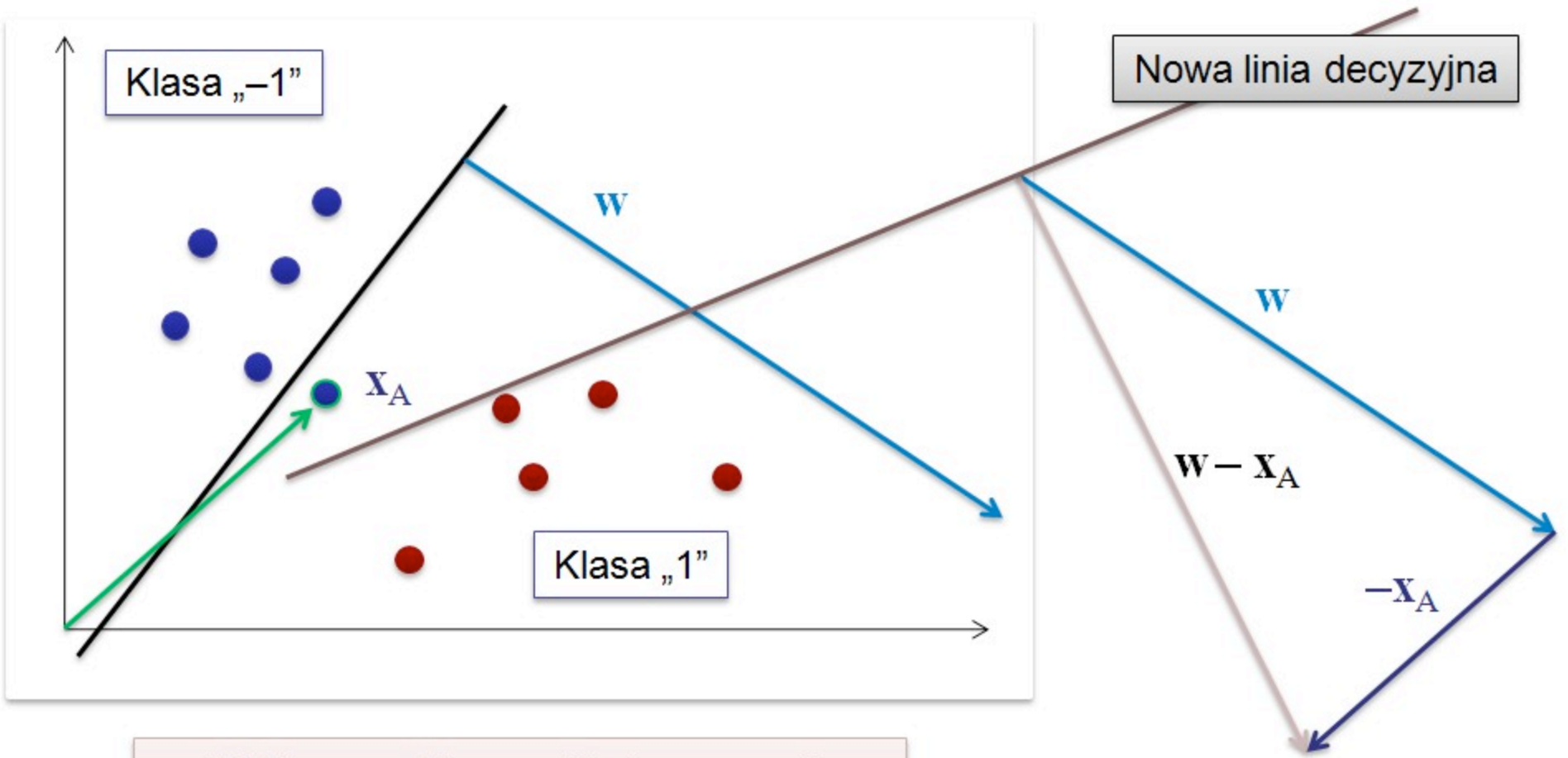
Perceptron



Regresja — estymowanie wartości prawdopodobieństwa przynależności do danej klasy
Reguła perceptronowa — „po prostu” wyznaczanie hiperpłaszczyzny decyzyjnej

Prezentacja multimedialna współfinansowana przez Unię Europejską w ramach Europejskiego Funduszu Społecznego

Uczenie perceptronu — modyfikacja wag



Wektor $\mathbf{x}_A$ z klasy „–1” niepoprawnie zaklasyfikowany do klasy „1”

Prezentacja multimedialna współfinansowana przez Unię Europejską w ramach Europejskiego Funduszu Społecznego

Reguła perceptronowa (*delta*)

- Przyjmij dowolnie początkowe wartości wag (≈0)
- Powtórz poniższe kroki, aż do momentu, gdy wszystkie wektory danych ze zbioru treningowego będą prawidłowo klasyfikowane:
  - Dla każdego wektora $\mathbf{x}_i$ ze zbioru treningowego:
    - Jeżeli wektor $\mathbf{x}_i$ jest nieprawidłowo klasyfikowany:
      - Jeżeli wektor $\mathbf{x}_i$ należy do klasy „1” to wykonaj $\mathbf{w} + \mathbf{x}_i$
      - Jeżeli wektor $\mathbf{x}_i$ należy do klasy „–1” to wykonaj $\mathbf{w} + \mathbf{x}_i$

- Można wykazać, że reguła jest zbieżna, jeśli tylko zadanie jest liniowo separowalne.
- Często potrzeba powtórzyć adaptację wag wiele razy dla tego samego zbioru wektorów treningowych. Pojedyncze przejście przez zbiór danych nazywa się *epoką*.
- Prędkość zbieżności uczenia można kontrolować stosując współczynnik długości kroku .

Prezentacja multimedialna współfinansowana przez Unię Europejską w ramach Europejskiego Funduszu Społecznego

Klasyfikator perceptronowy w zastosowaniu do zbioru *Iris*

```
=== Classifier model (full training set) ===
Sigmoid Node 0
    Inputs    Weights
    Threshold    -3.769559351070804
    Attrib sepallength    -1.32409409935167
    Attrib sepalwidth    3.517848070574029
    Attrib petallength    -4.735447298859189
    Attrib petalwidth    -4.3724179715078195
                          .
                          .
                          .
Class Iris-setosa
    Input
    Node 0
Class Iris-versicolor
    Input
    Node 1
Class Iris-virginica
    Input
    Node 2
```

Reguła perceptronowa

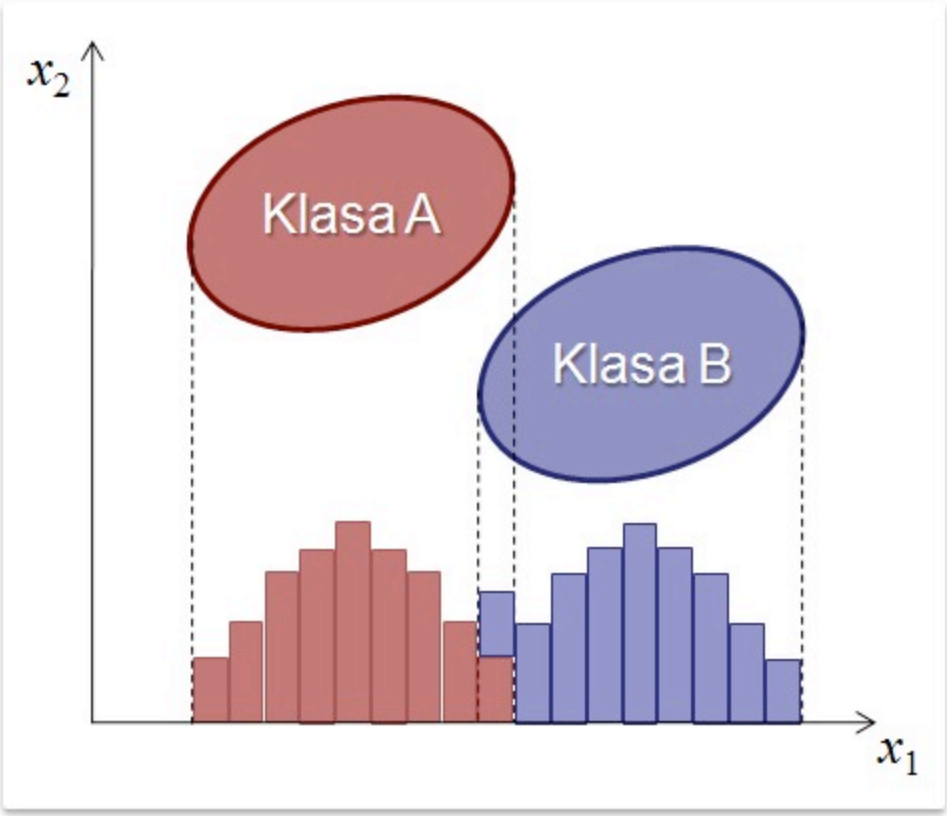
- W przypadku $K>2$ klas stosuje się K perceptronów połączonych w sieć. Jednocześnie tylko jeden neuron wyjściowy może być uaktywniony (=1).
- Perceptron stanowi punkt wyjścia do rozważań na temat sztucznych sieci neuronowych.
- Zamiast skokowej funkcji aktywacji można użyć tzw. funkcji *sigmoidalnej* . Wówczas do uczenia perceptronu stosuje się algorytm *wstecznej propagacji błędu*.

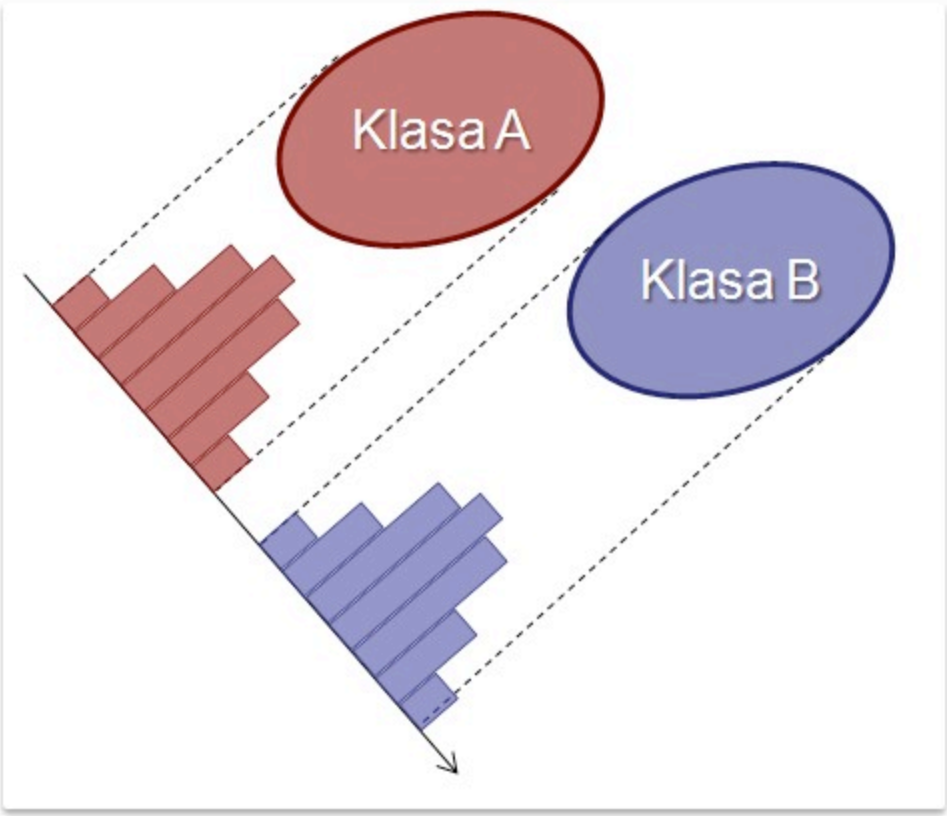
```
=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances         143               95.3333 %
Incorrectly Classified Instances         7                4.6667 %
```

Prezentacja multimedialna współfinansowana przez Unię Europejską w ramach Europejskiego Funduszu Społecznego

Liniowa analiza dyskryminacyjna





Zadanie liniowej analizy dyskryminacyjnej (LDA) polega na znalezieniu w przestrzeniu cech kierunku zapewniającego najlepszą możliwą separację klas.

Prezentacja multimedialna współfinansowana przez Unię Europejską w ramach Europejskiego Funduszu Społecznego

Zmienność międzygrupowa

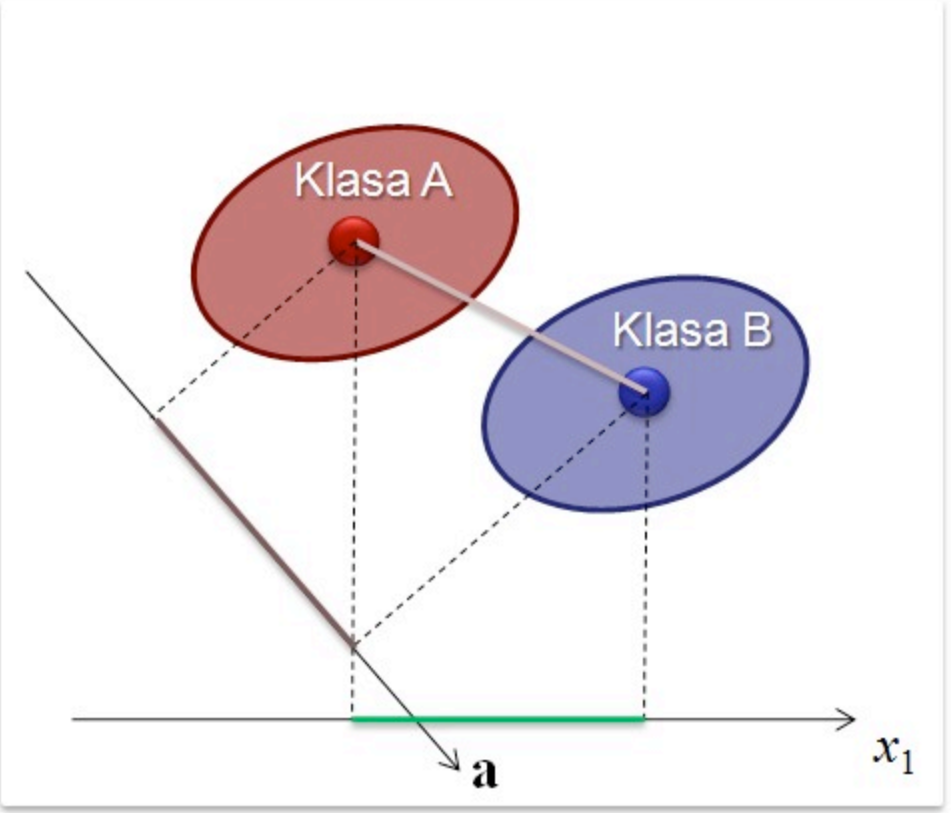
Wektor średni

$$\bar{\mathbf{x}} = \left[\bar{x}^{(1)}, \bar{x}^{(2)}, \ldots, \bar{x}^{(j)}, \ldots \bar{x}^{(N)}\right]$$

$$\bar{x}^{(j)} = \frac{1}{Q}\sum_{i=1}^{Q} x_i^{(j)}$$

Miara zmienności międzygrupowej

$$\left(\mathbf{a}^T\bar{\mathbf{x}}_A - \mathbf{a}^T\bar{\mathbf{x}}_B\right)^2$$



Odległość średnich wektorów klas wzdłuż kierunku **a**

Prezentacja multimedialna współfinansowana przez Unię Europejską w ramach Europejskiego Funduszu Społecznego

Zmienność wewnątrzgrupowa

Macierz kowariancji

$$\mathbf{S}_k = \frac{1}{Q_k - 1} \sum_{i=1}^{Q_k} (\mathbf{x}_i - \bar{\mathbf{x}})(\mathbf{x}_i - \bar{\mathbf{x}})^T$$

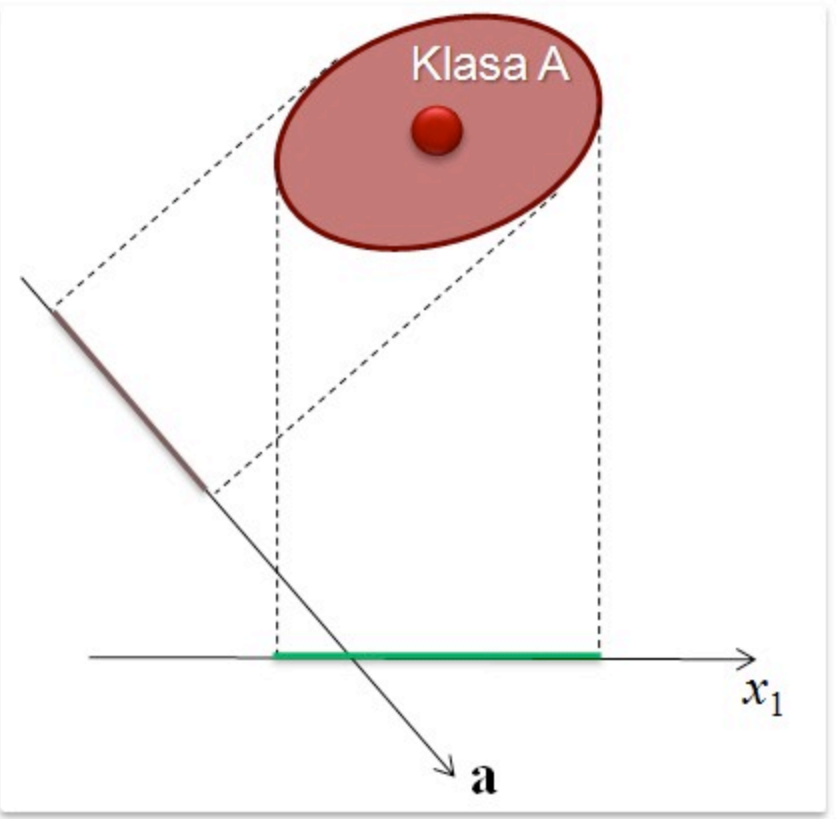
Pojedyncza klasa

$$\mathbf{W} = \frac{1}{Q - 2} \sum_{k=1}^{2} (Q_k - 1)\mathbf{S}_k$$

Wspólna dla obu klas

Miara zmienności wewnątrzgrupowej

$$\mathbf{a}^T \mathbf{W} \mathbf{a}$$



Rozproszenie wektorów wokół średnich wektorów klas w danym kierunku

Prezentacja multimedialna współfinansowana przez Unię Europejską w ramach Europejskiego Funduszu Społecznego

Analiza LDA a klasyfikacja

Zadanie LDA

Znaleźć taki kierunek $\hat{\mathbf{a}}$, który maksymalizuje wyrażenie

$$\frac{\left(\mathbf{a}^T\overline{\mathbf{x}}_A - \mathbf{a}^T\overline{\mathbf{x}}_B\right)^2}{\mathbf{a}^T\mathbf{W}\mathbf{a}}$$

Małe rozproszenie wewnątrz dobrze odseparowanych klas.

Klasyfikacja

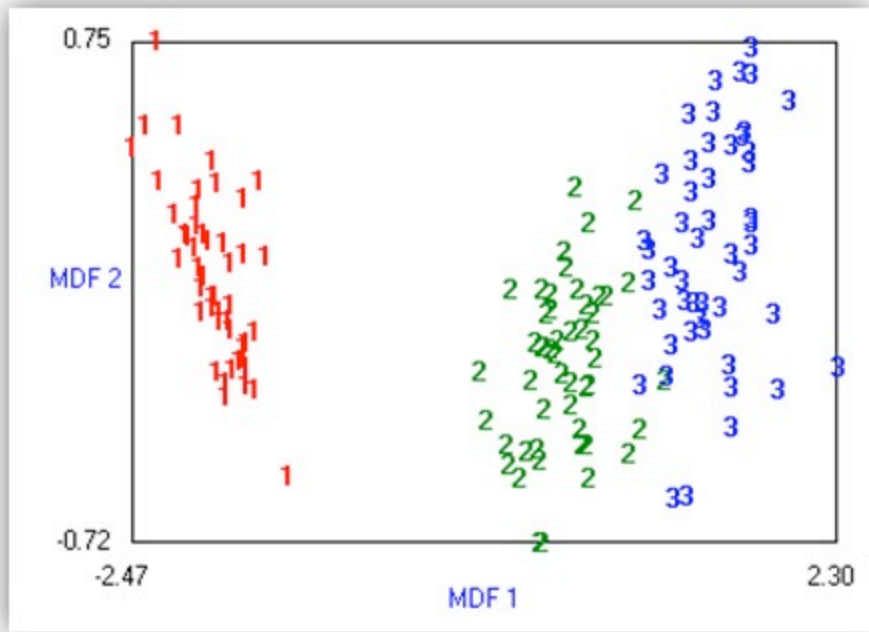
Nowy wektor $\mathbf{x}_N$ należy zaklasyfikować do A jeśli

$$\left|\hat{\mathbf{a}}^T\mathbf{x}_N - \hat{\mathbf{a}}^T\overline{\mathbf{x}}_A\right| < \left|\hat{\mathbf{a}}^T\mathbf{x}_N - \hat{\mathbf{a}}^T\overline{\mathbf{x}}_B\right|$$

Wektor testowy leży bliżej środka klasy A.

Prezentacja multimedialna współfinansowana przez Unię Europejską w ramach Europejskiego Funduszu Społecznego

LDA w zastosowaniu do zbioru *Iris*



MDF (ang. *most discriminative feature*) — kierunek najlepiej rozdzielający klasy

Dla *K* klas można otrzymać co najwyżej *K*-1 ortogonalnych kierunków MDF

* b11 report file [LDA analysis] <2009-01-12 22:43:42>
* Data file name: "IRIS.SEL"
* Selected features [4 out of 4]
sepal length [#1/#1]; p.mean= 5.84333E+000, p.std= 8.28066E-001
sepal width [#2/#2]; p.mean= 3.05400E+000, p.std= 4.33594E-001
petal length [#3/#3]; p.mean= 3.75867E+000, p.std= 1.76442E+000
petal width [#4/#4]; p.mean= 1.19867E+000, p.std= 7.63161E-001
Feature vector standardized: NO

Projection matrix for MDF:

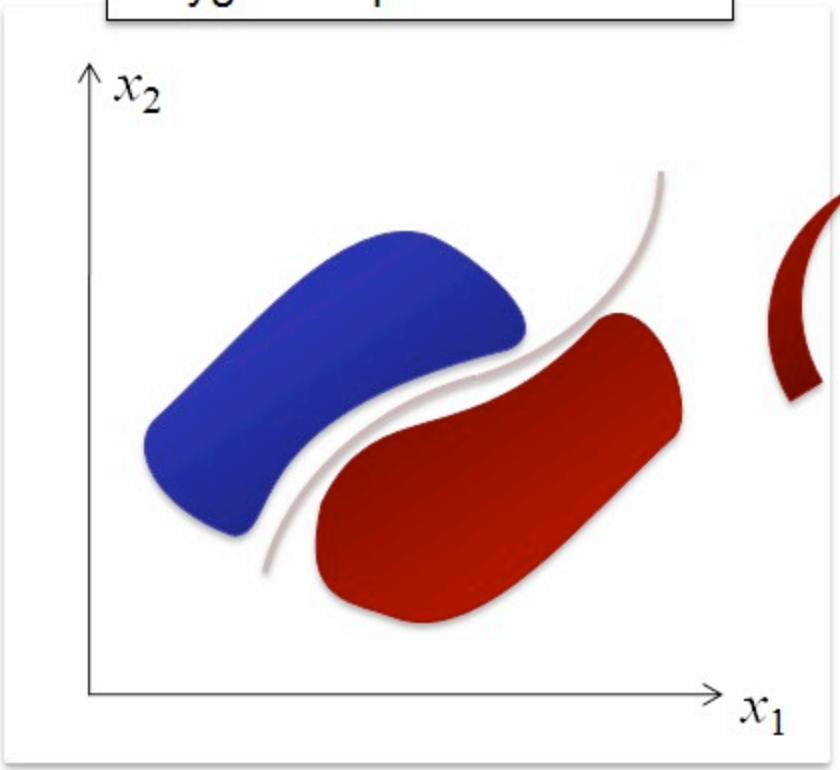
-2.04910E-001	8.98234E-003	-7.50750E-001	1.05258E-001
-3.87143E-001	5.88999E-001	4.27339E-001	-3.69821E-001
5.46482E-001	-2.54287E-001	4.42155E-001	-4.35665E-001
7.13785E-001	7.67032E-001	-2.41361E-001	8.13849E-001

LDA dimensionality: 2
Missclassified data vectors: 5/150 [or 3.33%]

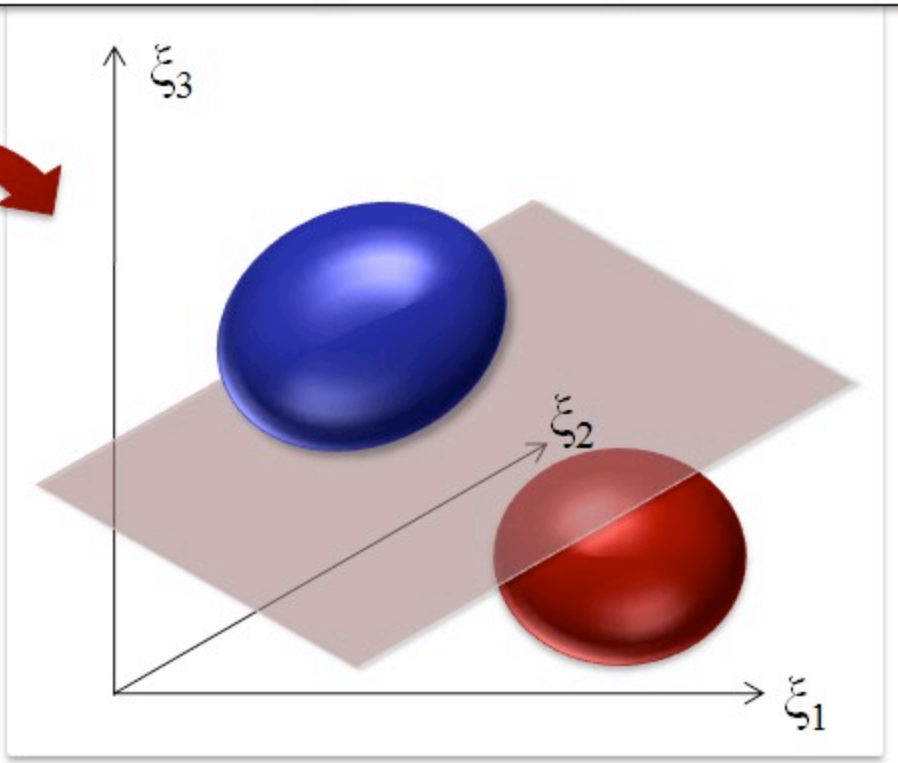
Prezentacja multimedialna współfinansowana przez Unię Europejską w ramach Europejskiego Funduszu Społecznego

Nieliniowe granice między klasami

Oryginalna przestrzeń cech



Przestrzeń cech po nieliniowej transformacji



Problem nieliniowej separacji klas można rozwiązać za pomocą klasyfikatorów liniowych po zastosowaniu nieliniowej transformacji przestrzeni cech.

Prezentacja multimedialna współfinansowana przez Unię Europejską w ramach Europejskiego Funduszu Społecznego

Nieliniowa transformacja przestrzeni cech

Przykładowa transformacja przestrzeni cech

$$\xi_1 = x_1^2 x_2 \quad \xi_2 = x_1 x_2^2 \quad \xi_3 = x_1^3$$

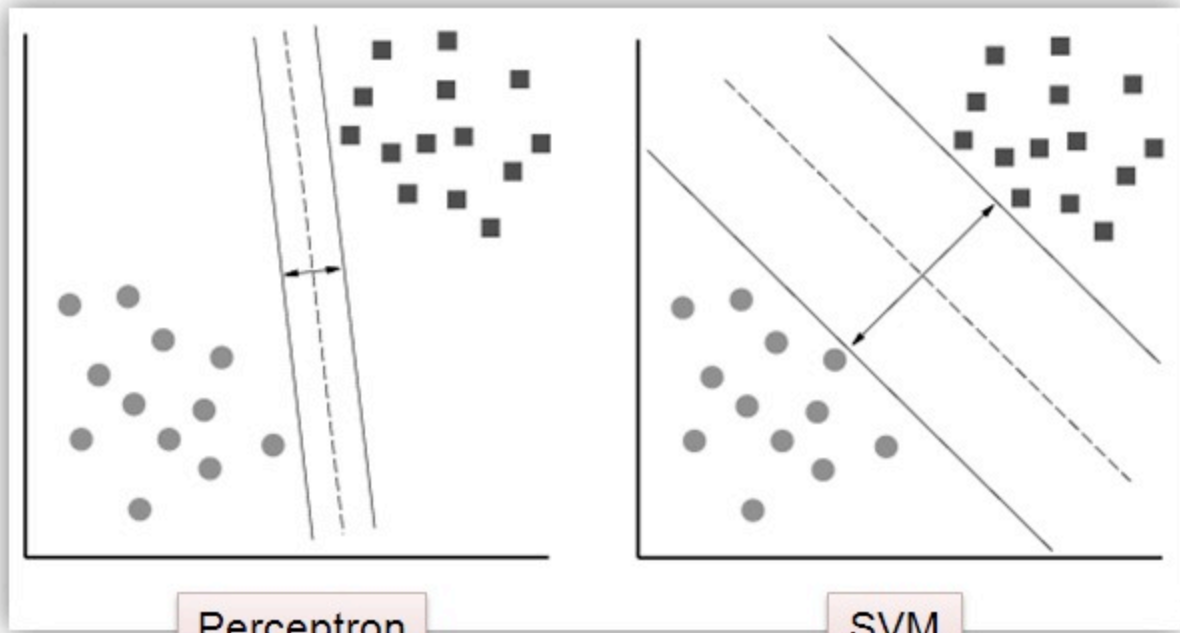
Często, aby uzyskać poprawną klasyfikację wymiar nowej przestrzeni cech musi być znacznie większy niż wymiar przestrzeni oryginalnej.

- duża złożoność obliczeniowa, zwłaszcza dla dużych wymiarów oryginalnej przestrzeni atrybutów
- niebezpieczeństwo nadmiernego dopasowania

Odpowiedzią na te problemy jest algorytm wektorów podpierających (ang. *support vector machines*)

Prezentacja multimedialna współfinansowana przez Unię Europejską w ramach Europejskiego Funduszu Społecznego

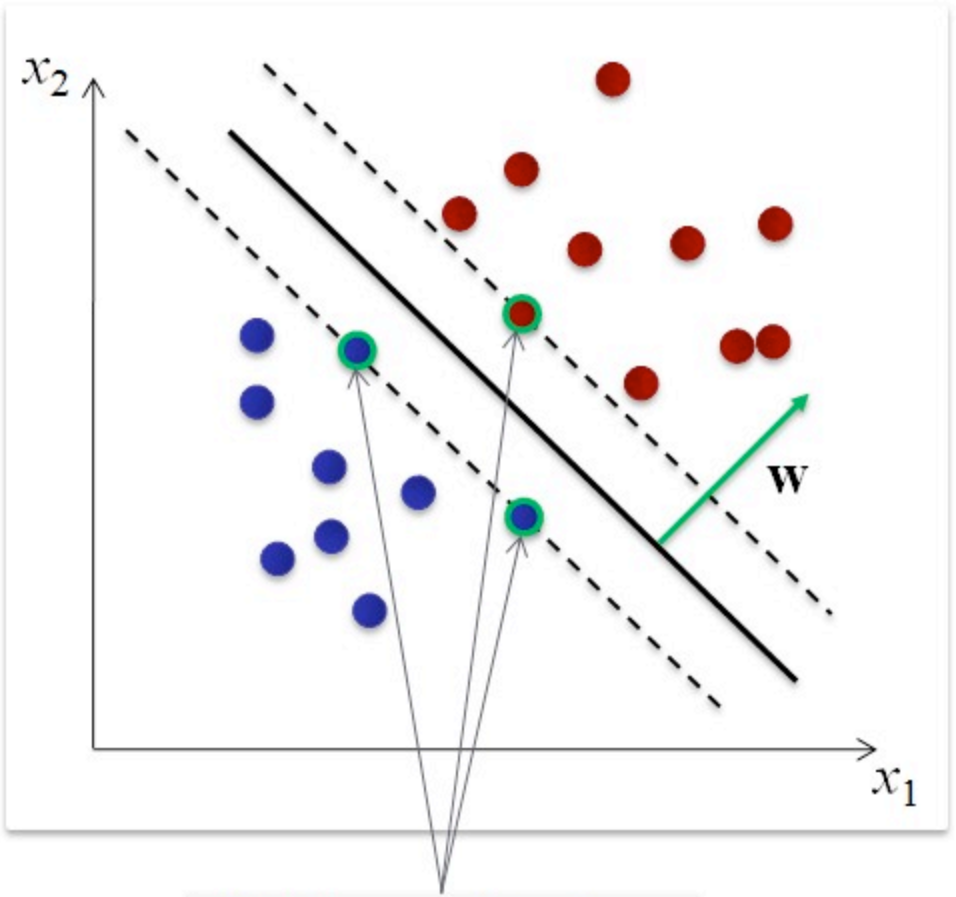
Algorytm *Support Vector Machines*



W algorytmie SVM konstruowana jest hiperpłaszczyzna decyzyjna w kierunku ortogonalnym do kierunku największego marginesu rozdzielającego klasy (ang. *optimal margin hyperplane*).

Prezentacja multimedialna współfinansowana przez Unię Europejską w ramach Europejskiego Funduszu Społecznego

Konstrukcja optymalnej hiperpłaszczyzny



Wektory podpierające

Równanie hiperpłaszczyzny

$$y(\mathbf{x}) = \mathbf{w}^T \mathbf{x} + w_0$$

$$y(\mathbf{x}) = b + \sum_{i-\text{indeks wektora podpierającego}} \alpha_i t_i \mathbf{x}_i^T \mathbf{x}$$

Współczynniki α_i znajdowane poprzez rozwiązanie zadania kwadratowej optymalizacja z ograniczeniami… (!)

Na szczęście istnieją gotowe implementacje → *sequential minimal optimization* (SMO)

Prezentacja multimedialna współfinansowana przez Unię Europejską w ramach Europejskiego Funduszu Społecznego

Funkcje jądra

Transformacja wektorów danych

$$\mathbf{x} \rightarrow \Phi(\mathbf{x})$$

Iloczyn skalarny w oryginalnej przestrzeni cech

$$\mathbf{x}_i^T \cdot \mathbf{x}_j$$

Iloczyn skalarny w przestrzeni transformowanej

$$\Phi^T(\mathbf{x}_i) \cdot \Phi(\mathbf{x}_j) \quad \Leftrightarrow \quad K(\mathbf{x}_i, \mathbf{x}_j)$$

K – funkcja jądra (ang. *kernel function)*

Wielomianowe funkcje jądra

$$K(\mathbf{x}_i, \mathbf{x}_j) = (\mathbf{x}_i^T \cdot \mathbf{x}_j)^n$$

$$K(\mathbf{x}_i, \mathbf{x}_j) = (\mathbf{x}_i^T \cdot \mathbf{x}_j + 1)^n$$

Radialna funkcja jądra

$$K(\mathbf{x}_i, \mathbf{x}_j) = \exp\left(-\frac{\|\mathbf{x}_i - \mathbf{x}_j\|^2}{c}\right)$$

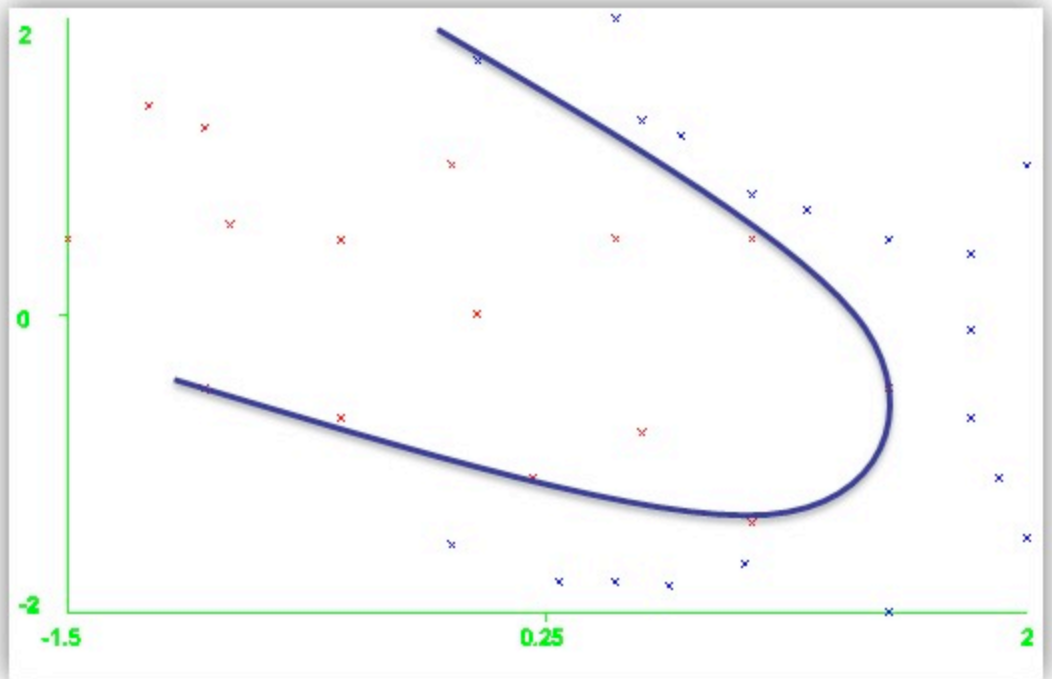
Prezentacja multimedialna współfinansowana przez Unię Europejską w ramach Europejskiego Funduszu Społecznego

Algorytm SVM dla przypadku klas nieliniowo separowalnych

Hiperpowierzchnia decyzjna

$$y(\mathbf{x}) = b + \sum_{i} \alpha_i t_i K(\mathbf{x}_i, \mathbf{x})$$

i – indeks wektora podpierającego



Zalety algorytm SVM

- Granica decyzyjna konstruowana jest w oparciu o niewielką liczbę wektorów podpierających
  - małe ryzyko nadmiernego dopasowania
  - odporność na wektory oddalone (ang. *outliers*)

- Dzięki zastosowaniu funkcji jądra możliwe jest rozwiązywania problemów nieliniowych bez zwiększania stopnia złożoności obliczeniowej.