Systemy uczące się
Metoda wektorów wspierających
wykład 2-3

Jerzy Stefanowski

Instytut Informatyki PP + PAN

2021 / update 2024

Poprzednia wersja przygotowana dla
Akademia Innowacyjnych Zastosowań Technologii Cyfrowych (AI-TECH) projekt finansowany z środków Programu Operacyjnego Polska Cyfrowa POPC.03.02.00-00-0001/20

Fundusze Europejskie
Polska Cyfrowa

Rzeczpospolita Polska

Unia Europejska
Europejski Fundusz
Rozwoju Regionalnego

Plan wykładu

1. Liniowa separowalność w statystycznej klasyfikacji
2. Podejścia klasyczne − Fisher Liniowa Analiz Dyskryminacyjna
3. LDA = sformułowanie probabilistyczne
4. Podstawy matematyczne metody SVM
5. Sformułowanie zadania optymalizacji

Kolejny wykład

1. Uogólnienie SVM (nie w pełni separowalne liniowo)
2. Funkcje jądrowe (tzw. kernal functions)
3. SVM dla danych z nieliniowymi granicami
4. Podsumowanie
5. Gdzie szukać więcej?

Przypomnienie

Poprzedni wykład:

Klasyfikatory liniowe g(**x**)=**w**^{T}**x**+b

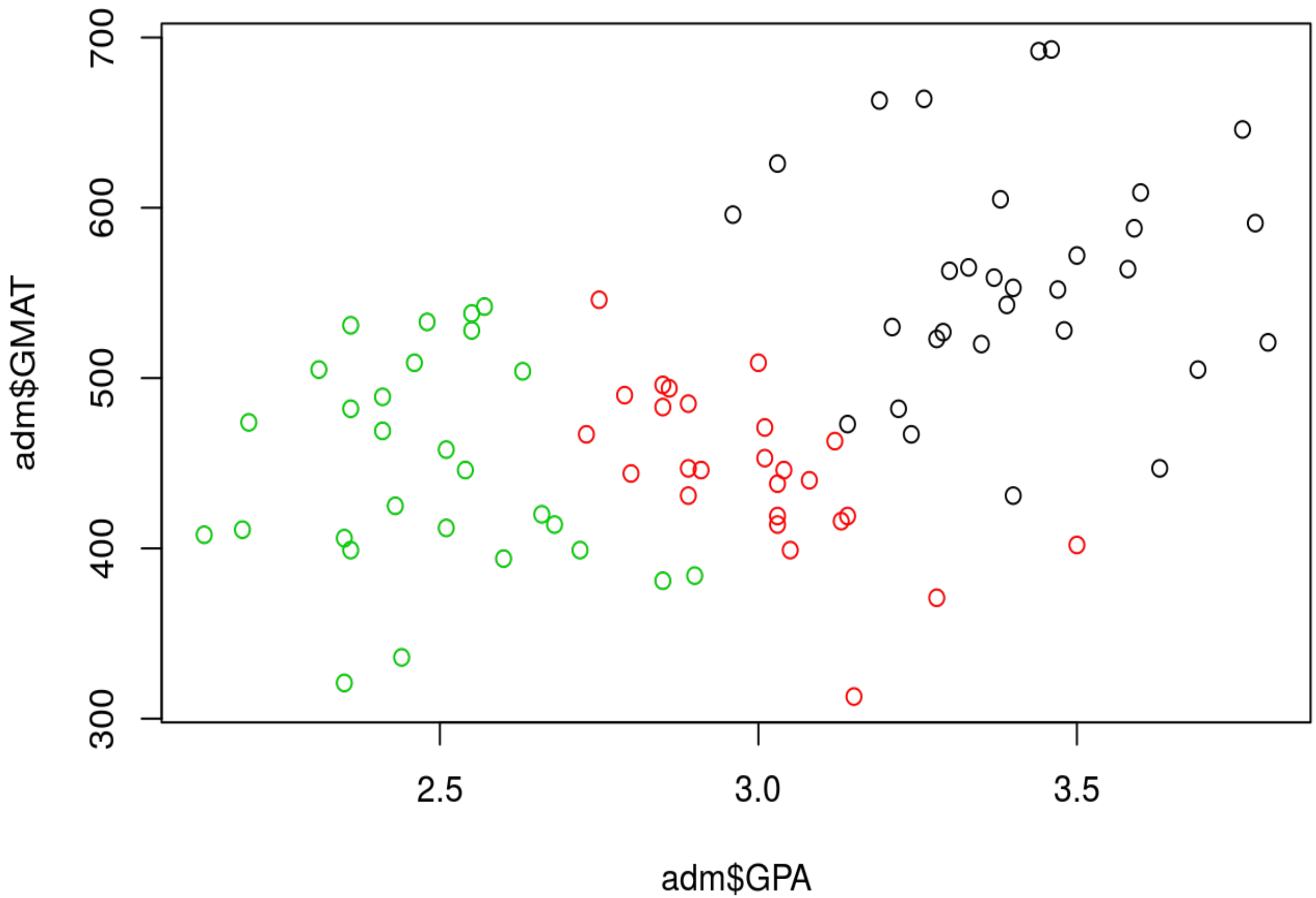
W wielowymiarowej przestrzeni danych *S* znajdują się przykłady **x** stanowiące próbę uczącą D, należące do dwóch klas

$$D = \left\{ \left(\mathbf{x}_i, c_i \right) \mid x_i \in R^p, c_i \in \{1, -1\} \right\}_{i=1}^{N}$$

Szukamy klasyfikatora pozwalającego na podział całej przestrzeni S na dwa rozłączne obszary odpowiadającej klasom {1,-1} oraz pozwalającego jak najlepiej klasyfikować nowe obiekty x do klas

Podejście opiera się na znalezieniu tzw. granicy decyzyjnej między klasami $\rightarrow$ funkcja g(x)

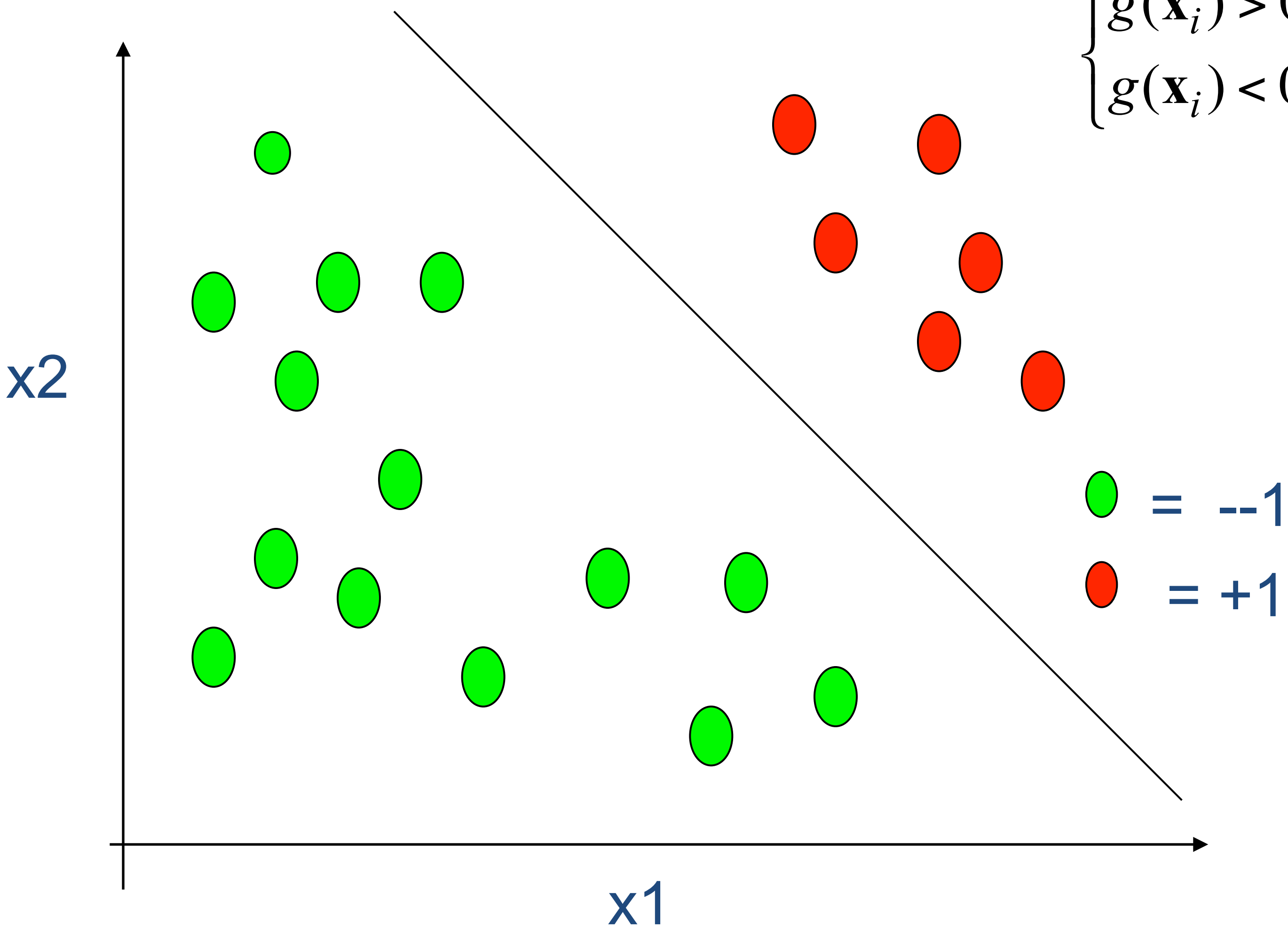
Przykład – dokumentacja R nt. Analizy dyskryminacyjnej



- Przyjęcie do amerykańskich uczelnii biznesowych − wybrane atrybuty wskaźniki GPA oraz GMAT oraz trzy klasy kandydatów (admit, notadmit, and borderline).
- 'http://www.biz.uiowa.edu/faculty/jledolter/DataMining/admission.csv'

Separowalność liniowa

$$\begin{cases} g(\mathbf{x}_i) > 0 & \mathbf{x}_i \in 1 \\ g(\mathbf{x}_i) < 0 & \mathbf{x}_i \in -1 \end{cases}$$



Sformułowanie problemu Fisher LDA

Cel

- Maksymalizuj odległość zrzutowanych średnich klas
- Minimalizuj wariancje wewnątrz klasową

- Odległość między rzutami średnich

$$(\mathbf{w}^T\bar{\mathbf{x}}_1 - \mathbf{w}^T\bar{\mathbf{x}}_2)^2$$

- Fisher założył, że obie klasy mają taką samą macierz kowariancji $S=S_1+S_2$. Dlatego wskaźnik zmienności wewnątrzgrupowej (wspólnej dla obu klas) zdefiniowany jest jako:

$$S_W = \frac{1}{n-2}\sum_{k=1}^{2}(n_k - 1)S_k$$

- Pamiętaj, że po rzutowaniu mamy $\mathbf{w}^T S_W \mathbf{w}$

Sformułowanie problemu Fiszerowskiej LDA

- W celu maksymalizacji odległości rzutów średnich klas i minimalizacji wariancji wewnątrzklasowej należy poszukiwać wektora w który maksymalizuje następujące wyrażenie:

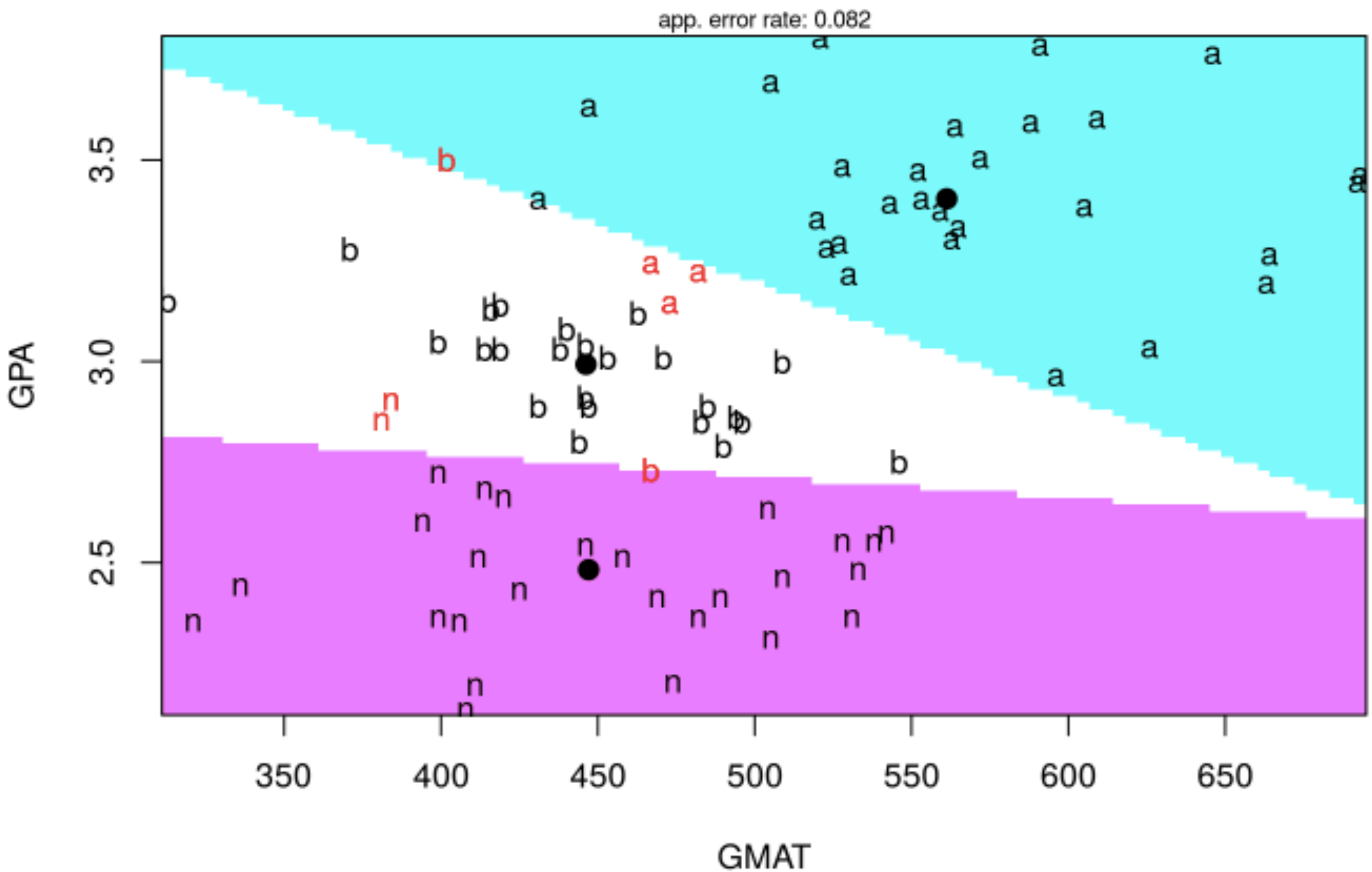
$$J(\mathbf{w}) = \frac{\left(\mathbf{w}^T \overline{\mathbf{x}}_{\mathbf{1}} - \mathbf{w}^T \overline{\mathbf{x}}_2\right)^2}{\mathbf{w}^T \mathbf{S}_{\mathbf{W}} \mathbf{w}}$$

- $\mathbf{S_w}$ − macierz kowariancji
- Ostatecznie liniowa funkcja dyskryminacyjna Fishera:

$$y = (\overline{\mathbf{x}}_1 - \overline{\mathbf{x}}_2)^T S_W^{-1} [\mathbf{x} - \frac{1}{2}(\overline{\mathbf{x}}_1 + \overline{\mathbf{x}}_2)]$$

- Dalej: uogólnienia na wiele klas;
- Quadratic Linear Disciminant Analysis
- Sformułowanie Bayesowskie − patrz wykłady na przedmiocie obieralnym inżynierskim

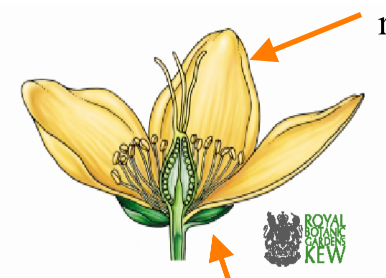
Dane o przyjęciach do uczelni biznesowych



Przykład – dane IRIS

R. Fisher iris data set

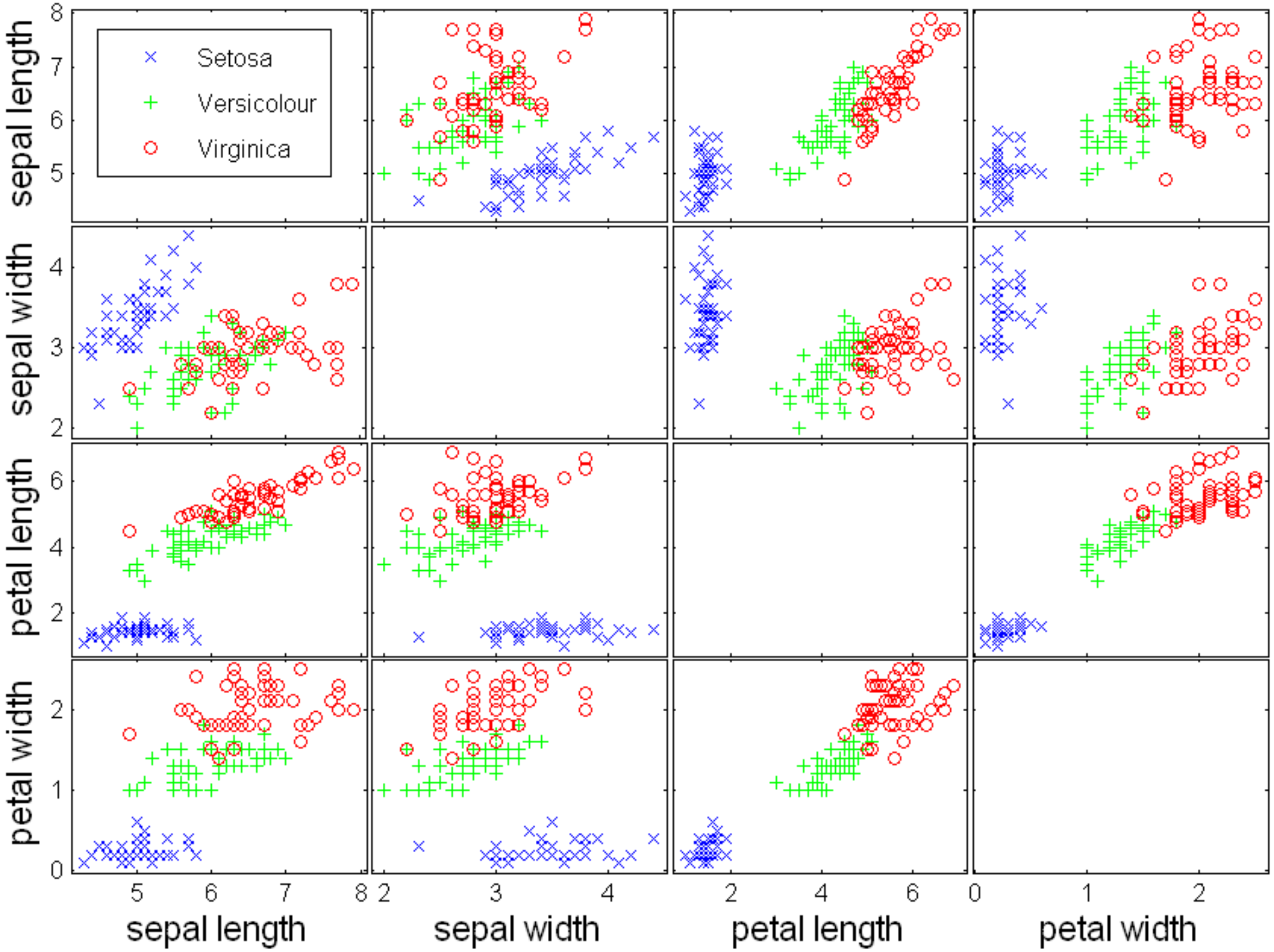
- 150 observations of 4 variables (length, width of petal and sepal)

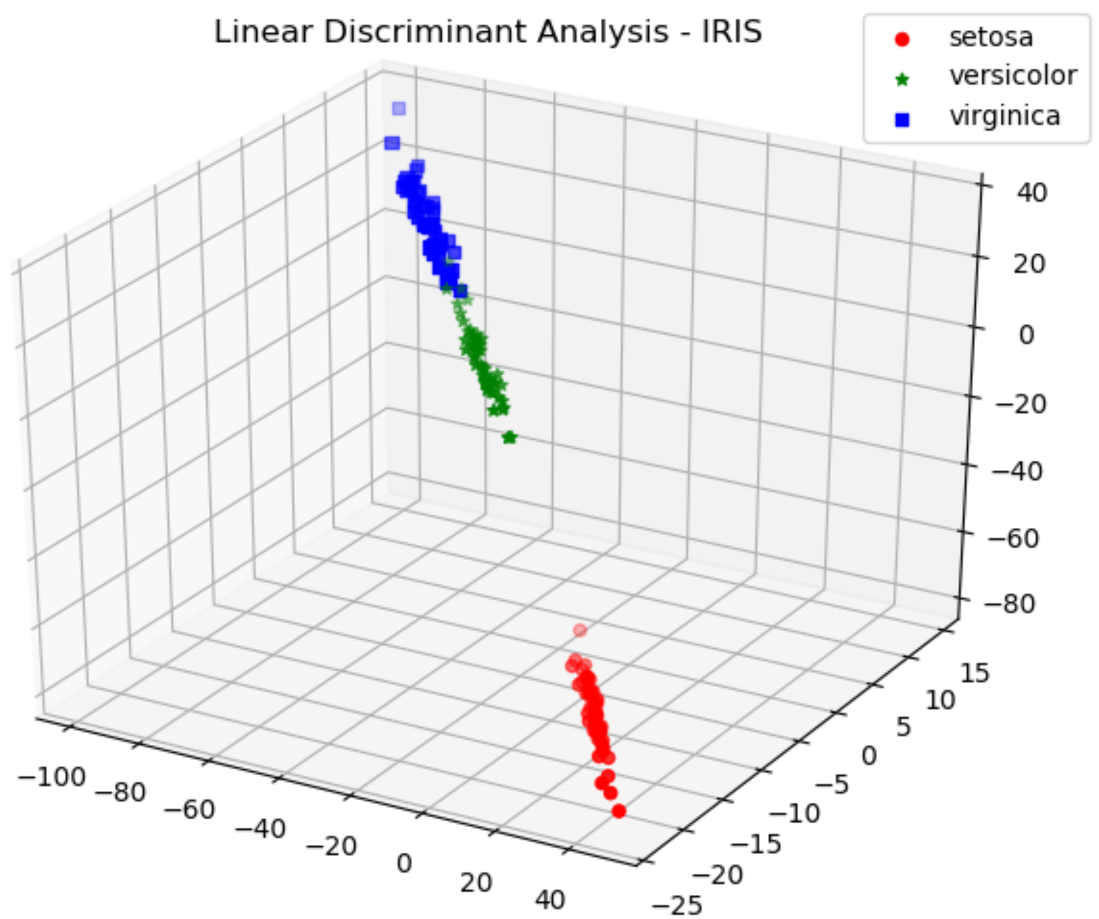


```
4 150
sepal_length
sepal_width
petal_length
petal_width
4.3  7.9  5
2.0  4.4  5
1.0  6.9  5
0.1  2.5  5
5.1    3.5    1.4    0.2
4.9    3      1.4    0.2
4.7    3.2    1.3    0.2
4.6    3.1    1.5    0.2
5      3.6    1.4    0.2
5.4    3.9    1.7    0.4
4.6    3.4    1.4    0.3
5      3.4    1.5    0.2
4.4    2.9    1.4    0.2
4.9    3.1    1.5    0.1
5.4    3.7    1.5    0.2
4.8    3.4    1.6    0.2
4.8    3      1.4    0.1
4.3    3      1.1    0.1
5.8    4      1.2    0.2
5.7    4.4    1.5    0.4
```

Challenge in visualization is to design the visualization to match the analytical task

Scatter Plot Array of Iris Attributes





Linear Discriminant Analysis - IRIS
setosa
versicolor
virginica
−100
−80
−60
−40
−20
0
20
40
15
10
5
0
−5
−10
−15
−20
−25
40
20
0
−20
−40
−60
−80

Porównanie rozwiązań LDA i QDA

Wybrany zbiór danych (za Hastie et al. Elements of Statistical Learning)

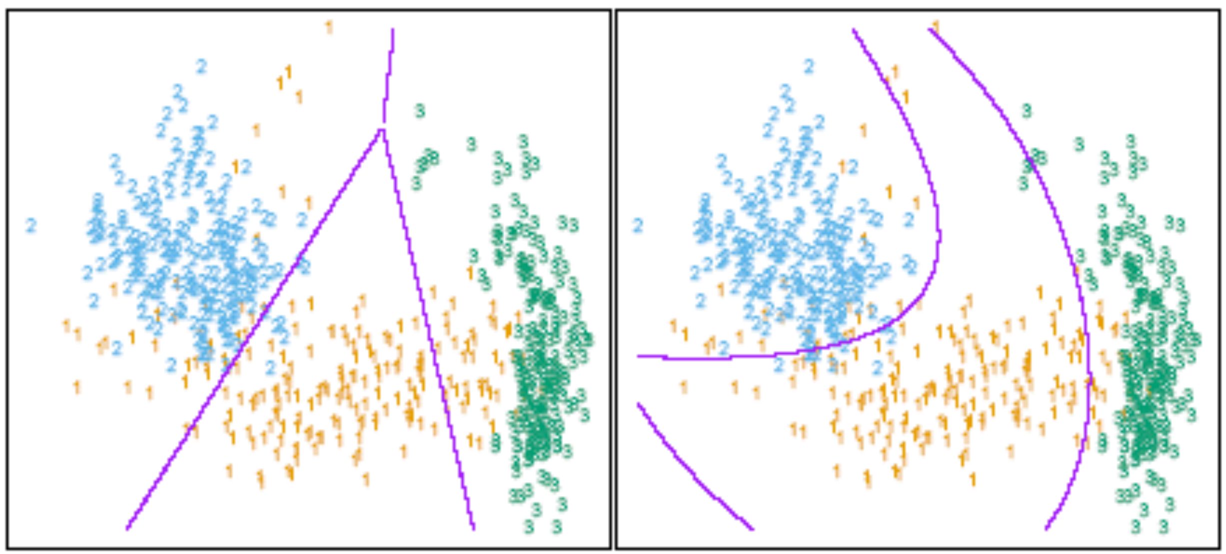


FIGURE 4.1. *The left plot shows some data from three classes, with linear decision boundaries found by linear discriminant analysis. The right plot shows quadratic decision boundaries. These were obtained by finding linear boundaries in the five-dimensional space $X_1, X_2, X_1X_2, X_1^2, X_2^2$. Linear inequalities in this space are quadratic inequalities in the original space.*

Wymogi stosowania modeli AD

- Zmienne wyrażone na skalach liczbowych
  - Specjalne podejścia dla zmiennych jakościowych (binaryzacja, model lokacyjny,…)
- Zmienne mają wielowymiarowy rozkład normalnych
- Macierze kowariancji dla poszczególnych klas są równe $\rightarrow$ jeśli nie, to bardziej złożone funkcje kwadratowe dyskryminujące.
- Problem doboru właściwych zmiennych.

Selekcja zmiennych

- W funkcji dyskryminującej uwzględniaj zmienne o dobrych właściwościach dyskryminujących
- Przykład kryterium jakości dyskryminacji:

$$\lambda = \frac{|S_W|}{|S_W + S_B|}$$

gdzie macierz zmienności wewnątrzklasowej

$$S_W = \frac{1}{n-k} \sum_{j=1}^{k} \sum_{i \in C_j} (x_i - \bar{x}_j)(x_i - \bar{x}_j)^T$$

a macierz zmienności międzyklasowej

$$S_B = \frac{1}{k-1} \sum_{j=1}^{k} n_j (\bar{x}_j - \bar{x})(\bar{x}_j - \bar{x})^T$$

Więcej

Przeczytaj literaturę

- T.Hastie, R.Tibshirani, J.Friedman: The Elements of Statistical Learning. Springer (zwłaszcza rozdz. 4) → poszukaj wersji elektronicznej pdf
- J.Koronacki, J.Ćwik: Statystyczne systemy uczące się (rozdz. 1 oraz o FDA w rozdz. 6)
- M.Krzyśko, W.Wołyński, T.Górecki,M.Skorzybut: Systemy uczące się. + wcześniejsze prace M.Krzyśko o analizie dyskryminacyjnej
- **Angielska Wikipedia „Linear discriminant analysis”**
- McLachlan, G. J. (2004). Discriminant Analysis and Statistical Pattern Recognition. Wiley.
- Duda, R. O.; Hart, P. E.; Stork, D. H. (2000). Pattern Classification (2nd ed.). Wiley

Separowalność liniowa - co dalej?

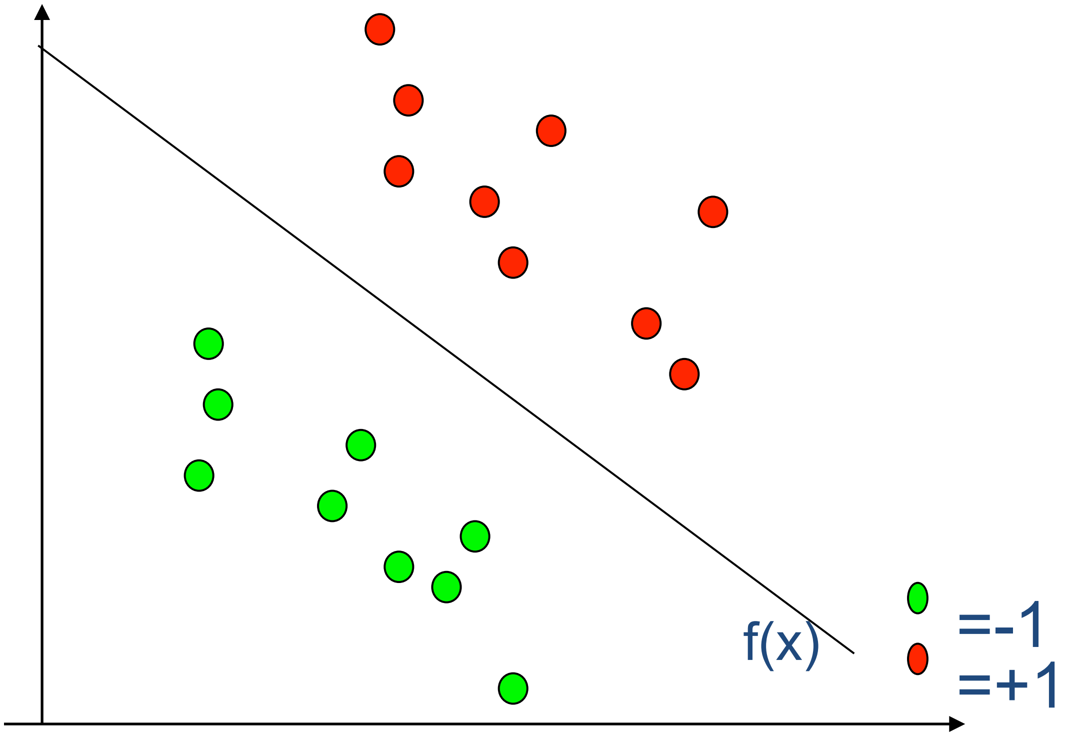
- Wiele rozwiązań – potencjalnie nieskoczenie
- Niektóre granice decyzyjne mogą być lepsze niż inne, z uwagi na pozycje wobec rozkładu przykładów z klas.
- Przejdźmy w stronę metody SVM (V.Vapnik) − podejście dyskryminacyjne (nie probabilistyczne) − specjalne podejście optymalizacji wektora wag modelu liniowego granicy decyzyjnej

SVM kontekst

Nowe spojrzenie na tworzenie klasyfikatorów liniowych

- Przed latami 80tymi
  - Popularność modeli liniowych (często w algorytmach, także analizy teoretyczne)
- Lata 80te
  - Początek popularności ANN, drzew decyzyjnych, innych metod heurystycznych (brak optymalności)
- Od lat 90tych
  - Nowe algorytmy oparte na teorii systemów uczących
  - Optymalność przy pewnych założeniach
  - Uogólnienie na problemy nieliniowe z wykorzystaniem tzw. kernelizacji

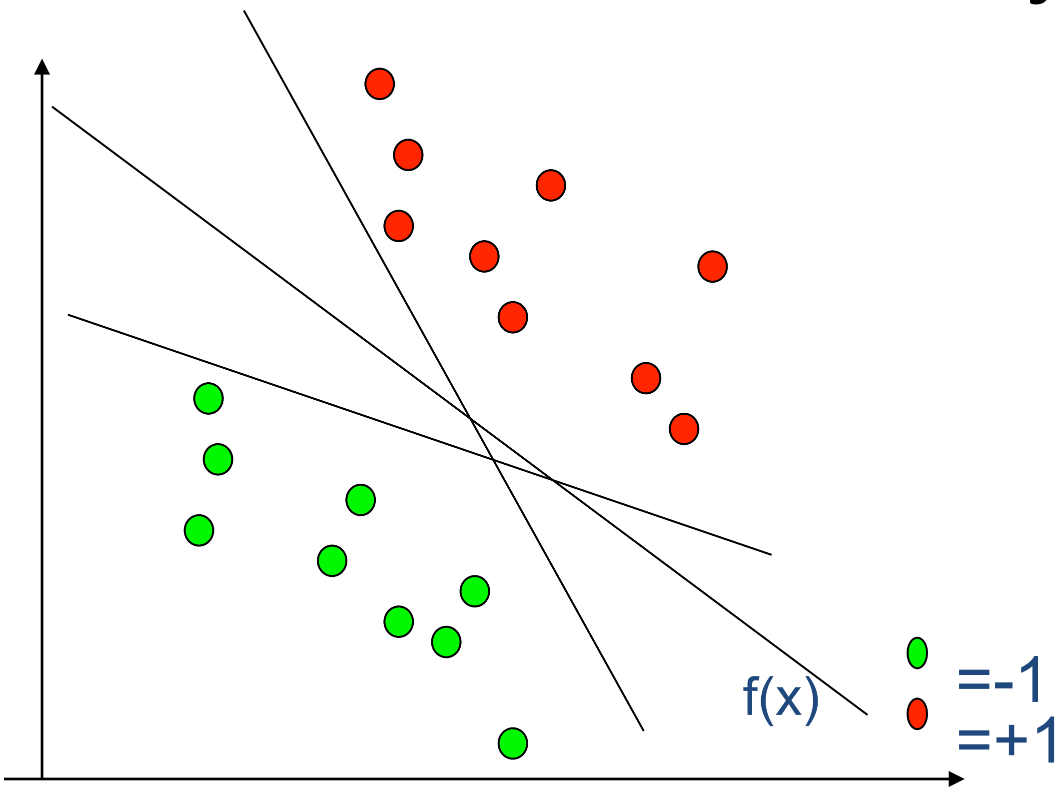
Support Vector Machines – V.Vapnik



Znajdź liniową hiperpłaszczyznę (ang. decision boundary) oddzielające obszary przykładów z dwóch różnych klas

Rysunek - jedno z możliwych rozwiązań, lecz pomyśl czy mogą być inne granice decyzyjne?

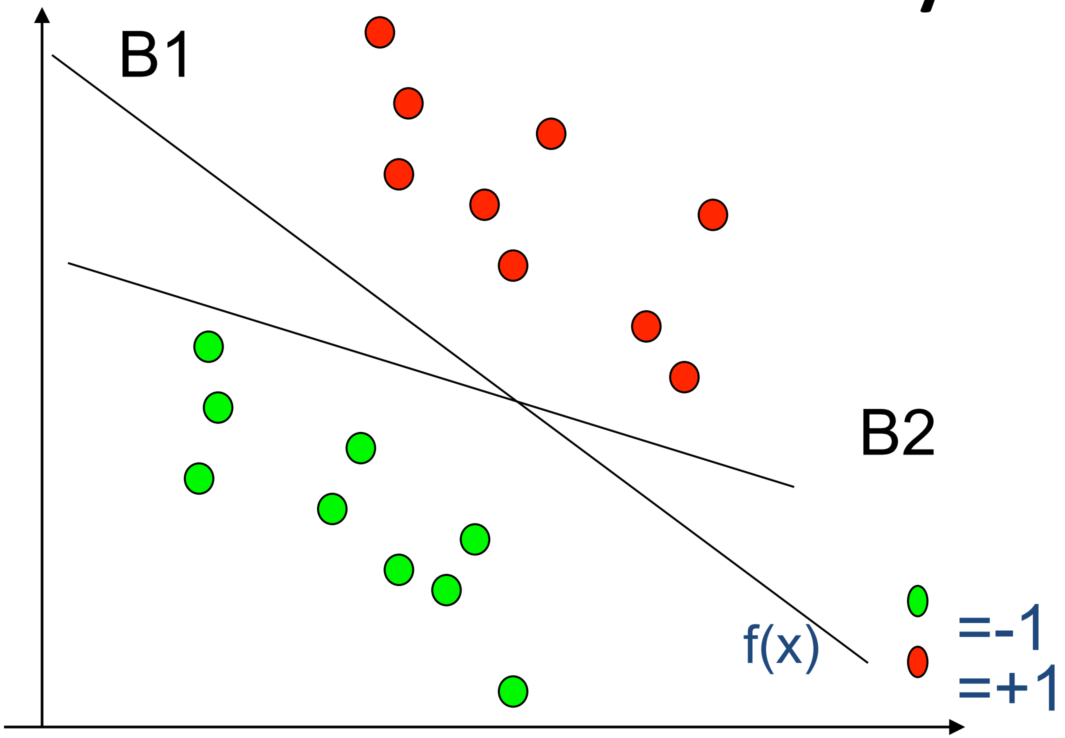
Liniowa wersja SVM



Dane: $<\mathbf{x}_i, y_i>$, $i=1,..,l$
$\mathbf{x}_i \in R^d$
$y_i \in \{-1,+1\}$

Istnienie nieskończenie wiele hiperpłaszczyzn separujących
Którą wybrać?
Być może rozważać wiele i „złożyć” je jako ważone podejście Bayesowskie? – nie jest to efektywne.
Poszukujemy jednej hiperpłaszczyzny!

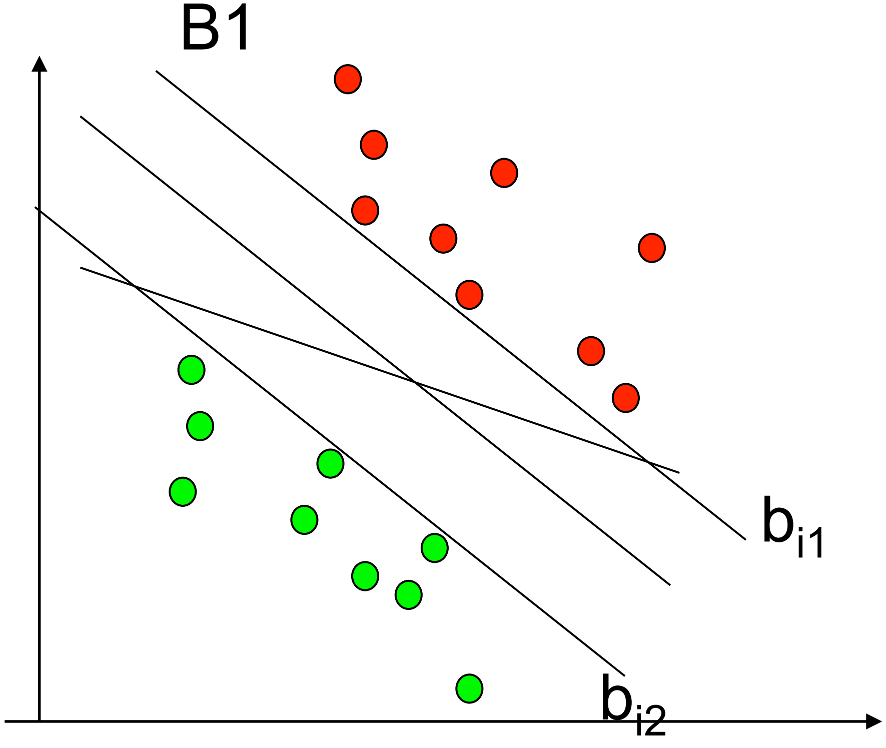
Liniowy SVM



Którą hiperpłaszczyznę B1 czy B2 – byśmy wybrali i dlaczego?

Uwagi o marginesie w SVM

- Hiperpłaszczyzny b_{i1} i b_{i2} są otrzymane przez równoległe przesuwanie hiperpłaszczyzny granicznej aż do pierwszych punktów z obu klas.
- Odległość między nimi – **margines** klasyfikatora liniowego
  - Odpowiednik odległości pomiędzy granicą a najbliższymi przykładami
- Jaki margines wybierać?



Uwagi o marginesie w SVM

- Hiperpłaszczyzny b_{i1} i b_{i2} są otrzymane przez równoległe przesuwanie hiperpłaszczyzny granicznej aż do pierwszych punktów z obu klas.
- Odległość między nimi – **margines** klasyfikatora liniowego
- Jaki margines wybierać?

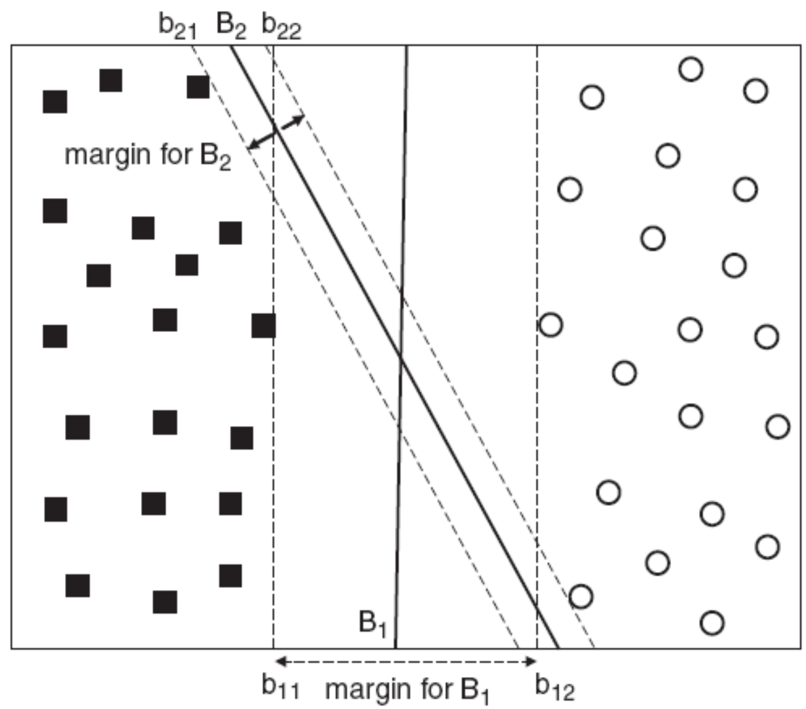


Figure 5.22. Margin of a decision boundary.

Węższe czy szersze marginesy?

- Szerszy margines $\rightarrow$ lepsze własności generalizacji, mniejsza podatność na ew. przeuczenie (overfitting)
- Wąski margines – mała zmiana granicy, radykalne zmiany klasyfikacji
- Bardziej formalnie tzw. wymiar VC, teoria Vapnik–Chervonenkis
- Znajdź hiperpłaszyczną, która maksymalizuje tzw. margines => B1 jest lepsze niż B2

- Uwaga − założenie w obszarze marginesu nie ma przykładów uczących (powinny być poza przesuniętymi hiperpłaszczyznami)

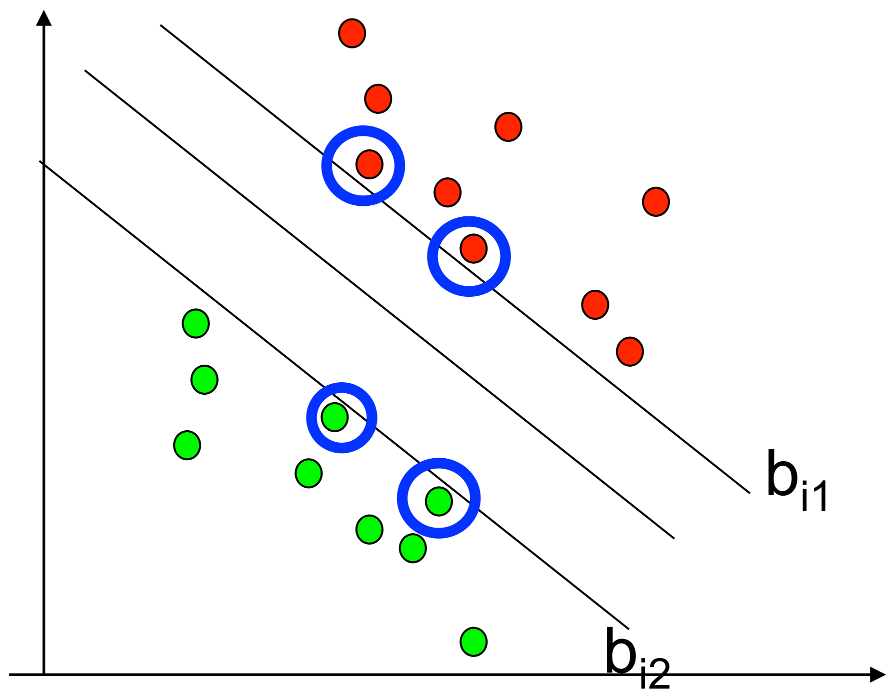
Teoria „Structural risk minimization"

- Oszacowanie górnej granicy błędu ze względu na błąd uczący R_e, liczbę przykładów N i tzw. model complexity h z prawdopodobieństwem 1-η „generalization error" nie przekroczy:

$$R \leq R_e + \varphi\left(\frac{h}{N}, \frac{\log(\eta)}{N}\right)$$

- Prace teoretyczne – *h* complexity dla modelu liniowego:
  - „Models with small margins have higher capacity - complexity because they are more flexible and can fit many training sets"
- Także „The hypothesis space with minimal **VC-dimension** according to SRM" / **teoria Vapnik–Chervonenkis**
- Reasumując modele o większej „complexity" mają gorsze oszacowanie błędu
- Dlatego wybieraj większy margines!

Wektory nośne (ang. support vectors)



Przykłady uczące wspierające przesunięte hiperpłaszczyzny (najbliższe granicy decyzyjnej) ; Ponadto będą pełnić specjalną rolę w rozwiązaniu zadania optymalizacji parametrów modelu.

Wektory nośne

- Punkty położone najbliżej granicy decyzyjnej
- Założenie: są najtrudniejsze do nauczenia się (dyskusyjne, ale podstawa metody SVM)
- Mają bezpośredni wpływ na znalezienie optymalnego rozwiązania i położenia hiperpłaszczyny (iloczyn wag w i cech opisu przykładu x)
  - Inne przykłady dużo mniejszy wpływ na rozwiążanie
- Optymalny margines – przypisuje największe wagi najistotniejszym cechom dla separowalności liniowej.

Liniowe SVM hiperpłaszczyzna graniczna

- Vapnik – poszukuj „maximal margin classifier”

$$\mathbf{w} \cdot \mathbf{x} + \mathbf{b} = 0$$

gdzie **w** i **b** są parametrami modelu

$$y = \begin{cases} 1 & \mathbf{w} \cdot \mathbf{x} + \mathbf{b} > 0 \\ -1 & \mathbf{w} \cdot \mathbf{x} + \mathbf{b} < 0 \end{cases}$$

- Parametry granicy wyznaczaj tak, aby maksymalne marginesy odpowiadały b_{i1} i b_{i2} i były miejscem geometrycznym punktów **x** spełniających warunki

$$b_{i1} \quad \mathbf{w} \cdot \mathbf{x} + \mathbf{b} = 1$$

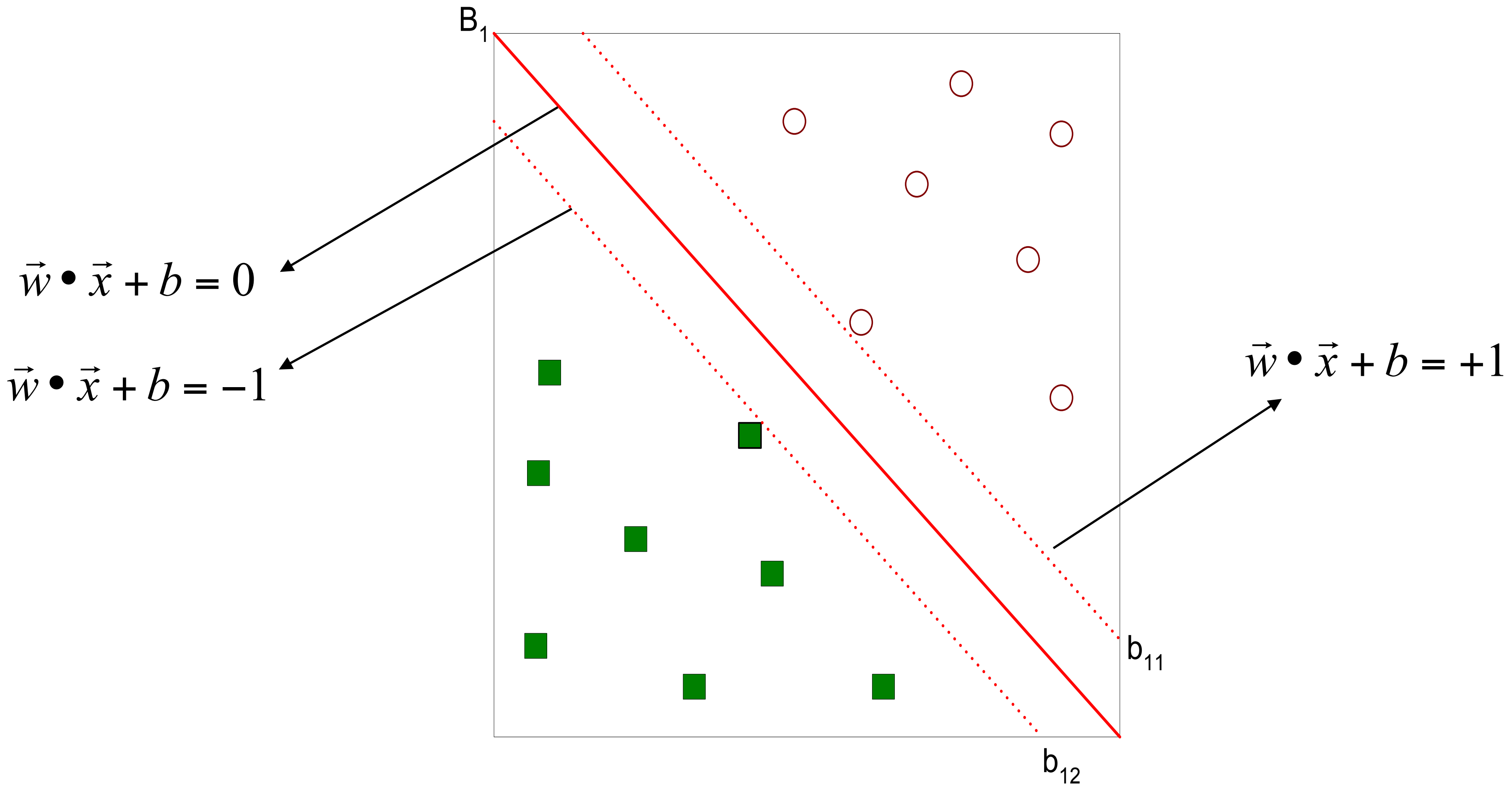
$$b_{i2} \quad \mathbf{w} \cdot \mathbf{x} + \mathbf{b} = -1$$

- **Margines** – odległość między płaszczyznami b_{i1} i b_{i2}

Komentarze

- hiperpłaszczyzny b_{i1} i b_{i2} muszą zawierać wektory nośne
- H − granica decyzyjna − „mediana” między nimi
- b+ najmniejsza odległość z H do najbliższych przykładów pozytywnych
- b- najmniejsza odległość z H do najbliższych przykładów negatywnych
- Tylko przesunięcie wektorów nośnych może zmienić położenie hiperpłaszczyzn
- Odległość między b_{i1} i b_{i2} (margines) może być przeformułowane na $2/||\mathbf{w}||$ = norma wektora wag

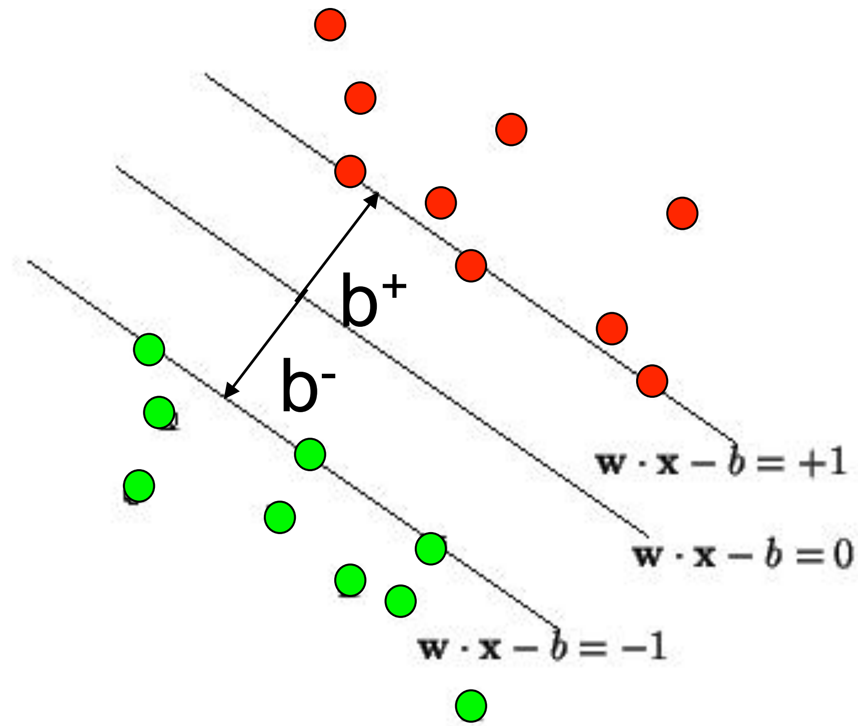
Poszukiwanie parametrów hiperpłaszczyzny



$$f(\vec{x}) = \begin{cases} 1 & \text{if } \vec{\mathrm{w}} \bullet \vec{\mathrm{x}} + \mathrm{b} \geq 1 \\ -1 & \text{if } \vec{\mathrm{w}} \bullet \vec{\mathrm{x}} + \mathrm{b} \leq -1 \end{cases}$$

L-SVM ilustracja i margines

$$f(\vec{x}) = \begin{cases} 1 & \text{if } \mathrm{w}\vec{\mathrm{x}} + \mathrm{b} \geq 1 \\ -1 & \text{if } \vec{\mathrm{w}}\vec{\mathrm{x}} + \mathrm{b} \leq -1 \end{cases}$$



Przekształcenia marginesu:

$\mathbf{wx_1}+\mathbf{b}$=+1 oraz $\mathbf{wx_2}+\mathbf{b}$=-1 =>

$\mathbf{w}(\mathbf{x_1}-\mathbf{x_2})$ +b − b = +1 −(1)

$\mathbf{w}(\mathbf{x_1}-\mathbf{x_2})$ = 2 =>

$$\frac{w}{\|w\|}(x_1 - x_2) = \frac{2}{\|w\|}$$

$$\textbf{margin} = \frac{2}{\|\mathbf{w}\|}$$

Ilustracje i sposób przekształceń

$$\textbf{margin} = \frac{2}{\|\mathbf{w}\|} \qquad \|\mathbf{w}\| \equiv \sqrt{w_1^2 + \ldots + w_p^2}$$

Cel: Maksymalizuj margines!

$$\frac{2}{\|\mathbf{w}\|} \longrightarrow \frac{\|\mathbf{w}\|}{2} \longrightarrow \frac{\|\mathbf{w}\|^2}{2}$$

maximize **minimize** **minimize**

L-SVM zadanie optymalizacji

Sformułowanie mat. problemu:

$$\min_{\mathbf{w}} = \frac{\| \mathbf{w} \|^2}{2}$$

– Przy warunkach ograniczających

$$y_i(w \cdot x_i + b) \geq 1 \quad i = 1,2,\ldots,N$$

– Jest to problem optymalizacji kwadratowej z liniowymi ogr. $\rightarrow$ uogólnione zadanie optymalizacji rozwiązywany metodą mnożników Lagrange'a (tak aby, np. nie dojść do **w** $\rightarrow$ 0)

LSVM – mnożniki Lagrange’a

- Minimalizuj funkcję Lagrange’a

$$L(w,b,\alpha)=\frac{1}{2}\|\mathbf{w}\|^2-\sum_{i=1}^{N}\alpha_i(y_i(\mathbf{wx}_i+b)-1)$$

- parametry $\alpha \geq 0$ mnożniki Lagrange’a
- Powinno się różniczkować *L* po **w** i *b* – nadal trudności w rozwiązaniu
- Przy przekształceniach wykorzystuje się ograniczenia Karush-Kuhn-Tucker na mnożniki:

$$\alpha_i \geq 0$$

$$\alpha_i[y_i(w\cdot x_i+b)-1]=0$$

- W konsekwencji α_i są niezerowe wyłącznie dla **wektorów nośnych x**, pozostałe są zerowe
- Rozwiązanie parametrów **w** i *b* zależy wyłącznie od wektorów nośnych!

LSVM – sformułowanie dualne

- Nadal zbyt wiele parametrów **w**, **b**, α do oszacowania
- Przechodzi się na postać *dualną zadania optymalizacji*
- Maksymalizuj L(α)

$$\sum_{i=1}^{N} \alpha_i - \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N} \alpha_i \alpha_j y_i y_j (\mathbf{x}_i \cdot \mathbf{x}_j)$$

- przy ograniczeniach

$$\alpha_i \geq 0, \ \forall i \quad \sum_{i=1}^{N} \alpha_i y_i = 0$$

- Rozwiązanie (α>0 dla i$\in$SV) ; b – odpowiednio uśredniane

$$\mathbf{w} = \sum_{i=1}^{N} \alpha_i y_i \mathbf{x}_i$$

- Hiperpłaszczyzna decyzyjna

$$\sum_{i=1}^{N} \alpha_i y_i \mathbf{x}_i \cdot \mathbf{x} + b = 0$$

Rozwiązanie LSVM

- Klasyfikacja – funkcja decyzyjna

$$f(x) = sign(\sum_{i=1}^{N} \alpha_i y_i \mathbf{x}_i \cdot \mathbf{x} + b)$$

- O ostatecznej postaci hiperpłaszczyzny decydują wyłącznie wektory nośne (α_i >0)
- Im większa wartość α_i , tym większy wpływ wektora na granicę decyzyjną
- Klasyfikacja zależy od iloczynu skalarnego nowego **x** z wektorami nośnymi $\mathbf{x}_i$ ze zbioru uczącego
- Pewne założenie metody – starać się zbudować klasyfikator liniowy używając możliwie minimalną liczbę wektorów z danych uczących (wektory nośne)

Przykład

x_1	x_2	y	Lagrange Multiplier
0.3858	0.4687	1	65.5261
0.4871	0.611	–1	65.5261
0.9218	0.4103	–1	0
0.7382	0.8936	–1	0
0.1763	0.0579	1	0
0.4057	0.3529	1	0
0.9355	0.8132	–1	0
0.2146	0.0099	1	0

Obliczmy wagi

$$w_1 = \sum_i \alpha_i y_i x_{i1} = 65.5621 \cdot 1 \cdot 0.3858 + 65.5621 \cdot (-1) \cdot 0.4871 = -6.64$$

$$w_2 = \sum_i \alpha_i y_i x_{i2} = 65.5621 \cdot 1 \cdot 0.4687 + 65.5621 \cdot (-1) \cdot 0.611 = -9.32$$

$$b' = 1 - (-6.64) \cdot 0.3858 - (-9.32)(0.4687) = 7.930$$

$$b'' = -1 - (-6.64) \cdot 0.4871 - (-9.32)(0.611) = 7.928$$

$$b = 0.5 \cdot (b' + b'') = 7.93$$

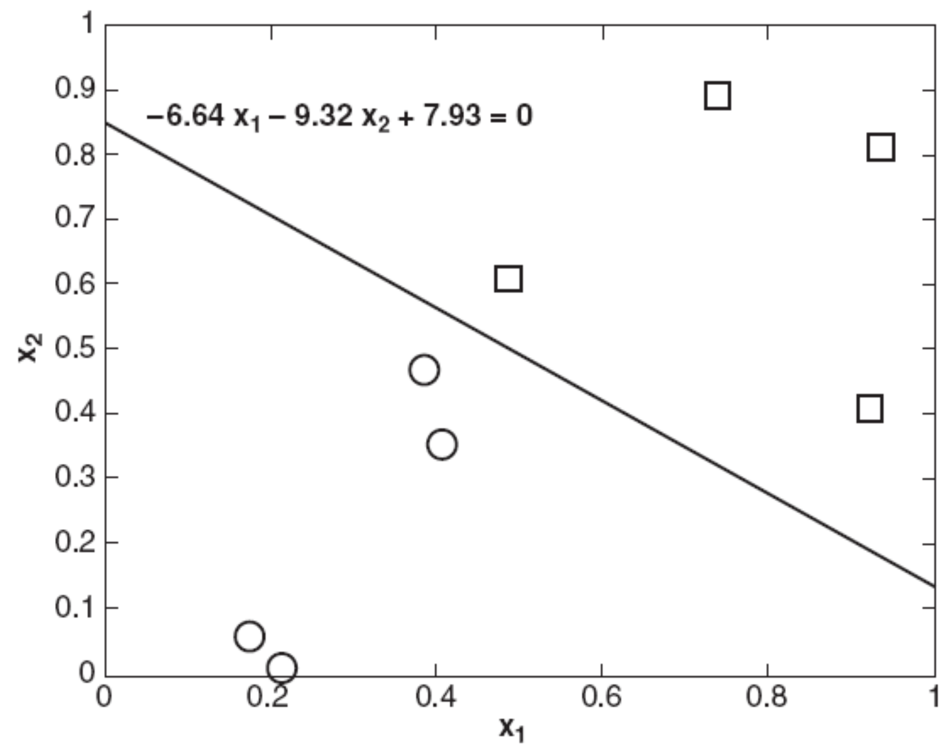
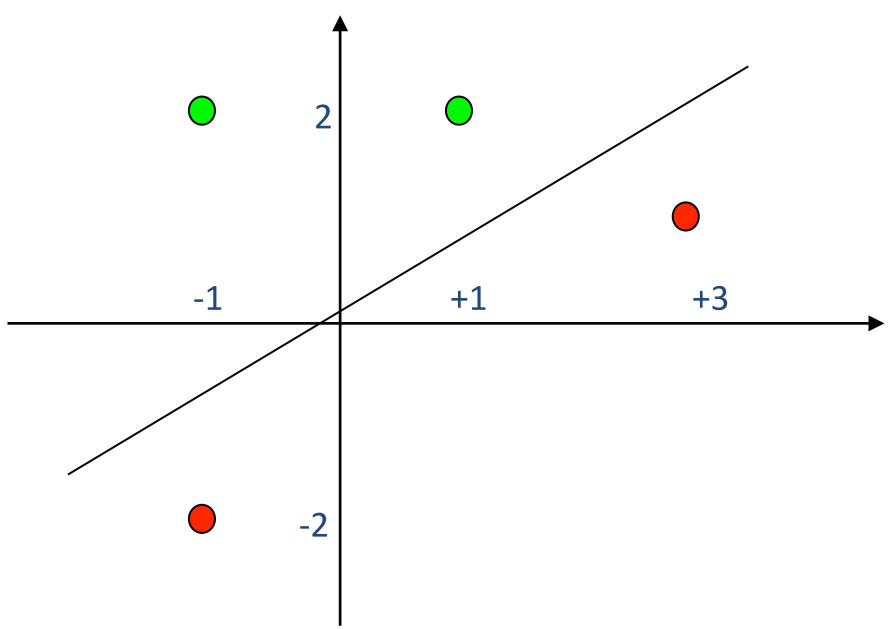


Figure 5.24. Example of a linearly separable data set.

Inny przykład

- Rozważmy 4 dwuwymiarowe (a1,a2,y) przykłady uczące: (1,2,-1); (-1,2,-1);(-1,-2;+1); (3,1,+1)
- Obliczenia współczynników Lagrange'a : α1=1/2 ; α2 = 0; α3=1/10 ; α4=2/5
- Ostateczny wektor wag w1=3/5 oraz w2=-4/5



Parę uwag podsumowujących

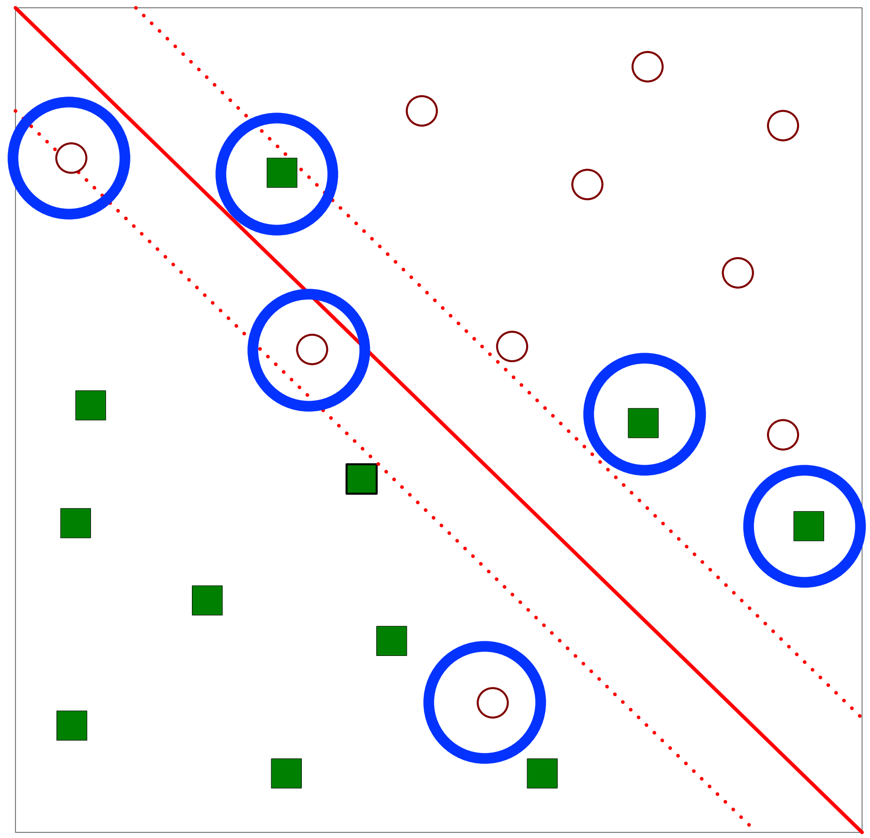
- Liniowy SVM − solidne matematycznie rozwiązanie problemu liniowo separowalnego
  - Podejście nazywane Hard margin SVM
  - Założenie o silnym znaczeniu przykładów brzegowych / inne przykłady we wnętrzu rozkładu klas nie mają takiego znaczenia (pomyśl na realnością i różnymi możliwymi rozkładami przykładów w klasach)
  - Związek z teoretycznymi pracami (np. wymiar VC)
- Użycie zmodyfikowanego sformułowanie zadania programowanie kwadratowego
  - Jedno globalne rozwiązanie
  - Algorytm optymalizacyjny
- Teoretycznie poszukują minimum globalnego a nie lokalnego (jak podejścia heurystyczne – MLP)

Dalsze pytania

- Liniowy SVM − praktycznie nie musi być konkurencyjny dla znanych wcześniej metod linowych (LDA, ANN, itd.)
- Otwarte pytania na kolejny wykład:
- Jak podejść do problemów nieseparowalnych (zmienne osłabiające i przeformułowanie zadania) − tzw. soft margin SVM
- Nieliniowe, trudne rozkłady przykładów
  - Dane odwzorowane (przy pomocy funkcji jądrowych) w nową przestrzeń cech – silna przewaga nad innymi metodami
  - W nowej przestrzeni dane powinny być liniowo separowalne

Support Vector Machines – dalej!

Co robić z LSVM gdy dane nie są w pełni liniowo separowalne?



Odnośniki do literatury anglojęzycznej

- "Statistical Learning Theory" by Vapnik − wymaga przygotowania matematycznego.
- C. J. C. Burges. A Tutorial on Support Vector Machines for Pattern Recognition. *Knowledge Discovery and Data Mining*, 2(2), 1998.
  - Czytelniejsza niż powyższa książka
- R. Berwick: An Idiot's guide to Support vector machines (SVMs) − łatwy wykład po ang.
- Książka "An Introduction to Support Vector Machines" by N. Cristianini and J. Shawe-Taylor
- Rozdziały w książce Hastie, Tibushirani, Friedman: Elements of statistical learning (dostępna online pdf).
- Po polsku: M.Krzyśko, T.Górecki i inni, książka pt. Systemy uczące się

Pytanie i komentarze?

Dalszy kontakt:

jerzy.stefanowski@cs.put.poznan.pl

http://www.cs.put.poznan.pl/jstefanowski/

Fundusze Europejskie
Polska Cyfrowa

Rzeczpospolita Polska

Unia Europejska
Europejski Fundusz Rozwoju Regionalnego