SYSTEMY UCZĄCE SIĘ

WYKŁAD 16. MASZYNA WEKTORÓW NOŚNYCH

Dr hab. inż. Grzegorz Dudek
Wydział Elektryczny
Politechnika Częstochowska

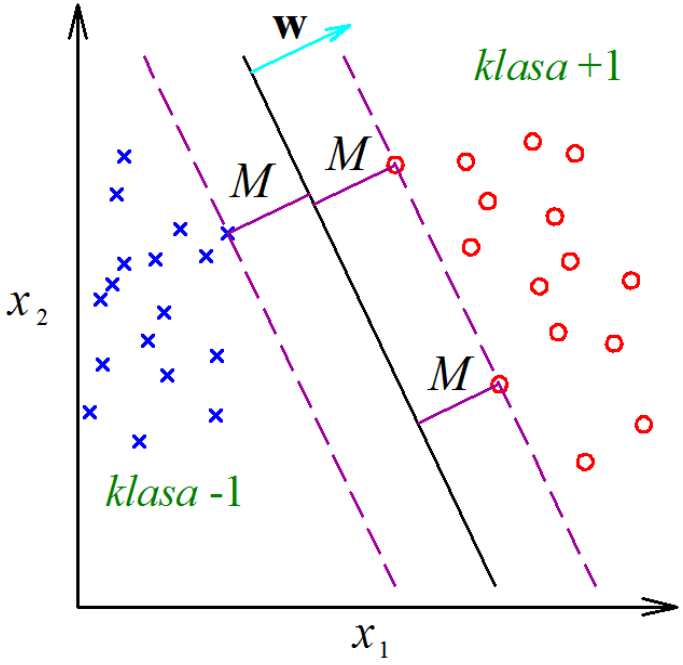
Częstochowa 2014

DEFINICJA PROBLEMU

Rozważmy problem klasyfikacji przykładów **x** do jednej z dwóch klas oznaczonych etykietami $y = +1$ lub $y = -1$.

Załóżmy, że chcemy znaleźć równanie płaszczyzny dyskryminacyjnej (decyzyjnej) separującej punkty z klasy +1 od punktów z klasy –1, ale takiej:

- która leży w środku pasma separującego te klasy i
- pasmo to ma maksymalną szerokość.



DEFINICJA PROBLEMU

Musimy więc znaleźć taką płaszczyznę leżącą pomiędzy obszarami obu klas, dla której odległość mierzona pomiędzy tą płaszczyzną a najbliższymi do niej punktami z obu klas (tzw. wektorami nośnymi/wspierającymi/podpierającymi) jest największa.

Odległość punktu od płaszczyzny wyraża wzór:

$$d = \frac{| w_1 x_1 + w_2 x_2 + \ldots + w_n x_n + w_0 |}{\sqrt{w_1^2 + w_2^2 + \ldots + w_n^2}} = \frac{| \mathbf{w}^T \mathbf{x} + w_0 |}{\|\mathbf{w}\|}$$

Wektor współczynników $\mathbf{w} = [w_1, w_{2,\ldots,} w_n]^T$ nazywamy wektorem normalnym płaszczyzny (jest on prostopadły do płaszczyzny). $\|\mathbf{w}\|$ to długość tego wektora.

Wartość bezwzględna w liczniku zapobiega ujemnym wartościom odległości dla punktów leżących po jednej stronie prostej. Jeśli przyjmiemy, że zachodzi to dla punktów z klasy –1, to odległość możemy zapisać:

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

DEFINICJA PROBLEMU

Żądamy, żeby odległość pomiędzy każdym punktem $\mathbf{x}_i$, a płaszczyzną dyskryminacyjną była nie mniejsza od pewnej wartości M zwanej marginesem:

$$\frac{y_i(\mathbf{w}^T\mathbf{x}_i + w_0)}{\|\mathbf{w}\|} \geq M \text{ , } \forall i \qquad (*)$$

Margines powinien być jak największy. Szukamy takich współczynników płaszczyzny dyskryminacyjnej, aby osiągnąć maksymalny margines. Takich rozwiązań jest nieskończenie wiele, ponieważ jest nieskończenie wiele wektorów normalnych do danej płaszczyzny (różnią się długością).

Narzućmy więc ograniczenie na długość wektora $\mathbf{w}$. Przyjmijmy, że ma ona być równa $\|\mathbf{w}\| = 1/M$. Teraz możemy zmienić cel zadania: zamiast szukać maksymalnego marginesu, szukamy minimalnej długości $\|\mathbf{w}\|$.

Przekształcamy nierówność (*):

$$\frac{y_i(\mathbf{w}^T\mathbf{x}_i + w_0)}{\|\mathbf{w}\|} \geq M \quad \rightarrow \quad \frac{y_i(\mathbf{w}^T\mathbf{x}_i + w_0)}{1/M} \geq M \quad \rightarrow \quad y_i(\mathbf{w}^T\mathbf{x}_i + w_0) \geq 1, \quad \forall i \quad (**)$$

FUNKCJA CELU

Problem optymalizacyjny zapiszemy następująco:

$$\min \frac{1}{2}\|\mathbf{w}\|^2 \quad \text{pod warunkiem} \quad y_i(\mathbf{w}^T\mathbf{x}_i + w_0) \geq 1, \quad \forall i$$

Jest to typowy problem optymalizacji kwadratowej (kwadratowa funkcja celu), który rozwiązuje się metodą Lagrange'a*. Metoda ta polega na wprowadzeniu funkcji Lagrange'a, która jest sumą oryginalnej funkcji celu i ograniczeń (dla każdego punktu) przemnożonych przez mnożniki $\alpha_i \geq 0$:

$$L_P(\mathbf{w}, w_0, \boldsymbol{\alpha}) = \frac{1}{2}\|\mathbf{w}\|^2 - \sum_{i=1}^{N} \alpha_i [y_i(\mathbf{w}^T\mathbf{x}_i + w_0) - 1] = \frac{1}{2}\|\mathbf{w}\|^2 - \sum_{i=1}^{N} \alpha_i y_i(\mathbf{w}^T\mathbf{x}_i + w_0) + \sum_{i=1}^{N} \alpha_i \quad (\#)$$

Funkcja jest minimalizowana ze względu na $\mathbf{w}$ i w_0 :

$$\frac{\partial L_P}{\partial \mathbf{w}} = 0 \quad \rightarrow \quad \mathbf{w} = \sum_{i=1}^{N} \alpha_i y_i \mathbf{x}_i \quad (\#\#)$$

$$\frac{\partial L_P}{\partial w_0} = 0 \quad \rightarrow \quad \sum_{i=1}^{N} \alpha_i y_i = 0$$

* Opis metody w: Kusiak J. i In.: Optymalizacja. PWN 2009, str. 121.

Podstawiamy wyniki tej optymalizacji do (#) i otrzymujemy (tzw. postać dualną funkcji *L*):

$$L_D = -\frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}\alpha_i\alpha_j y_i y_j \mathbf{x}_i^T \mathbf{x}_j + \sum_{i=1}^{N}\alpha_i$$

Teraz maksymalizujemy tę funkcję ze względu na α_i przy ograniczeniach: $\sum_{i=1}^{N}\alpha_i y_i = 0$ i $\alpha_i \geq 0,\ \forall i$.

W wyniku otrzymujemy zbiór mnożników α_i, przy czym tylko niektóre z nich mają wartości większe od zera. Pozostałe są wyzerowane. Te niezerowe mnożniki odpowiadają wektorom nośnym (punktom na których opiera się margines).

Mając wartości α_i z (##) możemy wyznaczyć wektor normalny płaszczyzny separującej:

$$\mathbf{w} = \sum_{i\in\Omega}\alpha_i y_i \mathbf{x}_i$$

gdzie Ω jest zbiorem indeksów wektorów nośnych.

REGUŁA DECYZYJNA SVM

Płaszczyzny przechodzące przez wektory nośne (ograniczające margines) mają postać:

$$y_i(\mathbf{w}^T\mathbf{x}_i + w_0) = 1, \quad i \in \Omega$$

Stąd:

$$w_0 = y_i - \mathbf{w}^T\mathbf{x}_i, \quad i \in \Omega$$

Zauważ, że współczynniki płaszczyzny dyskryminacyjnej zależą jedynie od wektorów nośnych i są niezależne od punktów leżących dalej (usunięcie tych punktów nie wpłynie na powierzchnię decyzyjną).

Przedstawiony algorytm klasyfikacji nosi nazwę maszyny wektorów nośnych (*Support Vector Machine*, SVM).

Ostatecznie reguła decyzyjna SVM przyjmuje postać:

$$f(\mathbf{x}) = \operatorname{sgn}(\mathbf{w}^T\mathbf{x} + w_0) = \operatorname{sgn}\left(\sum_{i\in\Omega} \alpha_i y_i (\mathbf{x}_i \cdot \mathbf{x}) + w_0\right)$$

gdzie: sgn(z) oznacza +1, gdy z > 0 i –1, gdy z < 0, a $(\mathbf{x}_i \cdot \mathbf{x})$ to iloczyn skalarny dwóch wektorów.

SVM DLA DANYCH NIESEPROWALNYCH LINIOWO

W przypadku danych nieseparowalnych liniowo poszukujemy płaszczyzny, która rozdziela obszary decyzyjne z najmniejszym błędem.

Wprowadzamy zmienne $\xi_i \geq 0$, które pozwolą osłabić ograniczenia. Możliwe są dwa rodzaje naruszenia ograniczeń:

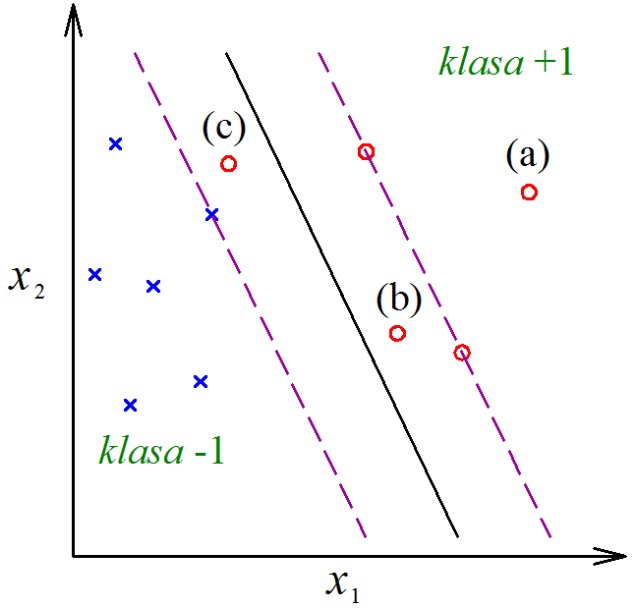
- przykład leży po niewłaściwej stronie powierzchni decyzyjnej
- przykład leży po właściwej stronie lecz na marginesie

Osłabiamy ograniczenia (**):

$$y_i(\mathbf{w}^T\mathbf{x}_i + w_0) \geq 1 - \xi_i, \quad \forall i \quad (@)$$

SVM DLA DANYCH NIESEPROWALNYCH LINIOWO

- gdy $\xi_i = 0$, $\mathbf{x}_i$ jest klasyfikowany poprawnie i leży poza marginesem lub na jego granicy (a),
- gdy $0 < \xi_i < 1$, $\mathbf{x}_i$ jest klasyfikowany poprawnie, ale leży na marginesie (b),
- gdy $\xi_i > 1$, $\mathbf{x}_i$ jest klasyfikowany błędnie (c).



SVM DLA DANYCH NIESEPROWALNYCH LINIOWO

Zdefiniujmy "miękki" błąd jako $\sum_{i=1}^{N} \xi_i$ i dodajmy go jako składnik karny do funkcji celu:

$$\frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i=1}^{N} \xi_i$$

Szukamy minimum tej funkcji przy ograniczeniach (@) oraz $\xi_i \geq 0$. Parametr $C \geq 0$ decyduje o "sile" kary.

Funkcja Lagrange'a przybiera postać:

$$L_P(\mathbf{w}, w_0, \boldsymbol{\alpha}, \boldsymbol{\mu}) = \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i=1}^{N} \xi_i - \sum_{i=1}^{N} \alpha_i [y_i(\mathbf{w}^T \mathbf{x}_i + w_0) - 1 + \xi_i] - \sum_{i=1}^{N} \mu_i \xi_i \quad (@@)$$

gdzie $\mu_i \geq 0$ jest mnożnikiem Lagrange'a dla ograniczenia związanego z wartością ξ_i.

Przyrównując pochodne L_P od $\mathbf{w}$, w_0 i ξ_i do zera otrzymujemy:

$$\frac{\partial L_P}{\partial \mathbf{w}} = 0 \quad \rightarrow \quad \mathbf{w} = \sum_{i=1}^{N} \alpha_i y_i \mathbf{x}_i$$

$$\frac{\partial L_P}{\partial w_0} = 0 \quad \rightarrow \quad \sum_{i=1}^{N} \alpha_i y_i = 0$$

$$\frac{\partial L_P}{\partial \xi_i} = 0 \quad \rightarrow \quad C - \alpha_i - \mu_i = 0$$

Ponieważ $\mu_i \geq 0$, to $0 \leq \alpha_i \leq C$. Podstawiając powyższe do (@@) otrzymamy postać dualną *L*:

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

przy ograniczeniach: $\sum_{i=1}^{N} \alpha_i y_i = 0$ i $0 \leq \alpha_i \leq C$.

Postać ta różni się od postaci rozważanej wcześniej dla liniowej separacji tylko tym, że α_i nie może być większe od *C*.

SVM DLA DANYCH NIESEPROWALNYCH LINIOWO

Maksymalizujemy L_D ze względu na mnożniki α_i i w wyniku otrzymujemy:

- $\alpha_i = 0$ dla punktów poprawnie klasyfikowanych leżących poza marginesem,
- $0 < \alpha_i < C$ dla punktów leżących na płaszczyznach granicznych marginesu,
- $\alpha_i = C$ dla punktów leżących na marginesie i błędnie klasyfikowanych.

W tym przypadku wektorami nośnymi nazywa się punkty, dla których $\alpha_i > 0$. Punkty te definiują wektor normalny **w** płaszczyzny decyzyjnej.

Dla punktów leżących na granicy marginesu mamy $\xi_i = 0$ i $y_i(\mathbf{w}^T\mathbf{x}_i + w_0) = 1$. Z tego równania możemy wyznaczyć w_0.

Parametr C dobiera się w kroswalidacji przyjmując jego wartości ze zbioru $[10^{-6}, 10^{-5}, ..., 10^{6}]$. Wartość C decyduje o kompromisie pomiędzy maksymalizacją marginesu a minimalizacją błędu. Zbyt duży prowadzi do przeuczenia modelu, zbyt mały skutkuje niedouczeniem.

Gdy dane nie są liniowo separowalne możemy:

- zastosować klasyfikator nieliniowy lub
- odwzorować dane nieliniowo do nowej przestrzeni, gdzie będą liniowo separowalne i użyć liniowego klasyfikatora[†]

Zdefiniujmy więc nowe przykłady $\mathbf{z} = [z_1, z_2, ..., z_m]$ przekształcając oryginalne przykłady za pomocą funkcji bazowej (wektorowej) $\boldsymbol{\varphi}(\mathbf{x})$:

$$\mathbf{z} = \boldsymbol{\varphi}(\mathbf{x}), \quad z_k = \varphi_k(\mathbf{x}), \quad k = 1,2,...,m$$

Przechodzimy z *n*-wymiarowej przestrzeni *X* do *m*-wymiarowej przestrzeni *Z* ($m > n$). Równania płaszczyzn dyskryminacyjnych mają w tych przestrzeniach postacie:

$$f(\mathbf{z}) = \mathbf{w}^T \mathbf{z},$$

$$f(\mathbf{x}) = \mathbf{w}^T \boldsymbol{\varphi}(\mathbf{x}) = \sum_{k=1}^{m} w_k \varphi_k(\mathbf{x})$$

[†] Podobne podejście zastosowaliśmy w aproksymacji funkcji - patrz wykład 6, slajd 16 (rozszerzona reprezentacja)

Pomijamy w zapisie w_0, przyjmując $z_1 = \varphi_1(\mathbf{x}) = 1$.

Funkcja celu ma taką sama postać jak poprzednio:

$$\frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i=1}^{N}\xi_i$$

ale ograniczenia są zdefiniowane w nowej przestrzeni:

$$y_i\mathbf{w}^T\boldsymbol{\varphi}(\mathbf{x}_i) \geq 1-\xi_i, \quad \forall i$$

Lagrangian ma postać:

$$L_P(\mathbf{w},\boldsymbol{\alpha},\boldsymbol{\mu}) = \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i=1}^{N}\xi_i - \sum_{i=1}^{N}\alpha_i[y_i\mathbf{w}^T\boldsymbol{\varphi}(\mathbf{x}_i) - 1 + \xi_i] - \sum_{i=1}^{N}\mu_i\xi_i$$

Z przyrównania pochodnych do zera otrzymamy:

$$\frac{\partial L_P}{\partial \mathbf{w}} = 0 \quad \rightarrow \quad \mathbf{w} = \sum_{i=1}^{N}\alpha_i y_i \boldsymbol{\varphi}(\mathbf{x}_i), \quad \frac{\partial L_P}{\partial \xi_i} = 0 \quad \rightarrow \quad C - \alpha_i - \mu_i = 0$$

Postać dualna lagrangianu:

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

przy ograniczeniach: $\sum_{i=1}^{N} \alpha_i y_i = 0$ i $0 \le \alpha_i \le C$.

Ideą maszyn jądrowych jest zastąpienie iloczynu skalarnego $\boldsymbol{\varphi}(\mathbf{x}_i)^T \boldsymbol{\varphi}(\mathbf{x}_j)$ funkcją jądrową $K(\mathbf{x}_i, \mathbf{x}_j)$ oryginalnych przykładów. Zamiast odwzorowywać przykłady **x** za pomocą funkcji **φ(x)** na przestrzeń *Z* i wyznaczać iloczyn skalarny w tej przestrzeni, stosujemy funkcję jądrową działającą bezpośrednio na przykładach **x** (*kernel trick*):

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

Płaszczyznę dyskryminacyjną możemy zapisać:

$$f(\mathbf{x}) = \mathbf{w}^T \boldsymbol{\varphi}(\mathbf{x}) = \sum_{i=1}^{N} \alpha_i y_i \boldsymbol{\varphi}(\mathbf{x}_i) \boldsymbol{\varphi}(\mathbf{x}) = \sum_{i=1}^{N} \alpha_i y_i K(\mathbf{x}_i, \mathbf{x}_j)$$

Klasyfikator wykorzystujący ten mechanizm nazywa się maszyną jądrową.

Najpopularniejsze funkcje jądrowe:

- wielomian stopnia *q*:

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

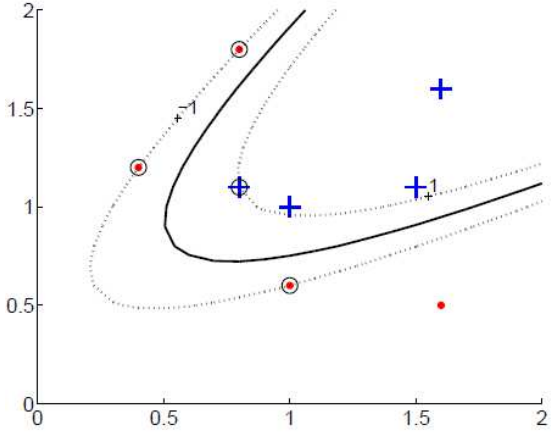
Dla $q = 2$ i $n = 2$:

$$K(\mathbf{x}_a, \mathbf{x}_b) = (\mathbf{x}_a^T \mathbf{x}_b + 1)^2 = (x_{a,1}x_{b,1} + x_{a,2}x_{b,2} + 1)^2 = 1 + 2x_{a,1}x_{b,1} + 2x_{a,2}x_{b,2} + 2x_{a,1}x_{b,1}x_{a,2}x_{b,2} + x_{a,1}^2 x_{a,2}^2 + x_{b,1}^2 x_{b,2}^2$$

Odpowiada to iloczynowi skalarnemu funkcji bazowych postaci:

$$\boldsymbol{\varphi}(\mathbf{x}) = [1, \sqrt{2}x_1, \sqrt{2}x_2, \sqrt{2}x_1x_2, x_1^2, x_2^2]^T$$

W takim przypadku powierzchnia decyzyjna w przestrzeni *X* ma postać:

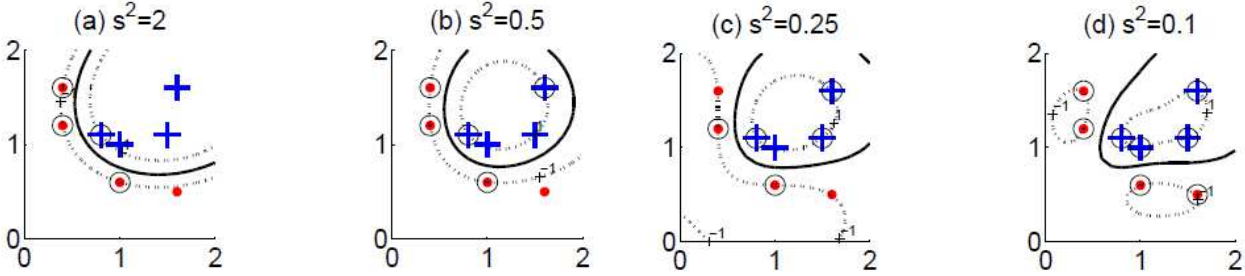


- radialna funkcja bazowa:

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

Parametr *s* - szerokość funkcji radialnej dobieramy w kroswalidacji.

Różne powierzchnie decyzyjne w zależności od parametru *s*:



- jądro sigmoidalne:

$$K(\mathbf{x}_i, \mathbf{x}) = \tanh(2\mathbf{x}^T\mathbf{x}_i + 1)$$

Funkcje jądrowe mierzą podobieństwo pomiędzy przykładami. Im przykłady są do siebie bardziej podobne tym wartość funkcji jądrowej jest większa (maksymalna dla identycznych przykładów).

WIELOKLASOWE MASZYNY JĄDROWE

Gdy liczba klas $K > 2$ konstruujemy K klasyfikatorów SVM. Każdy z nich separuje jedną klasę od pozostałych. Każdy więc tworzy powierzchnię dyskryminacyjną oddzielającą przykłady klasy l-tej (którym nadaje się etykietę +1) od przykładów z pozostałych klas (którym nadaje się etykietę –1):

$$f_l(\mathbf{x}) = \sum_{i \in \Omega_l} \alpha_i^l y_i K(\mathbf{x}_i, \mathbf{x}_j)$$

W trakcie klasyfikacji nowego przykładu wyznaczamy wartości funkcji decyzyjnych utworzonych przez wszystkie klasyfikatory SVM. Funkcja zwracająca największą wartość wskazuje klasę przykładu:

$$f_l(\mathbf{x}) = \arg\max_{l=1,2,\ldots,K} \left\{ \sum_{i \in \Omega_l} \alpha_i^l y_i K(\mathbf{x}_i, \mathbf{x}_j) \right\}$$

Alternatywnym podejściem jest konstrukcja $K(K–1)/2$ klasyfikatorów separujących klasy parami (każda z każdą).