7. Maszyny wektorów podpierających SVMs

dr inż. Urszula Libal

Politechnika Wrocławska

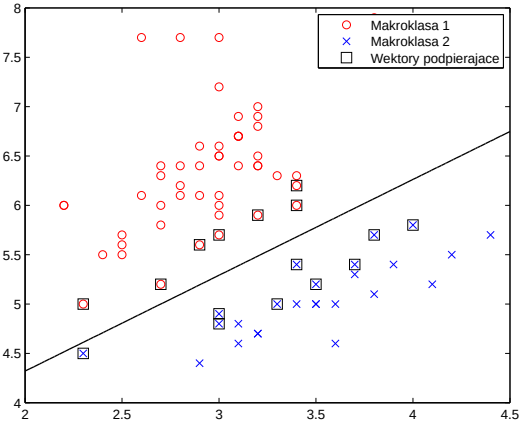
2015

1. Maszyny wektorów podpierających - SVMs

Maszyny wektorów podpierających (ang. *Support Vector Mashines,* SVMs) należą do grupy klasyfikatorów liniowych. Obiekt reprezentowany przez x jest klasyfikowany do jednej z dwóch klas 1 i -1 za pomocą liniowej funkcji dyskryminacyjna $\delta(\mathbf{x}) = \mathbf{w}^T\mathbf{x} + w_0$ w następujący sposób:

$$\Psi_{SVM}(\mathbf{x}) = \begin{cases} 1, & \text{jeżeli } \mathbf{w}^T\mathbf{x} + w_0 > 0 \\ -1, & \text{jeżeli } \mathbf{w}^T\mathbf{x} + w_0 < 0 \end{cases} . \tag{1}$$

Wektor $\mathbf{w} = (w_1, w_2, \ldots, w_D)$ oraz wyraz wolny w_0 są tak dobierane, aby jak najszerzej liniowo separować klasy, jeżeli jest to możliwe.



Rysunek 1. Wektory podpierające dla liniowego klasyfikatora SVMs.

Źródło: opracowanie własne

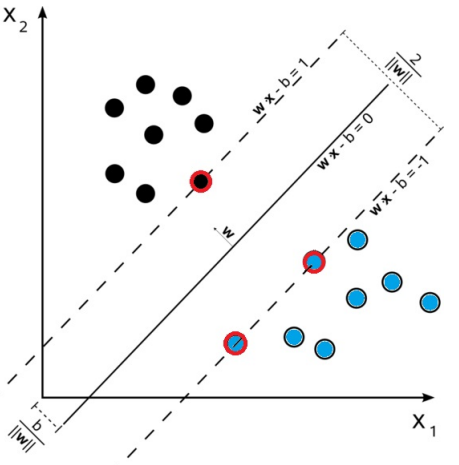
— Do treningu używamy ciągu uczącego $\{(\mathbf{x}_1, c_1), (\mathbf{x}_2, c_2), \ldots, (\mathbf{x}_N, c_N)\}$,
gdzie $\mathbf{x}_k$ to D-wymiarowy wektor cech, a c_k to jego klasa pochodzenia.

— Dane muszą zostać unormowane, co oznacza, że indeksy klas przyjmą wartości 1 i -1.

— Klasyfikator SVM dany wzorem (1) klasyfikuje zgodnie ze znakiem funkcji
dyskryminacyjnej $\delta(\mathbf{x}) = \mathbf{w}^T\mathbf{x} + w_0$.

— Wyznaczenie na podstawie ciągu uczącego parametrów w_i, $i = 0, 1, 2, \ldots, D$
stanowi zadanie optymalizacyjne.

2. Maksymalizacja marginesu



Rysunek 2. Hiperpłaszczyzna rozdzielająca o maksymalnym marginesie.

Źródło: [5]

Maksymalizacja marginesu polega na maksymalizacji odległości między wektorami podpierającymi a hiperpłaszczyzną rozdzielającą. Odległość między hiperpłaszczyzną $\delta(\mathbf{x}) = \mathbf{w}^T\mathbf{x} + w_0 = 0$ a pewnym wektorem $\mathbf{x}_n$ z ciągu uczącego ($n \in \{1, 2 \ldots, N\}$) wynosi

$$\frac{|\delta(\mathbf{x}_n)|}{||\mathbf{w}||}. \tag{2}$$

Maksymalizację wyrażenia (2) można sprowadzić do minimalizacji $||\mathbf{w}||$, lub równoważnie minimalizacji

$$\frac{1}{2}||\mathbf{w}||^2. \tag{3}$$

Bez zmniejszenia ogólności rozważań zakładamy, że marginesy będą postaci

$$\mathbf{w}^T\mathbf{x}+w_0=1 \quad \text{i} \quad \mathbf{w}^T\mathbf{x}+w_0=-1. \tag{4}$$

Wszystkie punkty z ciągu uczącego muszą się znaleźć poza pasem między marginesami, ale tak aby wszystkie punkty z klasy 1 były po odpowiedniej stronie marginesu $\mathbf{w}^T\mathbf{x}+w_0=1$, a wszystkie punkty z klasy -1 po odpowiedniej stronie marginesu $\mathbf{w}^T\mathbf{x}+w_0=-1$ (patrz rys. 2). Sprowadzamy ten warunek do nierówności

$$c_n\left(\mathbf{w}^T\mathbf{x}_n+w_0\right)\geq 1 \tag{5}$$

dla każdego wektora $\mathbf{x}_n$ z ciągu uczącego ($n\in\{1,2\ldots,N\}$), gdzie c_n to jego klasa pochodzenia.

Problem optymalizacyjny poszukiwania maksymalnego marginesu sprowadza się do

$$\begin{cases} \min_{\mathbf{w},w_0} \frac{1}{2}||\mathbf{w}||^2 \\ \text{przy } c_n\left(\mathbf{w}^T\mathbf{x}_n + w_0\right) \geq 1. \end{cases} \tag{6}$$

Metoda mnożników Lagrange'a polega na minimalizacji funkcji L

$$L(\mathbf{w}, w_0, \lambda) = \frac{1}{2}||\mathbf{w}||^2 - \sum_{n=1}^{N} \lambda_n \left\{c_n\left(\mathbf{w}^T\mathbf{x}_n + w_0\right) - 1\right\}. \tag{7}$$

Każdy mnożnik $\lambda_n \geq 0$ odpowiada jednemu wektorowi $\mathbf{x}_n$ z ciągu uczącego

($n \in \{1, 2 \ldots, N\}$).

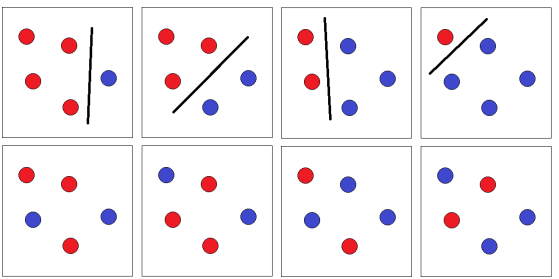
W celu minimalizacji funkcji L wyznaczamy jej pochodne i przyrównujemy je do zera

$$\begin{cases} \frac{\partial L(\mathbf{w}, w_0, \lambda)}{\partial \mathbf{w}} = 0, \\ \frac{\partial L(\mathbf{w}, w_0, \lambda)}{\partial w_0} = 0, \\ \frac{\partial L(\mathbf{w}, w_0, \lambda)}{\partial \lambda} = 0. \end{cases} \tag{8}$$

[przykład numeryczny dla D2]

3. Nieliniowe SVMs

Niestety nie zawsze klasyfikacja za pomocą maszyn wektorów podpierających SVMs (1) jest możliwa do przeprowadzenia. Może się zdarzyć, że klasy nie są liniowo separowalne.

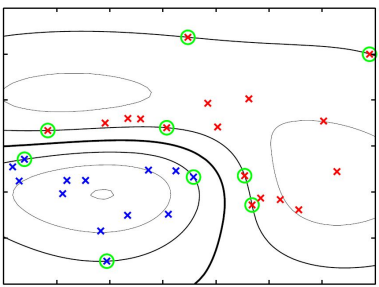


Rysunek 3. Liniowo separowalne oraz nieseparowalne klasy.

Źródło: opracowanie własne

W przypadku nierozdzielnych liniowo klas stosujemy trik z zastosowaniem funkcji jądrowych ϕ (kernel trick).

Inną postać przyjmuje funkcja dyskryminacyjna $\delta(\mathbf{x}) = \mathbf{w}^T \phi(\mathbf{x}) + w_0$.



Rysunek 4. Wektory podpierające dla nieliniowego klasyfikatora SVMs.

Źródło: [2]

Funkcja L w metodzie mnożników Lagrange'a również ulega zmianie

$$L(\mathbf{w}, w_0, \lambda) = \frac{1}{2}||\mathbf{w}||^2 - \sum_{n=1}^{N} \lambda_n \left\{ c_n \left(\mathbf{w}^T \phi(\mathbf{x}_n) + w_0 \right) - 1 \right\}. \tag{9}$$

Minimalizację L sprowadzamy do problemu dualnego.

Maksymalizujemy teraz $\widetilde{L}$

$$\widetilde{L}(\lambda) = \sum_{n=1}^{N} \lambda_n - \frac{1}{2} \sum_{n=1}^{N} \sum_{m=1}^{N} \lambda_n \lambda_m c_n c_m \kappa(\mathbf{x}_n, \mathbf{x}_m) \tag{10}$$

przy ograniczeniach

$$\lambda_n \geq 0, \quad n = 1, 2, \ldots, N, \tag{11}$$

$$\sum_{n=1}^{N} \lambda_n c_n = 0, \tag{12}$$

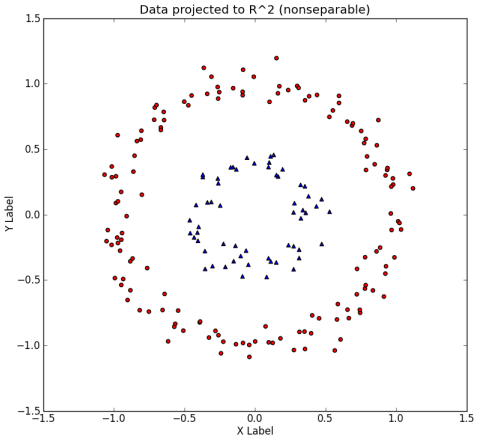
gdzie jądro κ przyjmuje postać

$$\kappa(\mathbf{x}_n, \mathbf{x}_m) = \phi(\mathbf{x}_n)^T \phi(\mathbf{x}_m). \tag{13}$$

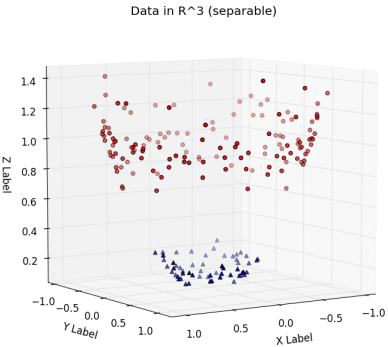
Funkcja jądrowa	Matematyczna forma $\kappa(\mathbf{x}, \mathbf{y})$
wielomianowa p-tego rzędu	$\kappa(\mathbf{x}, \mathbf{y}) = \left(\mathbf{x}^T\mathbf{y} + r\right)^p$
gaussowska (Radial Basis Function)	$\kappa(\mathbf{x}, \mathbf{y}) = \exp(-\frac{\|\mathbf{x}-\mathbf{y}\|^2}{2\sigma^2})$
sigmoidalna	$\kappa(\mathbf{x}, \mathbf{y}) = \tanh(\mathbf{x}^T\mathbf{y} + r)$

Jądrowa (nieliniowa) wersja klasyfikatora SVM to wtedy

$$\Psi_{kernel\,SVM}(\mathbf{x}) = \begin{cases} 1, & \text{jeżeli } \sum_{n=1}^{N} \lambda_n c_n \kappa(\mathbf{x}_n, \mathbf{x}) + w_0 > 0, \\ -1, & \text{jeżeli } \sum_{n=1}^{N} \lambda_n c_n \kappa(\mathbf{x}_n, \mathbf{x}) + w_0 < 0. \end{cases} \tag{14}$$



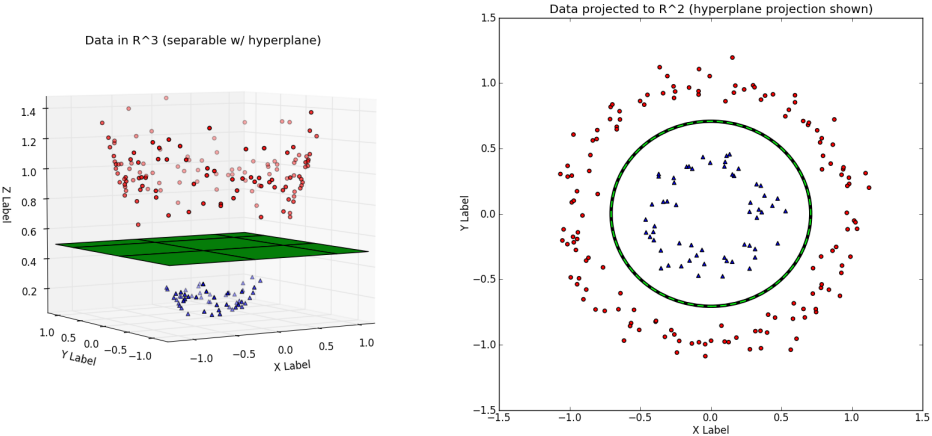
a)



b)

Rysunek 5. (a) Zbiór punktów nierozdzielny liniowo. (b) Ten sam zestaw danych przekształcony przez transformację $[x_1, x_2] \mapsto [x_1, x_2, x_1^2 + x_2^2]$.

Źródło: [6]



a) b)

Rysunek 6. Hiperpłaszczyzna rozdzielająca: (a) liniowa w $\mathbb{R}^3$, (b) nieliniowa w $\mathbb{R}^2$.

Źródło: [6]

Literatura

[1] A.R. Webb, K.D. Copsey, *Statistical Pattern Recognition*, 3rd ed., Wiley, (2011)

[2] C.M. Bishop, *Pattern Recognition and Machine Learning*, Springer Series: Information Science and Statistics (2006)

[3] M. Krzyśko, W. Wołyński, T. Górecki, M. Skorzybut, *Systemy uczące się. Rozpoznawanie wzorców, analiza skupień i redukcja wymiarowości*. WNT, Warszawa (2008)

[4] R.O. Duda, P.E. Hart, D.G. Stork, *Pattern Classification*, 2nd ed., Wiley, (2000)

[5] http://en.wikipedia.org/wiki/Support_vector_machine

[6] http://www.eric-kim.net/eric-kim-net/posts/1/kernel_trick.html