

UNIW. PRZYRODNICZY W WROCLAWIU



UNIwersytet PRZYRODNICZY WE WROCLAWIU

Data Mining Wykład 7

Maszyna Wektorów Nośnych (SVM)

Maszyna wektorów nośnych

- W przestrzeni danych (ang. measurement space) Ω znajdują się wektory danych x stanowiące próbkę uczącą D , należące do dwóch klas:

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

- Szukamy klasyfikatora pozwalającego na podział całej przestrzeni Ω na dwa rozłączne obszary odpowiadające klasom $\{1, -1\}$ oraz pozwalającego jak najlepiej klasyfikować nowe obiekty x do klas
- Podejście opiera się na znalezieniu tzw. granicy decyzyjnej między klasami $\rightarrow g(\mathbf{x})$

UNIwersytet PRZYRODNICZY WE WROCLAWIU

Separowalność liniowa

- Dwie klasy są liniowo separowalne, jeśli istnieje hiperpłaszczyzna H postaci $g(\mathbf{x})$

$$g(\mathbf{x}) = \mathbf{w}^T \mathbf{x} + b$$

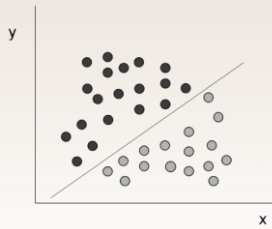
- przyjmująca wartości

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

- Jak poszukiwać takiej hiperpłaszczyzny granicznej?

UNIwersytet PRZYRODNICZY WE WROCLAWIU

Liniowa funkcja separująca

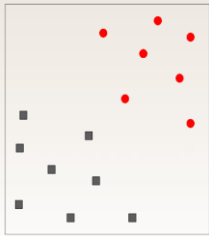


- Funkcja liniowa separująca
- Wyznacza podział przestrzeni na obszary odpowiadające dwóm klasom decyzyjnym.
- Oryginalna propozycja Fisher'a, ale tak że inne metody (perceptron, itp..)
- Uogólnienia dla wielu klas.

UNIWERSYTET PRZYRODNICZY WE WROCLAWIU

Support Vector Machine (SVM)

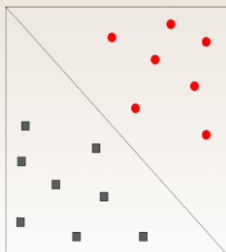
- Znajdź liniową hiperpłaszczyznę (decision boundary) oddzielającą obszary przykładów z dwóch różnych klas



UNIWERSYTET PRZYRODNICZY WE WROCLAWIU

Support Vector Machine (SVM)

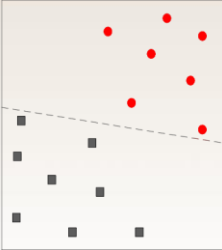
- Jedno z możliwych rozwiązań



UNIWERSYTET PRZYRODNICZY WE WROCLAWIU

Support Vector Machine (SVM)

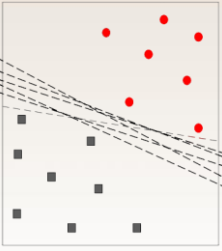
- Inne możliwe rozwiązanie



UNIWERSYTET PRZYRODNICZY WE WROCLAWIU

Support Vector Machine (SVM)

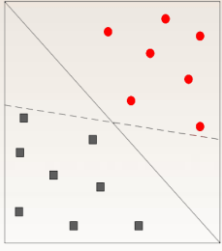
- Zbiór wielu możliwych rozwiązań



UNIWERSYTET PRZYRODNICZY WE WROCLAWIU

Support Vector Machine (SVM)

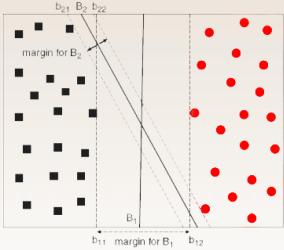
- Którą z hiperpłaszczyzn należy wybrać? B1 czy B2?
- Czy można to formalnie zdefiniować?



UNIWERSYTET PRZYRODNICZY WE WROCLAWIU

Margines

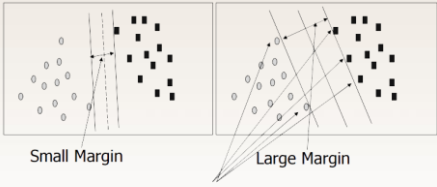
- Hiperpłaszczyzny b_{11} i b_{12} są otrzymane przez równoległe przesuwanie hiperpłaszczyzny granicznej aż do pierwszych punktów z obu klas.
- Odległość między nimi – margines klasyfikatora liniowego
- Jaki margines wybierać ?



UNIWERSYTET PRZYRODNICZY WE WROCLAWIU

Węższe czy szersze marginesy?

- Szerszy margines - lepsze własności generalizacji, mniejsza podatność na ew. przeuczenie (overfitting)
- Wąski margines – mała zmiana granicy, radykalne zmiany klasyfikacji



Support Vectors

UNIWERSYTET PRZYRODNICZY WE WROCLAWIU

Liniowe SVM hiperpłaszczyzna graniczna

- Vapnik – poszukuj „maximal margin classifier”

$$\mathbf{w} \cdot \mathbf{x} + \mathbf{b} = 0$$
 gdzie \mathbf{w} i \mathbf{b} są parametrami modelu

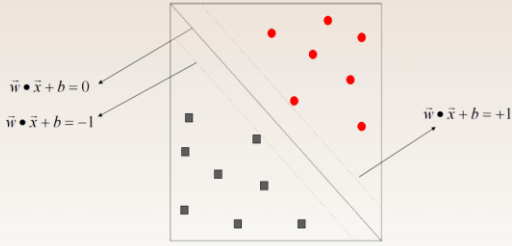
$$y = \begin{cases} 1 & \mathbf{w} \cdot \mathbf{x} + \mathbf{b} > 0 \\ -1 & \mathbf{w} \cdot \mathbf{x} + \mathbf{b} < 0 \end{cases}$$
- Parametry granicy wyznaczaj tak, aby maksymalne marginesy b_{11} i b_{12} były miejscem geometrycznym punktów \mathbf{x} spełniających warunki

$$b_{11} \quad \mathbf{w} \cdot \mathbf{x} + \mathbf{b} = 1$$

$$b_{12} \quad \mathbf{w} \cdot \mathbf{x} + \mathbf{b} = -1$$
- Margines – odległość między płaszczyznami b_{11} i b_{12}

UNIWERSYTET PRZYRODNICZY WE WROCLAWIU

Poszukiwanie parametrów hiperpłaszczyzny



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

UNIWERSYTET PRZYRODNICZY WE WROCLAWIU

Linear Support Vector Machines

- Sformułowanie problemu:

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

- Przy warunkach ograniczających

$$y_i(\mathbf{w} \cdot \mathbf{x}_i + b) \geq 1 \quad i = 1, 2, \dots, N$$

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

UNIWERSYTET PRZYRODNICZY WE WROCLAWIU

Linear Support Vector Machines

Minimalizuj funkcję Lagrange'a:

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

, gdzie parametry $\alpha \geq 0$ mnożniki Lagrange'a

Przy przekształceniach wykorzystuje się ograniczenia Karush-Kuhn-Tucker na mnożniki:

$$\alpha_i \geq 0$$

$$\alpha_i [y_i (\mathbf{w} \cdot \mathbf{x}_i + b) - 1] = 0$$

- W konsekwencji α_i są niezerowe wyłącznie dla wektorów nośnych \mathbf{x} , pozostałe są zerowe
- Rozwiązanie parametrów \mathbf{w} i b zależy wyłącznie od wektorów nośnych.

UNIWERSYTET PRZYRODNICZY WE WROCLAWIU

Linear Support Vector Machines – Duality Solution

Przy ograniczeniach:

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

Rozwiązanie ($\alpha > 0$ dla $i \in SV$):

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

Hiperpłaszczyzna decyzyjna:

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

UNIWERSYTET PRZYRODNICZY WE WROCLAWIU

LSVM - Klasyfikacja

Klasyfikacja – funkcja decyzyjna

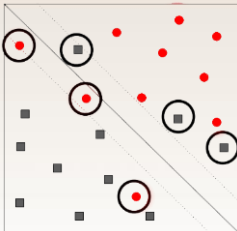
$$f(x) = \text{sign}(\sum_{i=1}^N \alpha_i y_i \mathbf{x}_i \cdot \mathbf{x} + b)$$

- O ostatecznej postaci hiperpłaszczyzny decydują wyłącznie wektory nośne ($\alpha_i > 0$)
- Im większa wartość α_i , tym większy wpływ wektora na granicę decyzyjną
- Klasyfikacja zależy od iloczynu skalarnego nowego \mathbf{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 treningowych (wektory nośne)

UNIWERSYTET PRZYRODNICZY WE WROCLAWIU

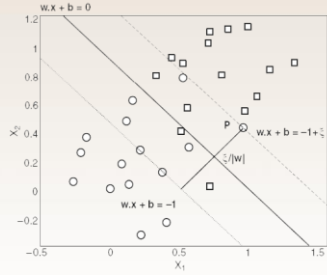
Niepełna liniowa separowalność

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



UNIWERSYTET PRZYRODNICZY WE WROCLAWIU

Zmienne dopełniające



UNIWERSYTET PRZYRODNICZY WE WROCLAWIU

Zmienne osłabiające - interpretacja

- Zmienne $\xi_i \geq 0$ (ang. Soft Margin) dobiera się dla każdego przykładu uczącego. Jej wartość zmniejsza margines separacji. (rodzaj „zwisu” punktu poza hiperpłaszczyzną nośną)
- Jeżeli $0 \leq \xi_i \leq 1$, to punkt danych (x_i, d_i) leży wewnątrz strefy separacji, ale po właściwej stronie
- Jeżeli $\xi_i > 1$, punkt po niewłaściwej stronie hiperpłaszczyzny i wystąpi błąd klasyfikacji
- Modyfikacja wymagań dla wektorów nośnych

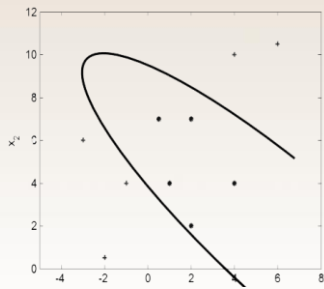
$$b_{11} \quad \mathbf{w} \cdot \mathbf{x} + \mathbf{b} = 1 - \xi$$

$$b_{12} \quad \mathbf{w} \cdot \mathbf{x} + \mathbf{b} = -1 + \zeta$$

UNIWERSYTET PRZYRODNICZY WE WROCLAWIU

Nonlinear Support Vector Machines

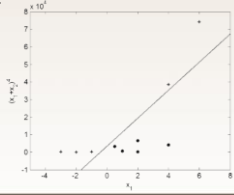
- Co zrobić gdy próby uczące powinny być nieliniowo separowalne?



UNIWERSYTET PRZYRODNICZY WE WROCLAWIU

Kernel Trick

- Transformacja do wysoce wielowymiarowej przestrzeni - tzw. Kernel Trick
- Kernel Trick - metoda mapowania obserwacji z pewnego zbioru S na przestrzeń unitarną V bez konieczności tworzenia explicite samego mapowania w nadziei, że nabiorą one tam sensownej struktury liniowej.



UNIWERSYTET PRZYRODNICZY WE WROCLAWIU

Kernel Trick - Przykład

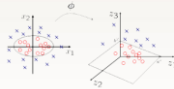
- Kernelem nazywamy funkcję $K(x, y)$, która dla $x, y \in S$ jest iloczynem skalarnym w pewnej przestrzeni V .
- Przykładowo mając mapowanie:

$$\phi : S \rightarrow V$$
- Kernelem jest po prostu:

$$K(x, y) = \langle \phi(x), \phi(y) \rangle_V$$

$$K(x, y) = (x - y)^2, x, y \in \mathbb{R}^2$$

$$\phi(x) = (x_1^2, \sqrt{2}x_1x_2, x_2^2)$$



UNIWERSYTET PRZYRODNICZY WE WROCLAWIU

Dlaczego Kernel Trick

- Dlaczego po prostu nie skonstruować mapowania i pracować na przestrzeni V zamiast S ?
 1. Złożoność obliczeniowa
 2. O wiele trudniej znaleźć dobre mapowanie niż dobry kernel
 3. Możliwość pracy na nieskończenie wymiarowych przestrzeniach.

UNIWERSYTET PRZYRODNICZY WE WROCLAWIU

Najczęściej używane 'Kernele'

- Kernel wielomianowy: $K(x, y) = (x \cdot y + 1)^p$
- Kernel Gaussowski: $K(x, y) = e^{-\|x-y\|^2/2\sigma^2}$
- Kernel sigmoidalny: $K(x, y) = \tanh(\kappa x \cdot y - \delta)$
- Kernel minimum (przecięcia histogramów):
 $K(x, y) = \sum_i \min(x_i, y_i)$
- Kernel logarytmiczny: $K(x, y) = -\log(\|x - y\|^d + 1)$
- Budowanie kerneli: suma, iloczyn, iloczyn przez stałą dodatnią

UNIWERSYTET PRZYRODNICZY WE WROCŁAWIU
