Metoda wektorów podpierających

Systemy uczące się - laboratorium

Mateusz Lango

Zakład Inteligentnych Systemów Wspomagania Decyzji
Wydział Informatyki i Telekomunikacji
Politechnika Poznańska

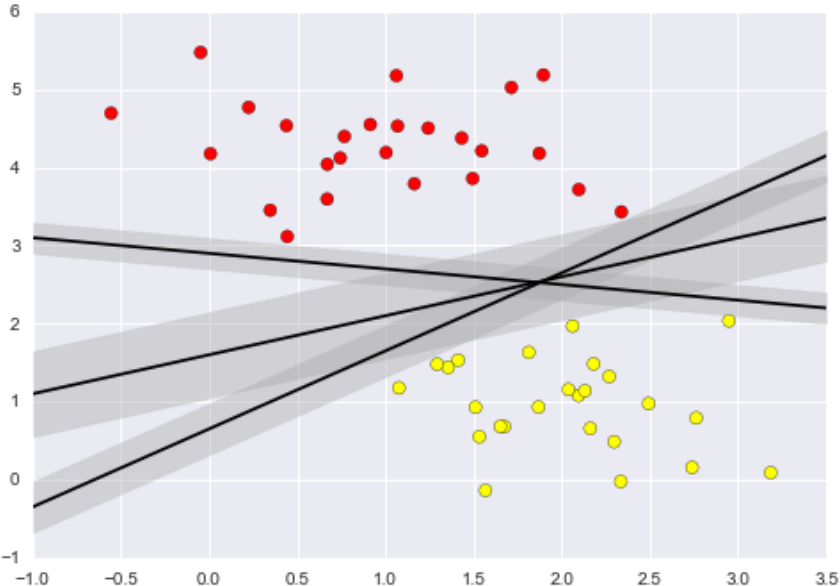
„Akademia Innowacyjnych Zastosowań Technologii Cyfrowych (AI Tech)”,
projekt finansowany ze środków Programu Operacyjnego Polska Cyfrowa POPC.03.02.00-00-0001/20

Fundusze Europejskie
Polska Cyfrowa

Rzeczpospolita Polska

Unia Europejska
Europejski Fundusz Rozwoju Regionalnego

Przypomnienie: Problem wyboru granicy



Zasada maksymalnego marginesu

Plan:

1. Formalne zdefiniowanie problemu wyboru granicy jako problem optymalizacyjny
2. Rozluźnienie problemu do sytuacji nieseparowalnej
3. Problem dualny i trik jądrowy
4. SVM dla dużych danych (oraz czym różni się SVM od regresji logistycznej)

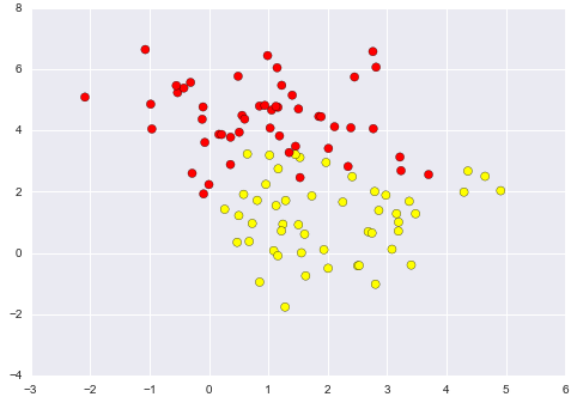
Przypomnienie: sformułowany problem optymalizacyjny

$$\min_{w,b} ||w||$$

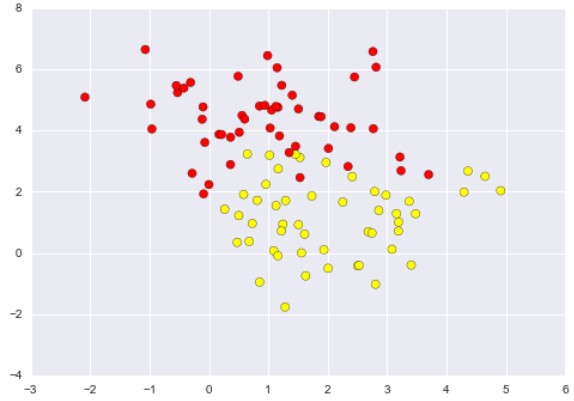
Przy ograniczeniach:

$$f(x_i) \geq 1 \quad \text{jeśli } y_i = 1 \qquad f(x_i) \leq -1 \quad \text{jeśli } y_i = -1$$

Co jeśli zbiór nie jest liniowo separowalny?



Co jeśli zbiór nie jest liniowo separowalny?



Dwa możliwe rozwiązania:

- Zmodyfikować definicję SVM, tak aby sobie z tym radziła
- Rozszerzyć przestrzeń cech, tak aby przestrzeń stała się liniowo separowalna

Soft-SVM

$$\min_{w,b,\xi} ||w|| + C \sum_{i=1}^{n} \xi_i$$

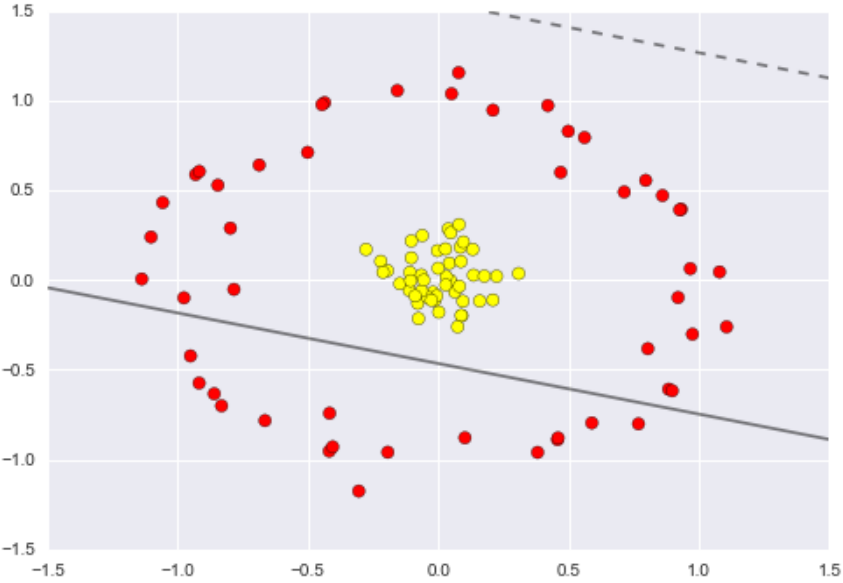
Przy ograniczeniach:

$$f(x_i) \geq 1 - \xi_i \quad \text{jeśli } y_i = 1$$

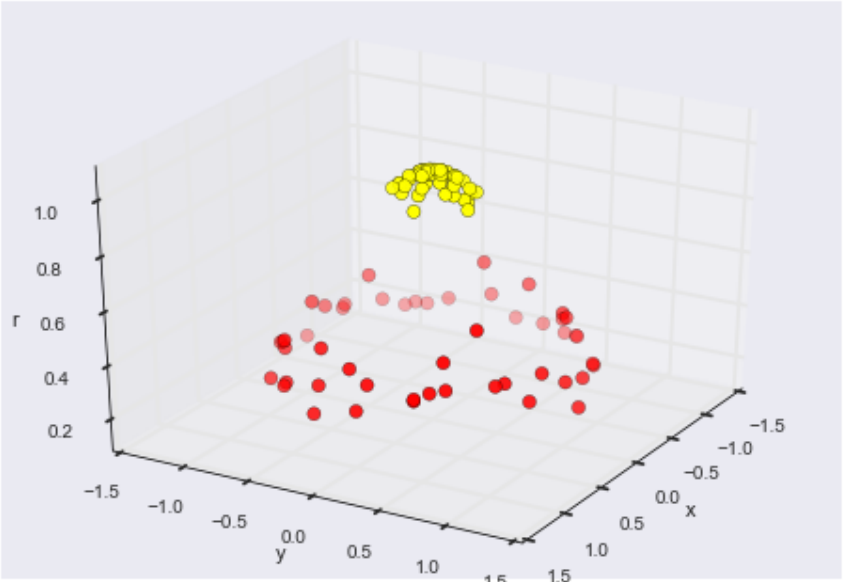
$$f(x_i) \leq -1 + \xi_i \quad \text{jeśli } y_i = -1$$

$$\xi_i \geq 0$$

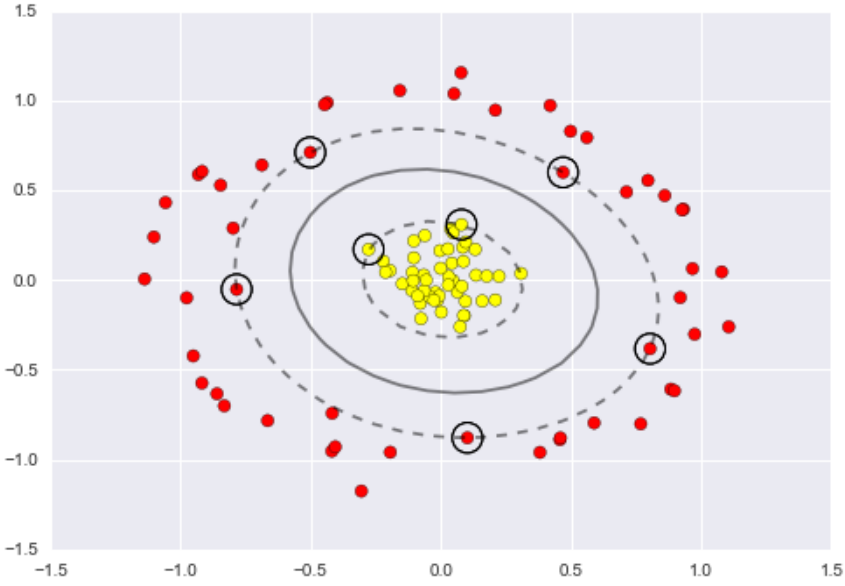
Dodawanie cech



Dodawanie cech



Dodawanie cech



Klasyfikatory liniowe a cechy wielomianowe

Theorem (Twierdzenie Weierstrassa)

Suppose f is a continuous real-valued function defined on the real interval $[a, b]$. For every $\epsilon > 0$, there exists a polynomial p such that for all x in $[a, b]$, we have $|f(x) - p(x)| < \epsilon$

- jest to (pośrednio) twierdzenie o uniwersalności dla regresji liniowej z dostateczną liczbą wielomianowych cech!
- problem praktyczny: dodawanie cech wielomianowych kosztuje nas czas i pamięć. Niestety liczba dodatkowych wielomianowych cech rośnie wykładniczo z wymiarowością problemu...

Zanim przejdziemy do dodawania cech: Problem dualny

$$\max_{\alpha} \sum_{i=1}^{N} \alpha_i - \frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{N} \alpha_i \alpha_j y_i y_j x_i^T x_j$$

Przy ograniczeniach:

$$\sum_{i=1}^{N} \alpha_i y_i = 0 \qquad \alpha_i \geq 0$$

Po zoptymalizowaniu równanie płaszczyzny można ew. obliczyć:

$$w = \sum_{i=1}^{N} \alpha_i y_i x_i \quad b = 1 - \min_{i:y_i=1} w^T x_i$$

(dodatkowo: dużo [większość?] α_i wynosi 0)

Zanim przejdziemy do dodawania cech: Problem dualny

$$\max_{\alpha} \sum_{i=1}^{N} \alpha_i - \frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{N} \alpha_i \alpha_j y_i y_j x_i^T x_j$$

Przy ograniczeniach:

$$\sum_{i=1}^{N} \alpha_i y_i = 0 \qquad \alpha_i \geq 0$$

W praktyce wektora wag nie wyznaczamy (gdyż, o czym za chwilę, możemy pracować nawet z ∞ liczbą cech):

$$f(x) = \sum_{i=1}^{N} \alpha_i y_i x_i^T x + b$$

(Za chwilę niezwykle ważna) właściwość: przy uczeniu i predykcji potrzebujemy tylko iloczynów $x_i^T x$

Zadania

Problem

Jaka jest interpretacja wartości współczynników Lagrange'a w formulacji dualnej problemu SVM? Odpowiedź uzasadnij odwołując się do warunków KKT.

Problem

W formulacji dualnej problemu SVM optymalizowane są zmienne, które zwykle oznaczamy jako α_i i nie są to wagi. W jaki sposób zatem znajdowana jest hiperpłaszczyzna separująca?

Problem

Przekształć problem (soft) SVM do problemu bez ograniczeń. Jakie są podobieństwa i różnice pomiędzy klasyfikatorem regresji logistycznej a klasyfikatorem SVM?

Zadania

Problem

Jaka jest interpretacja wartości współczynników Lagrange'a w formulacji dualnej problemu SVM? Odpowiedź uzasadnij odwołując się do warunków KKT.

Problem

W formulacji dualnej problemu SVM optymalizowane są zmienne, które zwykle oznaczamy jako α_i i nie są to wagi. W jaki sposób zatem znajdowana jest hiperpłaszczyzna separująca?

Problem

Przekształć problem (soft) SVM do problemu bez ograniczeń. Jakie są podobieństwa i różnice pomiędzy klasyfikatorem regresji logistycznej a klasyfikatorem SVM?

Zadania

Problem

Jaka jest interpretacja wartości współczynników Lagrange'a w formulacji dualnej problemu SVM? Odpowiedź uzasadnij odwołując się do warunków KKT.

Problem

W formulacji dualnej problemu SVM optymalizowane są zmienne, które zwykle oznaczamy jako α_i i nie są to wagi. W jaki sposób zatem znajdowana jest hiperpłaszczyzna separująca?

Problem

Przekształć problem (soft) SVM do problemu bez ograniczeń. Jakie są podobieństwa i różnice pomiędzy klasyfikatorem regresji logistycznej a klasyfikatorem SVM?

Dualny czy prymalny?

- Rozwiązanie problemu programowania kwadratowego z M zmiennymi ma złożoność obliczeniową rzędu $O(M^3)$
- W rozwiązaniu prymalnym mamy k zmiennych (liczba cech)
- W rozwiązaniu dualnym mamy n zmiennych (liczba przykładów)
- W sytuacji problemu wysokowymiarowego ze stosunkowo małą liczbą przykładów opłaca się stosować wersję dualną. W odwrotnej sytuacji: prymalną.
- Jednak rozwiązanie dualne pozwala na zastosowanie triku jądrowego i operowanie w przestrzeniach o nawet ∞ liczbie wymiarów!

Dodawanie cech

Rozważmy generację cech wielomianowych rzędu drugiego dla problemu z $k = 2$ cechami.

$$(x_1, x_2) \Rightarrow ?$$

Dodawanie cech

Rozważmy generację cech wielomianowych rzędu drugiego dla problemu z $k = 2$ cechami.

$$(x_1, x_2) \Rightarrow (x_1, x_2, x_1^2, x_2^2, x_1 x_2)$$

- Ile by było cech gdyby początkowo było $k = 1000$? Ok. pół miliona.
- Cechy rzędu trzeciego? Ok. 166 milionów.
- Nawet dla problemów o „rozsądnej" wymiarowości nie można tego zrobić w praktyce...

Trik jądrowy

Tak jak zauważyliśmy cały proces uczenia wymaga od nas obliczania iloczynu wektorowego pomiędzy przykładami:

$$x^T z$$

czyli zapisując nasze dodawanie cech jako funkcję $\phi()$ potrzebujemy obliczyć:

$$\phi(x)^T \phi(z)$$

Trik jądrowy

Tak jak zauważyliśmy cały proces uczenia wymaga od nas obliczania iloczynu wektorowego pomiędzy przykładami:

$$x^T z$$

czyli zapisując nasze dodawanie cech jako funkcję $\phi()$ potrzebujemy obliczyć:

$$\phi(x)^T \phi(z)$$

Trik jądrowy

Tak jak zauważyliśmy cały proces uczenia wymaga od nas obliczania iloczynu wektorowego pomiędzy przykładami:

$$x^T z$$

czyli zapisując nasze dodawanie cech jako funkcję $\phi()$ potrzebujemy obliczyć:

$$K(x, z) = \phi(x)^T \phi(z)$$

Trik jądrowy

Dla poprzedniego przykładu z ϕ:

$$(x_1, x_2) \Rightarrow (1, x_1, x_2, x_1^2, x_2^2, x_1 x_2)$$

otrzymujemy:

$$K(x, z) = \phi(x)^T \phi(z) =$$

Trik jądrowy

Dla poprzedniego przykładu z ϕ:

$$(x_1, x_2) \Rightarrow (1, \sqrt{2}x_1, \sqrt{2}x_2, x_1^2, x_2^2, \sqrt{2}x_1x_2)$$

otrzymujemy:

$$K(x, z) = \phi(x)^T \phi(z) =$$

Trik jądrowy

Dla poprzedniego przykładu z ϕ:

$$(x_1, x_2) \Rightarrow (1, \sqrt{2}x_1, \sqrt{2}x_2, x_1^2, x_2^2, \sqrt{2}x_1x_2)$$

otrzymujemy:

$$\begin{aligned} K(x, z) &= \phi(x)^T \phi(z) = \\ &= [1, \sqrt{2}x_1, \sqrt{2}x_2, x_1^2, x_2^2, \sqrt{2}x_1x_2]^T [1, \sqrt{2}z_1, \sqrt{2}z_2, ...] \end{aligned}$$

Trik jądrowy

Dla poprzedniego przykładu z ϕ:

$$(x_1, x_2) \Rightarrow (1, \sqrt{2}x_1, \sqrt{2}x_2, x_1^2, x_2^2, \sqrt{2}x_1x_2)$$

otrzymujemy:

$$\begin{aligned} K(x,z) &= \phi(x)^T\phi(z) = \\ &= [1, \sqrt{2}x_1, \sqrt{2}x_2, x_1^2, x_2^2, \sqrt{2}x_1x_2]^T[1, \sqrt{2}z_1, \sqrt{2}z_2, ...] \\ &= 1 + 2x_1z_1 + 2x_2z_2 + x_1^2z_1^2 + ... \end{aligned}$$

Trik jądrowy

Dla poprzedniego przykładu z ϕ:

$$(x_1, x_2) \Rightarrow (1, \sqrt{2}x_1, \sqrt{2}x_2, x_1^2, x_2^2, \sqrt{2}x_1x_2)$$

otrzymujemy:

$$\begin{aligned} K(x,z) &= \phi(x)^T \phi(z) = \\ &= [1, \sqrt{2}x_1, \sqrt{2}x_2, x_1^2, x_2^2, \sqrt{2}x_1x_2]^T [1, \sqrt{2}z_1, \sqrt{2}z_2, ...] \\ &= 1 + 2x_1z_1 + 2x_2z_2 + x_1^2z_1^2 + ... \\ &= (1 + x^T z)^2 \end{aligned}$$

Trik jądrowy

Dla poprzedniego przykładu z ϕ:

$$(x_1, x_2) \Rightarrow (1, \sqrt{2}x_1, \sqrt{2}x_2, x_1^2, x_2^2, \sqrt{2}x_1x_2)$$

otrzymujemy:

$$\begin{aligned} K(x,z) &= \phi(x)^T\phi(z) = \\ &= [1, \sqrt{2}x_1, \sqrt{2}x_2, x_1^2, x_2^2, \sqrt{2}x_1x_2]^T[1, \sqrt{2}z_1, \sqrt{2}z_2, ...] \\ &= 1 + 2x_1z_1 + 2x_2z_2 + x_1^2z_1^2 + ... \\ &= (1 + x^Tz)^2 \end{aligned}$$

Czas liniowy! Brak konieczności generacji tryliarda cech! Pomimo tego, że uzyskujemy ten sam wynik!

Trik jądrowy

- $K(x, z)$ to funkcja jądrowa
- pokazaliśmy trik dla jądra wielomianowego - istnieją też inne jądra (w tym takie mające ∞ liczbę cech)
- „zwykłe" SVM też używa jądra: jądro liniowe $K(x, z) = x^T z$
- jądra czasami interpretuje się jako „podobieństwo pomiędzy przykładami"
- $K(x, z) = x^T z$ to prawie (mocno nadużywając) korelacja Pearsona!
- więcej w ćwiczeniach

Trik jądrowy to nie tylko SVM

Problem

W jakich innych algorytmach można zastosować trik jądrowy?

- W domu warto zajrzeć do „Representer theorem".
- W ćwiczeniach: definicja popularnych jąder
- Tworzenie własnych jąder? W skrócie: K powinno być półdodatnio określone.
- Jak to uzyskać? Operacje zachowujące jądrowość (np. suma - analogia do operacji zachowujących wypukłość)

Trik jądrowy to nie tylko SVM

Problem

W jakich innych algorytmach można zastosować trik jądrowy?

- W domu warto zajrzeć do „Representer theorem".
- W ćwiczeniach: definicja popularnych jąder
- Tworzenie własnych jąder? W skrócie: K powinno być półdodatnio określone.
- Jak to uzyskać? Operacje zachowujące jądrowość (np. suma - analogia do operacji zachowujących wypukłość)

Trik jądrowy to nie tylko SVM

Problem

W jakich innych algorytmach można zastosować trik jądrowy?

- W domu warto zajrzeć do „Representer theorem".
- W ćwiczeniach: definicja popularnych jąder
- Tworzenie własnych jąder? W skrócie: K powinno być półdodatnio określone.
- Jak to uzyskać? Operacje zachowujące jądrowość (np. suma - analogia do operacji zachowujących wypukłość)

SVM a DNN[2]

Przez lata SVM były skutecznym przeciwnikiem głębokich sieci neuronowych (i nadal są w niektórych zastosowaniach).

- SVM przez krytyków DNN był traktowany jako „lepszy" perceptron
- SVM rozszerza wejście jako bardzo duża (nieskończona?) warstwa nieliniowych cech (bez adaptacji)
- SVM mają jedną warstwę uczonych wag
- SVM mają bardzo efektywną metodę unikania przeuczenia (margines!)
- Wynik SVM jest reprodukowalny (!) i lepiej przebadany teoretycznie
- Inny sposób patrzenia na SVM: każdy element zbioru uczącego jest „cechą" do której liczmy podobieństwo. Uczenie wag to wybór „cech" (przykładów uczących) i ich ważenie w funkcji podobieństwa.
- SVMy miały tak dobrą renomę, że na konferencjach Computer Vision uczenie się cech było traktowane jako *wadę!*[1].

[1] Słynny mail LeCun'a do edytorów CVPR: podejście trafniejsze, szybsze, z uczącymi cechami = reject???

[2] częściowo za slajdami G. Hintona „Neural Networks for Machine Learning"

Podsumowanie

- SVM to klasyfikator liniowy ...
- ale możemy wykorzystać trik jądrowy
- Problem wyboru dobrej funkcji jądrowej(+ jej parametryzacji)
- Problem wyboru stałej C
- Problem skalowalności optymalizacji (ćwiczenia)

Widzimy się za tydzień!

Fundusze Europejskie
Polska Cyfrowa

Rzeczpospolita Polska

Unia Europejska
Europejski Fundusz Rozwoju Regionalnego