Metoda wektorów podpierających

Systemy uczące się - laboratorium

Mateusz Lango

Zakład Inteligentnych Systemów Wspomagania Decyzji
Wydział Informatyki i Telekomunikacji
Politechnika Poznańska

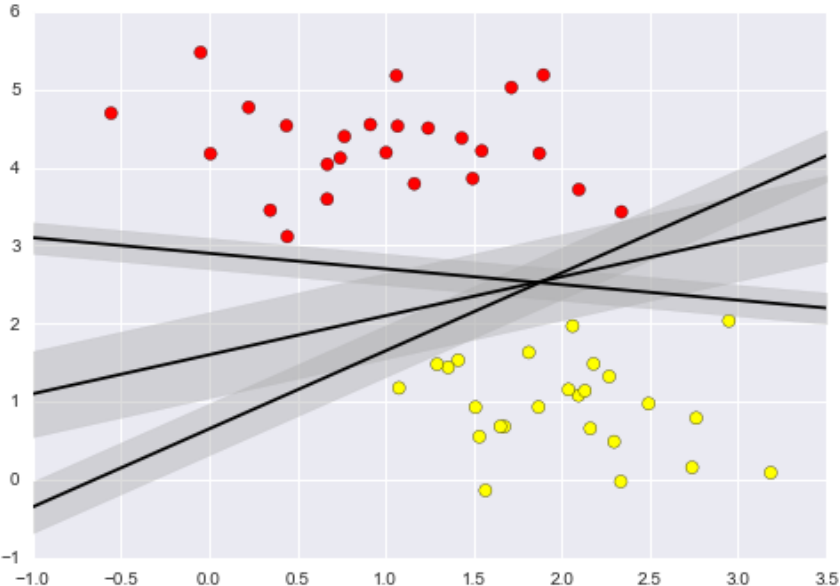
„Akademia Innowacyjnych Zastosowań Technologii Cyfrowych (AI Tech)”,
projekt finansowany ze środków Programu Operacyjnego Polska Cyfrowa POPC.03.02.00-00-0001/20

Fundusze Europejskie
Polska Cyfrowa

Rzeczpospolita Polska

Unia Europejska
Europejski Fundusz Rozwoju Regionalnego

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)

Zadanie

Problem

Naszkicuj na wykresie hiperpłaszczyznę separującą określoną wzorem $x_1 + 2x_2 - 3 = 0$.
Dorysuj również hiperpłaszczyzny określone wzorem: $x_1 + 2x_2 - 3 = 1$ *oraz* $x_1 + 2x_2 - 3 = -2$.
Zaznacz na wykresie punkty dla których $x_1 + 2x_2 - 3 > 0$.

Zadanie

Problem

Naszkicuj poniższe obserwacje na płaszczyźnie, a następnie:

- *zaznacz hiperpłaszczyznę separującą klasy o największym marginesie,*
- *podaj wzór na tę hiperpłaszczyznę,*
- *nanieś na wykres kierunek wektora wag.*

x_1	x_2	y
1	1	+1
3	2	+1
1	4	+1
2	4	-1
5	1	-1
6	3	-1
5	5	-1
5	4	-1

Zadanie

Problem

W zależności od liczby wymiarów d, jaka jest minimalna liczba obserwacji w zbiorze danych, aby można było określić unikalną hiperpłaszczyznę o maksymalnym marginesie?

Hiperpłaszczyzna separująca

- Jak zdefiniować hiperpłaszczyznę separującą dwie klasy?
- Jak obliczyć odległość pomiędzy hiperpłaszczyzną separującą a obserwacją?

Problem

Oblicz odległość poniższej obserwacji z klasy $-$

x_1	x_2	y
2	3	-1

do hiperpłaszczyzny określonej przez $w = [-4, 3]$ *i* $b = -2$,

- Jak obliczyć margines?

Przekształcenia problemu

$$\max_{w,b} \min_{i} \gamma_i = \frac{|f(x_i)|}{||w||}$$

Przy ograniczeniach:

$$f(x_i) \geq 0 \qquad \text{dla każdego przykładu z } y_i = 1$$

$$f(x_i) \leq 0 \qquad \text{dla każdego przykładu z } y_i = -1$$

Problem

Zdefiniuj problem optymalizacyjny dla następującego zbioru uczącego:

x_1	x_2	y
-1	7	+1
2	3	-1
4	2	+1

Przekształcenia problemu

$$\max_{w,b} \min_{i} \gamma_i = \frac{|f(x_i)|}{||w||}$$

Przy ograniczeniach:

$$f(x_i) \geq 0 \qquad \text{dla każdego przykładu z } y_i = 1$$

$$f(x_i) \leq 0 \qquad \text{dla każdego przykładu z } y_i = -1$$

Wyeliminujmy skomplikowaną funkcję celu $\min_i \gamma_i = \frac{|f(x_i)|}{||w||}$, która wybiera najbliższy do płaszczyzny punkt poprzez zastąpienie jej (sztuczną) zmienną γ^* oznaczającą odległość do najbliższego (ew. najbliższych) przykładu od płaszczyzny.

Przekształcenia problemu

$$\max_{w,b} \min_{i} \gamma_i = \frac{|f(x_i)|}{||w||}$$

Przy ograniczeniach:

$$f(x_i) \geq 0 \qquad \text{dla każdego przykładu z } y_i = 1$$

$$f(x_i) \leq 0 \qquad \text{dla każdego przykładu z } y_i = -1$$

Wyeliminujmy skomplikowaną funkcję celu $\min_i \gamma_i = \frac{|f(x_i)|}{||w||}$, która wybiera najbliższy do płaszczyzny punkt poprzez zastąpienie jej (sztuczną) zmienną γ^* oznaczającą odległość do najbliższego (ew. najbliższych) przykładu od płaszczyzny.

Przekształcenia problemu

$$\max_{w,b,\gamma^*} \gamma^*$$

$$f(x_i) \geq 0 \quad \text{jeśli } y_i = 1 \qquad f(x_i) \leq 0 \quad \text{jeśli } y_i = -1$$

$$\gamma^* \leq \gamma_i = \frac{|f(x_i)|}{||w||} \qquad \text{dla każdego i}$$

$$\gamma^* \leq \frac{|f(x_i)|}{||w||} \quad \Rightarrow \quad ||w||\gamma^* \leq |f(x_i)|$$

Ponieważ nie ma żadnych ograniczeń[1] co do $||w||$ to możemy je sobie wybrać w arbitralny sposób! W szczególności:

$$||w|| = \frac{1}{\gamma^*} \Rightarrow \gamma^* = \frac{1}{||w||}$$

[1] Płaszczyzna $w^T x + b = 0$ z $||w||$ oraz $cw^T x + cb = 0$ z $|c|||w||$ jest taka sama (jedyne co to inne $||w||$ wymaga innego b)

Przekształcenia problemu

$$\max_{w,b,\gamma^*} \gamma^*$$

$$f(x_i) \geq 0 \quad \text{jeśli } y_i = 1 \qquad f(x_i) \leq 0 \quad \text{jeśli } y_i = -1$$

$$\gamma^* \leq \gamma_i = \frac{|f(x_i)|}{||w||} \qquad \text{dla każdego i}$$

$$\gamma^* \leq \frac{|f(x_i)|}{||w||} \quad \Rightarrow \quad ||w||\gamma^* \leq |f(x_i)|$$

Ponieważ nie ma żadnych ograniczeń[1] co do $||w||$ to możemy je sobie wybrać w arbitralny sposób! W szczególności:

$$||w|| = \frac{1}{\gamma^*} \Rightarrow \gamma^* = \frac{1}{||w||}$$

[1] Płaszczyzna $w^T x + b = 0$ z $||w||$ oraz $cw^T x + cb = 0$ z $|c|||w||$ jest taka sama (jedyne co to inne $||w||$ wymaga innego b)

Przekształcenia problemu

$$\max_{w,b,\gamma^*} \gamma^*$$

$$f(x_i) \geq 0 \quad \text{jeśli } y_i = 1 \qquad f(x_i) \leq 0 \quad \text{jeśli } y_i = -1$$

$$\gamma^* \leq \gamma_i = \frac{|f(x_i)|}{||w||} \qquad \text{dla każdego i}$$

$$\gamma^* \leq \frac{|f(x_i)|}{||w||} \quad \Rightarrow \quad ||w||\gamma^* \leq |f(x_i)|$$

Ponieważ nie ma żadnych ograniczeń[1] co do $||w||$ to możemy je sobie wybrać w arbitralny sposób! W szczególności:

$$||w|| = \frac{1}{\gamma^*} \Rightarrow \gamma^* = \frac{1}{||w||}$$

[1] Płaszczyzna $w^T x + b = 0$ z $||w||$ oraz $cw^T x + cb = 0$ z $|c|||w||$ jest taka sama (jedyne co to inne $||w||$ wymaga innego b)

Przekształcenia problemu

$$\max_{w,b,\gamma^*} \gamma^*$$

$$f(x_i) \geq 0 \quad \text{jeśli } y_i = 1 \qquad f(x_i) \leq 0 \quad \text{jeśli } y_i = -1$$

$$\gamma^* \leq \gamma_i = \frac{|f(x_i)|}{||w||} \qquad \text{dla każdego i}$$

$$\gamma^* \leq \frac{|f(x_i)|}{||w||} \quad \Rightarrow \quad ||w||\gamma^* \leq |f(x_i)|$$

Ponieważ nie ma żadnych ograniczeń[1] co do $||w||$ to możemy je sobie wybrać w arbitralny sposób! W szczególności:

$$||w|| = \frac{1}{\gamma^*} \Rightarrow \gamma^* = \frac{1}{||w||}$$

[1] Płaszczyzna $w^T x + b = 0$ z $||w||$ oraz $cw^T x + cb = 0$ z $|c|||w||$ jest taka sama (jedyne co to inne $||w||$ wymaga innego b)

Przekształcenia problemu

$$\max_{w,b} \frac{1}{||w||}$$

Przy ograniczeniach:

$$f(x_i) \geq 0 \quad \text{jeśli } y_i = 1 \qquad f(x_i) \leq 0 \quad \text{jeśli } y_i = -1$$

$$\frac{1}{||w||} \leq \frac{|f(x_i)|}{||w||} \qquad \text{dla każdego i}$$

Przekształcenia problemu

$$\max_{w,b} \frac{1}{||w||}$$

Przy ograniczeniach:

$$f(x_i) \geq 0 \quad \text{jeśli } y_i = 1 \qquad f(x_i) \leq 0 \quad \text{jeśli } y_i = -1$$

$$1 \leq |f(x_i)| \qquad \text{dla każdego i}$$

Przekształcenia problemu

$$\max_{w,b} \frac{1}{||w||} \Rightarrow \min_{w,b} ||w||$$

Przy ograniczeniach:

$$f(x_i) \geq 0 \quad \text{jeśli } y_i = 1 \qquad f(x_i) \leq 0 \quad \text{jeśli } y_i = -1$$

$$1 \leq |f(x_i)| \qquad \text{dla każdego i}$$

Przekształcenia problemu

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

Przy ograniczeniach:

$$f(x_i) \geq 0 \quad \text{jeśli } y_i = 1 \qquad f(x_i) \leq 0 \quad \text{jeśli } y_i = -1$$

$$1 \leq |f(x_i)| \qquad \text{dla każdego i}$$

Przekształcenia problemu

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

Przy ograniczeniach:

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

Przekształcenia problemu

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

Przy ograniczeniach:

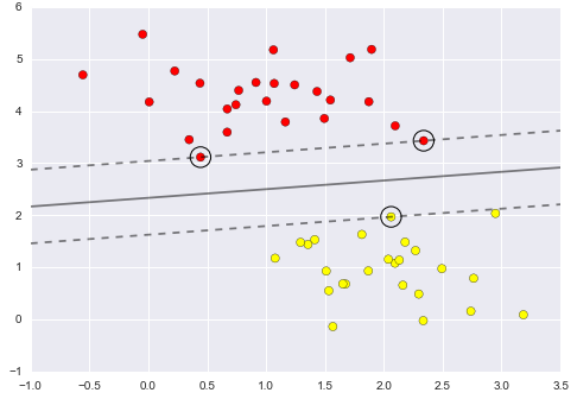
$$f(x_i) \geq 1 \quad \text{jeśli } y_i = 1 \qquad f(x_i) \leq -1 \quad \text{jeśli } y_i = -1$$

Problem

Zdefiniuj problem optymalizacyjny dla następującego zbioru uczącego:

x_1	x_2	y
-1	7	+1
2	3	-1
4	2	+1

Wektory podpierające



- Wektory podpierające to wektory, których odległość od indukowanej płaszczyzny jest równa marginesowi.
- Są to wektory, które „podpierają" płaszczyznę i uniemożliwiają jej zmianę położenia - od nich zależy granica decyzji.
- Można odrzucić wszystkie inne wektory, a wynik optymalizacji będzie ten sam!
- Można zmieniać położenie wszystkich inne wektorów (bez przekraczania przerywanej linii marginesu), a wynik optymalizacji będzie ten sam!
- Liczba wektorów podpierających jest miarą złożoności hipotezy.

Uogólnianie wiedzy przez SVM

Theorem (Vapnik)

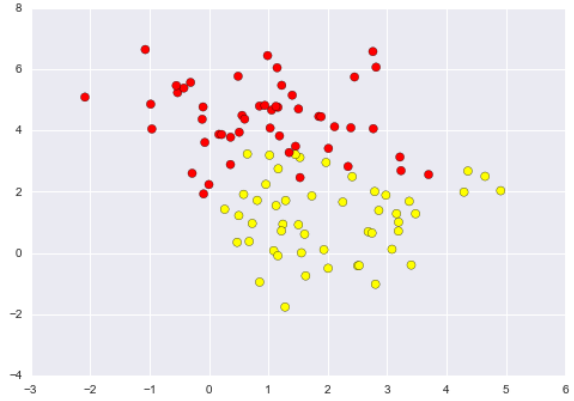
Jeśli zbiór uczący zawiera n przykładów rozdzielonych hiperpłaszczyzną o maksymalnym marginesie, wtedy oczekiwane (po zbiorach uczących) prawdopodobieństwo popełnienia błędu na zbiorze testowym jest ograniczone poprzez:

$$\mathbb{E}\, P(\hat{y} \neq y) \leq \mathbb{E}\left[\frac{m}{n}\right]$$

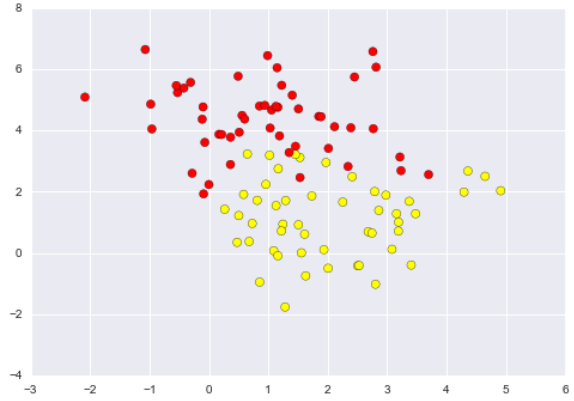
gdzie m to liczba wektorów podpierających.

Kilka analogicznych twierdzeń m.in. $\mathbb{E}\, P(\hat{y} \neq y) \leq \mathbb{E}\left[\frac{d}{n}\right]$ (d liczba wymiarów), które klasycznie były udowadniane dla innych metod. Tutaj: uogólnianie NIE zależy od liczby wymiarów (klątwa wymiarowości!).
W praktyce często liczba wektorów podpierających jest duża... ;(

Co jeśli zbiór nie jest liniowo separowalny?



Co jeśli zbiór nie jest liniowo separowalny?

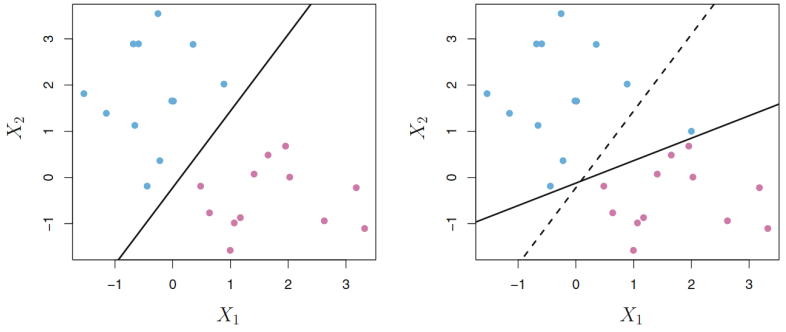


Dwa możliwe rozwiązania:

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

Soft-SVM – pomysł

- Pomysł: pozwólmy (niektórym) przykładom uczącym być w środku marginesu albo nawet po złej stronie hiperpłaszczyzny separującej
- Dlaczego?
  - Zbiór może być nieliniowy
  - Nawet jak jest liniowo separowalny: poprzednia definicja SVM gwarantuje nam 100% trafność! Jeden przykład z błędną etykietą mocno zmienia wynik!



- Jeśli jeden przykład mocno zmienia wynik – jesteśmy podatni na przeuczenie!

Soft-SVM

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

Przy ograniczeniach:

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

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

Soft-SVM

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

Przy ograniczeniach:

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

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

$$\xi_i \geq 0$$

Soft-SVM

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

Przy ograniczeniach:

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

$$\xi_i \geq 0$$

Soft-SVM

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

Przy ograniczeniach:

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

$$\xi_i \geq 0$$

Funkcja celu przyjmuje znaną w uczeniu maszynowym postać:

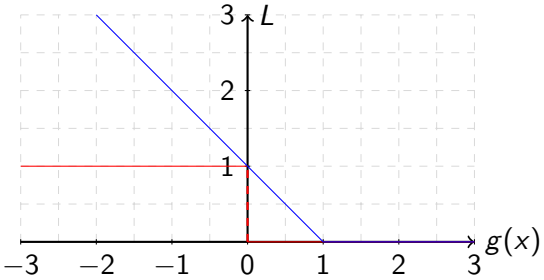
maksymalizuj prostotę modelu + minimalizuj błąd

Błąd zawiasowy

Przekształcając wzór dalej otrzymujemy:

$$\min_{w,b} \sum_{i=1}^{n} max(0, 1 - y_i g(x_i)) \underbrace{+\lambda||w||}_{\text{regularyzacja}}$$

przy czym $y_i \in \{-1, 1\}$.



- Błąd tej postaci nazywamy błędem zawiasowym (ang. *hinge loss*).
- Zauważ, że formulacja jest ogólna i za $g(x)$ można wstawić inny model wiedzy niż wyrażenie liniowe, uzyskując inny algorytm wykorzystujący tę funkcję błędu.

Uwaga terminologiczna

Wektory wspierające nie leżą już tylko „na marginesie" ale także „w środku marginesu" i po błędnej stronie granicy decyzji – wszystkie te wektory wpływają na wynik.

W niektórych pracach rozróżnia się pomiędzy:

- klasyfikatorem maksymalnego marginesu (ang. *maximal margin classifier, MMC*)
  $\Rightarrow$ nasza pierwsza formulacja problemu
- klasyfikatorem wektorów wspierających (ang. *support vector classifier, SVC*)
  $\Rightarrow$ wersja dla problemów nieliniowo separowalnych „soft-SVM"
- maszyną wektorów wspierających (ang. *support vector machine, SVM*)
  $\Rightarrow$ wersja z jądrami – przyszły tydzień ;)

Widzimy się za tydzień!