Uczenie maszynowe:
wykład 10

Paweł Cichosz

Model liniowy

Funkcja reprezentacji:

$$h(x) = \sum_{i=1}^{n} w_i a_i(x) + w_{n+1} = \sum_{i=1}^{n+1} w_i a_i(x) = \mathbf{w} \circ \mathbf{a}(x)$$

gdzie:

- $\mathbf{w}$ – wektor parametrów $w_1, w_2, \ldots, w_n, w_{n+1}$,
- $\mathbf{a}(x)$ – wektor wartości (ciągłych) atrybutów $a_1(x), a_2(x), \ldots, a_n(x), a_{n+1}(x)$, przy czym $a_{n+1}(x) \equiv 1$,
- $\circ$ – symbol iloczynu skalarnego.

Estymacja parametrów: minimalizacja *funkcji straty* proporcjonalnej do błędu średniokwadratowego na zbiorze trenującym:

$$E_{T,f}(h) = \frac{1}{2} \sum_{x \in T} (f(x) - h(x))^2$$

Gradientowa estymacja parametrów

Reguła spadku gradientu:

$$\mathbf{w} := \mathbf{w} + \beta \left(-\nabla_{\mathbf{w}} E_{T,f}(h)\right)$$

gdzie $0 < \beta < 1$ – rozmiar kroku.

Gradient błędu:

$$\nabla_{\mathbf{w}} E_{T,f}(h) = \sum_{x \in T} (f(x) - h(x))(-\nabla_{\mathbf{w}} h(x))$$

Gradient funkcji reprezentacji: $\nabla_{\mathbf{w}} h(x) = \mathbf{a}(x)$.

Gradientowa estymacja parametrów

Reguła aktualizacji parametrów:

$$\mathbf{w} := \mathbf{w} + \beta \sum_{x \in T} (f(x) - h(x))\mathbf{a}(x)$$

stosowana iteracyjnie.

Stochastyczny spadek gradientu (SGD): aktualizacja dla pojedynczych przykładów w kolejności losowej:

$$\mathbf{w} := \mathbf{w} + \beta (f(x) - h(x))\mathbf{a}(x)$$

Inicjalizacja parametrów: dowolna (np. 0 lub małe wartości losowe).

Kryteria stopu: mały błąd, mały spadek błędu, maksymalna liczba iteracji.

Bardziej zaawansowane metody gradientowe: np. metoda Newtona, *Adam*, *AdaGrad* – poza zakresem wykładu.

Przykład

x	a_1	a_2	f
1	1	1	2
2	1	2	5
3	1	3	8
4	2	1	3
5	2	2	4
6	2	3	5

$$\beta = 0.1$$

Inicjalizacja: $\mathbf{w} = [1, 1, 1]$.

$x = 1$:

$$h(x) = 3$$

$$w_1 = 1 + 0.1 \cdot (2 - 3) \cdot 1 = 0.9$$

$$w_2 = 1 + 0.1 \cdot (2 - 3) \cdot 1 = 0.9$$

$$w_3 = 1 + 0.1 \cdot (2 - 3) \cdot 1 = 0.9$$

$x = 2$:

$$h(x) = 3.6$$

$$w_1 = 0.9 + 0.1 \cdot (5 - 3.6) \cdot 1 = 1.04$$

$$w_2 = 0.9 + 0.1 \cdot (5 - 3.6) \cdot 2 = 1.18$$

$$w_3 = 0.9 + 0.1 \cdot (5 - 3.6) \cdot 1 = 1.04$$

. . . : kontynuacja pozostaje jako ćwiczenie.

Metoda najmniejszych kwadratów

- Wyznaczanie parametrów za pomocą zamkniętej formuły.
- Postulat dopasowania do danych trenujących jako nadokreślony układ równań liniowych:

$$a_1(x)w_1 + a_2(x)w_2 + \cdots + a_n(x)w_n + a_{n+1}(x)w_{n+1} = f(x)$$

  dla każdego $x \in T$, przy założeniu $n+1 < |T|$ (zwykle $n \ll |T|$).
- W zapisie wektorowo-macierzowym:

$$\mathbf{a}(T)\mathbf{w} = \mathbf{f}(T)$$

  gdzie:
  - $\mathbf{a}(T)$ – macierz $|T| \times (n+1)$ wartości atrybutów $a_1(x), \ldots, a_{n+1}(x)$ dla $x \in T$,
  - $\mathbf{f}(T)$ – wektor wartości $f(x)$ dla $x \in T$.
- Rozwiązanie przez pseudo-inwersję:

$$\begin{aligned}\mathbf{a}^\mathsf{T}(T)\mathbf{a}(T)\mathbf{w} &= \mathbf{a}^\mathsf{T}(T)\mathbf{f}(T)\\ \mathbf{w} &= \left(\mathbf{a}^\mathsf{T}(T)\mathbf{a}(T)\right)^{-1}\mathbf{a}^\mathsf{T}(T)\mathbf{f}(T)\end{aligned}$$

Przykład

x	a_1	a_2	f
1	1	1	2
2	1	2	5
3	1	3	8
4	2	1	3
5	2	2	4
6	2	3	5

$$\begin{bmatrix} 1 & 1 & 1 \\ 1 & 2 & 1 \\ 1 & 3 & 1 \\ 2 & 1 & 1 \\ 2 & 2 & 1 \\ 2 & 3 & 1 \end{bmatrix} \cdot \mathbf{w} = \begin{bmatrix} 2 \\ 5 \\ 8 \\ 3 \\ 4 \\ 5 \end{bmatrix}$$

$$\mathbf{w} = \left(\begin{bmatrix} 1 & 1 & 1 & 2 & 2 & 2 \\ 1 & 2 & 3 & 1 & 2 & 3 \\ 1 & 1 & 1 & 1 & 1 & 1 \end{bmatrix} \cdot \begin{bmatrix} 1 & 1 & 1 \\ 1 & 2 & 1 \\ 1 & 3 & 1 \\ 2 & 1 & 1 \\ 2 & 2 & 1 \\ 2 & 3 & 1 \end{bmatrix} \right)^{-1} \cdot \begin{bmatrix} 1 & 1 & 1 & 2 & 2 & 2 \\ 1 & 2 & 3 & 1 & 2 & 3 \\ 1 & 1 & 1 & 1 & 1 & 1 \end{bmatrix} \cdot \begin{bmatrix} 2 \\ 5 \\ 8 \\ 3 \\ 4 \\ 5 \end{bmatrix}$$

Reprezentacja liniowo-progowa

Wewnętrzna reprezentacja liniowa:

$$g(x) = \sum_{i=1}^{n} w_i a_i(x) + w_{n+1} = \mathbf{w} \circ \mathbf{a}(x)$$

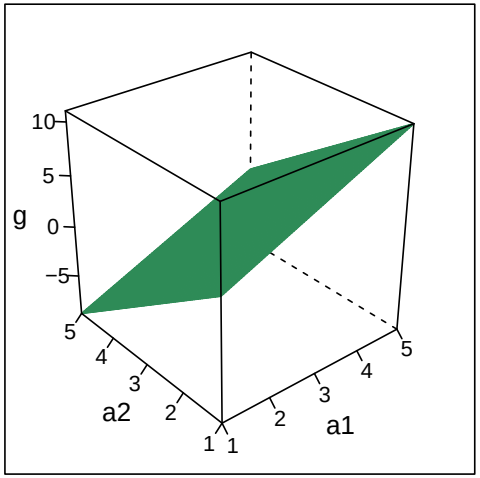
Zewnętrzna funkcja progowa:

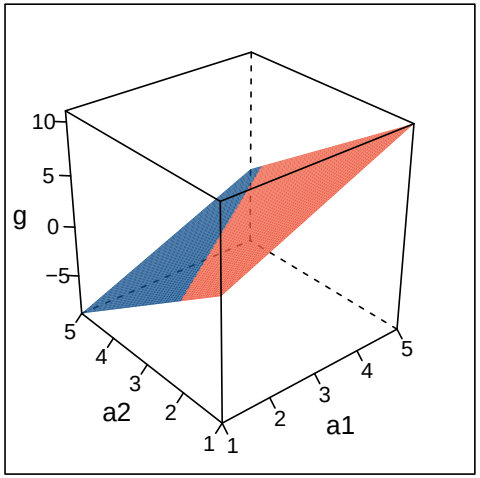
$$h(x) = \begin{cases} 1 & \text{jeśli } g(x) \geq 0 \\ 0 & \text{w przeciwnym przypadku} \end{cases}$$

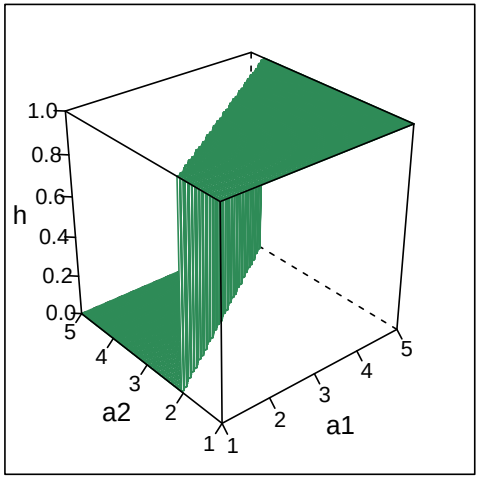
Granica decyzyjna: hiperpłaszczyzna o równaniu $g(x) = 0$ (punkty po stronie dodatniej klasyfikowane do klasy 1, punkty po stronie ujemnej klasyfikowane do klasy 0).

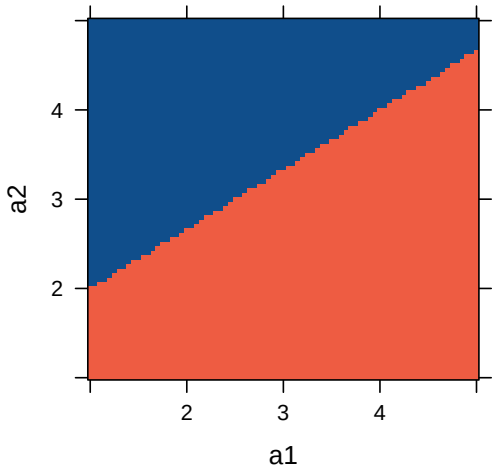
Wymiar VC: $n + 1$ – ograniczone ryzyko nadmiernego dopasowania.

Reprezentacja liniowo-progowa









Odległość od granicy decyzyjnej

Odległość ze znakiem:

$$\delta_{\mathbf{w}}(x) = \frac{\sum_{i=1}^{n} w_i a_i(x) + w_{n+1}}{\sqrt{\sum_{i=1}^{n} w_i^2}} = \frac{\mathbf{w} \circ \mathbf{a}(x)}{\|\mathbf{w}_{1:n}\|}$$

gdzie $\mathbf{w}_{1:n}$ – wektor parametrów $w_1, w_2, \ldots, w_n$ (bez w_{n+1}).

Odległość bezwzględna dla przykładów niepoprawnie klasyfikowanych:
$-c_-(x)\delta_{\mathbf{w}}(x)$, gdzie

$$c_-(x) = 2c(x) - 1 = \begin{cases} 1 & \text{jeśli } c(x) = 1 \\ -1 & \text{jeśli } c(x) = 0 \end{cases}$$

Funkcja decyzyjna: $g(x) = \mathbf{w} \circ \mathbf{a}(x)$ – proporcjonalna do odległości od granicy decyzyjnej, reprezentuje „ufność" predykcji modelu.

Algorytm prosty perceptron

Reguła aktualizacji parametrów:

$$\mathbf{w} := \begin{cases} \mathbf{w} + c_{-}(x)\mathbf{a}(x) & \text{jeśli } h(x) \neq c(x) \\ \mathbf{w} & \text{w przeciwnym przypadku} \end{cases}$$

Inicjalizacja: dowolna (np. 0).

Efekt aktualizacji: zmniejszenie odległości niepoprawnie klasyfikowanych przykładów od granicy decyzyjnej (uzasadnienie poza zakresem wykładu).

Tryb stosowania: iteracyjnie wielokrotnie dla wszystkich przykładów łącznie (wsadowy) lub dla pojedynczych przykładów (przyrostowy).

Kryterium stopu: poprawna klasyfikacja wszystkich przykładów.

Zbieżność: gwarantowana tylko jeśli w zbiorze trenującym klasy są liniowo separowalne.

Przykład

x	a_1	a_2	c
1	1	1	0
2	1	2	1
3	1	3	1
4	2	1	0
5	2	2	0
6	2	3	1

Inicjalizacja: $\mathbf{w} = [0.1, 0.1, 0.1]$.

$x = 1$:

$$g(x) > 0 \qquad h(x) = 1$$
$$w_1 = 0.1 + (-1) \cdot 1 = -0.9$$
$$w_2 = 0.1 + (-1) \cdot 1 = -0.9$$
$$w_3 = 0.1 + (-1) \cdot 1 = -0.9$$

$x = 2$:

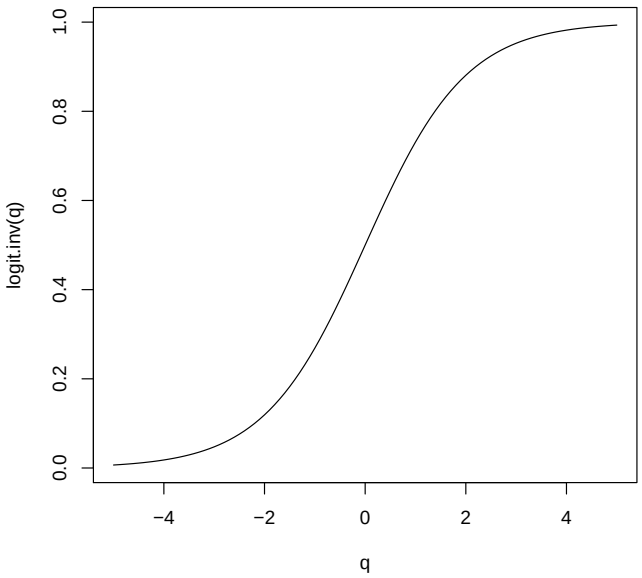
$$g(x) < 0 \qquad h(x) = 0$$
$$w_1 = -0.9 + 1 \cdot 1 = 0.1$$
$$w_2 = -0.9 + 1 \cdot 2 = 1.1$$
$$w_3 = -0.9 + 1 \cdot 1 = 0.1$$

. . . : kontynuacja pozostaje jako ćwiczenie.

Reprezentacja modelu regresji logistycznej

Wewnętrzna reprezentacja liniowa:

$$g(x) = \sum_{i=1}^{n} w_i a_i(x) + w_{n+1} = \mathbf{w} \circ \mathbf{a}(x)$$



Zewnętrzna logistyczna funkcja łącząca:

$$\text{logit}(p) = \ln \frac{p}{1-p}$$

$$\text{logit}^{-1}(q) = \frac{e^q}{e^q + 1} = \frac{1}{1 + e^{-q}}$$

$$\text{logit}^{-1}(g(x)) = \pi(x) = P(1|x)$$

Funkcja decyzyjna: $g(x) = \mathbf{w} \circ \mathbf{a}(x)$.

Estymacja parametrów regresji logistycznej

Maksymalizacja logarytmu wiarygodności:

$$\begin{aligned}\mathrm{LL}_{T,c}(\pi) &= \ln P(T, c; \pi) = \ln \prod_{x \in T} P(c(x)|x; \pi) \\ &= \ln \prod_{x \in T} \pi(x)^{c(x)} (1 - \pi(x))^{1-c(x)} \\ &= \sum_{x \in T} \big(c(x) \ln \pi(x) + (1 - c(x)) \ln(1 - \pi(x)) \big)\end{aligned}$$

Reguła wzrostu gradientu:

$$\mathbf{w} := \mathbf{w} + \beta \nabla_{\mathbf{w}} \mathrm{LL}_{T,c}(\pi)$$

Gradient logarytmu wiarygodności:

$$\nabla_{\mathbf{w}} \mathrm{LL}_{T,c}(\pi) = \sum_{x \in T} \big(c(x) - \pi(x) \big) \mathbf{a}(x)$$

(wyprowadzenie poza zakresem wykładu)

Atrybuty dyskretne

Kodowanie binarne: atrybut $a : X \to \{v_1, v_2, \ldots, v_k\}$ zastępowany przez $k-1$ atrybutów binarnych $a_1, a_2, \ldots, a_k$:

$$a_i(x) = \begin{cases} 1 & \text{jeśli } a(x) = v_i \\ 0 & \text{w przeciwnym przypadku} \end{cases}$$

- wartości v_i dla $i = 1, \ldots, k-1$ są reprezentowane przez $a_i(x) = 1$ i $a_j(x) = 0$ dla $j \neq i$,
- wartość v_k jest reprezentowana przez $a_j(x) = 0$ dla wszystkich $j = 1, \ldots, k-1$.

Klasyfikacja wieloklasowa

„1 kontra reszta” (one vs. rest, OvR): osobny model binarny dla każdej klasy (jedna klasa traktowana jako pozytywna, wszystkie pozostałe jako negatywne, predykcja przez wybór klasy o maksymalnej wartości funkcji decyzyjnej.

„1 kontra 1” (one vs. one, OvO): osobny model binarny dla każdej pary klas (jedna klasa traktowana jako pozytywna, druga jako negatywna), predykcja przez głosowanie.

Właściwości modeli liniowych

- Wyłącznie zależności przynajmniej w przybliżeniu liniowe lub liniowo separowalne.
- Dobór parametrów odporny na minima lokalne błędu.
- Efektywna metoda najmniejszych kwadratów (w przypadku regresji liniowej).
- Interpretowalne parametry modelu.
- Ograniczone ryzyko nadmiernego dopasowania.
- Możliwe bardziej zaawansowane warianty częściowo lub całkowicie pokonujące ograniczenia liniowości (np. regresja logistyczna, SVM/SVR).