

Лінійні моделі

У попередньому розділі ми встановили, що клас порогових поліномів на $\mathbb{R}$ має VC-вимірність ∞ , але представляється у вигляді об'єднання $\mathcal{H} = \bigcup_n \mathcal{H}_n$ класів поліномів обмеженого степеня, VC-вимірність яких скінченна.

Впр. Показати, що клас порогових поліномів на $\mathbb{R}^d$ має VC-вимірність $\infty \quad \forall d$.

Друге твердження мене здається вірним, але його довести складніше. Тут ми обмежилися поліномами степеня 1 - лінійними (точніше, афінними) ф-ціями d змінних. Їм відповідає добре відомий клас предикторів:

def. Лінійним предиктором у задачі ^{керованою} машинного навчання на $\mathbb{R}^d$ з областю визначення $\mathcal{Y}$ зветься композиція $\varphi \circ h_{w,b}: \mathbb{R}^d \rightarrow \mathcal{Y}$, де

$h_{w,b} : \mathbb{R}^d \rightarrow \mathbb{R} : x \mapsto \langle w, x \rangle + b = \sum_{i=1}^d w_i x_i + b$ —
арітметична ф-ція, що визначена вектором $w = (w_1, \dots, w_d) \in \mathbb{R}^d$
та числом $b \in \mathbb{R}$, а $\varphi : \mathbb{R} \rightarrow \mathcal{Y}$ — дефієна.

Рем. Тут $\mathcal{Y}$ це наприклад, скінч. множина класів (у задачі
класифікації) або $\mathbb{R}$ (у задачі регресії). Множину цих
ар. ф-цій позначатимемо $L_d = \{h_{w,b} \mid w \in \mathbb{R}^d, b \in \mathbb{R}\}$. При $b = 0$ ар.
ф-ція зветься однорідною (або просто лінійною; вник,
ар. ф-ції у заг. випадку мені часто звучать лінійними).

Насправді можна без обмеження загальності розглядати
лише однорідні ф-ції. Для цього введемо m зб. однорідні
координати на $\mathbb{R}^d$: будемо записувати m з коорд. $(x_1, \dots, x_d)$
як $\bar{x} = (1, x_1, \dots, x_d)$, і позначимо $\bar{w} = (b, w_1, \dots, w_d)$ (у
позн. як у деф.). Тоді
$$h_{w,b}(x_1, \dots, x_d) = \sum_{i=1}^d w_i x_i + b = \langle \bar{w}, \bar{x} \rangle = h_{\bar{w},0}(\bar{x}),$$

де $h_{\bar{w},0} \in L_{d+1}$ — однорідна.

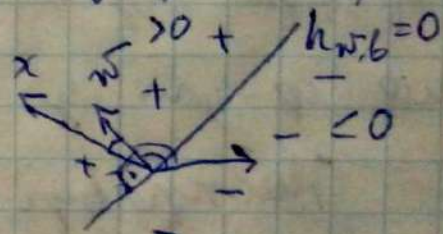
Лінійна бінарна класифікація

Для задані бінар. клас. розглянемо $Y = \{-1, 1\}$ і $\varphi := \text{sign}$ у понер. дєв., де для визначеності вважимо $\text{sign}(0) = 1$. Тоді, клас linomez має вигляд

$$HS_d = \text{sign} \circ L_d = \left\{ x \mapsto \text{sign} \left(\underbrace{\langle w, x \rangle + b}_{h_{w,b}(x)} \right) \right\}_{\substack{w \in \mathbb{R}^d \\ b \in \mathbb{R}}}$$

(або ан-по для однорізного списку). Тодімо точкам x присво

1 - це точки з "гогольного" навівпростору fig ліній-
ногому у $\mathbb{R}^d$ (прямої при $d=2$, площину при $d=3$) - тї, "в бік яких гвється" вектор



нормалі w го іперплєну $\langle w, x \rangle + b = 0$, тодімо ма
зних кут між w та вектором x гв. т. іперпл. го x
гострий (точніше, $\leq \frac{\pi}{2}$), а точка з літкою - лєткою
у протилєжнєму навівпросторі ("fig'скаву", кут тупий).

Тому цей клас зв. класом навівпросторів (halfspaces).

Оск. $h_{w,b}$ має $d+1$ гімєний параметр, оцінуваний є

наступне:

Тн. Клас наивирносногив HS_d на VC - величине $d+1$, а наивирносногив, ако задати ознаковану φ -функцию (може бити интервалу или прох. через 0) - d .

► Показано $d+1$ вектор $e_0 = (0, \dots, 0)$, $e_1 = (1, 0, \dots, 0)$, ..., $e_d = (0, \dots, 0, 1)$.
Нека h - дов. φ -функция з вредн. $y \in \{\pm 1\}$: $h(e_i) = y_i$, $i = \overline{0, d}$.
Пока $h_{w, b}$, где $b = y_0$, а $w = (y_1 - y_0, \dots, y_d - y_0)$ приймає те
же значення: $h_{w, b}(e_0) = b = y_0$ і $h_{w, b}(e_i) = y_i - y_0 + y_0 = y_i$ $\forall i = \overline{1, d}$; а отже і кожна з sign: $sign \circ h_{w, b} = h$. Тк. ч. HS_d
розб. дану множину: $VCdim(HS_d) \geq d+1$. Для ознаковану
мож само з множиною $\{e_1, \dots, e_d\}$ і $b = y_0 = 0$, може іт
 VC -вир. $\geq d$.

Одину зверну до ст. довести у ознаковану базу.
Дімо, $\forall d+2$ вектор $\{c_1, \dots, c_{d+2}\}$ якщо $\exists HS_d$ розд-
вача цю множину, то ознаковану φ -функ на $\mathbb{R}^{d+1}$ розд-вача

$\sigma \{ \bar{c}_1, \dots, \bar{c}_{d+1} \}$, де кожен $\bar{c} \in \text{Rem}$ вище означає го-
писування 1 першого коорд.

Т.ч. достатньо довести, що однор. $d+1$ -ви $\bar{c}_i$ на $\mathbb{R}^d$ не розли-
вають простору $\{ \bar{c}_i \}_{i=1}^{d+1}$. $\nearrow$ це не так, $\sqrt{\text{можливо } \exists \{ \bar{c}_i \}_{i=1}^{d+1}, \text{ що розд.}}$ В сфери $\dim \mathbb{R}^d = d$
і деб. лінійності, $\exists \{ \lambda_i \}_{i=1}^{d+1} \subset \mathbb{R} : \text{lin. комб. } \sum_{i=1}^{d+1} \lambda_i \bar{c}_i = 0$,
і не всі λ_i є нульові.

Положено $y_i := \text{sign } \lambda_i$, $i = \overline{1, d+1}$ (де, нагадаємо, $\text{sign } 0 = 1$). За

прин. про розбиття, $\exists$ ^{sign} $h_{w,0} : C_i \mapsto y_i$. Можна вбачити,

що $\langle w, c_i \rangle \neq 0 \forall i$. Оскільки, $\forall i$ $\text{sign } \langle w, c_i \rangle = y_i = \text{sign } \lambda_i$. Погі

$$0 = \langle w, \sum_{i=1}^{d+1} \lambda_i c_i \rangle \stackrel{=0}{=} \sum_{i=1}^{d+1} \lambda_i \langle w, c_i \rangle > 0,$$

То не всі $\lambda_i = 0$ ∇ . П.ч. VC-вим. $\leq d \Rightarrow = d$ для однор.
ф-цій, а отже $VCdim(HS_d) \leq d+1 \Rightarrow = d+1$. $\triangle$

Тож з загальних фактів маємо:

Сол. Клас наївних гіпотез на $\mathbb{R}^d \in PAC$ навчання для
ERM алгоритма зі складністю вибірки $\leq \frac{C}{\epsilon} \left((d+1) \ln \frac{1}{\epsilon} + \ln \frac{1}{\delta} \right)$
і еквівалентно PAC навчання для ERM однієї з кл. вибірки
 $\leq \frac{\Phi}{\epsilon^2} \left(d+1 + \ln \frac{1}{\delta} \right)$ (для деяких універсальних $C, \Phi > 0$),

Відемо, що $\sqrt{ERM}$ -навчання для цієї задачі є
обчислювально важким. Натомість наведемо два $\sqrt{ERM}$ -підходи
до "звичайного" PAC навчання, тобто з припущеннями
реалізованості, та оцінимо їх обчисл. складність (які
ми розподатимо складніші алгоритми для цієї задачі –
SVM). Тут прип. реалізованості означає, що $\exists h_{w,v}$, що
"правильно" розділяє точки (зокрема, у навчальних даних)

Зв'язок 1

і тому зветься це умовою розділювості (separability, separability). Щоб обійти цю умову, можна замінити 0-1 втрами (станд. для задач класифікації), розкладаючи інші ф-ції втрам (або замінити задачу клас. на лінійну регресію, див. нижче). Цим вважатимемо умову розділювості виконаною.

Отже, дані ^{навчальні} дані $S = \{(x_i, y_i)\}_{i=1}^m$. Треба знайти $h_{w,b} \in H_S$ з нульовою експрименою помилкою, тобто таку, що $h_{w,b}(x_i) = y_i \forall i$. Як пояснено вище, ф-цію завжди можна вважати однорідною, тобто $b = 0 \forall i$ $\text{sign} \langle w, x_i \rangle = y_i$. За прип. реалізованості (розділювості), такий $w \in \mathbb{R}^d$ існує. Також, трохи "повдурснувши" цю, можна вважати, що $\langle w, x_i \rangle \neq 0 \forall i$, а отже умову можна записати як $y_i \langle w, x_i \rangle > 0 \forall i$ (в однор. коорд. не може бути $x_i = 0$).

$$y_i \langle w, x_i \rangle > 0 \quad \forall i = \overline{1, m}$$

Третій підхід: лінійне програмування. Нагадаємо, що deb. (Оптимізація) задачею лінійного програмування (linear programming, LP) на $\mathbb{R}^d$ зветься знаходження вектора $w \in \mathbb{R}^d$, що максимізує (або мінімізує) лінійну ^{цільову} ф-цію $w \mapsto \langle u, w \rangle$ та задовольняє умови $Aw \geq v$ для деякої $m \times d$ матриці A і вектора $v \in \mathbb{R}^m$, тобто набору m лінійних нерівностей.

Тл.ч., задача LP задається векторами $u \in \mathbb{R}^d$, $v \in \mathbb{R}^m$ і m -ю A . Добре відомі ефективні методи розв'язування таких задач (симплекс-метод), а саме поліноміальні за m, d і розіграні типу дійсних чисел (float) (якщо умови нецелочисливі).

У нас умови $y_i \langle w, x_i \rangle > 0$ можна переписати наст. чином. За умовою розділюваності $\exists w^* \in \mathbb{R}^d : y_i \langle w^*, x_i \rangle > 0 \quad \forall i$. "Нормуємо" його: $w := \frac{w^*}{\min_{i=1, m} y_i \langle w^*, x_i \rangle}$. Тоді $\forall i$: $y_i \langle w, x_i \rangle = \frac{y_i \langle w^*, x_i \rangle}{\min_{i=1, m} y_i \langle w^*, x_i \rangle} \geq 1$, тобто $\exists w$ з такою умовою.

Почему того можна шукати як розв'язок задачі LP з умовами $Aw \geq b$, де рядки $m \times d$ -і a_i A — це координати x_i , помножені на y_i , і $b = (1, \dots, 1)$. При цьому y_i — це коефіцієнти, яку треба максимізувати, взятими просто нульовою: $u = 0$. Це означає, що для розв'язку LP поверне w , що задов. умовам.

Другий підхід: перцептрон Розенблатта. Це ітеративний алгоритм для знаходження w . Він базується послідовністю $w^{(1)}, w^{(2)}, \dots \in \mathbb{R}^d$, де $w^{(1)} = 0$, і, якщо $\exists i = \overline{1, m}$ таке, що $y_i \langle w^{(t)}, x_i \rangle \leq 0$, то алгоритм "наближає" $w^{(t)}$ до правильного значення на цьому прикладі: $w^{(t+1)} = w^{(t)} + y_i x_i$ (для першого такого i). При цьому

$$y_i \langle w^{(t+1)}, x_i \rangle = y_i \langle w^{(t)}, x_i \rangle + y_i^2 \langle x_i, x_i \rangle = y_i \langle w^{(t)}, x_i \rangle + |x_i|^2$$

тобто кожен крок дійсно збільшує початок.

У випадку псевдокода:

$$y_i \langle w^{(t)}, x_i \rangle,$$

$$w := 0$$

while $\exists i : y_i \langle w, x_i \rangle \leq 0 \Leftarrow$ (мима безпеки перше місце i)

$$w := w + y_i x_i$$

конец w

Коректність викон. цього алгоритма забезпечується теоремою (що таке розв. єдиними обчислювальну складність):

ТН. Нехай $S = \{(x_i, y_i)\}_{i=1}^m$ - розділені навч. дані. Позначмо

$$B := \min \{ \|w\| \mid y_i \langle w, x_i \rangle \geq 1, i = \overline{1, m} \}, \quad R := \max \{ \|x_i\| \}_{i=1}^m.$$

Тоді перекриття зупиниться після $\leq (RB)^2$ операцій, і

для його результату w $y_i \langle w, x_i \rangle > 0 \quad \forall i$, маємо

$\forall w \in \text{ERM}$ - алгоритм для задачі класифікації.

Рем. Тут, вище і далі $\|\cdot\|$ - евкл. норма: $\|x\|^2 = \sum_{i=1}^d x_i^2$. Існування

w з ф-ли для B забезпечується умовою розділюваності, як

було показано вище при обговоренні LP-метода. Тому вага.

$$0 < B < +\infty.$$

$\Rightarrow$ Если перемножить T неравенств, полученных м.ч. $w^{(T+1)}$.
 Тогда w^* вектор з осн. B (что $\exists$ в силу нормированности):
 $y_i \langle w^*, x_i \rangle \geq 1 \forall i$, $i | w^* | = B \leq$ норма $\forall$ вектора з
 осн. B вект. $\forall$ w $\langle w^{(t)}, x_i \rangle \leq 0$ (первое нерав.) Тогда

$$w^{(t+1)} = w^{(t)} + y_i x_i,$$

откуда

$$|w^{(t+1)}|^2 = |w^{(t)}|^2 + 2y_i \langle w^{(t)}, x_i \rangle + y_i^2 |x_i|^2 \leq |w^{(t)}|^2 + R^2$$

так как $|w^{(T+1)}|^2 \leq |w^{(T)}|^2 + R^2 \leq |w^{(T-1)}|^2 + 2R^2 \leq \dots \leq |w^{(1)}|^2 + TR^2 = TR^2$

Значит не,

$$\langle w^*, w^{(t+1)} \rangle = \langle w^*, w^{(t)} \rangle + y_i \langle w^*, x_i \rangle \geq \langle w^*, w^{(t)} \rangle + 1,$$

таким образом $\langle w^*, w^{(T+1)} \rangle \geq \langle w^*, w^{(T)} \rangle + 1 \geq \dots \geq \langle w^*, w^{(1)} \rangle + T = T$.

Оск. $|w^*| = B$, за пер. K - B - M . м.ч.

$$1 \geq \frac{\langle w^{(T+1)}, w^* \rangle}{|w^{(T+1)}| |w^*|} \geq \frac{T}{\sqrt{TR^2} \cdot B} = \frac{\sqrt{T}}{RB} \Rightarrow \sqrt{T} \leq RB \Rightarrow T \leq R^2 B^2.$$

($\cos \angle(w^{(T+1)}, w^*)$)

Все означает, что критерий не может быть выполнен за $R^2 B^2$,

значит, что для заданного n и $T \leq R^2 B^2$ не будет $\exists i$:
 $y_i \langle w^{(T+1)}, x_i \rangle \leq 0$. Все i пока не пер. $w = w^{(T+1)}$. Δ

Мінімальна регресія

В задачі регресії множина $Y = \mathbb{R}$, тому для лін. предиктора у деф. Терено $\varphi = \text{id} : h = h_{w,b}$ - φ -гія лін. регресії (гіпотеза) $\mathbb{R}^d \rightarrow \mathbb{R}$, а L_d - клас таких гіпотез.

Згадаємо, що тут "акісна" φ -гія втрачає ℓ_{0-1} не є корисною, тому розглянемо "кількісний" : $\ell_{sq}(h, (x, y)) = (h(x) - y)^2$ - квадратичне віхи, або абсолютну втрачу $\ell(h, (x, y)) = |h(x) - y|$.

Впр. Показати, що знаходження ERM-гіпотези для лін. регр. і φ -гії абс. втрачу $\ell(h, (x, y)) = |h(x) - y|$ можна реалізувати як задачу лін. програмування.

Далі розглядаємо саме ℓ_{sq} . Для неї ERM на $S = \{(x_i, y_i)\}_{i=1}^m$ це мінімізація значення

$$L_S(h) = \frac{1}{m} \sum_{i=1}^m (h(x_i) - y_i)^2 -$$

середньоквадратичної помилки (віхи, mean square error).

На віхи віхи задачі лін. клас., тут m вже не можна

оцінити через VC-випірність. Один зі способів - використання дискретизації, тобто те, що клас фактично скінч., і оцінку на складність вид. для адгностичної PAC навч. (див. розділ про рівномірну збігність), але для цього потрібна була б φ -гія витрат $0 \leq l \leq 1$. (least squares)

Розглянемо метод найменших квадратів - класичний спосіб розв'язку задачі лін. регресії. Як і раніше, не обм. загальність, шукаємо однорідні $h_{w,0}$ для $w \in \mathbb{R}^d$, згідно з ERM з усією класі на навч. вибірці $S = \{(x_i, y_i)\}_{i=1}^m$ для l_{sq} :

$$(h_{w,0})_S \in \argmin_w \frac{1}{m} \sum_{i=1}^m (h_{w,0}(x_i) - y_i)^2 = \argmin_w \frac{1}{m} \sum_{i=1}^m (\langle w, x_i \rangle - y_i)^2.$$

Щоб досягти мінімуму, w повинна бути критичною т. $\forall \bar{y} \in \mathbb{R}^d$ φ -гії $\mathbb{R}^d \rightarrow \mathbb{R}$, тобто її значення повинні бути 0:

$$\forall \bar{y} \stackrel{\text{т.д.}}{=} 0 = \frac{\partial}{\partial w_{\bar{y}}} \left(\frac{1}{m} \sum_{i=1}^m (\langle w, x_i \rangle - y_i)^2 \right) = \frac{2}{m} \sum_{i=1}^m (\langle w, x_i \rangle - y_i) x_{i,\bar{y}},$$

де $x_{i,\bar{y}}$ - $\bar{y}$ -та коорд. x_i !

$$\sum_{i=1}^m \langle w, x_i \rangle x_i = \sum_{i=1}^m y_i x_i$$

$$\forall \bar{j} = \overline{1, d} \quad \sum_{\substack{i=1, m \\ k = \overline{1, d}}} x_{i, \bar{j}} x_{i, k} w_k = \sum_{i=1}^m y_i x_{i, \bar{j}}$$

У матричній формі це виглядає як $Aw = b$, де A — $d \times d$ м-ца $\left(\sum_{i=1}^m x_{i, \bar{j}} x_{i, k} \right)_{\bar{j}, k=1}^d = \sum_{i=1}^m x_i x_i^T$ (якщо x_i записувати як вектор-стовпчик), $b = \sum_{i=1}^m y_i x_i \in \mathbb{R}^d$.

Впр. A неінвертна $\Leftrightarrow \{x_i\}_{i=1}^m$ — повна система у $\mathbb{R}^d$.

$\mathbb{R}^d = \text{span} \{x_i\}_{i=1}^m$. Більш точно, $\text{span} \{x_i\}_{i=1}^m$ — це образ лінійного $w \mapsto Aw$.

Якщо A інвертна ($\det A \neq 0 \Rightarrow \exists A^{-1}$), то $w = A^{-1}b$ дає потрібний розв'язок. У іншому випадку, оскільки A симетрична ($\sum_i x_{i, \bar{j}} x_{i, k} = \sum_i x_{i, k} x_{i, \bar{j}}$), з лінійної алгебри відомо, що вона має вигляд $A = V \mathcal{D} V^T$, де $V \in O(d)$ (ортогональна, тобто $V V^T = \overset{(\equiv V^T V)}{I_d}$), а $\mathcal{D}$ — діагональна (сум. з власних значень A). Позначимо через $\hat{\mathcal{D}}$ діаг. м-цю:

$$\hat{\mathcal{D}} := \left(\hat{d}_{i\bar{j}} := \begin{cases} 0, & i \neq \bar{j} \\ 0, & i = \bar{j}, d_{ii} = 0 \\ \frac{1}{d_{ii}}, & i = \bar{j}, d_{ii} \neq 0 \end{cases} \right)_{i, \bar{j} = \overline{1, d}},$$

$$\hat{A} := V \hat{D} V^T \quad ; \quad \hat{w} = \hat{A} b, \quad \Pi \log i$$

$$A \hat{w} = A \hat{A} b = V \hat{D} V^T V \hat{D} V^T b = [V \in O(d)] = V \hat{D} V^T b = V.$$

$$\left(\begin{array}{l} 0, \quad i \neq \bar{i} \text{ adı } d_{ii} = 0 \\ 1, \quad d_{ii} \neq 0 \end{array} \right)_{i, \bar{i} = 1, d} V^T b = \begin{bmatrix} \text{norm. } v_i - i\text{-min} \\ \text{составная } V \end{bmatrix} = \sum_{i: d_{ii} \neq 0} v_i v_i^T b = b.$$

Definirano, n -ga $\hat{I} := \begin{pmatrix} 0, i \neq j \text{ ali } d_{ii} = 0 \\ 1, d_{ii} \neq 0 \end{pmatrix}_{i,j=1,d}$ mat. vnačin. $\hat{I} = \hat{I}^2$
 (do tega gora z 0 na 1 na gori.), i $V \hat{I}$ - je n -ga z i -max
 strobnarib V na i -max, ambo $d_{ii} \neq 0$, i nujroboac strobnar. na
 i-max nujroboac; nam samo $\hat{I} V^T = (V \hat{I})^T$ - je n -ga z i -max
 strobnar. V je $d_{ii} \neq 0$, uo zaručani su i -mi pragu, i nujroboac
 pragu. Torej $\forall j, k = 1, d$ (j, k)-min ex-m $V \hat{I} V^T = (V \hat{I})(\hat{I} V^T)$ -
 je $\sum_{i: d_{ii} \neq 0} v_{ji} v_{ki} = \sum_{i: d_{ii} \neq 0} (v_i v_i^T)_{jk}$ (gub. buse an-min omuc

A). Знаємо з Вопр., що і образом лін. оператора з n -ого $\sum_{i: d_{ii} \neq 0} v_i v_i^T \in \text{span} \{v_i\}_{i=1, d_{ii} \neq 0}^d$. Але це ж саме можна сказати і про $A = VDV^T$: n -га D діагональна і має ненульові ел-ти на діаг. для тих же i . Тому образ таким же, як у оператора з n -ого A (за Вопр.): $\text{span} \{v_i\}_{i=1, d_{ii} \neq 0}^d = \text{span} \{x_i\}_{i=1}^n$.
 Оскільки $b = \sum_{i=1}^m y_i x_i$ належить до простору з цих просторів, він належить і до простору: $b = \sum_{i: d_{ii} \neq 0} \lambda_i v_i$. Тому

$$\sum_{i: d_{ii} \neq 0} v_i v_i^T b = \sum_{\substack{i: d_{ii} \neq 0 \\ j: d_{jj} \neq 0}} \lambda_j v_i v_i^T v_j = \left[\begin{array}{l} \text{умова ортого-} \\ \text{нальності: } v_i^T v_j = \\ \delta_{ij} = \begin{cases} 1, & i=j \\ 0, & i \neq j \end{cases} \end{array} \right] = \sum_{i: d_{ii} \neq 0} \lambda_i v_i = b.$$

(Сол.)

Отже, $A\hat{w} = b$, що означає, що $h_{\hat{w}, 0}$ - розв'язок ERM-задачі для лін. регресії (або $h_{\hat{w}, 0}$, де $\hat{w} = A^{-1}b$ у випадку невикористання). А) Як бачимо, цю знаходження зводиться до стандартних задач обчислювальної лінійної алгебри, для яких існують ефективні алгоритми.

Реш. Лінійну регресію (як і лін. біл. класифікацію)