Упередженість (відхилення) та складність класу гіпотез

У def. PAC навчаності розподіли $\mathcal{D}$ були обмежені припущенням реалізовності $\min_{h \in \mathcal{H}} L_{\mathcal{D},f}(h) = 0$ – це вираження того, що клас гіпотез $\mathcal{H}$ гарно підходить до задачі ML. У def. агностичної PAC навчаності замість цього маємо оцінку вигляду $L_{\mathcal{D}}(h_S) \leq \min_{h \in \mathcal{H}} L_{\mathcal{D}}(h) + \varepsilon$ – це означає, що помилка може бути великою для "погано сумісних" $\mathcal{D}$ і $\mathcal{H}$. У дійсно, ми показали, що для задачі бінарної клас. найменшу помилку дає для конкретного $\mathcal{D}$ байєсівський предиктор $f_{\mathcal{D}}$, і що "поганий" вибором $\mathcal{D}$ (коли мітка не залежить від даних) її можна зробити будь-якою $\leq \frac{1}{2}$ (це дуже погано).

Підійдемо тепер з іншого боку: покажемо, що у цій задачі $\forall$ алгоритма ML $\exists$ розподіл $\mathcal{D}$ що не є "поганим" взагалі насправді (мітка визначається даними), але поганий саме для цього алгоритма. Це значить, що не існує універсальних алгоритмів ML: вибір алг. повинен базуватися на наших знаннях про задачу ML (тобто про можливі $\mathcal{D}$). Це пояснює назву!

Th. (про неіснування безкоштовного обіду, No-Free-Lunch (NFL) theorem).
Нехай A - деякий алгоритм бінарної класифікації на множині X, а m - довільне натуральне $< \frac{|X|}{2}$. Тоді існує розподіл $\mathcal{D}$ на $X \times \{0,1\}$ з наступними властивостями:

- $L_{\mathcal{D}}(f) = 0$ для деякої $f : X \to \{0,1\}$;
- результат застосування A до навчальних даних S з m ел-тів $X \times \{0,1\}$, що i.i.d. відповідно до $\mathcal{D}$, має помилку $\geq \frac{1}{8}$ зі ймовірністю $\geq \frac{1}{7}$:

$$\underset{S\sim\mathcal{D}^m}{P}\left(L_{\mathcal{D}}(A(S)) \geq \frac{1}{8}\right) \geq \frac{1}{7}.$$

Rem. Тут пом. $L_{\mathcal{D}}$ - для 0-1 втрати l_{0-1}, як завжди у задачах класифікації. Перша умова означає, що $\forall$ клас гіп. $\mathcal{H}$, що містить f, задов. припущенню реалізовності. Тому, наприклад, $\forall$ скінч. $\mathcal{H} \ni f$ є PAC навчаним, і $\forall$ $ERM_{\mathcal{H}}$ - алгоритм працює краще за A (для відп. виборів ε і δ у def. PAC навчаності).

$\triangleright$ Нагадаємо загальний факт з теорії ймовірностей!

<u>Лг.</u> (нерівність Маркова). Для $\forall$ випадкової величини $X \geq 0$ і $a > 0$
$$P(X \geq a) \leq \frac{E(X)}{a}.$$

$\triangleright$ Ф-ція $x \mapsto P(X \geq x)$, оч., незростаюча, а її інтеграл - мат. очікування X за ф-лою залишк. сум:
$$E(X) = \int_0^{+\infty} P(X \geq x)\,dx \geq \int_0^a P(X \geq x)\,dx \geq \int_0^a P(X \geq a)\,dx = a\,P(X \geq a).\ \blacktriangle(\text{Лг.})$$

<u>Сон</u> $\forall$ вип. вел. $0 \leq X \leq 1$ і $a \in (0,1)$ $P(X > a) \geq \frac{E(X) - a}{1-a}$.

$\triangleright$ $P(X > a) = 1 - P(X \leq a) = 1 - P(1 - X \geq 1 - a) \geq [1 - X \geq 0,$ попер. Лг.$] \geq 1 - \frac{E(1-X)}{1-a} = \left[\begin{matrix}\text{лін. } E,\\ E(1) = 1\end{matrix}\right] = 1 - \frac{1 - E(X)}{1-a} = \frac{E(X) - a}{1-a}.\ \blacktriangle$ (<u>Сон</u>).

Застосовуючи <u>Сон</u> до вип. величини $S \mapsto L_{\mathcal{D}}(A(S)) \in [0,1]$ на $(\mathcal{X} \times \{0,1\})^m$ отримаємо друге тв. <u>Тh</u> за умови $\underset{S \sim \mathcal{D}^m}{E}\left(L_{\mathcal{D}}(A(S))\right) \geq \frac{1}{4}$. (для $a = \frac{1}{8}$), тому її й будемо перевіряти.

За умовою <u>Тh.</u>, $\exists$ $C \subset X$ з $2m$ ел-тів, і $\forall S$ з m ел-тів т.ч. містить не більше половини ел-тів C. Ідея: обрати $\mathcal{D}$ і f так, що $A(S)$ дає велику помилку на іншій половині.

Перелічимо усі можливі $f: C \to \{0,1\}$: $f_1, \dots, f_T$, де $T = 2^{2m}$. $\forall i$ визначимо розподіл на $\mathcal{X} \times \{0,1\}$, що перетворює його на дискретний ймовірнісний простір:

$$\mathcal{D}_i(\{(x,y)\}) = \begin{cases} \frac{1}{|C|} = \frac{1}{2m}, & x \in C \text{ і } y = f_i(x); \\ 0 & \text{у інш. вип.} \end{cases}$$

Покажемо, що для даного A $\exists$ таке i, що $\underset{S \sim \mathcal{D}_i^m}{E}(L_{\mathcal{D}_i}(A(S))) \geq \frac{1}{4}$.

За зауваженням вище, такий $\mathcal{D}_i$ буде задов. тв. теореми для $\forall f: \mathcal{X} \to \{0,1\}$ такої, що $f|_C = f_i$ (бо $L_{\mathcal{D}_i}(f) = 0$ за побудовою).

За виглядом $\mathcal{D}_i$, вибірка S з m ел-тів $\mathcal{X}$, що i.i.d. за $\mathcal{D}_i$, міститься у C зі ймов. 1. Тому далі вважаємо $S \subset C$.

Таких S усього $K = (2m)^m$, перенумеруємо їх: $S_1, \dots, S_K$. $\forall j = \overline{1,K}$, $i = \overline{1,T}$ розставимо мітки у $\underbrace{S_j}_{=(x_{j1}, \dots, x_{jm})}$ відп. до f_i, отримавши $S_j^i = \{(x_{j1}, f_i(x_{j1})), \dots, (x_{jm}, f_i(x_{jm}))\}$. Якщо розподілом є $\mathcal{D}_i$, то навч. дані мають вигляд $S_1^i, \dots$, або S_K^i зі ймов. $\frac{1}{K}$ (кожний), тому

$$\underset{S \sim \mathcal{D}_i^m}{E}(L_{\mathcal{D}_i}(A(S))) = \frac{1}{K} \sum_{j=1}^{K} L_{\mathcal{D}_i}(A(S_j^i)).$$

Найбільше за i знач. цих очікувань не перевищує середнє арифм.:

$$\max_{i=\overline{1,T}} \mathop{\mathbb{E}}_{S\sim\mathcal{D}_i^m}\left(L_{\mathcal{D}_i}(A(S))\right) \geq \frac{1}{T}\sum_{i=1}^{T} \mathbb{E}\left(L_{\mathcal{D}_i}(A(S))\right) = \left[\begin{matrix}\text{попер. рівн.,}\\ \text{зміна порядку додав.}\end{matrix}\right] =$$

$$= \frac{1}{K}\sum_{j=1}^{K}\frac{1}{T}\sum_{i=1}^{T} L_{\mathcal{D}_i}(A(S_j^i)) \geq \left[\begin{matrix}\text{сер. ар.} \geq\\ \text{найменше}\end{matrix}\right] \geq \min_{j=\overline{1,K}} \frac{1}{T}\sum_{i=1}^{T} L_{\mathcal{D}_i}(A(S_j^i)) =$$

Тут $p \geq m$.

$\forall j$ нехай $\{y_1, \ldots, y_p\}$ - ел-ти C, що не містяться у S_j.

Наступні оцінки вірні $\forall j$, тому у $=$ можемо вважати, що вже обрали найменше:

$(m \leq p)$

$$= \frac{1}{T}\sum_{i=1}^{T}\frac{1}{2m}\sum_{x\in C} 1_{\{A(S_j^i)(x)\neq f_i(x)\}} \geq \frac{1}{T}\sum_{i=1}^{T}\frac{1}{2p}\sum_{l=1}^{p} 1_{\{A(S_j^i)(y_l)\neq f_i(y_l)\}} \geq$$

$$\geq \left[\begin{matrix}\text{змінюємо пор. додав.,}\\ \text{сер. ар.} \geq \min\end{matrix}\right] \geq \frac{1}{2}\min_{l=\overline{1,m}} \frac{1}{T}\sum_{i=1}^{T} 1_{\{A(S_j^i)(y_l)\neq f_i(y_l)\}} =$$

$\forall i=\overline{1,T}$ $\exists!\ i'=\overline{1,T}$: f_i збігається з $f_{i'}$ всюди на C крім y_l (для фікс. l) : $f_i(c)=f_{i'}(c)$ при $c\neq y_l$ і $f_i(y_l)\neq f_{i'}(y_l)$. При цьому $S_j^i = S_j^{i'}$, бо $y_l \notin S_j$. Т.ч., числа від 1 до T розбиваються на пари (i, i'), в $\forall$ з яких

$$1_{\{A(S_j^i)(y_l)\neq f_i(y_l)\}} + 1_{\{A(S_j^{i'})(y_l)\neq f_{i'}(y_l)\}} = 1,$$

тому сер. ар. перед ⊖ дор. $\frac{1}{2}$ $\forall l$ і ⊜ $\frac{1}{4}$, що й потрібно $\triangle$

Упр. Узагальнити формулювання і доведення на випадок $m < \frac{|X|}{k}$, $k \in \mathbb{N}$

Rem. У одному з Ex. вище ми бачили, що брати в якості $\mathcal{H}$ в ERM-алгоритмі клас усіх ф-цій $X \to \{0,1\}$ погано, бо може привести до перенавчання. Тепер ми можемо формально довести "поганість" цього класу:

Сл. Клас усіх $h: X \to \{0,1\}$ у задачі класифікації на нескінченній X не є PAC навчаним.

▷ (припустимо, що) це не так, тобто для цього класу $\mathcal{H}$ визначене $m_{\mathcal{H}}: (0,1)^2 \to \mathbb{N}$ з def. PAC навч. і алгоритм A. Нехай $m \geq m_{\mathcal{H}}(\varepsilon, \delta)$, де $\varepsilon < \frac{1}{8}$ і $\delta < \frac{1}{7}$. Застосуємо до A і m NFL-теорему: отримаємо $\mathcal{D}$ і f (що гарантує виконання прип. реалізовності). Тоді $\mathbb{P}_{S \sim \mathcal{D}^m}\left(L_{\mathcal{D}}(A(S)) \geq \frac{1}{8}\right) \geq \frac{1}{7}$ за Th. і $\mathbb{P}_{S \sim \mathcal{D}^m}\left(L_{\mathcal{D}}(A(S)) \leq \varepsilon < \frac{1}{8}\right) \geq 1-\delta > \frac{6}{7}$ за def. PAC навчаності, але сума цих ім. повинна бути ≤ 1 ↯ (протиріччя) $\triangle$

Rem. Звичайно, для скінч. X клас усіх гіп. скінченний (як і усі $\mathcal{H}$, що містяться у ньому), тому формально є PAC навчаним. Втім, для нього маємо оцінку

$$m_{\mathcal{H}}(\varepsilon,\delta) \le \left\lceil \frac{1}{\varepsilon}\ln\frac{|\mathcal{H}|}{\delta}\right\rceil = \left\lceil \frac{1}{\varepsilon}\ln\frac{2^{|\mathcal{X}|}}{\delta}\right\rceil = \left\lceil \frac{\ln\frac{1}{\delta}+|\mathcal{X}|\ln 2}{\varepsilon}\right\rceil,$$

що для малих ε і δ суттєво перевищує $|\mathcal{X}|$! Т.ч., цей клас у будь-якому випадку не є корисним.

Із задачею вибору "правильного" $\mathcal{H}$ для даної задачі пов'язане т. зв. розкладення помилки:

<u>def.</u> Нехай у задачі ML на множині Z обрано вибірку S з ел-тів i.i.d. відн. до розподілу $\mathcal{D}$ на Z, застосовано $ERM_{\mathcal{H}}$-алгоритм для деякого класу гіпотез $\mathcal{H}$ і отримано гіпотезу h_S. Тоді її

– <u>помилкою апроксимації</u> (approximation error) зветься $\varepsilon_{app} := \min_{h\in\mathcal{H}} L_{\mathcal{D}}(h)$;

– <u>помилкою оцінки</u> (estimation error) зветься $\varepsilon_{est} := L_{\mathcal{D}}(h_S) - \varepsilon_{app}$;

де $L_{\mathcal{D}}$ обчислюється відн. до ф-ції втрат $\ell: \mathcal{H} \times Z \to \mathbb{R}_+$.

<u>Rem.</u> Тобто $L_{\mathcal{D}}(h_S) = \varepsilon_{app} + \varepsilon_{est}$, де $\varepsilon_{app} \geqslant 0$ не залежить від S (визначається класом $\mathcal{H}$). Зокрема, у задачі клас., $\varepsilon_{app} = 0$, якщо вик. припущення реалізовності, і $\varepsilon_{app} \geqslant L_{\mathcal{D}}(f_{\mathcal{D}})$, де $f_{\mathcal{D}}$ - байєсівський предиктор для $\mathcal{D}$, як бачили вище. З іншого боку, $\varepsilon_{est} \geqslant 0$ - це помилка, що оцінюється зверху у <u>вв.</u> (агностичної) PAC навчаності незалежно від $\mathcal{D}$. Вибір $\mathcal{H}$ визначається спів-відношенням між цими помилками:

- малеиькі ε_{app} означають великі (складні) $\mathcal{H}$. Зокрема, у задачі клас. клас усіх гіпотез дає найменше можливе знач. $\varepsilon_{app} = L_{\mathcal{D}}(f_{\mathcal{D}})$, але, як показано вище, великі ε_{est} - не є PAC навчаним (принаймні практично). Це проблема <u>перенавчання</u>: маленька упередженість, занадто велика складність.

- маленьких ε_{est} можна досягти для маленьких (простих) $\mathcal{H}$ в силу їх (аг.)PAC навчаності та відн. малих оцінок на $m_{\mathcal{H}}$ (порядка $\ln|\mathcal{H}|$). Але вони можуть мати великі ε_{app} - не просто не містити (у задачі клас.) $f_{\mathcal{D}}$, а й бути "далекими" від нього. Це проблема недонавчання (underfitting): жодна з гіпотез нездатна описати $\mathcal{D}$.

В силу сказаного, вибір $\mathcal{H}$ зазвичай характеризують як компроміс між упередженістю (відхиленням) і складністю (bias-complexity tradeoff).

Ex. Розглянемо вибір кількості параметрів d у задачі типу "класифікації фруктів": бінарної класифікації з $X \subset \mathbb{R}^d$, де $\mathcal{H}_d$ складається з паралелепіпедів $[a_1, b_1] \times \ldots \times [a_d, b_d]$ (тобто їхніх характеристичних ф-цій). Більш точно, нехай $X \subset \mathbb{R}^D$, де D - найбільша можл. кількість параметрів, а ф-ції з $\mathcal{H}_d$ задаються $[a_1, b_1] \times \ldots \times [a_d, b_d] \times \mathbb{R} \times \ldots \times \mathbb{R}$, $d < D$. Тоді $\varepsilon_{app} = \min\limits_{h \in \mathcal{H}_d} L_{\mathcal{D}}(h)$ зменшується зі збільшенням d, бо клас гіпотез зростає (характер цього зменш. залежать від $\mathcal{D}$). З іншого боку, як було показано на практиці, усі ці класи є PAC навчаними, і $m_{\mathcal{H}_d}(\varepsilon, \delta) \leq$ $\leq \left\lceil \frac{2d}{\varepsilon} \ln \frac{2d}{\delta} \right\rceil$. З результатів наст. розділу випливатиме, що вони також агностично PAC навчані з $m_{\mathcal{H}_d}(\varepsilon, \delta) \leq C \frac{2d + \ln \frac{1}{\delta}}{\varepsilon^2}$ для деякого $C \in \mathbb{R}$. Це означає, що для досягнення тієї ж $\varepsilon_{est} = \varepsilon$ зі ймов. $\geq 1 - \delta$, нам потрібно брати більші вибірки S для більших d.

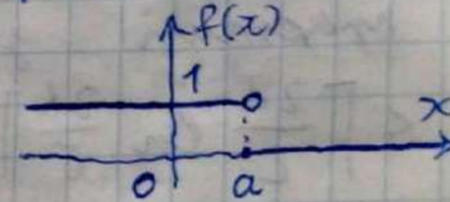
VC-вимірність та критерій навчаності

Перелічимо деякі приклади PAC навчаних і ненавчаних класів:

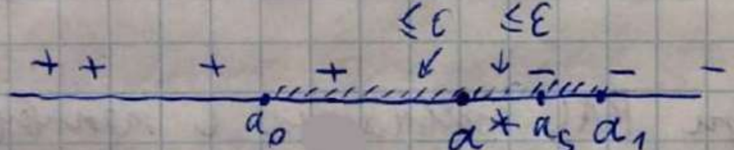
Ex. 1. Усі скінч. класи $\mathcal{H}$ є PAC навчаними для $m_{\mathcal{H}}(\varepsilon, \delta) \leq$ $\leq \left\lceil \frac{1}{\varepsilon} \ln \frac{|\mathcal{H}|}{\delta} \right\rceil$ і агностично PAC навчаними для $m_{\mathcal{H}}(\varepsilon, \delta) \leq \left\lceil \frac{2}{\varepsilon^2} \ln \frac{2|\mathcal{H}|}{\delta} \right\rceil$ (див. вище).

2. Клас усіх відображень $\mathcal{X} \to \{0, 1\}$ для задачі бін. клас. на нескін. $\mathcal{X}$ не є PAC навчаним (а отже й агностично) (попер. розділ).

3. (Нескінченний) клас порогових (threshold) ф-цій



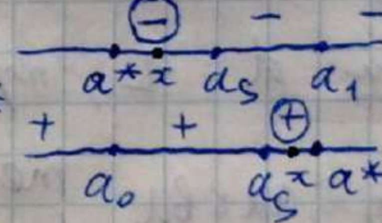
$\mathcal{H} = \{h_a := 1_{\{x<a\}}\}_{a\in\mathbb{R}} \in$ PAC навчаним для задачі бін. клас. на $X = \mathbb{R}$. Дійсно, нехай $\varepsilon, \delta \in (0,1)$. Умова реалізовності означає, що $\exists\, a^* \in \mathbb{R}$ таке, що $x < a^*$ майже для усіх x з $f(x)=1$ (де f — "правильна" ф-ція міток) і $x \geq a^*$ майже для усіх x з $f(x)=0$, зокрема $\forall$ вибірка S містить усі "+" зліва від a^* і "−" — справа зі ім. 1:



"−": $f(x)=0$,
"+": $f(x)=1$.

Розглянемо алг., що повертає найменше з x таких, що $f(x)=0$ (тобто $(x,0)\in S$), зокрема $a_S \geq a^*$; або $\forall\, a_S$, що більше за усі x з $f(x)=1$ (тобто $(x,1)\in S$), якщо S не містить $(x,0)$. Оч., для нього $L_S(h_{a_S}) = 0$, тобто він $ERM_{\mathcal{H}}$ (та лінійний за $m = |S|$).

Нехай $(a_0, a_1) \ni a^*$ такий, що (a_0, a^*) і (a^*, a_1) мають міри ε (або принаймні $\le \varepsilon$; a_0 може бути $-\infty$, а a_1 — $+\infty$). Тоді для $a_S \in (a_0, a_1)$ точки $(x,y) \in X \times \{0,1\}$, де $h_{a_S}(x) \neq f(x)$ - це або $(x,0)$ з $x < a_S$ і $x \geq a^*$ при $a_S \geq a^*$, або $(x,1)$ з $x \geq a_S$ і $x < a^*$ при $a_S < a^*$ (крім точок міри 0). В будь-якому разі їх міра, тобто помилка $L_{\mathcal{D},f}(h_{a_S}) \le \varepsilon$.



Тому $\mathcal{D}^m(\{S \mid L_{\mathcal{D},f}(h_{a_S}) > \varepsilon\}) \le [\text{моном.}] \le \mathcal{D}^m(\{S = \{(x_i,y_i)\}_{i=1}^m \mid \forall i\,(x_i \notin (a_0, a^*) \wedge y_i = 1) \vee (x_i \notin [a^*, a_1) \wedge y_i = 0)\}) \le [\text{субад.}] \le$

$\le \mathcal{D}^m(\{S \mid \forall i\; x_i \notin (a_0, a^*)\}) + \mathcal{D}^m(\{S \mid \forall i\; x_i \notin [a^*, a_1)\}) \le [\text{моном.}] \le$

це завжди вірно при $y_i = 0$ це завжди так при $y_i = 1$

$\le \mathcal{D}^m((\mathbb{R} \setminus (a_0, a^*))^m) + \mathcal{D}^m((\mathbb{R} \setminus (a^*, a_1))^m) = 2(1-\varepsilon)^m < 2e^{-\varepsilon m} \le$

Як завжди, при $m \geq \frac{1}{\varepsilon} \ln \frac{2}{\delta}$ це $\le \delta$, тому $\mathcal{D}^m(\{S \mid L_{\mathcal{D},f}(h_{a_S}) \le \varepsilon\}) \geq 1 - \delta$, що демонструє PAC навч. для $m_{\mathcal{H}}(\varepsilon, \delta) \le \lceil \frac{1}{\varepsilon} \ln \frac{2}{\delta} \rceil$.

4. Аналогічним чином на практиці показали, що (неск.) клас характеристичних ф-цій d-вимірних куль $\mathcal{H} = \{1_{\{|x| \le r\}}\}_{r \geq 0} \in$ PAC навч. для $X \subset \mathbb{R}^d$ (у

задачі бін. клас.), і $m_{\mathcal{H}}(\varepsilon,\delta) \le \lceil \frac{1}{\varepsilon} \ln \frac{1}{\delta} \rceil$ (незалежно від d)

5. Для (неск.) класу ч-ф d-вим. паралелепіпедів (відрізків при $d=1$, прямокутників при $d=2$) $\mathcal{H} = \{ 1_{[a_1,b_1]\times\ldots\times[a_d,b_d]} \}_{a_1\le b_1,\ldots,a_d\le b_d}$ теж маємо PAC навчаність у задачі бін. клас. на $\mathcal{X} \subset \mathbb{R}^d$, але тепер $m_{\mathcal{H}}(\varepsilon,\delta) \le \lceil \frac{2d}{\varepsilon} \ln \frac{2d}{\delta} \rceil$ залежить від d (теж показали на практиці).

Як бачимо, нескінченні класи бувають як PAC навч., так і ні, а відповідна складність вибірки може залежати від "кількості параметрів" класу. Щоб узагальнити ці спостереження, нам знадобиться наступна серія понять, що мотивується доведенням NFL-теореми у попер. розділі. Знову розглядаємо задачу бін. клас. на $\forall$ множині $\mathcal{X}$.

Def. Обмеженням класу гіпотез $\mathcal{H}$, тобто деякої множини відобр. $h: \mathcal{X} \to \{0,1\}$, на скінченну підмножину $C \subset \mathcal{X}$ зветься множина обмежень $h|_C$ для усіх $h \in \mathcal{H}$. Ототожнимо її з множиною

$$\mathcal{H}_C = \{(h(c_1), \dots, h(c_m)) \mid h \in \mathcal{H}\} \subset \{0,1\}^{|C|},$$

де $C = \{c_1, \dots, c_m\}$.

<u>def.</u> Говорять, що клас гіпотез $\mathcal{H}$ <u>розбиває</u> (shatters) скінч. множину $C \subset \mathcal{X}$, якщо будь-яка ф-ція $C \to \{0,1\}$ є обмеженням якоїсь ф-ції з $\mathcal{H}$, тобто $\mathcal{H}_C = \{0,1\}^{|C|}$.

<u>Rem.</u> У цьому випадку $|\mathcal{H}_C| = 2^{|C|}$. Повторюючи доведення NFL-теореми отримуємо:

<u>Cor.</u> Нехай $\mathcal{H}$ - клас гіпотез бін. клас. на мн. $\mathcal{X}$, що розбиває деяку $C \subset \mathcal{X}$ з $2m$ ел-тів. Тоді $\forall$ алгоритма A, що повертає гіпотези з $\mathcal{H}$ (зокрема, $ERM_{\mathcal{H}}$-алг.), $\exists$ розподіл $\mathcal{D}$ на $\mathcal{X} \times \{0,1\}$ такий, що

- $L_{\mathcal{D}}(h) = 0$ для деякої $h \in \mathcal{H}$;
- $\mathbb{P}_{S \sim \mathcal{D}^m}\left(L_{\mathcal{D}}(A(S)) \geq \frac{1}{8}\right) \geq \frac{1}{7}$.

Тобто у цьому випадку $\mathcal{H}$ неможливо "гарно" навчати на m-ел-тних тренувальних вибірках.

<u>def.</u> <u>Вимірність Вапніка-Червоненкіса</u> (VC-dimension) класу $\mathcal{H}$ гіпотез бінарної класифікації на мн. $\mathcal{X}$ зветься найб. $\overbrace{\text{потужність}}^{|C| \text{ скінч.}}$ множини $C \subset \mathcal{X}$, яку розбиває $\mathcal{H}$. Позн. $VC\dim(\mathcal{H})$.

<u>Rem.</u> Тобто $VC\dim(\mathcal{H}) = d$ означає, що $\exists\, C \subset \mathcal{X} : |C| = d$ і $\mathcal{H}$ розб. C; й $\forall\, C \subset \mathcal{X}$ з $|C| > d$ $\mathcal{H}$ не розб. C. (дост. $|C| = d+1$) Зокрема, $VC\dim(\mathcal{H}) = \infty$ означає існування $\underbrace{C}_{\text{скінч.}}$ довільної потужності, що розбивається $\mathcal{H}$.

<u>Pr.</u> Якщо $VC\dim(\mathcal{H}) = \infty$, то $\mathcal{H}$ не є PAC навчаним.

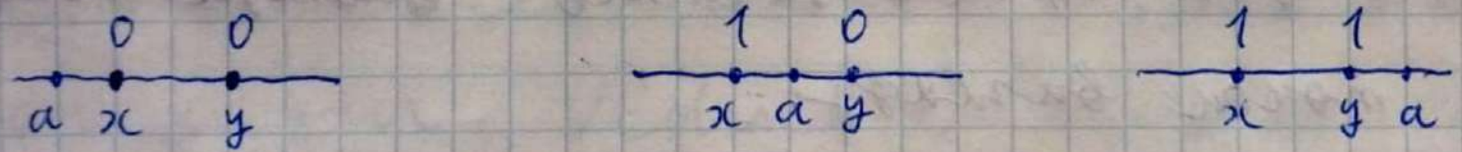
▷ Випливає з попер. <u>Cor.</u> ан-но до доведення ненавчаності класу усіх гіпотез (<u>Cor.</u> з NFL-теореми) вище. △

<u>Ex. 1.</u> Якщо $\mathcal{H}$ скінч., то $\forall\, C$ $|\mathcal{H}_C| \leq |\mathcal{H}|$, тому C не може бути розбита при $|\mathcal{H}| < 2^{|C|}$. Т.ч., $VC\dim(\mathcal{H}) \leq$ $\leq \log_2 |\mathcal{H}|$, але може бути суттєво меншою. Наприклад,

у <u>Ex.3.</u> ми показано, що VCdim $= 1$ для класу порогових ф-цій. Якщо обирати пороги зі скінч. множини, тобто $\mathcal{H} = \{ 1_{\{x < a_i\}} \}_{i=1}^{n}$, то $|\mathcal{H}| = n$, але так само $\mathrm{VCdim}(\mathcal{H}) = 1$.

<u>2.</u> Клас усіх $\mathcal{X} \to \{0, 1\}$ розбиває $\forall\, C \subset \mathcal{X}$, тому для нього VCdim $= |\mathcal{X}|$, зокрема ∞ для неск. $\mathcal{X}$ (і тоді він ненавчаний, як бачили).

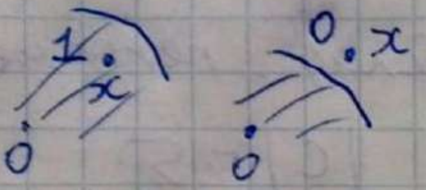
<u>3.</u> Клас $\mathcal{H} = \{ h_a := 1_{\{x<a\}} \}_{a \in \mathbb{R}}$ порогових ф-цій має VC-вимірність 1. Дійсно, $\forall\, C = \{x\}$ $h_{x-1}(x) = 0$ і $h_{x+1}(x) = 1$, але для $C = \{x, y\}$, де $x < y$, $\nexists\, a$ такого, що $h_a(x) = 0$, $h_a(y) = 1$:



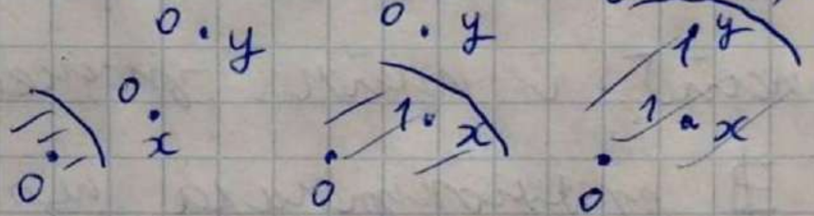
Це вірно і якщо обирати a зі скінч. множини, як у <u>Ex. 1.</u> (тоді вже й не всі 1-ел-ні множини розбиваються, але деякі – так).

(або нуль)

4. Клас х.ф. кругів $\mathcal{K} = \{1_{\{|x|\le r\}}\}_{r \ge 0}$ теж має VC-вим. 1 з тих же міркувань: для 1-точкової $C = \{x\}$ $(\frac{+}{0})$

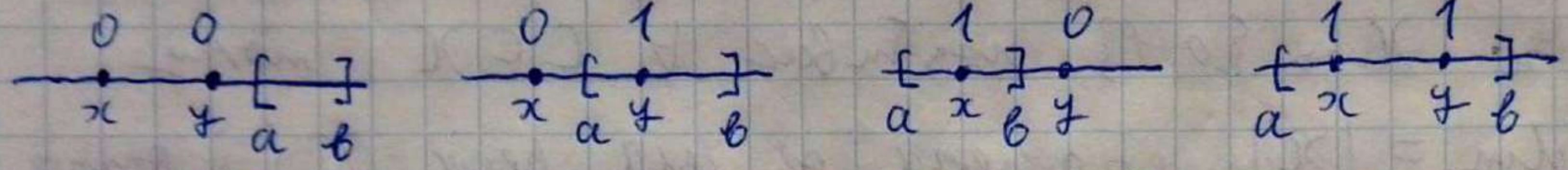


маємо 0 при $|x| > r$ і 1 при $r \ge |x|$, а для двоточкової $\{x, y\}$, $|x| \le |y|$ неможливо отримати $(0,1)$:



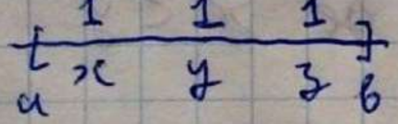
5. Клас $\mathcal{K} = \{ 1_{[a_1,b_1] \times \ldots \times [a_d,b_d]} \}$ х.ф. d-вим. паралелі-

недів має VC-вим. $2d$. Дійсно, при $d=1$, тобто для відрізків, завжди можна розбити мн. з 2 ел-тів $C = \{x, y\}$, $x < y$:



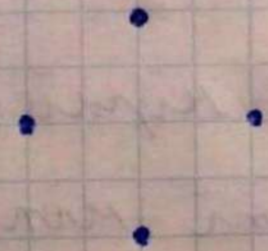
,

але ніколи - з 3 $C = \{x, y, z\}$, $x < y < z$: неможливо отримати значення $(1,0,1)$, бо $h(x)=1 \Rightarrow a \le x$, $h(z)=1 \Rightarrow z \le b$, але тоді $y \in [a,b] \Rightarrow h(y) = 1$.

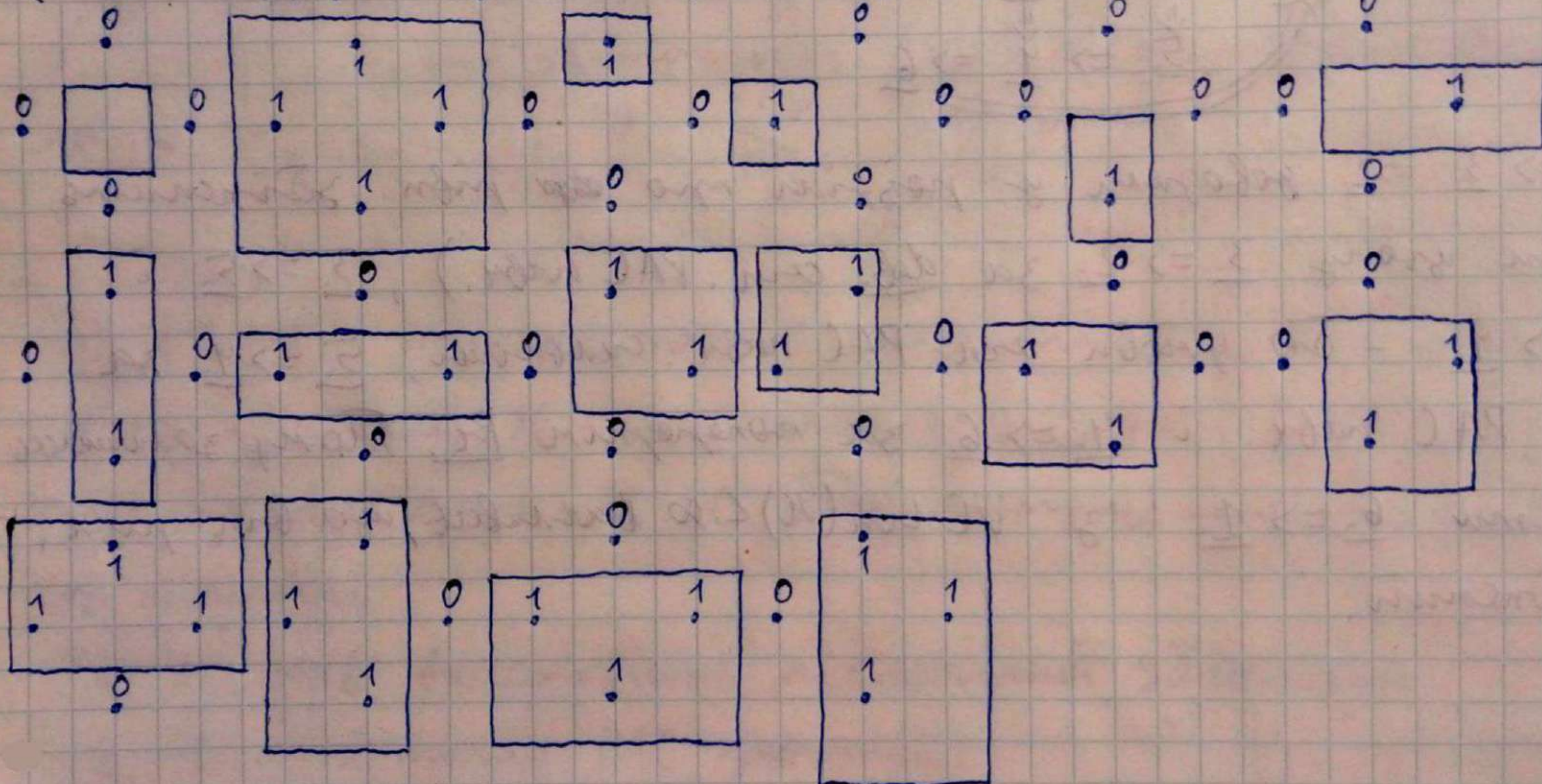


Для $d=2$, тобто для прямокутників:

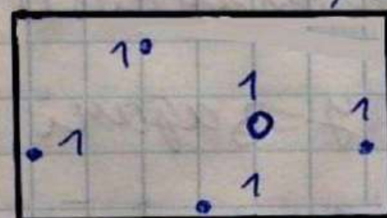
множину з 4 точок вигляду



(наприклад, з коорд. $(\pm 1, 0)$, $(0, \pm 1)$) можна розбити:



(це так не для $\forall$ мн. з 4 точок: наприклад, неможливо отримати $\overset{0}{\bullet}\,\overset{1}{\bullet}\,\overset{0}{\bullet}\,\overset{1}{\bullet}$ – доведіть це!). Якщо ж $|C| = 5$, виділимо точки з найб. і найм. значеннями кожної з координат. Тоді $\nexists$ прямокутника, що у цих ≤ 4 точках дає 1 (містить їх), а у п'ятій – 0 (не містить):



Впр. Довести це формально і узагальнити на довільне d.

Разом з цим, $VCdim(\mathcal{H})$ не завжди є "кількістю параметрів" $\mathcal{H}$ як у Ex. вище:

Впр. Показати, що клас відобр. $\mathbb{R} \to \{0,1\}$

$$\mathcal{H} = \{x \mapsto \lceil \sin \theta x \rceil\}_{\theta \in \mathbb{R}},$$

де вважаємо $\lceil -1 \rceil = 0$, має VC-вимірність ∞, хоча й залежить від 1 параметра.