

Ром Аннино (сестра и сын Гид, класирана)

можна використовувати не лише з лінійними моделями. Так, у задачі поліноміальної регресії степеня n на $\mathbb{R}$ треба знайти $p: \mathbb{R} \rightarrow \mathbb{R}$ -поліном степеня n . Для цього можна його представити у вигляді $h_{w,0} \circ f$, де $f: \mathbb{R} \rightarrow \mathbb{R}^{n+1}: x \mapsto (1, x, \dots, x^n)$: $p(x) = w_0 + w_1 x + \dots + w_n x^n = \langle w, f(x) \rangle$, і далі шукати $w = (w_0, \dots, w_n)$ методом найменших квадратів.

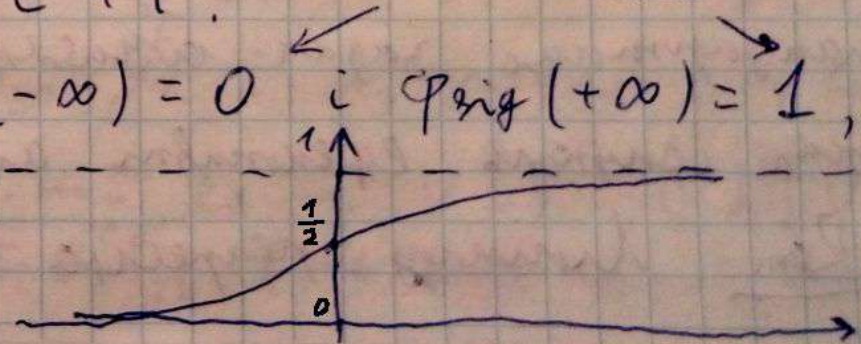
Логістична регресія.

Путем $\mathcal{Y} = [0, 1]$, і клас іпотез має вигляд $\Phi \circ L_d = \{ \Phi \circ h_{w,b} \}$ для деякої фіксованої (строго зростаючої) $\Phi: \mathbb{R} \rightarrow [0, 1]$. В задачі лог. регр. використовують сигмоїд, або логістичну Φ -цію (і звичайно беруть $b=0$):

$$\Phi_{\text{sig}}(z) := \frac{1}{1+e^{-z}} = \frac{e^z}{e^z + 1} > 0$$

Зокрема, $\Phi'_{\text{sig}} = \frac{e^{-z}}{(1+e^{-z})^2} > 0$, $\Phi_{\text{sig}}(-\infty) = 0$ і $\Phi_{\text{sig}}(+\infty) = 1$, тобто її графік:

$$(\Phi_{\text{sig}}(0) = \frac{1}{2}).$$



Лог. регресію інтерпретують як "нейонну" бінарну класифікацію: замість мітки 0 чи 1 вона повертає ймовірність отримати 1. Як бачимо, для великих $|h_{w,0}(x)|$ її передбачення близькі до рез-ту роботи наівотраперу (із заміною -1 на 0). Тому визначимо ф-цію втрам як для бінар. клас. - для $Y = \{\pm 1\}$, $\ell: (\text{Fsig} \circ L_d) \times \mathbb{R}^d \times \{\pm 1\} \rightarrow \mathbb{R}_+$.
 Для гіпотези $x \mapsto \text{Fsig}(h_{w,0}(x)) = \frac{1}{1+e^{-\langle w, x \rangle}}$ ма $y = 1$ ℓ повина бути великою для малих значень ф-ції (гіпотези) $\Leftrightarrow$ для великих $1 - \frac{1}{1+e^{-\langle w, x \rangle}} = \frac{e^{-\langle w, x \rangle}}{1+e^{-\langle w, x \rangle}} = \frac{1}{1+e^{\langle w, x \rangle}} \Leftrightarrow$ для великих $1+e^{-\langle w, x \rangle} = 1+e^{-y\langle w, x \rangle}$.
 Ано, при $y = -1$ ℓ велика для великих $\frac{1}{1+e^{-\langle w, x \rangle}} \Leftrightarrow$ мале $\frac{1}{1+e^{\langle w, x \rangle}} \Leftrightarrow$ великих $1+e^{\langle w, x \rangle} = 1+e^{-y\langle w, x \rangle}$. Це пояснює означення логістичної ф-ції втрам:

$$\ell_{\log}(\text{Fsig} \circ h_{w,0}, (x, y)) := \ln(1 + e^{-y\langle w, x \rangle}), \quad y \in \{\pm 1\}.$$

Зокрема, вона опукла, тому ERM може бути згайнена ставк. для таких ф-цій методами (губ. галі).

Boosting

Ідея Boosting (boosting) – "підсилювати" (to boost) "слабкий", але обчислювально ефективний алгоритм ML, щоб отримати "сильніший", але менш ефективний. Як побачимо, статистичні властивості цих алгоритмів корисні, а алгоритм Boosting дійсно гарантує збільшення обчисл. ефективності.

Наст. означення являє собою ще одну модифікацію PAC навчання:
Def. Нехай $\gamma \in (0, \frac{1}{2})$. Клас гіпотез $\mathcal{H}$ у задачі бінарної класифікації ($\forall x \in \mathcal{X} : y = \{\pm 1\}$) зветься γ -слабко навчанням (γ -weak learnable), якщо для нього існує γ -слабкий алгоритм A (γ -weak learner): $\exists$ ф-ція $m_A : (0, 1) \rightarrow \mathbb{N}$ така, що $\forall \delta \in (0, 1)$ $\forall$ розподілу $\mathcal{D}$ на $\mathcal{X}$ та ф-ції розстановки міток $f : \mathcal{X} \rightarrow \{\pm 1\}$ таке, що $\mathcal{H}$ задов. припущенню реалізованості, зі ймовірністю $\geq 1 - \delta$

за вибіркою S з m ек-тів, що i.i.d. від $\mathcal{D}$ і помічені за допомогою f , A повертає $h_S = A(S)$ таку, що $L_{\mathcal{D}, f}(h_S) \leq \frac{1}{2} - \gamma$.

Рем. Подібно вимагається, щоб за параметром точності $\epsilon = \frac{1}{2} - \gamma$ (фіксованим) A повертав гіпотезу "правильну" (на γ) країні за випадковий вибір. При цьому з NFL-теорему або з оцінки $m_K(\frac{1}{2} - \gamma, \delta) \geq C \frac{VCdim(K) + \ln \frac{1}{\delta}}{\frac{1}{2} - \gamma}$ у більшій версії основної теореми випливає:

Сол. Якщо K не є PAC навчанням, тоді $VCdim(K) = \infty$, то він не є і γ -слабо навчанням $\forall \gamma$.

Рем. Подібно слабка навчальність потрібна не для послідовної статистичної обробки на K , а для обчислювальної ефективності. При цьому алгоритм A може шукати наступним чином. Нехай B - якийсь інший клас гіпотез $\mathcal{H}$ цієї задачі (базовий (base) клас

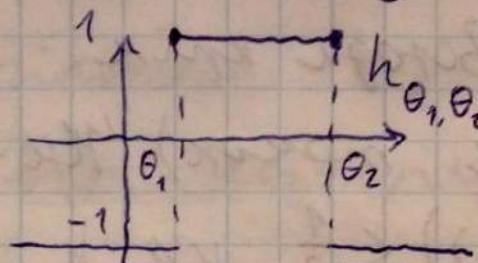
интерес) , где одно и то же представление ERM₁₃ -
аннотации. Другое bin ∈ Y - сканам анн. где Y

(зауважимо, що у деф. ми не вимагаємо $h_S \in \mathcal{H}$),
тобто повертає $h_S = A(S)$ за вибірками, що розв'язали
ф-ції з $\mathcal{H}$ (у деф. ми можемо вважати $f \in \mathcal{H}$ з
точністю до мн. точок міри 0 в силу прип. реалізов-
ності), з помилкою $\leq \frac{1}{2} - \gamma$, в силу деф., $\mathcal{H} \in$
 γ -слабко навчанням.

Ек. Нехай $\mathcal{H}$ - клас відрізків зі знаком (вони
ще зветься 3-елементними класифікаторами (3-piece

classifiers):

$$\mathcal{H} := \{ h_{\theta_1, \theta_2, b} : x \mapsto \begin{cases} b, & x \in [\theta_1, \theta_2] \\ -b, & x \notin [\theta_1, \theta_2] \end{cases} \mid \theta_1 < \theta_2 \in \mathbb{R}, b = \pm 1 \}$$

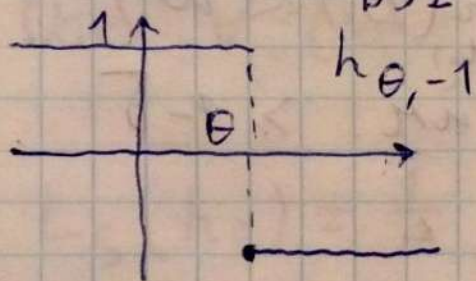


На практиці ми говоримо, що $VCdim(\mathcal{H}) = 3$ (це пояснює назву).

Класи $\mathcal{B}$ - клас "переріз", або переріз ріменя (decision stump):

$$(h_{+\infty, b} \equiv -b, h_{-\infty, b} \equiv b)$$

$$\mathcal{B} = \mathcal{H}_{DS1} = \{ h_{\theta, b} : x \mapsto \text{sign}(x - \theta) \cdot b \mid \theta \in \mathbb{R} \cup \{\pm\infty\}, b = \pm 1 \}$$



$VCdim(\mathcal{H}_{DS1}) = 2$ (бо по суті це клас

парів/пропорцій при $d=1$; це неважко

перевірити безпосередньо). Т.ч., обидві ці класи PAC навч. за допомогою ERM-алг.

При цьому $\mathcal{B}$ задов. умові попер. Рем. для $\gamma = \frac{1}{12}$.

Дійсно, нехай $\mathcal{D}$ і f задов. прип. реал. для $\mathcal{H}$, тоді

$f = h_{\theta_1, \theta_2, b}$ зі ім. 1. Оберемо той з 3 можливих

$(-\infty, \theta_1), [\theta_1, \theta_2], (\theta_2, +\infty)$, міра $\mathcal{D}$ якої $\leq \frac{1}{3}$
 (такий існує, бо сума іхніх мір дор. 1), Тоді
 $\exists h_{\theta, b} \in \mathcal{B}$, що збігається з f в точці крім
 цього проміжку (і можливо скінч. кільк. точок). На-
 приклад, для $f = h_{\theta_1, \theta_2, 1} \in \mathcal{D}((-\infty, \theta_1)) \leq \frac{1}{3}$ це $h_{\theta_2, -1}$
 (див. рис. вище; Впр. Перевірити для інших випадків).

Тоді $L_{\mathcal{D}}(h_{\theta, b}) \leq \frac{1}{3}$. Оскільки за основною теоремою цієї
 клас априорно PAC навчання ($VCdim(\mathcal{B}) < \infty$),
^{що зам. від} $\exists m_0(\delta) : \forall m \geq m_0$, $\mathcal{D}$ і f^* , $S \sim \mathcal{D}^m$ зі ім. $> 1-\delta$

$$L_{\mathcal{D}}(h_S) \leq \frac{1}{3} + \frac{1}{12} = \frac{5}{12} = \frac{1}{2} - \frac{1}{12}$$

$\forall \text{ERM}_{\mathcal{B}}$ алг. A і $h_S = A(S)$, що й потрібно.

Тепер покажемо, що існує обчислювально ефектив-
 ний $\text{ERM}_{\mathcal{B}}$ -алгоритм. Зробимо це для більш загаль-
 ного випадку нехвів рівень на $\mathbb{R}^d$: $\bigcup \{\pm \infty\}$

$$\mathcal{H}_{DS} := \left\{ \underbrace{h_{\theta, \theta, b}}_{\mathbb{R}^d} : x \mapsto b \cdot \text{sign}(\theta - x_i) \right\} \theta \in \mathbb{R}, i=1, d, b = \pm 1$$

Розглянемо нині іншим загальну помилку (це буде важливо, зокрема, для гл. AdaBoost). Нехай Φ - (дискретний) тип. розподіл на S , тобто $\Phi = (\Phi_1, \dots, \Phi_m)$, де всі $\Phi_i \in [0, 1]$ і $\sum_{i=1}^m \Phi_i = 1$. Визначимо помилку $\text{sign. } \Phi$

як мат. очікування φ - y_i $\frac{1}{(bun. \text{ ver.})}$ $h(x) \neq y$: $x \mapsto \begin{cases} 1, & h(x) \neq y \\ 0, & h(x) = y \end{cases}$ $\left(= 1_{\{(x,y) | h(x) \neq y\}} \right)$

$$L_{\Phi}(h) := \sum_{i=1}^m \Phi_i \cdot 1_{h(x_i) \neq y_i} = \sum_{\substack{i=1, m: \\ h(x_i) \neq y_i}} \Phi_i$$

Зокрема, $L_{\Phi}(h) = L_S(h)$ для $\Phi = (\frac{1}{m}, \dots, \frac{1}{m})$ - "збалансана" одн. помилка. Будемо мінімізувати L_{Φ} на $S = \{(x_i, y_i)\}_{i=1}^m$ серед функцій з $b=1$ (та довільним θ і певним коорд. j):

$$h_S \in \argmin_{\substack{j=1, d \\ \theta \in \mathbb{R} \cup \{\pm \infty\}}} \left(\sum_{\substack{i=1, m: \\ y_i=1}} \Phi_i \cdot 1_{x_{i,j} > \theta} + \sum_{\substack{i=1, m: \\ y_i=-1}} \Phi_i \cdot 1_{x_{i,j} \leq \theta} \right) = L_{\Phi}(h_{j,\theta,1})$$

(немає повертає -1) $\uparrow$ $(\text{повертає } 1)$

де $\forall i$ $x_i = (x_{i,1}, \dots, x_{i,d})$. Нехай $j=1, d$. Відсортуємо $\{x_i\}$ за неспаданням j -мої коорд.: $x_{1,j} \leq x_{2,j} \leq \dots \leq x_{m,j}$. Тоді

$$\Theta_j := \left\{ x_{1,j} - 1, \frac{x_{1,j} + x_{2,j}}{2}, \frac{x_{2,j} + x_{3,j}}{2}, \dots, \frac{x_{m-1,j} + x_{m,j}}{2}, x_{m,j} + 1 \right\}$$

Потім, що пошук за усіма $\theta \in \mathbb{R}$ можна для
 даною $\bar{j}$ замінити на пошук $\theta \in \Theta_{\bar{j}}$: завжди $\exists \theta' \in \Theta_{\bar{j}}$
 з тими ж значеннями $h_{\bar{j}, \theta', 1}$ на S , що й $h_{\bar{j}, \theta, 1}$. Проти
 підрахунок сум для усіх $\theta \in \Theta_{\bar{j}}$ дасть час $O(dm^2)$ (точніше,
 $O(d([\text{час сортування}] + m^2))$, але це менс same). Потім, що,
 якщо позначити через $F(\theta)$ помилку для $h_{\bar{j}, \theta, 1}$ (що ми міні-
 мізуємо), то $F(\theta) = L_{\mathcal{D}}(h_{\bar{j}, \theta, 1})$ і ми $\min_{\theta \in \Theta_{\bar{j}}} F(\theta)$ з $\bar{j}$ -max
 точок S лежить лише $x_{i, \bar{j}}$ (тобто $\theta \in (x_{i, \bar{j}}, x_{i+1, \bar{j}})$, $\theta' \in$
 $\in (x_{i, \bar{j}}, x_{i+1, \bar{j}})$, то

$$F(\theta') = \begin{cases} F(\theta) - \mathcal{D}_i, & y_i = 1 \\ F(\theta) + \mathcal{D}_i, & y_i = -1 \end{cases} = F(\theta) - y_i \mathcal{D}_i$$

тому не потрібно кожен раз перевираховувати $F(\theta)$: на
 кожному з m кроків достатньо просто додати $-y_i \mathcal{D}_i$.
 Це означає, що час виконання - $O(d([\text{час сортування}] + m))$,
 наприклад, $O(dm)$ для сортування за розрядами (для фікс.
 типу даних). Отже, $ERM_{\mathcal{D}_S}$ дійсно обчислювально ефективний.

Впр. Записати цей алгоритм у вигляді псевдокода (або не псевдо-). Дав. відповідь у [1].

AdaBoost (Adaptive Boosting)

Це алгоритм бінарної класифікації, що базується на основі "слабкого" класифікатора WL (weak learner), який на кожній ітерації t здійснює ERM відносно розподілу $\mathcal{D}^{(t)}$ на навчальній вибірці $S = \{(x_i, y_i)\}_{i=1}^m$, як у алгоритмі для певних рішень вище. Після цього ми оповіємо розподіл (починаючи з однорідного $\mathcal{D}^{(1)} = (\frac{1}{m}, \dots, \frac{1}{m})$), як показано у псевдокоді:

$\mathcal{D}^{(1)} := (\frac{1}{m}, \dots, \frac{1}{m})$
for $t = 1, T$ ← кількість кроків
 $h_t: X \rightarrow \{+1\} \rightarrow h_t := WL(S, \mathcal{D}^{(t)})$
 $\epsilon_t := L_{\mathcal{D}^{(t)}}(h_t) = \sum_{i=1}^m \mathcal{D}_i^{(t)} \mathbb{1}_{h_t(x_i) \neq y_i}$ ← помилка h_t
then $\rightarrow w_t := \frac{1}{2} \ln \left(\frac{1}{\epsilon_t} - 1 \right)$
 $\mathcal{D}_i^{(t+1)} := \frac{\mathcal{D}_i^{(t)} e^{-w_t y_i h_t(x_i)}}{\sum_{j=1}^m \mathcal{D}_j^{(t)} e^{-w_t y_j h_t(x_j)}}$ ← новий розподіл
for $i = 1, m$
 $h_S := \text{sign} \left(\sum_{t=1}^T w_t h_t \right)$ ← результат.

Перш за все, перевіримо, що цей алг. ефективно здійснює ERM:

Тн. Пусть для неравных S и $\forall t$ на t -м
 этапе алгоритма AdaBoost слабый алгоритм возвращает
 гипотезу h_t с ошибкой $\epsilon_t \leq \frac{1}{2} - \gamma$ (где $\gamma \in (0, \frac{1}{2})$;
 sign. разности $\mathcal{D}^{(t)}$ на S). Тогда алгоритм возвращает
 h_S с ошибкой $L_S(h_S) \leq e^{-2\gamma^2 T}$ после T итераций.

$\Rightarrow \forall t$ помним $f_t := \sum_{p=1}^t w_p h_p$; м.ч., $h_S = \text{sign } f_T$.
 Также помним $Z_t := \frac{1}{m} \sum_{i=1}^m e^{-y_i f_t(x_i)}$ (где $S = \{x_i, y_i\}_{i=1}^m$).

Оск. $\forall (x, y)$ $\frac{1}{h(x) \neq y} = \begin{cases} 0, & h(x) = y \\ 1, & h(x) \neq y \end{cases} \leq e^{-y h(x)}$

$$L_S(h_S) = \frac{1}{m} \sum_{i=1}^m \frac{1}{h_S(x_i) \neq y_i} \leq \frac{1}{m} \sum_{i=1}^m e^{-y_i h_S(x_i)} \leq Z_T, \text{ так как}$$

последнее известно

$$e^{-2\gamma^2 T} \geq Z_T = \frac{Z_T}{Z_{T-1}} \cdot \frac{Z_{T-1}}{Z_{T-2}} \cdots \frac{Z_2}{Z_1} \cdot \frac{Z_1}{Z_0},$$

где вначале $Z_0 = 1$ (и $f_0 = 0$). Т.ч. известно,

$$\text{что } \frac{Z_{t+1}}{Z_t} \leq e^{-2\gamma^2} \quad \forall t = 0, T-1.$$

Зем. $\forall t, i = 1, m$.

$$Q_i^{(t+1)} = \frac{e^{-y_i f_t(x_i)}}{\sum_{j=1}^n e^{-y_j f_t(x_j)}}$$

⇒ Induktion für t !

$t=1$. $\forall i = \overline{1, m}$ права часть:

$$\frac{e^{-y_i f_1(x_i)}}{\sum_j e^{-y_j f_1(x_j)}} = \frac{e^{-y_i w_1 h_1(x_i)}}{\sum_j e^{-y_j w_1 h_1(x_j)}} = \left[\text{перен. на } \frac{1}{m} \right] = \frac{\frac{1}{m} e^{-y_i w_1 h_1(x_i)}}{\sum_j \frac{1}{m} e^{-y_j w_1 h_1(x_j)}} =$$

$$= \frac{\mathcal{Q}_i^{(1)} e^{-w_1 y_i h_1(x_i)}}{\sum_j \mathcal{Q}_j^{(1)} e^{-w_1 y_j h_1(x_j)}} = \mathcal{Q}_i^{(2)}$$

$t-1 \rightarrow t$. $\forall i = \overline{1, m}$, $t \geq 2$:

$$\mathcal{Q}_i^{(t+1)} = \frac{\mathcal{Q}_i^{(t)} e^{-w_t y_i h_t(x_i)}}{\sum_{j=1}^m \mathcal{Q}_j^{(t)} e^{-w_t y_j h_t(x_j)}} = \left[\begin{array}{l} \text{перен. инд. } i \\ \text{знаменателя} \\ \text{суммирования} \end{array} \right] = \frac{e^{-(y_i f_{t-1}(x_i) + y_i w_t h_t(x_i))}}{\sum_j e^{-(y_j f_{t-1}(x_j) + y_j w_t h_t(x_j))}} =$$

$$= \frac{e^{-y_i f_t(x_i)}}{\sum_j e^{-y_j f_t(x_j)}}, \text{ что и требовалось. } \triangle (\text{лем.})$$

Once, $\forall t = 0, T-1$ (from $t=0$ we know, so $F_0 = 0$: $Z_0 = 1$):

$$\frac{Z_{t+1}}{Z_t} = \frac{\sum_{i=1}^m e^{-y_i} f_{t+1}(x_i)}{\sum_{j=1}^m e^{-y_j} f_t(x_j)} = \frac{\sum_{i=1}^m e^{-y_i} f_t(x_i) e^{-y_i} w_{t+1} f_{t+1}(x_i)}{\sum_{j=1}^m e^{-y_j} f_t(x_j)} = [\text{lem}] =$$

$$= \sum_{i=1}^m \mathbb{Q}_i^{(t+1)} e^{-y_i} w_{t+1} h_{t+1}(x_i) = \sum_{\substack{i=1, m: \\ h_{t+1}(x_i) = y_i}} \mathbb{Q}_i^{(t+1)} e^{-w_{t+1}} + \sum_{\substack{i=1, m: \\ h_{t+1}(x_i) \neq y_i}} \mathbb{Q}_i^{(t+1)} e^{w_{t+1}} =$$

$$= \left[\frac{\text{deb. } w_{t+1} i}{\epsilon_{t+1}} \right] = (1 - \epsilon_{t+1}) \sqrt{\frac{\epsilon_{t+1}}{1 - \epsilon_{t+1}}} + \epsilon_{t+1} \sqrt{\frac{1 - \epsilon_{t+1}}{\epsilon_{t+1}}} = 2 \sqrt{\epsilon_{t+1} (1 - \epsilon_{t+1})} \leq$$

$$\leq \left[\begin{array}{l} \epsilon_{t+1} \leq \frac{1}{2} - \delta \text{ за условие;} \\ f(x) = x(1-x) \text{ не убывает на } [0, \frac{1}{2}]; \\ f'(x) = 1 - 2x \end{array} \right] \leq 2 \sqrt{(\frac{1}{2} - \delta)(\frac{1}{2} + \delta)} = \sqrt{1 - 4\delta^2} < [4\delta^2 < 1] <$$

$$< \sqrt{e^{-4\delta^2}} = e^{-2\delta^2}, \text{ что и требовалось. } \triangle$$

Rem. 3 описи в базисе (из значения $t+1$ на t) можно

$$\sum_{j=1}^m \mathbb{Q}_j^{(t)} e^{-y_j w_t h_t(x_j)} = 2 \sqrt{\epsilon_t (1 - \epsilon_t)}. \text{ Тогда аналогично, что}$$

$$\text{формула } L_{\mathbb{Q}^{(t+1)}}(h_t) = \sum_{i=1}^m \mathbb{Q}_i^{(t+1)} \mathbb{1}_{h_t(x_i) \neq y_i} = [\text{deb. } \mathbb{Q}_i^{(t+1)}] =$$

$$= \frac{1}{2 \sqrt{\epsilon_t (1 - \epsilon_t)}} \sum_{\substack{i=1, m: \\ h_t(x_i) \neq y_i}} \mathbb{Q}_i^{(t)} e^{-w_t y_i h_t(x_i)} = \left[\frac{\text{deb.}}{\epsilon_t} \right] \frac{\epsilon_t e^{w_t}}{2 \sqrt{\epsilon_t (1 - \epsilon_t)}} = \frac{\sqrt{\epsilon_t (1 - \epsilon_t)}}{2 \sqrt{\epsilon_t (1 - \epsilon_t)}} = \frac{1}{2}.$$

Тогда мы на конкретный шаг определим $\mathbb{Q}^{(t)}$ так, что бы значение не убывало.

на початку кожної наступної ітерації точки x_i , на якое попередня спробувала з помилкою, мають більші ймовірності. Експоненціальність за T означає, що ERM може здійснювати успішно навіть для малих γ (тобто слабких WL). В кінці кожної ітерації (з малими ϵ_t) мають більші ваги w_t .
Реш. На кожній кроці AdaBoost здійснює ваги WL та $O(m)$ операцій, тому загальна маємо $O(T([\text{складність}]_{WL} + m))$. Наприклад, для неохів рівень це $O(Tmd)$ за поперед. прикладом. Тобто, це "підсилення" дійсно залишається обчислювано ефективно.

Реш. Задаємо, що WL має помилку $\epsilon_t \leq \frac{1-\gamma}{2}$ зі ім. $1-\delta$. Тоді за субадитивністю маємо ім. більшої помилки $\leq T\delta$, а отже успіху $> 1-T\delta$. Оскільки за дов. γ -слабкої навч. δ можна звести малим за допомогою великого m (логістично за $\frac{1}{\delta}$ для фікс. ϵ_t для $V(\dim < \infty)$), параметр впевності AdaBoost $\frac{(1-T\delta)}{m}$ можемо залишати високим.

Більше того, можна запропонувати процедуру "Бустінга" ^{задати M на $\mathbb{Z}$} з
 впевності. А саме, нехай A - алгоритм ()
 наступною властивістю: $\exists \delta_0 \in (0, \frac{1}{2})$ і $m_\epsilon : (0, 1) \rightarrow \mathbb{N}$ такі,
 що $\forall \epsilon \in (0, 1)$ і $\forall m \geq m_\epsilon(\epsilon)$ $\forall \mathcal{D}$ на $\mathbb{Z}$ зі ім. $\geq 1 - \delta_0$ для
 $S \sim \mathcal{D}^m$ $L_{\mathcal{D}}(A(S)) \leq \min_{h \in \mathcal{H}} L_{\mathcal{D}}(h) + \epsilon$ (для деякого класу
 ім. $\mathcal{H}$) Також фіксуємо $\delta = \delta_0$ у дов. алг. PAC навч. Нехай
 тепер $\delta \in (0, 1)$ довільне, і $k := \lceil \frac{\ln \delta}{\ln \delta_0} \rceil$. Розділимо S на
 $k+1$ частин, де перші k -по $m_\epsilon(\frac{\epsilon}{2})$. Застосуємо A до
 кожної з цих частин. Оскільки S і.і.д., $L_{\mathcal{D}}$ на якій рез-мас
 роботи A буде $> \min_{h \in \mathcal{H}} L_{\mathcal{D}}(h) + \frac{\epsilon}{2}$ зі ім. $< \delta_0^k \leq \delta_0^{\frac{\ln \delta}{\ln \delta_0} + 1} =$
 $= \delta \delta_0 < \frac{\delta}{2}$ (консно і для інших δ_0 за рахунок k). Переміш,
 розглянемо ці k результатів як новий клас $\mathcal{H}'$ і
 зрештою ERM візн. цього. Тоді за однією для рівн.
 збінності (тут ще припускаємо $\ell \in [0, 1]$), потрібно
 ще $\lceil \frac{2}{(\frac{\epsilon}{2})^2} \ln \frac{2|\mathcal{H}'|}{\delta/2} \rceil = \lceil \frac{8}{\epsilon^2} \ln \frac{4k}{\delta} \rceil$ точок S , щоб отримати

помилку $\frac{\epsilon}{2}$ зі ім. $\geq 1 - \frac{\delta}{2}$, а отже загалом
 зі ім. $\geq 1 - \delta$ за субадитивністю. Т.ч., отримали PAC
 навчання для $\mathcal{H}$ зі складністю (для намов авт. A'):

$$P_{S \sim \mathcal{D}^m} (L_{\mathcal{D}}(A'(S)) \leq \min_{h \in \mathcal{H}} L_{\mathcal{D}}(h) + \epsilon) \leq \min_{h \in \mathcal{H}} L_{\mathcal{D}}(h) + \frac{\epsilon}{2} \leq \min_{h \in \mathcal{H}} L_{\mathcal{D}}(h) + \frac{\epsilon}{2} + \frac{\epsilon}{2} = \min_{h \in \mathcal{H}} L_{\mathcal{D}}(h) + \epsilon \geq 1 - \delta.$$

Реш. Чи гарантує AdaBoost PAC навчання, якщо $\mathcal{H}$

гарантує? Оскільки це задача fin. клас., можемо просто оцінити
 VC-випірність fin. класу $\mathcal{H}$ напрям. Кожна $\mathcal{H}$ належить
 до базового класу $\mathcal{B}$: $h_t \in \mathcal{B} \forall t$. Тоді за побудовою рез-ти
 роботи T кроків AdaBoost - це лінійна комб. h_t з коеф.
 w_t , тобто, її комб. з sign:

$$h_S \in L(\mathcal{B}, T) := \left\{ x \mapsto \text{sign} \sum_{t=1}^T w_t h_t(x) \right\} w_t \in \mathbb{R}^T, h_t \in \mathcal{B} \forall t.$$

(композиція $x \mapsto (h_1(x), \dots, h_T(x))$ з одновимірним напрям.).

Реш. Якщо у задачі fin. клас. $3 \leq \text{VCdim}(\mathcal{B}) < \infty$ і $T \geq 3$,

то

$$\text{VCdim}(L(\mathcal{B}, T)) \leq T(\text{VCdim}(\mathcal{B}) + 1)(3 \ln T(\text{VCdim}(\mathcal{B}) + 1) + 2).$$

► Нехай C розд. $L(B, T)$, $|C| = m$.

За лемою Загера, кінц. об'ємності B на C , можна розстановити різні індекси з B $|B_C| \leq \left(\frac{em}{d}\right)^d$,
 де $d := VCdim(B)$. Тоді для T індекс з B маємо
 $\leq \left(\frac{em}{d}\right)^{dT}$ варіантів. Тоді само, оскільки $VCdim$ класу
 ознак невідомо і R^T ген. T за непер. розподілом,
 обираючи n , загальн. маємо $\leq \left(\frac{em}{d}\right)^{dT} \cdot \left(\frac{em}{T}\right)^T < [T, d+3] <$
 $< m^{dT} m^T = m^{T(d+1)}$
клас. усіх класів розд.
 вона розбивається, $2^m \leq m^{(d+1)T}$, тоді
 $m \leq \ln m \cdot \frac{(d+1)T}{\ln 2}$.

За лем. 4, у довільній оцінці T_n , необхід. умовою є $\in$

$$m \leq 2^{\frac{(d+1)T}{\ln 2}} \ln \frac{(d+1)T}{\ln 2} < \left[\frac{\frac{2}{\ln 2} < 3}{\ln \ln 2} - \frac{1}{2} \right] < 3(d+1)T \left(\ln(d+1)T + \frac{1}{2} \right)$$

що і потрібно. $\triangle$

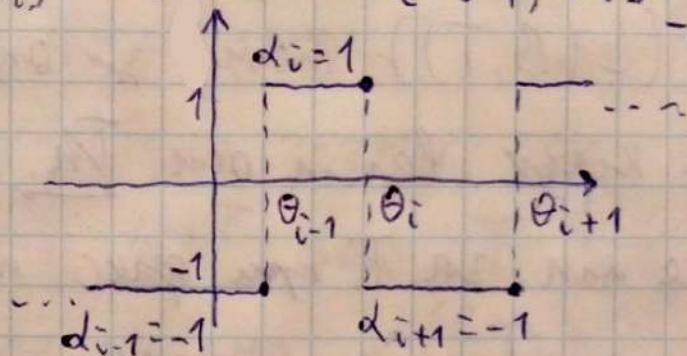
Сол. Якщо базовий клас B (ан.) PAC навчаний, то $\forall T$
 $L(B, T)$ (ан.) PAC навчаний.

Рем. Згідно з цим Рл., при зростанні T (найменше лінійно за T)
зростає $VCdim(L(B, T))$, тому з оцінки $m_{\eta}(\epsilon, \delta) \geq \frac{C}{\epsilon} (VCdim(L(B, T)) + \ln \frac{1}{\delta})$ у кільк. версії осн. Тл. випл., що помилка оцінки
 $E_{\text{est}} = \epsilon$ (найменше лін. за T при фікс. m і δ) зростає. З іншого
боку, за Тл. вище, L_S при цьому експоненційно спадає (клас
краще навчається). Дати ми покажемо, що помилка аппроксимації
 $E_{\text{app}} = \min_{h \in L(B, T)} L_D(h)$ теж спадає. Тл.ч., зі зрост. T у
АдаBoost збільшується складність і зменшується упередженість,
зі спад. T - навчати. Тому цей параметр можна використати
для "керування" дан компромісом (bias-complexity tradeoff).
Еск. Наведено конкретний приклад збільшення VC -вимірності
за рахунок переходу до лін. комбінацій (ї дано оцінку
знизу). Нехай знову $B = \mathcal{H}_{DS1}$ (їого VC -вим. початково,

2), а G_T - класс кусково постоянных классификаторов из T признаков:

$$G_T := \left\{ g_{\{\theta_i\}, \{\alpha_i\}} : x \mapsto \sum_{i=1}^T \alpha_i \cdot 1_{(\theta_{i-1}, \theta_i]} \right\} \quad \alpha_i = \pm 1, \quad -\infty = \theta_0 < \theta_1 < \dots < \theta_{T-1} < \theta_T = +\infty,$$

подобно ф-ции знака



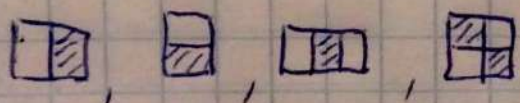
Положим $G_T \subset L(\mathcal{H}_{DS1}, T)$. Действительно, напр. для ф-ции $g \in G_T$ описанного выше вида берем $\alpha_i = (-1)^i$ (иные $\{\alpha_i\}$ будут нам $\leq$ зря знать и для нас все же проще), где T парное, полагаем

$$w_1 = \frac{1}{2} \quad \text{и} \quad w_t = (-1)^t \quad \text{при} \quad t \geq 2 \quad \text{и} \quad \text{получим}$$

$$x \mapsto \text{sign} \sum_{t=1}^T w_t \text{sign}(x - \theta_{t-1}) = \begin{cases} \frac{1}{2} \text{sign}(x - (-\infty)) + \text{sign}(x - \theta_1) - \text{sign}(x - \theta_2) + \dots = \\ \text{sign} \left(\frac{1}{2} - 1 + 1 - \dots + (-1)^{T-1} \right) = \text{sign} \left(-\frac{1}{2} \right) = -1, & x < \theta_1 \\ \text{sign} \left(\frac{1}{2} + 1 + 1 - \dots + (-1)^{T-1} \right) = \text{sign} \left(\frac{3}{2} \right) = 1, & \theta_1 < x < \theta_2 \\ \text{sign} \left(\frac{1}{2} + 1 - 1 - \dots + (-1)^{T-1} \right) = \text{sign} \left(-\frac{1}{2} \right) = -1, & \theta_2 < x < \theta_3 \\ \dots \end{cases}$$

П.ч., все g (Впр., что подобно для нечетного T). Отсюда,

$$G_T \subset L(\mathcal{H}_{DS1}, T) \Rightarrow \text{VCdim}(L(\mathcal{H}_{DS1}, T)) \geq \text{VCdim } G_T = T.$$

Ця функція оцінка має на ім'я z з параметром T .
Ес. Алгоритм Віона-Джонса (P. Viola, M. Jones, 2001) за-
стосовує AdaBoost у задачі розпізнавання облич на
 24×24 черно-білих зображеннях. Подібно $x = \mathbb{R}^{24^2}$, $y = 1$
означає, що зображення є обличчям, -1 - ні. Клас B
складається тут з σ -її вигляду $f \circ g$, де $f \in \mathcal{H}_{DS1}$,
а $g: \mathbb{R}^{24^2} \rightarrow \mathbb{R}$ рахує за прямокутником у зображенні
(сторони якого паралельні сторонам зобр.) і одного з 4
масок  різницю сум значень пікселів у
"чорній" та "білій" частині (тобто W здійснює ERM у цій класі).