Тоді коректно визначена $h: X \to \{0,1\} : x \mapsto \begin{cases} h_n(x), & x \in X_n, \\ & n = \overline{1,\infty}; \\ 0 & \text{у ін. вип.} \end{cases}$

При цьому $h \in \mathcal{H}$ і $\forall n \; h \notin \mathcal{H}_n$, бо $h|_{X_n} = h_n$ не є обм. ф-єю з $\mathcal{H}_n$ за побудовою ↯ △

мінімізація

Приклад SRM: довжина опису

Поки що у наших прикладах вагових ф-цій w значення $w(n)$ не були явно пов'язані з класами $\mathcal{H}_n$ (крім прикладу з поліномами, де, скажімо, для $w(n) = \frac{6}{(\pi n)^2}$ це значення спадає з $n = VCdim(\mathcal{H}_n)$ – степінь полінома + 1). Побудуємо приклад, де w визначена класами.

Нехай $\mathcal{H}$ – (не більш ніж) зліченний. Тоді його можна предст. як зліченне об'єднання одноточкових класів: $\mathcal{H} = \bigcup_{\wedge} \underbrace{\{h_n\}}_{=\mathcal{H}_n}$. За Рч. про рівн. збіжність скінч. класів, усі $\{h_n\}$ рівн. збіжні з $m^{UC}_{\{h_n\}}(\varepsilon, \delta) \le \frac{1}{2\varepsilon^2} \ln \frac{2}{\delta}$ ∀n За умови, що ф-ціонал втрат даної задачі $ML \in \ell : \mathcal{H} \times Z \to [0,1]$

(наприклад, для задач класифікації з $\ell = \ell_{0-1}$). Тоді в силу (*) $\varepsilon_n(m,\delta) \le \sqrt{\frac{1}{2m}\ln\frac{2}{\delta}}$, тому SRM можна реалізувати як

$$h_S \in \operatorname*{argmin}_{h\in\mathcal{H}}\left(L_S(h) + \sqrt{\frac{1}{2m}\underbrace{\ln\frac{2}{w(h)\cdot\delta}}_{= -\ln w(h) + \ln\frac{2}{\delta}}}\right),$$

де тепер вважаємо $w: \mathcal{H} \to [0,1] : h_n \to w(n)$ (бо тепер $\mathcal{H}$ у нас відповідності з $\mathbb{N}$). Визначимо тепер $w(h)$ як відповідності до <u>довжини опису</u> h:

<u>Опр.</u> Нехай Σ - <u>алфавіт</u>, тобто деяка скінченна множина символів. <u>Рядок</u> (string) - це будь-яка послідовність символів з Σ; позн. через $|\sigma|$ довжину рядка σ, а через Σ^* - множину усіх скінченних рядків з символів Σ.

<u>Мова опису</u> (description language) на множині $\mathcal{H}$ (скажімо, на класі гіпотез задачі ML) - це будь-яке (ін'єктивне) відображення $d: \mathcal{H} \to \Sigma^*$. $\forall\, h \in \mathcal{H}$ його образ $d(h)$ тоді зветься <u>описом</u> h, а $|h| := |d(h)|$ - <u>довжиною опису</u> в мові d. Мова опису d зветься <u>префіксною</u> (вільною від пре-

фіксів, prefix-free), якщо $\forall\ h \neq h' \in \mathcal{H}$ жоден з рядків $d(h)$, $d(h')$ не є початком (префіксом) іншого (зокр., не ін'єкт.).

Ex. Наприклад, для двійкового коду $\Sigma = \{0,1\}$, а $\Sigma^* = \{0,1\}^*$ – усі скінченні рядки вигляду $(0,1,1,1,0,..)$. Прикладом префіксного $d \in$ код Хаффмана (а для $\Sigma = \{0,..,9\}$ – телефонні коди країн). Для класу гіпотез $\mathcal{H}$ конкретною реалізацією d може бути, наприклад, запис $h \in \mathcal{H}$ як програми мовою C++ та застосування до її тексту gzip (що містить кодування за Хаффманом, а отже реалізує префіксну мову). Тоді $d: \mathcal{H} \to \{0,1\}^*$ префіксна.

Пр. (нерівність Крафта) $\forall$ зліченної підмножини рядків $S \subset \{0,1\}^*$, жоден з яких не є префіксом іншого, зокрема, для $S = d(\mathcal{H})$, де $d: \mathcal{H} \to \{0,1\}^*$ префіксна,

$$\sum_{\sigma \in S} \frac{1}{2^{|\sigma|}} \leq 1.$$

▷ Один зі способів доведення: побудуємо на S ймовір-

ністний розподіл (як на дискретному ім. просторі):
$\forall \sigma \in S$ нехай $P(\sigma)$ - це ймовірність отримати σ, підкидаючи кожного разу монетку і записуючи відп. біт (0 або 1) у рядок. В силу префіксності, цей процес може привести нас до одного $\sigma \in S$ (або до нескінченого) зі ім.
$P(\sigma) = \frac{1}{2^{|\sigma|}}$. Т.ч., $\sum_{\sigma \in S} \frac{1}{2^{|\sigma|}} = \sum_{\sigma \in S} P(\sigma) \leq 1$. ▲

Rem. В силу нер-сті Крафта, $w : h \mapsto \frac{1}{2^{|h|}}$, де $|h| = |d(h)|$ визначається відп. до деякої префіксної мови опису d, може бути обрана як вагова ф-ція на $\mathcal{H}$: довші за описом гіпотези мають меншу вагу. Отримана вище оцінка на ε_n тоді набуває вигляду

$$\varepsilon_n(m, w(h)\cdot\delta) \leq \sqrt{\frac{1}{2m}\left(-\ln w(h) + \ln\frac{2}{\delta}\right)} = \sqrt{\frac{1}{2m}\left(\underset{|h|\ln 2}{\ln 2^{|h|}} + \ln\frac{2}{\delta}\right)} < \sqrt{\frac{1}{2m}\left(|h| + \ln\frac{2}{\delta}\right)}.$$

Тоді з одного з Rem. вище отримуємо:

Cor. Нехай $\mathcal{H}$ - клас гіпотез ML на Z з ф-зією втрат $L : \mathcal{H} \times Z \to [0,1]$, $\mathcal{H}$ зліченний, і $d : \mathcal{H} \to \{0,1\}^*$ - префіксна мова його опису.

Тоді $\forall\ \delta \in (0,1)$, $m \in \mathbb{N}$ і розподілу $\mathcal{D}$ на Z зі вм. $\geq 1-\delta$ для вибірки S з m ел-тів, що i.i.d. відн. до $\mathcal{D}$,

$$\forall h \in \mathcal{H} \quad L_{\mathcal{D}}(h) \leq L_S(h) + \sqrt{\frac{1}{2m}\left(|h| + \ln\frac{2}{\delta}\right)}.$$

Це, у свою чергу, мотивує наст. версію SRM:

<u>Зав.</u> Говорять, що алгоритм A (ML для задачі на Z з ф-цією втрат $0 \leq \ell \leq 1$) здійснює <u>мінімізацію довжини опису</u> (minimum description length, MDL) у зліченному класі гіпотез $\mathcal{H}$ з префіксною мовою опису $d: \mathcal{H} \to \{0,1\}^*$, якщо $\forall\ m \in \mathbb{N}$ і $\delta \in (0,1)$ за наяв. вибіркою S з m ел-тів він повертає гіпотезу

$$h_S = A(S) \in \operatorname*{argmin}_{h \in \mathcal{H}} \left(L_S(h) + \sqrt{\frac{1}{2m}\left(|h| + \ln\frac{2}{\delta}\right)} \right).$$

<u>Кот.</u> Як і у загальному SRM, тут виникає упередженість на користь h з малою $|h|$: гіпотези з коротким описом заслуговують на більшу довіру (принцип бритви Оккама).

Втім, це поняття має сенс лише в рамках обраної

мови d і суттєво залежить від мови: вибір правильної d є тут важливою частиною розуміння конкретної задачі ML (разом з вибором класу H і ф-ції втрат l).

Послідовність алгоритмів ML

У свою чергу, <u>деф.</u> неперервної навченості можна ще послабити, дозволивши складності вибірки залежати не лише від ε, δ і $h \in H$, але й від розподілу D на Z:

<u>деф.</u> Нехай задача ML визначена областю визначення Z, класом гіпотез H і ф-єю втрат l, а P - якась множина ймовірнісних розподілів на Z. Алгоритм A для цієї задачі зветься <u>послідовним</u> (несуперечним, consistent) відповідно до H і P, якщо існує ф-ція $m_H^{con}: (0,1)^2 \times H \times P \to N$ така, що для будь-яких $\varepsilon, \delta \in (0,1)$, $h \in H$ і $D \in P$, якщо $m \geq m_H^{con}(\varepsilon, \delta, h, D)$, то зі ймовірністю $\geq 1-\delta$ A повертає за вибіркою з m

ел-тів, що і.і.d. відн. до $\mathcal{D}$ гіпотезу $A(S)$ з рахунок

$$L_{\mathcal{D}}(A(S)) \leqslant L_{\mathcal{D}}(h) + \varepsilon.$$

Зокрема, якщо $\mathcal{P}$ тут – сукупність усіх ім. розподілів на Z, то A зветься <u>універсально послідовним</u> (universally consistent) для $\mathcal{H}$.

<u>Зам.</u> Нагадаємо, що тут на Z визначена деяка б-алгебра, і усі $\mathcal{D} \in \mathcal{P}$ повинні бути визначені на ній (тоді "усі розподіли" – усі, що визначені на цій б-алг.). З <u>доб.</u> вимог:

<u>Соч.</u> Якщо клас гіпотез $\mathcal{H}$ (у задачі ML на Z з ф-цією втрат ℓ) рівномірно навчальний, то будь-який алгоритм A, що задов. умові рівн. навчальності $\mathcal{H}$, є універсально послідовним для $\mathcal{H}$.

<u>Зам.</u> При цьому обмеження (універсальної) послідовності дуже слабке: існують алгоритми з цією властивістю для очевидно "поганих" класів, наприклад, для класу

усіх бін. класифікаторів на $\mathcal{X} = \mathbb{N}$ (тобто ф-цій $h: \mathbb{N} \to \{0,1\}$), що, як ми показали у Th. вище, не є перевн. навчання!

Ex. Нехай $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$, де $\mathcal{X}$ ($\overset{(\leq)}{\text{злічена}}$), а $\mathcal{Y}$ скінченна, тобто маємо задачу класифікації на $\mathcal{X}$ (з зов. числом класів), $\ell = \ell_{0-1}$. Розглянемо алгоритм "запам'ятовування" навчальної вибірки:

$$\text{MEMORIZE}(S) := \left(h_S : x \mapsto \begin{cases} \underset{y \in \mathcal{Y}}{\text{argmax}} \, |\{(x,y) \in S\}|, & \text{якщо } (x,y) \in S \text{ для деякого } y; \\ y_0 \text{ у іншому випадку.} \end{cases} \right),$$

де y_0 - якась фіксована мітка. Це узагальнення прикладу перенавченої гіпотези для бін. клас., що було наведений раніше. Такі гіпотези, як бачили, можуть давати дуже високі $L_{\mathcal{D}}(h_S)$ для багатьох $\mathcal{D}$, а отже цей алгоритм очевидно "поганий". Але він усіх. погн. для (теж "поганою") класу усіх $h: \mathcal{X} \to \mathcal{Y}$. Дійсно, покажемо це для випадку, коли у ост. вигл. $\mathcal{X}$ ($\exists$ "правильна" $f: \mathcal{X} \to \mathcal{Y}$).

Нехай $\mathcal{D}$ - довільний розподіл на зліч. $\mathcal{X}$. Цей простір дискр., тобто $\mathcal{D}$ визначений своїми значеннями на 1-точкових підм. $\mathcal{X}$, і $\sum_{x \in \mathcal{X}} \mathcal{D}(\{x\}) = 1$ - абсолютно збіжний ряд (≥ 0). За власт. таких рядів, його ел-ти можна переставити, впорядкувавши за спаданням, що відп. деякому впорядкуванню $\mathcal{X}$: $\mathcal{X} = \{x_n\}_{n=1}^{\infty}$, і $\mathcal{D}(\{x_1\}) \geq \mathcal{D}(\{x_2\}) \geq \dots \geq \mathcal{D}(\{x_n\}) \geq \mathcal{D}(\{x_{n+1}\}) \geq \dots$ Крім того, у збіжного $\sum_{n=1}^{\infty} \mathcal{D}(\{x_n\}) = 1$ посл. залишкових сум прямує до 0: $\sum_{n=N}^{\infty} \mathcal{D}(\{x_n\}) \xrightarrow[N \to \infty]{} 0$. Осн. вона незростаюча (бо усі $\mathcal{D}(\{x_n\}) \geq 0$) це означає, що $\forall\, \varepsilon > 0\ \exists\, N \in \mathbb{N}: \varepsilon > \sum_{n=N}^{\infty} \mathcal{D}(\{x_n\}) = [\text{адит.}] = \mathcal{D}(\{x_n\}_{n=N}^{\infty})$.

Якщо тепер вибірка $S \in \mathcal{X}^m$ містить усі x_n для $n = \overline{1, N-1}$, то т.ч. $\mathcal{D}(\{x \mid x \notin S\}) < \varepsilon$. За побудовою $A = \text{MEMORIZE}$, тоді й $L_{\mathcal{D}}(A(S)) < \varepsilon$, бо "пам'ятовими" можуть бути лише точки за межами S. Тс.ч.,

$$P_{S \sim \mathcal{D}^m}\left(L_{\mathcal{D}}(A(S)) > \varepsilon\right) \leq \mathcal{D}^m\left(\{S \mid \exists\, n = \overline{1, N-1} : x_n \notin S\}\right) \leq [\text{субад.}] \leq$$

$$\leq \sum_{n=1}^{N-1} \mathcal{D}^m(\{S \mid x_n \notin S\}) = \sum_{n=1}^{N-1} \mathcal{D}(X \setminus \{x_n\})^m = \sum_{n=1}^{N-1} \left(1 - \mathcal{D}(\{x_n\})\right)^m \quad (\leq)$$

Позначимо $\eta := \min_{n=\overline{1,N-1}} \mathcal{D}(\{x_n\})$ $\in (0,1)$. Як і N, це значення залежить від $\mathcal{D}$ і ε. Тоді

$$(\leq) \sum_{n=1}^{N-1} (1-\eta)^m = (N-1)(1-\eta)^m < (N-1)e^{-m\eta}$$

Для $m \geq \frac{1}{\eta} \ln \frac{N-1}{\delta}$ це $\leq \delta$. Т.ч., $\forall h \in \mathcal{H}$

$$P_{S \sim \mathcal{D}^m} \left(L_{\mathcal{D}}(A(S)) \leq \varepsilon \leq L_{\mathcal{D}}(h) + \varepsilon \right) \geq 1 - \delta$$

для $m \geq m_{\mathcal{H}}^{con}(\varepsilon, \delta, \mathcal{D})$, де $m_{\mathcal{H}}^{con}(\varepsilon, \delta, \mathcal{D}) \leq \left\lceil \frac{1}{\eta} \ln \frac{N-1}{\delta} \right\rceil$ не залежать від $h \in \mathcal{H}$. Звідси й маємо утв. } залежать від $\mathcal{D}, \varepsilon$

послідовність.

Тут за рахунок залежності $m_{\mathcal{H}}^{con}$ від $\mathcal{D}$ вдалося забезпечити, щоб (з ім. $\geq 1-\delta$) S містила усі ел-ти X з дост. великими $\mathcal{D}$.

Звичайно, це вже ніяке не "навчання". Цей алгоритм послідовний лише у тому сенсі, що збільшення S гарантовано зменшує помилку.

Ex. Прикладом послідовного буде "алгоритм", що, скажімо, завжди повертає одну й ту саму h або обирає її випадково незалежно від S.