По-перше, замість пари з розподілу $\mathcal{D}$ на $\mathcal{X}$ та $f: \mathcal{X} \to \mathcal{Y}$ розглянемо т. зв. об'єднаний розподіл $\underbrace{\mathcal{D}\text{ — ймовірнісну}}_{\text{(joint distribution)}}$ міру на $\mathcal{X} \times \mathcal{Y}$ (де $\mathcal{X}$ та $\mathcal{Y}$ тепер $\forall$ такі, що на $\mathcal{X} \times \mathcal{Y}$ $\exists$ стр. ймовірнісного пр-ру).

Тобто $\forall A \subset \mathcal{X} \times \mathcal{Y}$ (більш точно, A з відп. σ-алгебри) $\mathcal{D}(A)$ — це ймовірність того, що випадково обрана пара з точки даних $x \in \mathcal{X}$ та мітки $y \in \mathcal{Y}$ належить до A: $\mathcal{D}(A) = \underset{(x,y) \sim \mathcal{D}}{P}\left((x,y) \in A\right)$.

Зокрема, $\mathcal{D}$ визначає:

— маргінальний (marginal) розподіл $\mathcal{D}_x$ на $\mathcal{X}$: для $A \subset \mathcal{X}$ це ім. того, що випадкова $x \in \mathcal{X}$ належить до A незалежно від мітки:

$$\mathcal{D}_x(A) = \underset{x \sim \mathcal{D}_x}{P}(x \in A) = \underset{(x,y) \sim \mathcal{D}}{P}\left((x,y) \in A \times \mathcal{Y}\right) = \mathcal{D}(A \times \mathcal{Y}).$$

У Ex. з фруктами це, як раніше, ім. потрапити у певну область за кольором і твердістю.

— умовна ймовірність (conditional probability) того, що $y \in B$, для деякої $B \subset Y$ за умови $x \in A$:

$$\underset{(x,y)\sim \mathcal{D}}{P}(y \in B \mid x \in A) = \frac{\mathcal{D}(A\times B)}{\mathcal{D}_x(A)} = \frac{\mathcal{D}(A\times B)}{\mathcal{D}(A\times Y)}.$$

Зокрема, для одноточкової $B = \{y\}$ це ймов. отримати мітку y для усіх точок даних $x \in A$: $\underset{(x,y)\sim \mathcal{D}}{P}(y \mid x \in A) = \frac{\mathcal{D}(A\times \{y\})}{\mathcal{D}_x(A)}$.

Якщо ж і $A = \{x\}$ одноточкова, це ймов. отримати y для цієї x (у Ex. — того, що фрукт з кольором і твердістю x буде смачним (при $y = 1$) чи ні ($y = 0$)). Формально це $\underset{(x,y)\sim \mathcal{D}}{P}(y \mid x) = \frac{\mathcal{D}(\{(x,y)\})}{\mathcal{D}_x(\{x\})}$. Для недискретного $(X, \mathcal{D}_x)$ (скажімо, $X \subset \mathbb{R}^d$ з однорідним чи норм. $\mathcal{D}_x$) це може бути невизначеним $\left(\frac{0}{0}\right)$, тоді будемо (не дуже формально) розуміти це як границю $\lim\limits_{A \to \{x\},\, x \in A} \frac{\mathcal{D}(A\times\{y\})}{\mathcal{D}_x(A)}$. Далі вважаємо $\underset{(x,y)\sim \mathcal{D}}{P}(y \mid x)$ визначеним всюди, де це буде потрібно.

(напр., кулі)

Знову ж, на практиці усі простори дискретні, тому можемо вважати, що точки мають ненульові міри.

Ex. Для попередньої моделі з $(\mathcal{D}, f)$ $\mathcal{D}_X$ — це і є (старий) $\mathcal{D}$,

а $P(y \in B \mid x) = \begin{cases} 1, & f(x) \in B; \\ 0, & f(x) \notin B, \end{cases}$ зокр. $P(y \mid x) = \begin{cases} 1, & f(x) = y; \\ 0, & f(x) \neq y; \end{cases}$

$\forall\, A \subset \mathcal{X} \times \mathcal{Y}$

$$\underset{\substack{\uparrow \\ \text{у новому} \\ \text{сенсі}}}{\mathcal{D}}(A) = \underset{\substack{\uparrow \\ \text{у старому} \\ \text{сенсі}}}{\mathcal{D}}\big(\{x \in \mathcal{X} \mid (x, f(x)) \in A\}\big).$$

Зар. Чи визначений $\mathcal{D}$ марг. розп. $\mathcal{D}_X$ та/або умовними ймн. у загальному випадку?

Т.ч., тепер мітки y, взагалі кажучи, не визначені однозначно за x (лише з деякою ймв.), що реалістичніше.

По-друге, уточнимо def. помилки. Для задач класифікації, тобто коли $\mathcal{Y}$ скінченна, це можна зробити так:

def. (Справжньою) помилкою гіпотези $h: \mathcal{X} \to \mathcal{Y}$ для задачі класифікації з розподілом $\mathcal{D}$ на $\mathcal{X} \times \mathcal{Y}$ будемо називати

$$L_{\mathcal{D}}(h) := \mathcal{D}\left(\{(x,y) \in \mathcal{X} \times \mathcal{Y} \mid h(x) \neq y\}\right) \in [0,1]$$

Rem. Для моделі з $(\mathcal{D}, f)$ це те ж саме, що $L_{\mathcal{D},f}(h)$ (чому?), а припущення реалізовності означала б, що $L_{\mathcal{D}}(h^*) = 0$ для деякої h^* з обраного класу гіпотез $\mathcal{H}$. Втім, тут ми не будемо його застосовувати, бо у загальному випадку $L_{\mathcal{D}}(h)$ для $h \in \mathcal{H}$ можуть бути обмежені знизу:

Ex. Байєсівським оптимальним предиктором (гіпотезою) для задачі класифікації з розп. $\mathcal{D}$ на $\mathcal{X} \times \mathcal{Y}$, де $\mathcal{Y} = \{0,1\}$, зв.

$$f_{\mathcal{D}}: \mathcal{X} \to \mathcal{Y} \; : \; f_{\mathcal{D}}(x) := \begin{cases} 1, & P_{(x,y)\sim\mathcal{D}}(y = 1 \mid x) \geq \frac{1}{2}; \\ 0 & \text{у інш. вип.} \end{cases}$$

Pr. Б.о.п. має найменшу справжню помилку серед усіх гіпотез: $L_{\mathcal{D}}(f_{\mathcal{D}}) \underset{\text{мн.}}{\leq} L_{\mathcal{D}}(h) \quad \forall\, h: \mathcal{X} \to \mathcal{Y}$.

▷ $\forall\, h$ $L_{\mathcal{D}}(h)$ – міра точок (x,y), де $h(x) \neq y$. У нас:

$$L_{\mathcal{D}}(h) = \mathcal{D}\left(\{(x,0) \mid h(x) = 1\} \cup \{(x,1) \mid h(x) = 0\}\right) = [\text{адит.}] =$$

$= \mathcal{D}(\{(x,0) \mid h(x)=1\}) + \mathcal{D}(\{(x,1) \mid h(x)=0\}) =$ (Тут і далі $\int_A f(x)\,d\mu(x)$ – інт. Лебега)

Використаємо відповідь попер. Впр. для $Y = \{0,1\}$:

$$= \int_{\{x \mid h(x)=1\}} P(0|x)\,d\mathcal{D}_X(x) + \int_{\{x \mid h(x)=0\}} P(1|x)\,d\mathcal{D}_X(x) \geq [P(0|x)+P(1|x)=1\ \forall x] \geq$$

$$\geq \int_{\{x \mid h(x)=1\}} \min\{P(1|x), 1-P(1|x)\}\,d\mathcal{D}_X(x) + \int_{\{x \mid h(x)=0\}} \min\{P(1|x), 1-P(1|x)\}\,d\mathcal{D}_X(x) =$$

$= [\text{адитивність інтеграла}] = \int_X \min\{P(1|x), 1-P(1|x)\}\,d\mathcal{D}_X(x)$.

Якщо ж $h = f_{\mathcal{D}}$, то $h(x)=1 \Leftrightarrow P(1|x) \geq \frac{1}{2}$ і $h(x)=0 \Leftrightarrow P(1|x) < \frac{1}{2}$. Оск. $\int_A f\,d\mu = \int_X 1_A f\,d\mu$, де, нагадаємо, характеристична функція $1_A(x) = \begin{cases} 1, & x \in A \\ 0, & x \notin A \end{cases}$; замість нер-сті $\geq$ вище можемо записати:

$$L_{\mathcal{D}}(f_{\mathcal{D}}) = \ldots = \int_X 1_{\{x \mid P(1|x) \geq \frac{1}{2}\}}(x)(1-P(1|x))\,d\mathcal{D}_X(x) + \int_X 1_{\{x \mid P(1|x) < \frac{1}{2}\}}(x) \cdot P(1|x)\,d\mathcal{D}_X(x) = [\text{лінійність інтеграла}] = \int_X \left(1_{\{P(1|x) \geq \frac{1}{2}\}}(x)(1-P(1|x)) + 1_{\{P(1|x) < \frac{1}{2}\}}(x) \cdot P(1|x)\right) d\mathcal{D}_X(x) =$$

Помітимо, що ф-ція під інтегралом – це $1-P(1|x) \leq P(1|x)$ при $P(1|x) \geq \frac{1}{2}$ і $P(1|x) < 1-P(1|x)$ при $P(1|x) < \frac{1}{2}$, тобто $= \int_X \min\{P(1|x), 1-P(1|x)\}\,d\mathcal{D}_X(x) \leq L_{\mathcal{D}}(h)$

за все доведення, що й потрібно. ◭

<u>Rem.</u> Згідно з цим <u>Pr.</u> та його доведенням, якщо, наприклад, $P(1|x) = \frac{1}{4}$ $\forall x$ (тобто чверті точок присвоєна мітка 1 незалежно від точки), то $\forall h$

$$L_{\mathcal{D}}(h) \geq L_{\mathcal{D}}(f_{\mathcal{D}}) = \int\limits_{X} \frac{1}{4}\, d\mathcal{D}_X = \frac{1}{4}\mathcal{D}_X(X) = \frac{1}{4},$$

тобто припущення реалізовності невірне $\forall \mathcal{H}$.

Нагадаємо, що наша мета — мінімізувати $L_{\mathcal{D}}(h)$ для $h \in \mathcal{H}$. При цьому просто $h = f_{\mathcal{D}}$ покласти не можна, бо алгоритм не "знає" $\mathcal{D}$, лише навчальні дані $S = \{(x_i, y_i)\}_{i=1}^{m} \in (X \times Y)^m$. Як і раніше, застосовуємо парадигму $ERM_{\mathcal{H}}$: для $h \in \mathcal{H}$ $\underbrace{\text{мінімізуємо } L_S(h)}_{\text{емпіричну помилку}}$, що визначена для задачі класифікації як раніше:

$$L_S(h) = \frac{1}{m}\left|\{i = \overline{1,m} \mid h(x_i) \neq y_i\}\right| \in [0,1].$$

Але тепер ми не можемо сподіватися опустити $L_{\mathcal{D}}(h)$ нижче $\min\limits_{h' \in \mathcal{H}} L_{\mathcal{D}}(h')$ (взагалі кажучи, і.в.), що, як ми покажем, У заг. випадку > 0. Тепер можна дати нове означення навчаності $\mathcal{H}$, [illegible] узагальнивши $L_{\mathcal{D}}$ і L_S на довільні $\mathcal{Y}$.

<u>Def.</u> Нехай задача ML моделюється ймовірнісним простором $\mathcal{Z}$, а $\mathcal{H}$ – клас гіпотез ($\forall$ множина). Тоді <u>(узагальненою) функцією втрат</u> (loss function) зв. $\forall$ відображення $\ell : \mathcal{H} \times \mathcal{Z} \to \mathbb{R}_+$ таке, що $\forall\, h \in \mathcal{H}$ $\ell(h, \cdot)$ (тобто відобр. $z \in \mathcal{Z} \mapsto \ell(h, z) \in \mathbb{R}_+$) є вимірною ф-цією на $\mathcal{Z}$.

<u>Rem.</u> Нагадаємо, що $\mathbb{R}_+ = [0, +\infty)$ і що вимірна ф-ція $\to \mathbb{R}_+$ – та, для якої прообрази вимірних множин вимірні ($\Leftrightarrow \ell(h, \cdot)^{-1}([0, a])$ вимірна у $\mathcal{Z}$ $\forall\, a$).

Іншими словами, $\ell(h, \cdot)$ – <u>випадкова величина</u> (невід'ємна) на $\mathcal{Z}$. Для наших задач керованого ML $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$ з розподілом $\mathcal{D}$, як вище, але у випадку некерованого ML може бути й $\mathcal{Z} = \mathcal{X}$ (немає міток).

<u>Ex. 1.</u> Для задач класифікації розглядають зазвичай

т.зв. <u>0-1 втрату</u> (0-1 loss). Тут Y скінченна, $h: X \to Y$, і

$$l_{0-1}(h,(x,y)) := \begin{cases} 1, & h(x) \neq y; \\ 0, & h(x) = y. \end{cases} \quad \left(\text{При цьому } l_{0-1}(h,\cdot) = 1_{\{(x,y) \mid h(x) \neq y\}}.\right)$$

<u>Впр.</u> Перевірити умову вимірності.

<u>2.</u> Якщо $Y = \mathbb{R}$, то зазвичай говорять про задачу <u>регресії</u> (regression). Для такої задачі та $h: X \to Y$ типовим вибором є <u>квадратична втрата</u> (square loss)

$$l_{sq}(h,(x,y)) := (h(x)-y)^2.$$

Це можна узагальнити й на $Y \subset \mathbb{R}^d$: тоді це квадрат евкл. відстані $\sum_{i=1}^{d} (h(x)^i - y^i)^2$ (метод найменших квадратів).

<u>def.</u> (Справжньою) помилкою (ризиком) гіпотези $h \in \mathcal{H}$ для функції втрат l і розподілу ім. $\mathcal{D}$ на Z зветься математичне очікування

$$L_{\mathcal{D}}(h) := \mathop{E}_{z \sim \mathcal{D}} l(h,z).$$

Емпіричною $\underbrace{\text{помилкою}}_{\text{(ризиком)}}$ h $\underbrace{\text{на } S = \{z_i\}_{i=1}^{m} \in Z^m}_{\text{навчальних даних}}$ зветься середнє значення

$$L_S(h) := \frac{1}{m} \sum_{i=1}^{m} l(h, z_i).$$

Rem. Нагадаємо, що мат. очікування (expectation) вип. величини (у даному випадку $\ell(h,\cdot)$) – це інтеграл $\underset{z\sim\mathcal{D}}{\mathbb{E}}\ell(h,z)=\int_Z \ell(h,z)\,d\mathcal{D}(z)$ (що існує в силу вимірності).

Зокрема, середнє значення – це мат. оч. для скінченного простору S і рівномірного розподілу (коли кожна z_i має ім. $\frac{1}{m}$): $L_S(h)=L_{\underset{(\text{на } S)}{\mathcal{D}_{\text{рівн.}}}}(h)$.

Ex. 1. Для 0-1 втрати:

$$L_{\mathcal{D}}(h)=\underset{(x,y)\sim\mathcal{D}}{\mathbb{E}}\ell_{0\text{-}1}(h,\underbrace{\quad}_{(x,y)})=\int_{X\times Y}1_{\{(x,y)\mid h(x)\neq y\}}\,d\mathcal{D}=\mathcal{D}(\{(x,y)\mid h(x)\neq y\})$$

– та, що вище, і так само для L_S (чому?).

2. Для квадратичної втрати:

$$L_{\mathcal{D}}(h)=\underset{(x,y)\sim\mathcal{D}}{\mathbb{E}}\ell_{sq}(h,(x,y))=\underset{(x,y)\sim\mathcal{D}}{\mathbb{E}}(h(x)-y)^2,$$

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

$$(\ell_{sq}"(h,(x_i,y_i)))$$

Тепер можемо переписати def. PAC навчаності:

def. Для задачі ML на просторі Z з функцією втрат ℓ клас гіпотез $\mathcal{H}$ (на якому ℓ визначена) зветься __агностично PAC навчаним__, якщо існує ф-ція $m_{\mathcal{H}} : (0,1)^2 \to \mathbb{N}$ і „алгоритм ML, що для $\forall\, \varepsilon, \delta \in (0,1)$ та розподілу $\mathcal{D}$ на Z (такому, що зберігається умова вимірності на ℓ) за навчальною вибіркою S з $m \geq m_{\mathcal{H}}(\varepsilon, \delta)$ елементів, що i.i.d. відповідно до $\mathcal{D}$, зі ймовірністю не меншою за параметр впевненості $1-\delta$ повертає гіпотезу $h_S \in \mathcal{H}$, для якої ризик $L_{\mathcal{D}}(h_S)$ обмежений зверху значенням $\min\limits_{h \in \mathcal{H}} L_{\mathcal{D}}(h) + \varepsilon$, де ε - параметр точності:

$$\mathcal{D}^m(\{S \mid L_{\mathcal{D}}(h_S) \leq \min_{h \in \mathcal{H}} L_{\mathcal{D}}(h) + \varepsilon\} \geq 1-\delta.$$

__Rem.__ "Агностично" тут означає без "віри" у припущення реалізовності (якщо ж воно вірне, то $\min\limits_{h \in \mathcal{H}} L_{\mathcal{D}}(h) = 0$). Як бачимо, це означення працює для різних типів керованого навч. і навіть для некерованого! Як на пам'ятне, далі вважатимемо, що $m_{\mathcal{H}}(\varepsilon, \delta)$ - найменше, для якого виконана умова - __складність вибірки__.

Впр. Показати, що $\forall$ агностично PAC навчаний клас для задачі класифікації є PAC навчаним (для того ж алгоритма).

① Достатня умова агностично PAC навчаності – рівномірна збіжність

Згадаємо, що алгоритм ML, що реалізує $ERM_{\mathcal{H}}$, мінімізує емп. ризик $L_S(h)$ у класі $h \in \mathcal{H}$; наша ж задача – мінімізувати справжній ризик $L_{\mathcal{D}}(h)$. Розглянемо S, для яких вони близькі:

def. Навчальна вибірка S зветься ε-репрезентативною для $\varepsilon > 0$, задачі ML з простором Z, розподілом ім. $\mathcal{D}$, класом гіпотез $\mathcal{H}$ і ф-цією втрат ℓ, якщо $\forall h \in \mathcal{H}$

$$|L_S(h) - L_{\mathcal{D}}(h)| \le \varepsilon.$$

Рч. Якщо (в умовах попер. def.) S є $\frac{\varepsilon}{2}$-репрезентативною, то $\forall$ гіпотези $h_S \in ERM_{\mathcal{H}}(S)$

$$L_{\mathcal{D}}(h_S) \leq \min_{h\in\mathcal{H}} L_{\mathcal{D}}(h) + \varepsilon.$$

▷ Згадаємо, що умова означає $h_S \in \arg\min_{h\in\mathcal{H}} L_S(h)$, тобто $L_S(h_S) \leq L_S(h)$ $\forall h \in \mathcal{H}$. Тому

$$L_{\mathcal{D}}(h_S) \leq \left[\tfrac{\varepsilon}{2}\text{-репр.}\right] \leq L_S(h_S) + \frac{\varepsilon}{2} \leq L_S(h) + \frac{\varepsilon}{2} \leq \left[\tfrac{\varepsilon}{2}\text{-репр.}\right] \leq L_{\mathcal{D}}(h) + \varepsilon,$$

$\forall h \in \mathcal{H}$. Переходячи справа до min, отримаємо потрібне. ◁

def. Говорять, що клас гіпотез $\mathcal{H}$ має властивість рівномірної збіжності (uniform convergence) для задачі ML з простором Z і ф-цією втрат ℓ, якщо $\exists$ така ф-ція $m^{UC}_{\mathcal{H}} : (0,1)^2 \to \mathbb{N}$, що $\forall \varepsilon, \delta \in (0,1)$ і розподілу $\mathcal{D}$ на Z (такого, що вик. умова вимірності на ℓ) навчальна вибірка з $m \geq m^{UC}_{\mathcal{H}}(\varepsilon,\delta)$ елементів, що i.i.d. відповідно до $\mathcal{D}$, є ε-репрезентативною (для $Z, \mathcal{D}, \mathcal{H}, \ell$) зі ймовірністю $\geq 1-\delta$.

Rem. Тут зазвичай теж беруть $m^{UC}_{\mathcal{H}}(\varepsilon,\delta)$ найменше можливе натуральне. [illegible] "Рівномірно" означає незалежність вибірки від $\mathcal{D}$ або конкретного $h \in \mathcal{H}$. З означень і попер. Рч. очевидно випливає: $(\forall \varepsilon, \delta)$

Сл. Якщо (в умовах попер. def.) клас гіпотез $\mathcal{H}$ має власт. рівномірної збіжності з деякою $m^{UC}_{\mathcal{H}}$, то він агностично PAC навчаний, причому складність вибірки $m_{\mathcal{H}}(\varepsilon,\delta) \le m^{UC}_{\mathcal{H}}(\frac{\varepsilon}{2},\delta)$, а алгоритмом може бути будь-який, що здійснює $ERM_{\mathcal{H}}$.

Ця умова дозволяє узагальнити (тобто посилити) доведену раніше PAC навчаність скінч. класів на агностичний випадок:

Рч. Будь-який скінченний клас гіпотез $\mathcal{H}$ має властивість рівн. збіжності для задачі ML з довільним Z і $\ell : \mathcal{H} \to Z \to [0,1]$, причому

$$m^{UC}_{\mathcal{H}}(\varepsilon,\delta) \le \left\lceil \frac{1}{2\varepsilon^2} \ln \frac{2|\mathcal{H}|}{\delta} \right\rceil, \quad (\forall \varepsilon, \delta)$$

а отже агностично PAC навчаний для $m_{\mathcal{H}}(\varepsilon,\delta) \le m^{UC}_{\mathcal{H}}(\frac{\varepsilon}{2},\delta)$

і $\forall$ алгоритма з $ERM_{\mathcal{H}}$.

Rem. Тут буде суттєвою обмеженість ф-ції втрат: $\ell(h, z) \leq 1 \;\; \forall\, h, z$ (це вірно й для $a \leq \ell(h,z) \leq b$, як буде видно з доведення, але оцінка $m_{\mathcal{H}}^{UC}$ може змінитися).

$\triangleright$ В силу попер. Cor., дост. довести саме рівн. збіжність (і отримати оцінку $m_{\mathcal{H}}^{UC}$). Початок практично той же, що у дов. Cor. про скінченні $\mathcal{H}$ вище: щоб оцінити зверху ймовірність отримати ε-репр. S, розглянемо множину вибірок, що такими не $\in$: за def.,

$$\{S \mid S - \text{не } \varepsilon\text{-репр.}\} = \{S \mid \exists h \in \mathcal{H} : |L_S(h) - L_{\mathcal{D}}(h)| > \varepsilon\} =$$

$$= \bigcup_{h \in \mathcal{H}} \{S \mid |L_S(h) - L_{\mathcal{D}}(h)| > \varepsilon\},$$ тому за субадитивністю ймовірність

$$\mathcal{D}^m(\{S \mid S - \text{не } \varepsilon\text{-репр.}\}) \leq \sum_{h \in \mathcal{H}} \mathcal{D}^m(\{S \mid |L_S(h) - L_{\mathcal{D}}(h)| > \varepsilon\}) \leq$$

Розглянемо $S \mapsto L_S(h)$ для фікс. $h \in \mathcal{H}$ як випадкову величину (на $\mathcal{Z}^m$ з розп. ім. $\mathcal{D}^m$) її мат. оч.:

$$\mathop{\mathbb{E}}_{S \sim \mathcal{D}^m}(L_S(h)) = \mathop{\mathbb{E}}_{S = \{z_i\}_{i=1}^m \sim \mathcal{D}^m}\left(\frac{1}{m}\sum_{i=1}^{m} \ell(h, z_i)\right) = \left[\begin{matrix}\text{лінійність} \\ \text{мат. оч.}\end{matrix}\right] =$$

$$= \frac{1}{m}\sum_{i=1}^{m} \mathop{E}_{S\sim\mathcal{D}^m} \left(\ell(h, z_i)\right) = \begin{bmatrix} (z_1,\dots,z_m)\mapsto \ell(h,z_i) - \\ \text{ф-ія від 1 з коорд.,} \\ \text{власт. } \mathcal{D}^m \text{ і інтеграла,} \\ \mathcal{D}(Z)=1 \end{bmatrix} = \frac{1}{m}\sum_{i=1}^{m} \mathop{E}_{z\sim\mathcal{D}} \left(\ell(h,z)\right) =$$

$= [\underline{\text{def.}}\, L_{\mathcal{D}}] = L_{\mathcal{D}}(h)$ (пор. також з однією з задач практики).

Тобто $S \mapsto |L_S(h) - L_{\mathcal{D}}(h)|$ – відхилення вип. величини від її мат. очікування. Його можна оцінити за допомогою наступної версії закону великих чисел:

$\underline{\text{Тh.}}$ (нерівність Хьоффдінга концентрації міри). Нехай $\{\theta_i\}_{i=1}^{m}$ – i.i.d. випадкові величини на імов. пр-рі Z, і $E(\theta_i) = \mu$ $\forall i$. Якщо крім того $P(\theta_i \in [a,b]) = 1$ для деяких $a<b$ і $\forall i$, то $\forall \varepsilon > 0$

$$P\left(\left|\frac{1}{m}\sum_{i=1}^{m}\theta_i - \mu\right| > \varepsilon\right) \le 2e^{-\frac{2m\varepsilon^2}{(b-a)^2}}.$$

▷ Див. додаток B у [1] або курс мат. статистики. ◁

Тобто середні значення наближаються до мат. очікування так, що різницю можна оцінити. У нас для $i=\overline{1,m}$ $\theta_i = \ell(h, z_i)$, а отже $\frac{1}{m}\sum_{i=1}^{m}\theta_i = L_S(h)$ і $\mu = L_{\mathcal{D}}(h)$.

П.ч. за пер. $\mathcal{H}$. (при $a=0, b=1$),

$\circledcirc{\leq} \sum_{h\in\mathcal{H}} 2e^{-2m\varepsilon^2} = 2|\mathcal{H}|\,e^{-2m\varepsilon^2} \leq$,

а отже при $m \geqslant \frac{1}{2\varepsilon^2}\ln\frac{2|\mathcal{H}|}{\delta}$

$\leq 2|\mathcal{H}|\,e^{-\ln\frac{2|\mathcal{H}|}{\delta}} = \delta$.

Тому $\underbrace{\mathcal{D}^m(\{S \mid S-\varepsilon\text{-репр.}\}) > 1-\delta}_{\text{для таких } m}$, що й потрібно. $\blacktriangle$

<u>Rem.</u> Ця оцінка працює і для (теоретично) нескінченних класів на практиці в силу дискретизації. Дійсно, нехай ел-ти $\mathcal{H}$ задаються d дійсними параметрами (у наших прикладах з лекцій та практики: $d=2$ для прямокутників, довільне $d\in\mathbb{N}$ для паралелепіпедів, $d=1$ для кругів або багатокутн. куль або для порогових ф-цій $h(x)=\operatorname{sign}(x-a)$ з пар. a), які задаються реально, скажімо, 64 бітами кожен. Тоді $|\mathcal{H}|\leq 2^{64d}$, отже за Рп. $\mathcal{H}$ агностично PAC навч. для ($\forall\, \varepsilon, \delta$)

$$m_{\mathcal{H}}(\varepsilon,\delta) \leq m^{UC}_{\mathcal{H}}\left(\frac{\varepsilon}{2},\delta\right) \leq \left\lceil \frac{2}{\varepsilon^2}\ln\frac{2|\mathcal{H}|}{\delta}\right\rceil \leq \left\lceil \frac{2}{\varepsilon^2}\ln\frac{2\cdot 2^{64d}}{\delta}\right\rceil = \left\lceil \frac{2\ln\frac{2}{\delta} + 128d\cdot \overset{\frac{1}{\vee}}{\widehat{\ln 2}}}{\varepsilon^2}\right\rceil.$$