Machine Learning: Theory & Algorithms I

Mathematics of Machine Learning

Математика машинного навчання

<u>Література</u>

1. S. Shalev-Schwartz, S. Ben-David. Understanding Machine Learning.
2. M. Mohri, A. Rostamizadeh, A. Talwalkar. Foundations of Machine Learning.
3. D.J. MacKay. Information Theory, Inference, and Learning Algorithms.

Машинне навчання (ML) – область інформатики (більш точно, т.зв. штучного інтелекту), що досліджує <u>статистичні</u> алгоритми, які здатні модифікувати себе з використанням відомих даних ("навчатися") і робити змістовні висновки з раніше невідомих даних ("застосовувати набути знання").

Застосування: у задачах, де побудова явного алгоритму (на противагу статистичному, "традиційний" підхід штучного інтелекту – імітація людського мислення) дуже складна (як правило, практично неможлива). Зокрема,

– задачі, з якими добре справляються люди або тварини: розпізнавання образів, мови, інші задачі класифікації (виявлення спаму), переклад, спілкування природними мовами, водіння транспортних засобів;

– задачі, навпаки, за межами людських можливостей: аналіз великих об'ємів даних у науці, медицині, бізнесі, знаходження прихованих закономірностей.

З іншого боку, відмінність від статистики – фокус на алгоритмах, відсутність наперед визначеної гіпотези про розподіл даних.

Наші задачі у цьому курсі:

1. Побудувати загальні математичні моделі машинного навчання та дослідити їхні властивості.

2. Застосувати цей загальний підхід до конкретних алгоритмів ML.

Більш точно, нас цікавитиме перш за все кероване статистичне пакетне пасивне навчання. Значення цих слів:

Кероване (з учителем, supervised) – зразки даних (training data), на яких алгоритм навчається, містять суттєву інформацію, що відсутня у даних, до яких він застосовується і яку потрібно відновити (мітки класифікації, розшифровки, переклади, продовження фраз...) на відміну від некерованого (без учителя, unsupervised), де навчальні та тестові (робочі) дані не відрізняються (наприклад, задачі кластеризації)

Статистичне у цьому контексті означає, що навчальні дані згенеровано деяким випадковим процесом, на відміну від ситуації "вчителя", що допомагає чи заважає.

Пакетне (batch learning) — алгоритм застосовується до тестових даних лише після обробки суттєвої кількості навчальних, на відміну від онлайн-навчання, де ці процеси можуть відбуватися одночасно.

Пасивне — алгоритм лише сприймає інформацію, що йому надають, на відміну від активного, коли він може робити запити до оточення під час навчання.

Початкова математична модель ML

Елементи моделі:

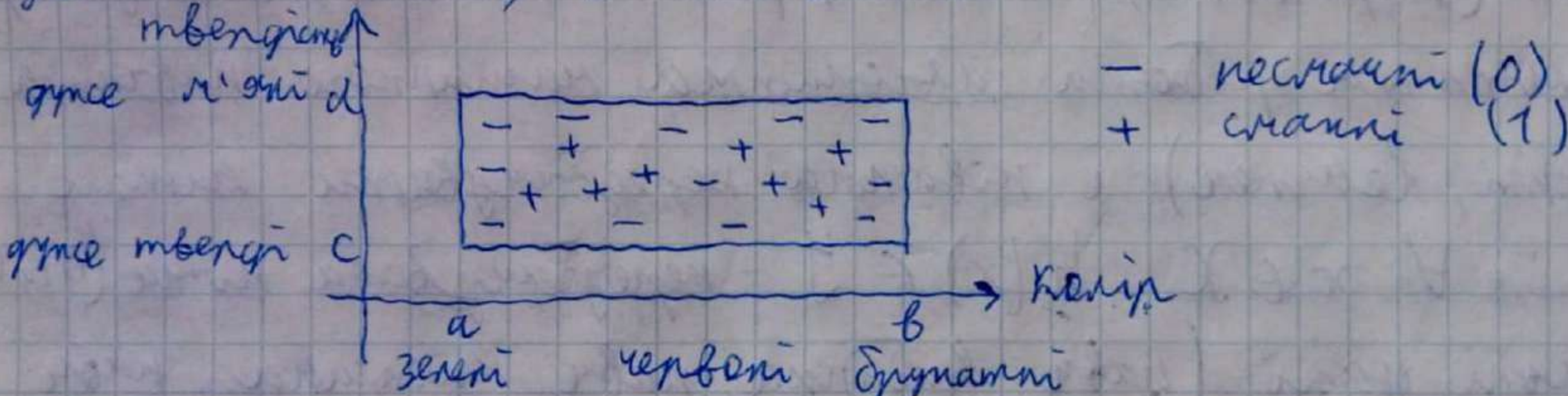
- Область визначення $\mathcal{X}$ (domain set) — довільна множина об'єктів, що потрібно класифікувати. Часто вважають $\mathcal{X} \subset \mathbb{R}^d$; у цьому випадку координати $x \in \mathcal{X}$ зв.

його ознаками чи характеристиками (features), тобто x - вектор ознак.

- <u>Множина міток</u> (область значень) $\mathcal{Y}$ (label set) - тут розглянуто задачу бінарної класифікації, тобто на 2 класи, тоді $\mathcal{Y} = \{0, 1\}$ (часто беруть $\mathcal{Y} = \{-1, +1\}$), де 0 і 1 позначають належність $x \in \mathcal{X}$ до одного з класів.
- <u>Гіпотеза</u> (предиктор, правило передбачення, класифікатор) $h: \mathcal{X} \to \mathcal{Y}$ (hypothesis, predictor, prediction rule, classifier), що є результатом роботи <u>алгоритму</u> машинного навчання A (algorithm, learner) і повинна передбачувати мітки ел-тів $\mathcal{X}$: $\forall x \in \mathcal{X}$ $h(x) \in \mathcal{Y}$ - передбачувана мітка (0 чи 1)
- <u>Навчальні дані</u> (навч. вибірка, навч. множина, навч. приклади) S, що отримує алгоритм (training data, set, sample). Оск. навчання <u>кероване</u>, S тут є послідовністю $S = \{(x_1, y_1), \ldots, (x_m, y_m)\}$ m пар з точок $x_i \in \mathcal{X}$ та їхніх міток $y_i \in \mathcal{Y}$, що ми вважаємо правильними. Для <u>пакетного</u>

пасивного навчання гіпотеза повинна визначатися алгоритмом за S : $h = A(S)$.

Ex. Задача визначення смачності невідомого фрукта за 2 ознаками: $X = [a, b] \times [c, d] \subset \mathbb{R}^2$ - прямокутник, де координати фрукта $x \in X$ означають його колір та твердість (у деякому числовому вираженні), $Y = \{0, 1\}$, де 0 - несмачні, 1 - смачні:



Навчальні дані $\{(x_i, y_i)\}_{i=1}^{m}$ - результат збирання m випадкових фруктів, визначення їх ознак та того, чи є вони смачними. h - функція, що за ознаками визначає чи фрукт смачний. Ключовими є 2 наступних питання:

Що ми припускаємо про побудову S?

Ост. навчання статистичне, будемо припускати, що у $S = \{(x_i, y_i)\}_{i=1}^{m}$ точки $x_i \in \mathcal{X}$ обираються випадковим чином згідно з імовірнісним розподілом $\mathcal{D}$ на $\mathcal{X}$ (probability distribution), що моделює властивості оточення (у Ex. - саду, де ростуть фрукти).

Це означає, що $(\mathcal{X}, \mathcal{F}, \mathcal{D})$ - імовірнісний простір, тобто

- задана σ-алгебра $\mathcal{F}$ підмножин $\mathcal{X}$ (вимірних) - деяка сукупність підмн. $A \subset \mathcal{X}$, що
  - містить $\mathcal{X}$: $\mathcal{X} \in \mathcal{F}$;
  - замкнена відн. доповнень: $A \in \mathcal{F} \Rightarrow \mathcal{X} \setminus A \in \mathcal{F}$;
  - замкнена відн. ($\leq$)злічених об'єднань: $\{A_i\}_{i=1}^{\infty} \subset \mathcal{F} \Rightarrow \bigcup_{i=1}^{\infty} A_i \in \mathcal{F}$ (де замість ∞ може бути скінч. n). (звідси вона замкнена й відн. перетинів за правилом де Моргана.)

Зокрема, для $\mathcal{X} \subset \mathbb{R}^d$ $\mathcal{F}$ можна вважати борелівською (тобто такою, що містить усі відкр., замкнені множини, усі інші $\leq$ зліч. перетини та об'єднання). Далі ми, Як правило, не будемо згадувати $\mathcal{F}$ явно, просто говорячи про $\mathcal{D}(A)$ для $A \subset X$.

– задана ймовірнісна міра $\mathcal{D} : \mathcal{F} \to [0,1]$ з властивостями:

- $\mathcal{D}(X) = 1$
- зліченної адитивності: $\forall$ $(\leq)$ злічен. набору підмножин $\{A_i\}_{i=1}^{\infty} \subset \mathcal{F}$ (тут і далі замість ∞ може бути скінч. n), що попарно не перетинаються: $A_i \cap A_j = \varnothing$ для $\forall i \neq j$

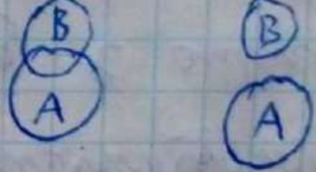
$$\mathcal{D}\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} \mathcal{D}(A_i)$$

При цьому підмн. $A \in \mathcal{F}$ звуть подіями, а $\mathcal{D}(A) \in [0,1]$ – ймовірністю події A. Вона вимірює, з якою ймовірністю випадково обране $x \in X$ належить до A. Інші власт. $\mathcal{D}$:

- монотонність: $A \subset B \Rightarrow \mathcal{D}(A) \leq \mathcal{D}(B)$

(бо $\mathcal{D}(B) = \mathcal{D}(A \cup (B \setminus A)) = \mathcal{D}(A) + \mathcal{D}(B \setminus A) \geq \mathcal{D}(A)$).

· субадитивність: $\mathcal{D}\left(\bigcup_{i=1}^{\infty} A_i\right) \le \sum_{i=1}^{\infty} \mathcal{D}(A_i)$ для $\forall$ набору підмн., де замість ∞ знову може бути n. (це звуть обмеженістю об'єднання – union bound).



$\mathcal{D}(A \cup B) \le \mathcal{D}(A) + \mathcal{D}(B)$

· $\underbrace{\mathcal{D}(\mathcal{X} \setminus A) = 1 - \mathcal{D}(A)}_{}$ $\forall A$, зокр. $\mathcal{D}(\emptyset) = 0$ (чому?). Ex. Для $\mathcal{X} \subset \mathbb{R}^d$ прикладами є однорідний розподіл: $\mathcal{D}(A) := \frac{V(A)}{V(\mathcal{X})}$, де V – об'єм (міра Лебега) на $\mathbb{R}^d$, та нормальний.

Отже, обираємо $x_1, \dots, x_m \in \mathcal{X}$ згідно з $\mathcal{D}$ випадковим чином. Також тут будемо припускати існування "правильної" функції класифікації (labeling function) $f: \mathcal{X} \to \mathcal{Y}$, що визначає за ознаками $x \in \mathcal{X}$ мітку $y = f(x) \in \mathcal{Y}$, і до якої ми повинні "наблизити" (див. нижче) гіпотезу h. Тоді $\forall i$ визначимо $y_i := f(x_i)$ й таким чином побудуємо $S = \{(x_i, y_i = f(x_i))\}_{i=1}^{m}$ (Ex.: смачність однозначно визначена кольором і твердістю).

Т.ч., S однозн. визначено набором $\{x_i\}_{i=1}^{m} \in \mathcal{X}^m$, де $x_i \in \mathcal{X}$ незалежно та ідентично (однаково) розподілені (independently and identically distributed, i.i.d.):

кожен з них обирається з $\mathcal{X}$ незалежно від інших згідно з $\mathcal{D}$. Тобто на $\mathcal{X}^m$ розглядаємо добуток ймовірнісних мір $\mathcal{D}^m := \underbrace{\mathcal{D}\times\ldots\times\mathcal{D}}_{m}$. Ця міра задана на $\mathcal{F}^m$ (найменшій σ-алг. на $\mathcal{X}^m$, що містить усі $A_1\times\ldots\times A_m$, де $A_i\in\mathcal{F}\ \forall i$) й однозначно визначена умовою $\mathcal{D}^m(A_1\times\ldots\times A_m) = \mathcal{D}(A_1)\cdot\ldots\cdot\mathcal{D}(A_m)$. Далі ми будемо за її допомогою вимірювати множини, що складаються з вибірок S (строго кажучи, $S\in(\mathcal{X}\times\mathcal{Y})^m$, але оскільки в $S=\{(x_i,y_i)\}_{i=1}^m$ $y_i=f(x_i)$, тому бієктивно відповідає $\{x_i\}_{i=1}^m\in\mathcal{X}^m$; множину цих наборів ми й вимірюємо).

Як ми виміряємо успіх, тобто "якість" гіпотези h?

Тут ми маємо "правильну" $f:\mathcal{X}\to\mathcal{Y}$, з якою потрібно порівняти h: чим вони "далі", тим гірше.

def. Помилкою гіпотези h (справжньою помилкою,

помилкою узагальнення, (справжнім) ризиком) ((true) error, generalization error, (true) risk) зветься

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

тобто ймовірність того, що випадково обрана точка x буде мати неправильну мітку $h(x)$ від h (на відміну від правильної $f(x)$).

<u>Rem.</u> $L_{\mathcal{D},f}(h)$ дозволяє оцінити "успішність" h, але не може бути використана при її знаходженні, бо залежить від $\mathcal{D}$ і f, а алгоритм A їх не "знає", лише S. Але він може обчислити наступне:

<u>def.</u> <u>Помилкою навчання</u> гіпотези h на навчальних даних $S = \{(x_i, y_i)\}_{i=1}^m$ (емпіричною помилкою, емпіричним ризиком) (training error, empirical error, emp. risk) зветься

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

(де $|\cdot|$ означає кільк. ел-тів скінченної множини), тобто частка точок $\overbrace{\text{з}}^{(x_i, y_i)}$ навч. даних S, для яких h повертає неправильну (у порівнянні з відомою y_i) мітку $h(x_i)$.

Тепер ми можемо визначити основний елемент (навчальну парадигму) алгоритма:

<u>def.</u> Говорять, що алгоритм маш. навч. A здійснює <u>мінімізацію емпіричного ризику</u> (empirical risk minimization, ERM), якщо за навчальними даними S він повертає гіпотезу h з найменшим значенням $L_S(h)$:

$$A(S) = h \in \operatorname*{argmin}_h L_S(h) := \{h : X \to Y \mid L_S(h) \leq L_S(h') \ \forall h' : X \to Y\}$$

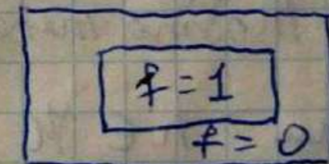
<u>Тобто ми обираємо h, що найкраще описує S.</u>

<u>Rem.</u> Цей (дуже природній) підхід крім складностей у реалізації (треба перебрати усі відображення $h : X \to Y$) має суттєвий принциповий недолік: він може давати h з дуже високим $L_{\mathcal{D},f}(h)$. Ця традиційна для ML проблема зветься <u>перенавчання</u> (overfitting).

<u>Ex.</u> Для $Y = \{0, 1\}$ розглянемо гіпотезу

$$h_S(x) := \begin{cases} y_i, & \exists\, i=\overline{1,m}: x = x_i; \\ 0 & \text{у іншому випадку}. \end{cases} \qquad (*)$$

За побудовою $L_S(h_S) = 0$ $\forall$ $S = \{(x_i, y_i)\}_{i=1}^{m}$. Але якщо при цьому X нескінченна, а $\mathcal{D}$ такий, що скінч. множини мають міру 0, то $L_{\mathcal{D},f}(h_S) = \mathcal{D}(\{x \mid f(x) = 1\})$, бо h_S приписує 0 майже усім точкам (міри 1). Це може бути дуже високе значення. Наприклад, у Ex. з фруктами, однорідним $\mathcal{D}$ і f, що задається умовою



, де площа внутр. пр-кутника у 2 рази менша за площу зовнішнього, $L_{\mathcal{D},f}(h_S) = \frac{1}{2}$ $\forall$ S, тобто h_S "нічого не знає" про $\mathcal{D}$ і f – бо надто добре підходить до S.

<u>Стандартний розв'язок проблеми перенавчання – ERM з індуктивною упередженістю</u>

Ідея: попередньо (до отримання S) фіксуємо <u>клас гіпотез</u>, тобто підмножину $\mathcal{H}$ у множині усіх $h: X \to Y$.

Def. Говорять, що алгоритм маш. навч. A здійснює ERM з індуктивною упередженістю (inductive bias), в напрямку класу гіпотез $\mathcal{H}$, якщо за навчальними даними S він повертає гіпотезу h з найменшим значенням $L_S(h)$ серед $h \in \mathcal{H}$:

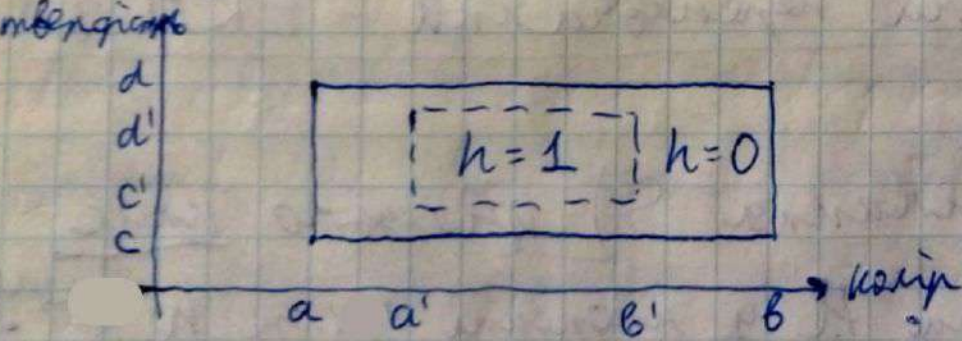
$$A(S) = h \in \underset{h \in \mathcal{H}}{\operatorname{argmin}} L_S(h) := \{ h \in \mathcal{H} \mid L_S(h) \le L_S(h') \ \forall h' \in \mathcal{H} \}$$

Rem. Тобто тепер треба найкращим чином описати S, використавши $h \in \mathcal{H}$. Далі цей клас алгоритмів позначатимемо $ERM_{\mathcal{H}}$:

$h_S \in ERM_{\mathcal{H}}(S)$ озн. $\exists A \in ERM_{\mathcal{H}} : h_S = A(S)$.

Вибір $\mathcal{H}$ залежить від конкретної задачі ML.

Ex. 1. В задачі про смачність фруктів природно очікувати,



що смачними будуть такі, твердість і колір яких – з певних проміжків, тобто $\mathcal{H}$ – це клас ф-цій, що визначені прямокутниками зі сторонами, що паралельні осям координат:

$$h(x) = \begin{cases} 1, & x \in [a', b'] \times [c', d']; \\ 0, & x \in [a, b] \times [c, d] \setminus [a', b'] \times [c', d'] \end{cases}, \text{ де}$$

$[a', b'] \subset [a, b]$, $[c', d'] \subset [c, d]$. Це дійсно "гарний" клас (див. нижче).

<u>Ex. 2.</u> Втім, не усі класи є корисними. Нехай $\mathcal{H}$ – клас т.зв. порогових (thresholded) поліномів від координат точок $X \subset \mathbb{R}^d$:

$$h(x) = \begin{cases} 1, & P(x) \geq 0; \\ 0, & P(x) < 0, \end{cases}$$

де P – поліном.

<u>Впр.</u> Побудувати за $S = \{(x_i, y_i)\}_{i=1}^m$ поліном P_S такий, що відповідна гіпотеза h_S – та ж, що у $(*)$.

Це означає, що вибір класу $\mathcal{H}$ призведе до перенавчання.

<u>Ex. 3</u> Нехай клас $\mathcal{H}$ скінченний: $|\mathcal{H}| < \infty$. На практиці усі вони є такими, бо алгоритм, що генерує h, повинен

закінчити роботу за скінч. час, є дискретизація (наприклад, $\mathcal{H}$ з <u>Ex. 1.</u> перетворюється на скінч., якщо брати a', b', c', d' зі стандартних типів дійсних чисел у певній мові програмування) etc. Втім, $|\mathcal{H}|$ може бути дуже великим.

Показано, що будь-який скінч. $\mathcal{H}$ – "гарний" у певному сенсі. Для цього, щоправда, знадобиться ще одне припущення:

<u>def.</u> Говорять, що клас гіпотез $\mathcal{H}$ задовольняє <u>припущення реалізовності</u> (realizability assumption), якщо існує $h^* \in \mathcal{H}$ така, що $L_{\mathcal{D},f}(h^*) = 0$.

<u>Rem.</u> Це й означає, що цей $\mathcal{H}$ добре підходить для даної задачі. Отже, $f(x) = h^*(x)$ для x з тої множини X міри 1, тобто для "майже кожного" $x \in X$: це так зі ймовірністю 1. Зокрема, для "майже кожної" вибірки S, тобто зі ймовірністю 1, $L_S(h^*) = 0$, бо $h^*(x_i) = f(x_i) = y_i \ \forall (x_i, y_i) \in S$. Це означає, що для $\forall h_S \in ERM_{\mathcal{H}}(S)$, тобто для $\forall$ результата

роботи алгоритма з $ERM_{\mathcal{H}}$ на цій вибірці, $L_S(h_S)=0$ (де необов'язково $h_S = h^*$: важливо, що найменша помилка $=0$).
Але для нас важливи справжня помилка $L_{\mathcal{D},f}(h_S)$. Нехай $\varepsilon \in (0,1)$. Яка "кількість" вибірок S є поганою, тобто $L_{\mathcal{D},f}(h_S) > \varepsilon$? ($\varepsilon$ тут зветься <u>параметром точності</u> (accuracy))
Такі є завжди (наприклад, може просто m раз повторитися одна й та сама пара (x,y)). Відкинувши підмножину S міри 0, можемо вважати, що для таких S $L_S(h_S)=0$, але h_S погана, тобто належить до підмножини

$$\mathcal{H}_B := \{h \in \mathcal{H} \mid L_{\mathcal{D},f}(h) > \varepsilon\}.$$

П.ч. (крім підмножини міри 0, можливо)

$$\{S \mid \exists h_S \in ERM_{\mathcal{H}}(S) : L_{\mathcal{D},f}(h_S) > \varepsilon\} \subset \{S \mid \exists h \in \mathcal{H}_B : L_S(h)=0\}$$

Тому в силу монотонності

$$\mathcal{D}^m(\{S \mid \exists h_S \in ERM_{\mathcal{H}}(S) : L_{\mathcal{D},f}(h_S) > \varepsilon\}) \leq \mathcal{D}^m(\{S \mid \exists h \in \mathcal{H}_B : L_S(h)=0\})$$

Представимо цю другу множину як об'єднання

$\bigcup_{h\in\mathcal{H}_B}\{S \mid L_S(h)=0\}$ за усіма ел-тами $\mathcal{H}_B$:

$\ominus\ \mathcal{D}^m\left(\bigcup_{h\in\mathcal{H}_B}\{S \mid L_S(h)=0\}\right) \leq [\text{субадитивність } \mathcal{D}^m] \leq \sum_{h\in\mathcal{H}_B}$

$\mathcal{D}^m(\{S \mid L_S(h)=0\}) \ominus$

Нагадаємо, що суми тут скінченні. $L_S(h)=0$ для $S=\{(x_i,y_i)\}_{i=1}^m$ означає, що $h(x_i)=\underbrace{y_i}_{=f(x_i)}\ \forall i=\overline{1,m}$, тобто це декартовий добуток множин: $\{S \mid L_S(h)=0\} = \{x_1\in\mathcal{X} \mid h(x_1)=f(x_1)\}\times\ldots\times\{x_m \mid h(x_m)=f(x_m)\}$

$\ominus \sum_{h\in\mathcal{H}_B}\mathcal{D}^m\left(\prod_{i=1}^m\{x_i\in\mathcal{X} \mid h(x_i)=f(x_i)\}\right) = [\text{власт. } \mathcal{D}^m] = \sum_{h\in\mathcal{H}_B}\prod_{i=1}^m \mathcal{D}$

$(\{x_i\in\mathcal{X} \mid h(x_i)=f(x_i)\}) \leq$

Але за означенням $\mathcal{H}_B$ кожна з цих ймовірностей не перевищує $1-\varepsilon$:

$\leq \sum_{h\in\mathcal{H}_B}\prod_{i=1}^m(1-\varepsilon) = |\mathcal{H}_B|(1-\varepsilon)^m \leq |\mathcal{H}|(1-\varepsilon)^m <$

Оск. для $\varepsilon\in(0,1)$ $e^{-\varepsilon} = 1-\varepsilon+\overbrace{\frac{1}{2}\varepsilon^2-\frac{1}{6}\varepsilon^3}^{>0}+\ldots^{>0} > 1-\varepsilon$,

$< |\mathcal{H}| e^{-\varepsilon m} \le$

Нехай тепер m достатньо велике (тобто навч. дані S містять дост. велику кількість точок), а саме $m \geq \frac{1}{\varepsilon} \ln \frac{|\mathcal{H}|}{\delta}$ для деякого $\delta \in (0,1)$. Тоді

$\le |\mathcal{H}| e^{-\ln \frac{|\mathcal{H}|}{\delta}} = \delta$,

тобто ще раз:

$$\mathcal{D}^m \{ S \mid \exists h_S \in ERM_{\mathcal{H}}(S) : L_{\mathcal{D},f}(h_S) > \varepsilon \} < \delta \quad -$$

ймовірність того, що за випадково обраною навчальною вибіркою S алгоритм $ERM_{\mathcal{H}}$ видасть поганий результат h_S, менша за δ, отже зі ймовірністю $\geq (1-\delta)$ (що зветься параметром впевненості (довіри, confidence parameter) цей результат буде добрим, тобто $L_{\mathcal{D},f}(h_S) \le \varepsilon$. При цьому оцінка на m не залежить від $\mathcal{D}, f$, лише від $\mathcal{H}$, ε та δ (та від припущення про реалізовність). Т.ч., ми довели загальний результат:

Сен. Нехай $\mathcal{H}$ - скінченний клас гіпотез у задачі ML з довільним $\mathcal{X}$ і $\mathcal{Y} = \{0, 1\}$, $\varepsilon, \delta \in (0,1)$ і $m \in \mathbb{N}$ таке, що $m \geq \frac{1}{\varepsilon} \ln \frac{|\mathcal{H}|}{\delta}$. Тоді для будь-яких розподілу ймовірностей $\mathcal{D}$ на $\mathcal{X}$ і $f: \mathcal{X} \to \mathcal{Y}$ таких, що $\mathcal{H}$ задовольняє припущенню реалізовності ($L_{\mathcal{D},f}(h^*) = 0$ для деякої $h^* \in \mathcal{H}$) зі ймовірністю не меншою за $1-\delta$ алгоритм $ERM_{\mathcal{H}}$ за вибіркою S з m ел-тів $\mathcal{X}$, що є i.i.d. за $\mathcal{D}$, поверне гіпотезу h_S, для якої $L_{\mathcal{D},f}(h_S) \leq \varepsilon$:

$$\mathcal{D}^m(\{S \mid \forall h_S \in ERM_{\mathcal{H}}(S)\ L_{\mathcal{D},f}(h_S) \leq \varepsilon\}) \geq 1-\delta.$$

PAC навчання

Попередній розділ мотивує наступне загальне означення:

Озн. Для задачі ML з довільною областю визначення $\mathcal{X}$ і множиною міток $\mathcal{Y} = \{0,1\}$ клас гіпотез $\mathcal{H} \subset \{h: \mathcal{X} \to \mathcal{Y}\}$ зветься <u>ймовірно приблизно коректно навчаним</u> (probably approximately correctly learnable, PAC learnable), якщо існує функція $m_{\mathcal{H}} : (0,1)^2 \to \mathbb{N}$ і алгоритм ML, що для будь-яких $\varepsilon, \delta \in (0,1)$, розподілу ймовірностей $\mathcal{D}$ на $\mathcal{X}$ та функції класифікації $f : \mathcal{X} \to \mathcal{Y}$ таких, що $\mathcal{H}$ задовольняє припущенню реалізовності ($L_{\mathcal{D},f}(h^*) = 0$ для деякої $h^* \in \mathcal{H}$), за навчальною вибіркою S з $m \geq m_{\mathcal{H}}(\varepsilon, \delta)$ елементів, що незалежно та ідентично розподілені відповідно до $\mathcal{D}$ та помічені за допомогою f, зі ймовірністю не меншою за параметр впевненості $1-\delta$ повертає гіпотезу $h_S \in \mathcal{H}$, для якої справжня помилка $L_{\mathcal{D},f}(h_S)$ не більша за параметр точності ε: (approximately correct) (probably)

$$\mathcal{D}^m \{ S \mid L_{\mathcal{D},f}(h_S) \leq \varepsilon \} \geq 1-\delta.$$

Rem.

Тут "ймовірно" відноситься до $\geq 1-\delta$, а "приблизно коректно" – до $\leq \varepsilon$.

Тепер можна переписати попередній наслідок:

<u>Cor.</u> Будь-який скінченний клас гіпотез $\mathcal{H}$ для будь-якої задачі ML є PAC навчанним, де $m_{\mathcal{H}}(\varepsilon, \delta) \leq \leq \left\lceil \frac{1}{\varepsilon} \ln \frac{|\mathcal{H}|}{\delta} \right\rceil$, а алгоритм – з $ERM_{\mathcal{H}}$.

Впр. Показати, що клас $\mathcal{H}$ гіпотез, що визначені прямокутниками (див. Ex.1. вище) $\in$ PAC навчаним для задачі "класифікації фруктів", тобто з $X = [a,b]\times[c,d] \subset \mathbb{R}^2$, причому $m_{\mathcal{H}}(\varepsilon,\delta) \le \left\lceil \frac{4}{\varepsilon} \ln\frac{4}{\delta} \right\rceil$. Узагальнити це на довільну вимірність d.

Rem. Ф-ція $m_{\mathcal{H}}$ це зветься складністю вибірки (sample complexity). Її часто визначають більш точно як найменше натуральне число, для якого виконана умова з def. PAC навчаності, саме тому вище з'явилися нерівності $\le$ та округлення $\lceil\cdot\rceil$ (перехід до найближчого справа натурального числа).

Недоліки цієї моделі:

1. Обмеженість лише бінарною класифікацією: було б добре розглядати задачі й для інших Y, зокрема, більших скінченних або $\mathbb{R}$. Але тоді потрібно

буде змінити й означення помилки.

2. Існування "правильного" класифікатора $f: X \to Y$: в реальності мітка y не обов'язково однозн. визначена точкою x, наприклад, два фрукти однакового кольору і твердості можуть бути смачним і несмачним через інші фактори. Але озн. помилки залежить і від f.

3. Припущення реалізовності: теж часто насправді невірне: наприклад, у випадку прямокутників завжди є невелика, але не міри 0 частка точок з міткою 1 за межами $\forall$ пр-ка і з 0 – всередині. Це ще не означає, що цей клас гіпотез H – поганий.

Щоб прибрати ці недоліки, потрібно змінити модель у частині побудови S (залишивши вимогу статистичності та i.i.d.) та формулу обчислення справжньої помилки: