Машинное обучение. Bias-complexity tradeoff. VC-размерность

Алексей Колесов

Белорусский государственный университет

18 сентября 2019 г.

Краткое содержание предыдущих лекций

- **probably approximately correct learning** — с наперёд заданной (probably) вероятностью найдётся (approximately correct) гипотеза с разумной ошибкой
- **uniform convergence** — true и empirical risk для любой гипотезы отличается несильно
- конечный класс гипотез является PAC-изучаемым с помощью ERM-алгоритма

Вопросы

- есть ли универсальный алгоритм обучения?
- бывают ли бесконечные PAC-learnable классы?
- какие классы гипотез PAC-learnable?
- как оценить выборочную сложность класса гипотез?

Содержание

Вопрос о существовании универсального алгоритма обучения

- во избежание проблемы *переобучения* можно использовать ограничение класса гипотез H
- нужно ли?
- есть ли алгоритм и размер выборки m, что для любого D алгоритм найдёт хорошую гипотезу (относительно δ и ϵ)?

No free lunch theorem

No free lunch theorem

Пусть A — любой алгоритм машинного обучения для задачи бинарной классификации и 0-1 функции потерь над пространством X. Пусть m — число, меньшее, чем $|X|/2$. Тогда при размере выборки m будет существовать распределение D, такое что:

- найдётся функция $f : X \to \{0, 1\}$, такая что $L_D(f) = 0$
- с вероятностью не меньшей $\frac{1}{7}$ выполняется, что $L_D(A(S)) \geqslant \frac{1}{8}$

Доказательство No free lunch theorem: обозначения

- пусть $C \subseteq X$, $|C| = 2m$.
- $f_1, \ldots, f_T$ — функции из C в $\{0, 1\}$
- $T = 2^{2m}$
- $D_i(x, y) = \begin{cases} 1/|C| & \text{если } y = f_i(x) \\ 0 & \text{иначе} \end{cases}$
- $L_{D_i}(f_i) = 0$

Доказательство No free lunch theorem: план

Докажем, что любого алгоритма A:

$$\max_{i \in [T]} \mathop{\mathbb{E}}_{S \sim D_i^m} [L_{D_i}(A(S))] \geqslant 1/4$$

Это означает, что для любого алгоритма A, который принимает выборку размера m из $X \times \{0, 1\}$ найдётся размечающая функция f и распределение D над $X \times \{0, 1\}$, такое что, хоть $L_D(f) = 0$, но

$$\mathop{\mathbb{E}}_{S \sim D^m} [L_D(A(S))] \geqslant 1/4$$

Из этого:

$$P[L_D(A(S)) \geqslant 1/8] \geqslant 1/7$$

Доказательство No free lunch theorem: обозначения

Хотим:

$$\max_{i \in [T]} \mathop{\mathbb{E}}_{S \sim D_i^m} [L_{D_i}(A(S))] \geqslant 1/4$$

Обозначим:

- $k = (2m)^m$ — количество выборок размера m из C $(S_1, \ldots, S_k)$.
- S_j — j-я из выборок $(S_j = (x_1, \ldots, x_m))$
- S_j^i — j-я выборка, размеченная функцией f_i $(S_j^i = ((x_1, f_i(x_1)), \ldots, (x_m, f_i(x_m))))$

Доказательство No free lunch theorem

Хотим:

$$\max_{i \in [T]} \mathop{\mathbb{E}}_{S \sim D_i^m} [L_{D_i}(A(S))] \geqslant 1/4$$

По определению:

$$\mathop{\mathbb{E}}_{S \sim D_i^m} [L_{D_i}(A(S)) = \frac{1}{k} \sum_{j=1}^{k} L_{D_i}(A(S_j^i))$$

Доказательство No free lunch theorem

Имеем:

$$\mathop{\mathbb{E}}_{S\sim D_i^m}[L_{D_i}(A(S)) = \frac{1}{k}\sum_{j=1}^{k} L_{D_i}(A(S_j^i))$$

Распишем:

$$\max_{i\in[T]} \frac{1}{k}\sum_{j=1}^{k} L_{D_i}(A(S_j^i)) \geqslant \frac{1}{T}\sum_{i=1}^{T}\frac{1}{k}\sum_{j=1}^{k} L_{D_i}(A(S_j^i)) \tag{1}$$

$$= \frac{1}{k}\sum_{j=1}^{k}\frac{1}{T}\sum_{i=1}^{T} L_{D_i}(A(S_j^i)) \tag{2}$$

$$\geqslant \min_{j\in[k]} \frac{1}{T}\sum_{i=1}^{T} L_{D_i}(A(S_j^i)) \tag{3}$$

Доказательство No free lunch theorem: обозначения 2

Исследуем:

$$\min_{j\in[k]} \frac{1}{T}\sum_{i=1}^{T} L_{D_i}(A(S_j^i))$$

Зафиксируем $j \in [k]$ **и обозначим**:

- $S_j = (x_1, \ldots, x_m)$
- $v_1, \ldots, v_p$ — объекты из C, которые не встречаются в S_j

Заметим:

$$L_{D_i}(h) = \frac{1}{2m}\sum_{x\in C} 1_{[h(x)\neq f_i(x)]} \quad (4)$$

$$\geqslant \frac{1}{2m}\sum_{r=1}^{p} 1_{[h(v_r)\neq f_i(v_r)]} \quad (5)$$

$$\geqslant \frac{1}{2p}\sum_{r=1}^{p} 1_{[h(v_r)\neq f_i(v_r)]} \quad (6)$$

Доказательство No free lunch theorem

Исследуем:

$$\min_{j\in[k]} \frac{1}{T}\sum_{i=1}^{T} L_{D_i}(A(S_j^i))$$

Распишем:

$$\frac{1}{T}\sum_{i=1}^{T} L_{D_i}(A(S_j^i)) \geqslant \frac{1}{T}\sum_{i=1}^{T}\frac{1}{2p}\sum_{r=1}^{p} 1_{[A(S_j^i)(v_r)\neq f_i(v_r)]} \tag{7}$$

$$= \frac{1}{2p}\sum_{r=1}^{p}\frac{1}{T}\sum_{i=1}^{T} 1_{[A(S_j^i)(v_r)\neq f_i(v_r)]} \tag{8}$$

$$= \frac{1}{2}\cdot\min_{r\in[p]}\frac{1}{T}\sum_{i=1}^{T} 1_{[A(S_j^i)(v_r)\neq f_i(v_r)]} \tag{9}$$

Доказательство No free lunch theorem

Можно показать, что:

$$\frac{1}{T}\sum_{i=1}^{T} 1_{[A(S_j^i)(v_r)\neq f_i(v_r)]} = \frac{1}{2}$$

А значит:

$$\max_{i\in[T]} \mathop{\mathbb{E}}_{S\sim D_i^m}[L_{D_i}(A(S))] \geqslant \min_{j\in[k]} \frac{1}{T}\sum_{i=1}^{T} L_{D_i}(A(S_j^i)) \tag{10}$$

$$\geqslant \frac{1}{2}\cdot \min_{r\in[p]} \frac{1}{T}\sum_{i=1}^{T} 1_{[A(S_j^i)(v_r)\neq f_i(v_r)]} \tag{11}$$

$$\geqslant \frac{1}{2}\cdot\frac{1}{2} = \frac{1}{4} \tag{12}$$

Априорное знание

- для успеха обучения ограничиваем класс гипотез
- пусть H — все функции из X в Y — полное отсутствие априорного знания
- из No FLT — **любой** алгоритм будет ошибаться с таким **классом**

No FLT для универсума функций

Пусть X — бесконечный домен и H — множество всех функций из X в $\{0, 1\}$. Тогда H не является PAC-learnable классом

Как всё-таки учиться?

- предположения о D
- ограничение H
- как выбрать H?

Bias-Complexity tradeoff

$$L_D(h_S) = \epsilon_{\text{app}} + \epsilon_{\text{est}} + \epsilon_{\text{bayes}}$$

- ϵ_{bayes} — ошибка оптимального байесовского классификатора
- $\epsilon_{\text{app}} = \min\limits_{h \in H} L_D(h) - \epsilon_{\text{bayes}}$ — ошибка аппроксимации (насколько H подходит задаче)
- $\epsilon_{\text{est}} = L_D(h_S) - \min\limits_{h \in H} L_D(h)$ — упущенное качество на данном H (насколько хорошо решили задачу при данном H)

Bias-Complexity tradeoff

- чем богаче класс, тем выше ошибка $\epsilon_{\text{est}} \Rightarrow$ переобучение
- чем беднее, тем выше $\epsilon_{\text{app}} \Rightarrow$ недообучение
- где остановиться?

Итоги

- нет алгоритма, который работает всегда
- необходимо использование априорного знания
- **можно** ограничивать H
- тогда нужно решать bias-complexity tradeoff

Вопросы на понимание

Как согласуются:

- ERM-алгоритм над конечным классом H — PAC-learnable в случае гипотезы реализуемости и No FLT?
- ERM-алгоритм над конечным классом H — agnostic PAC-learnable и No FLT?

Содержание

Вопросы

- какие классы PAC-learnable?
- конечные — да (см. 1-ю лекцию)
- класс всех функций — нет
- бывают ли бесконечные PAC-learnable классы?
- что влияет на PAC-learnability?
- как оценить выборочную сложность?

План

- покажем, что бывают бесконечные PAC-learnable классы
- введём VC-размерность — характеристику всех learnable-классов
- приведём примеры вычисления VC-размерности
- докажем связь VC-размерности и PAC-learnability

Бесконечные класс могут быть PAC-learnable

Зададим семейство пороговых (threshold) функций:

$$H = \{h_\alpha : \alpha \in \mathbb{R}\}, \text{ где } h_\alpha(x) = 1_{[x<\alpha]}$$

Пример бесконечного PAC-learnable класса

$H = \{h_\alpha : \alpha \in \mathbb{R}\}$ является PAC-learnable с ERM-алгоритмом, причём выборочная сложность $m_H(\epsilon, \delta) \leqslant \lceil log(2/\delta)/\epsilon \rceil$

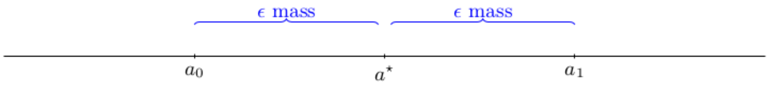
Доказательство

Пусть α^* — такая, что:

$$L_D(h_{\alpha^*}) = 0$$

Найдём α_0 и α_1, такие что:

$$\mathop{\mathbb{P}}_{x\sim D}[x \in (\alpha_0, \alpha^*)] = \mathop{\mathbb{P}}_{x\sim D}[x \in (\alpha^*, \alpha_1)] = \epsilon$$



Кроме того, пусть

$$b_0 = \max\{x : (x, 1) \in S\}$$
$$b_1 = \min\{x : (x, 0) \in S\}$$

Доказательство

Имеем (b_S — ERM-гипотеза):

$$b_S \in (b_0, b_1)$$
$$b_0 \geqslant \alpha_0 \text{ и } b_1 \leqslant \alpha_1 \Rightarrow L_D(h_S) \leqslant \epsilon$$

Значит:

$$\mathbb{P}_{S\sim D^m}[L_D(h_s) > \epsilon] \leqslant \mathbb{P}_{S\sim D^m}[b_0 < \alpha_0 \vee b_1 > \alpha_1]$$
$$\mathbb{P}_{S\sim D^m}[L_D(h_s) > \epsilon] \leqslant \mathbb{P}_{S\sim D^m}[b_0 < \alpha_0] + \mathbb{P}_{S\sim D^m}[b_1 > \alpha_1]$$

$b_0 < \alpha_0$ значит, что все объекты не попали в (α_0, α^*)

Доказательство

С какой вероятностью все объекты выборки не попали в (α_0, α^*)?

$$\mathbb{P}_{S\sim D^m}[b_0 < \alpha_0] = \mathbb{P}_{S\sim D^m}[\forall (x, y) \in S, x \notin (\alpha_0, \alpha^*)] = (1 - \epsilon)^m \leqslant e^{-\epsilon m}$$

Мотивация

- конечность $|H|$ — лишь **достаточное** условие для PAC-изучаемости
- для упрощения рассуждений будем предполагать гипотезу реализуемости
- если класс гипотез не ограничен $\Rightarrow$ No FLT (всегда можно выбрать плохую f)
- в PAC-изучаемом сценарии можно выбирать лишь из $h : L_D(h) = 0, \ h \in H$!

Ограничение класса гипотез на множество

Ограничение класса гипотез на множество

Пусть H — семейство функций из X в $\{0,1\}$ и $C = (c_1, \dots, c_m) \subset X$. Ограничением H на C называется семейство функций из C в $\{0,1\}$, заданное таким образом:

$$H_C = \{(h(c_1), \dots, h(c_m)) : h \in H\}$$

где каждая функция представляется как вектор из значений на каждом объекте

«Разукрашивание» множества

«Разукрашивание» (shattering) множества

Семейство гипотез H «разукрашивает» (shatters, размечает всеми способами) множество C, если H_C состоит из всех функций из C в $\{0, 1\}$, т.е. $|H_C| = 2^{|C|}$

Например, семейство пороговых функций:

- разукрашивает $C = \{c_1\}$
- не разукрашивает $C = \{c_1, c_2\}$, где $c_1 < c_2$

Следствие о чрезмерной разукрашиваемости

Следствие о чрезмерной разукрашиваемости

Пусть H — семейство гипотез из X в $\{0,1\}$, m — размер тренировочной выборки. Пусть существует $C \subset X$ размера $2m$, который разукрашиваем с помощью H. Тогда для любого алгоритма A найдётся распределение D над $X \times \{0,1\}$ и функция $h \in H$, такая что $L_D(h) = 0$, там не менее с вероятностью как минимум $\frac{1}{7}$ выполняется $L_D(A(S)) \geqslant \frac{1}{8}$.

VC-размерность

VC-размерность

VC-размерность (размерность Вапника-Червоненкиса) семейства H (обозначается VCdim(H)) — максимальный размер множества $C \subset H$, которое может быть разукрашено с помощью H. Если H может разукрасить произвольно большое множество, то говорят, что VCdim(H) $= \infty$

Следствие о бесконечной VC-размерности

Если VCdim(H) $= \infty$, то H не является PAC-изучаемым

Общий план

Для доказательства, что $\mathrm{VCdim}(H) = d$ нужно:

- доказать, что найдётся C размера d, что его можно разукрасить с помощью H
- доказать, что любое множество размера $d+1$ не может быть разукрашено с помощью H

Пороговые функции

$$H = \{h_\alpha : \alpha \in \mathbb{R}\}, \text{ где } h_\alpha(x) = 1_{[x<\alpha]}$$

- можем разукрасить одну точку $\Rightarrow$ VCdim$(H) \geqslant 1$
- не можем разукрасить две точки $\Rightarrow$ VCdim$(H) < 2$
- VCdim$(H) = 1$

Интервалы

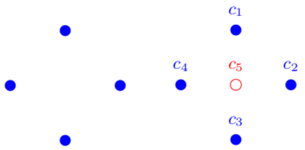
$$H = \{h_{\alpha,\beta} : \alpha, \beta \in \mathbb{R}\}, \text{ где } h_{\alpha,\beta}(x) = 1_{[x \in (\alpha,\beta)]}$$

- можем разукрасить $C = \{1, 2\}$
- не можем разукрасить $C = (c_1, c_2, c_3)$ $(c_1 \leqslant c_2 \leqslant c_3)$
- $\text{VCdim}(H) = 2$

Прямоугольники со сторонами, параллельными осям координат

$$h_{a_1,b_1,a_2,b_2}((x_1,x_2)) = \begin{cases} 1 & \text{если } a_1 \leqslant x_1 \leqslant b_1 \text{ и } a_2 \leqslant x_2 \leqslant b_2 \\ 0 & \text{иначе} \end{cases}$$

- можем разукрасить картинку с четырьмя точками
- не можем разукрасить с пятью
- $\mathrm{VCdim}(H) = 4$



Конечные классы гипотез

- $|H_C| \leqslant |H| \Rightarrow C$ не может быть разукрашиваемым, если $|H| < 2^{|C|}$
- $\mathrm{VCdim}(H) \leqslant \log_2 |H|$
- может быть намного меньше

VCdim и количество параметров

- во всех примерах количество параметров совпадало с VCdim
- $H = \{h_\theta(x) = \lceil 0.5\sin(\theta x)\rceil : \theta \in \mathbb{R}\}$ задаётся одним параметром, но $\text{VCdim}(h) = \infty$ (см. домашнее задание)

Фундаментальная теорема PAC-изучаемости

Фундаментальная теорема PAC-изучаемости

Пусть H — семейство гипотез из X в $\{0, 1\}$ и мы используем $0 - 1$ функцию потерь. Тогда следующие утверждения эквивалентны:

1. H обладает свойством равномерной сходимости
2. H агностически PAC-изучаемый с ERM-алгоритмом
3. H агностически PAC-изучаемый
4. H PAC-изучаемый
5. H PAC-изучаемый с ERM-алгоритмом
6. $\mathrm{VCdim}(H) < \infty$

Фундаментальная теорема PAC-изучаемости

Фундаментальная теорема PAC-изучаемости

Пусть H — семейство гипотез из X в $\{0, 1\}$ и мы используем $0-1$ функцию потерь. Кроме того, пусть $\mathrm{VCdim}(H) = d < \infty$. Тогда существуют константы C_1, C_2, такие что:

1. H обладает свойством равномерной сходимости, причём:
$$C_1 \frac{d + \log(1/\delta)}{\epsilon^2} \leqslant m_H^{UC} \leqslant C_2 \frac{d + \log(1/\delta)}{\epsilon^2}$$
2. H агностически PAC-изучаемый, причём:
$$C_1 \frac{d + \log(1/\delta)}{\epsilon^2} \leqslant m_H \leqslant C_2 \frac{d + \log(1/\delta)}{\epsilon^2}$$
3. H PAC-изучаемый, причём:
$$C_1 \frac{d + \log(1/\delta)}{\epsilon} \leqslant m_H \leqslant C_2 \frac{d \log 1/\epsilon + \log(1/\delta)}{\epsilon^2}$$

План доказательства

- докажем, что если $\mathrm{VCdim}(h) < \infty$, то размер ограничения на произвольное C невелико, а именно, что $|H_C|$ растёт полиномиально с $|C|$ (Sauer's lemma, лемма Саура)
- покажем, что если $|H_C|$ растёт полиномиально с $|C|$, то класс H обладает свойством равномерной сходимости

Функция роста

Функция роста

Пусть H — семейство гипотез. Тогда τ_H (функция из $\mathbb{N}$ в $\mathbb{N}$) называется **функцией роста** (growth function) и определяется:

$$\tau_H(m) = \max_{C \subset X : |C| = m} |H_C|$$

- $\mathrm{VCdim}(H) = d < \infty \Rightarrow \tau_H(m) = 2^m$ при $m \leqslant d$
- что, если $m > d$?

Лемма Саура

Пусть H — семейство гипотез и $\mathrm{VCdim}(H) \leqslant d < \infty$. Тогда для всех m, $\tau_H(m) \leqslant \sum_{i=0}^{d} C_m^i$.

Например, при $m > d + 1$ выполняется $\tau_H(m) \leqslant (em/d)^d$

Доказательство леммы Саура

Хотим:

$$\tau_H(m) \leqslant \sum_{i=0}^{d} C_m^i$$

Докажем, что для любого $C = (c_1, \ldots, c_m)$:

$$\forall H,\ |H_C| \leqslant |\{B \subseteq C : H \text{ разукрашивает } B\}|$$

Действительно:

$$|\{B \subseteq C : H \text{ разукрашивает } B\}| \leqslant \sum_{i=0}^{d} C_m^i$$

Доказательство леммы Саура

Хотим доказать, что для любого $C = (c_1, \ldots, c_m)$:

$$\forall H,\ |H_C| \leqslant |\{B \subseteq C : H \text{ разукрашивает } B\}|$$

Доказательство по индукции. База:

При $m = 1$ обе части равны 1 или 2

Доказательство леммы Саура

Имеем, что для $k < m$ для любого C, такого что $|C| = k$:

$$\forall H, |H_C| \leqslant |\{B \subseteq C : H \text{ разукрашивает } B\}|$$

Докажем, что для $k = m$ для любого C, выполняется что $|C| = k$:

- зафиксируем H и $C = \{c_1, \ldots, c_m\}$
- пусть $C' = \{c_2, \ldots, c_m\}$
- $Y_0 = \{(y_2, \ldots, y_m) : (0, y_2, \ldots, y_m) \in H_C \vee (1, y_2, \ldots, y_m) \in H_C\}$
- $Y_1 = \{(y_2, \ldots, y_m) : (0, y_2, \ldots, y_m) \in H_C \wedge (1, y_2, \ldots, y_m) \in H_C\}$
- $|H_C| = |Y_0| + |Y_1|$

Доказательство леммы Саура

Имеем:

$$Y_0 = \{(y_2, \ldots, y_m) : (0, y_2, \ldots, y_m) \in H_C \vee (1, y_2, \ldots, y_m) \in H_C\}$$

Легко проверить, что $Y_0 = H_{C'}$.
Распишем:

$$|Y_0| = |H_{C'}| \leqslant |\{B \subseteq C' : H \text{ разукрашивает } B\}| = |\{B \subseteq C : c_1 \notin B \wedge H \text{ разукрашивает } B\}|$$

Доказательство леммы Саура

Имеем:

$$Y_1 = \{(y_2, \dots, y_m) : (0, y_2, \dots, y_m) \in H_C \wedge (1, y_2, \dots, y_m) \in H_C\}$$

Введём $H' \subset H$:

$$H' = \{h \in H : \exists h' \in H \text{ т.ч. } (1 - h'(c_1), h'(c_2), \dots, h'(c_m)) = (h(c_1), h(c_2), \dots, h(c_m))\}$$

Имеем:

$$|Y_1| = |H'_{C'}| \leqslant |\{B \subseteq C' : H' \text{ разукрашивает } B\}| = |\{B \subseteq C' : H' \text{ разукрашивает } B \cup \{c_1\}\}| = |\{B \subseteq C : c_1 \in B \wedge H' \text{ разукрашивает } B\}| \leqslant |\{B \subseteq C : c_1 \in B \wedge H \text{ разукрашивает } B\}|$$

Доказательство леммы Саура

Имеем:

- $|H_C| = |Y_0| + |Y_1|$
- $|Y_0| \leqslant |\{B \subseteq C : c_1 \notin B \wedge H \text{ разукрашивает } B\}|$
- $|Y_1| \leqslant |\{B \subseteq C : c_1 \in B \wedge H \text{ разукрашивает } B\}|$

А значит:

$$|H_C| = |Y_0| + |Y_1| \leqslant |\{B \subseteq C : c_1 \notin B \wedge H \text{ разукрашивает } B\}| + |\{B \subseteq C : c_1 \in B \wedge H \text{ разукрашивает } B\}| = |\{B \subseteq C : H \text{ разукрашивает } B\}|$$

Равномерная сходимость для классов с «малой» функцией роста

Равномерная сходимость для классов с «малой» функцией роста

Пусть H — семейство гипотез и τ_H — функция роста. Тогда для любого D и $\delta \in (0,1)$ с вероятностью не меньше $1-\delta$ выполняется:

$$|L_D(h) - L_S(h)| \leqslant \frac{4 + \sqrt{\log\left(\tau_H(2m)\right)}}{\delta\sqrt{2m}}$$

Доказательство

Имеем:

- $m > d \Rightarrow \tau_H(2m) \leqslant (2em/d)^d$
- (с высокой вероятностью)
  $$|L_D(h) - L_S(h)| \leqslant \frac{4 + \sqrt{\log\left(\tau_H(2m)\right)}}{\delta\sqrt{2m}}$$

Получается:

$$|L_D(h) - L_S(h)| \leqslant \frac{4 + \sqrt{d\log\left(2em/d\right)}}{\delta\sqrt{2m}}$$

Содержание

Итоги

- доказали, что нет универсального алгоритма обучения
- произвели декомпозицию ошибку классификатора
- рассмотрели bias-complexity tradeoff
- ввели понятие VC-размерности
- доказали, что для бинарной классификации возможно обучение с помощью ERM-алгоритма и только в случае конечной VCdim

Литература

- Shai Shalev-Shwartz and Shai Ben-David — Understanding Machine Learning: From theory to algorithms (главы 5-6)
- https://en.wikipedia.org/wiki/VC_dimension