Машинное обучение. Неравномерная изучаемость. Выбор модели

Алексей Колесов

Белорусский государственный университет

8 октября 2019 г.

Содержание

Равномерная изучаемость

Класс гипотез H называют **вероятно приблизительно верно изучаемым** (probably approximately correct learnable) если существует такая функция $m_H : (0,1)^2 \to \mathbb{N}$ и алгоритм, такой что

- для любых $\epsilon, \delta \in (0,1)$
- для любого распределения D над X
- для любой функции $f : X \to \{0,1\}$

если выполняется предположение о реализуемости, то если мы выполним алгоритм на выборке из $m \geqslant m_H(\epsilon, \delta)$ независимых одинаково распределённых элементов из D и размеченных f, то алгоритм вернёт гипотезу $h \in H$ такую, что с вероятностью как минимум $1 - \delta$, выполняется $L_{D,f}(h) \leqslant \epsilon$

Ослабление равномерной изучаемости

- H — PAC-изучамемый $\iff$ $\mathrm{VCdim}(H) < \infty$
- можно ли ослабить определение?
- полезно ли ослаблять такое ограничение?

Неравномерная изучаемость

Гипотеза h называется (ϵ, δ)-**конкурентной с гипотезой** h' ((ϵ, δ)-competitive), если $\mathbb{P}[L_D(h) \leqslant L_D(h') + \epsilon] > 1 - \delta$

Класс гипотез H называют **неравномерно изучаемым** (nonuniform learnable) если существует такая функция $m_H^{NUL} : (0,1)^2 \times H \to \mathbb{N}$ и алгоритм, такой что

- для любых $\epsilon, \delta \in (0,1)$
- для любой $h' \in H$
- для любого распределения D над X

если мы выполним A на выборке из $m \geqslant m_H^{NUL}(\epsilon, \delta, h')$ независимых элементов из D, то с вероятностью как минимум $1 - \delta$, выполняется $L_D(A(S)) \leqslant L_D(h') + \epsilon$

Характеризация классов с неравномерной изучаемостью

Критерий неравномерной изучаемости

Класс гипотез H является неравномерно изучаемым, тогда и только тогда, когда H — объединение не более чем счётного множества PAC-изучаемых классов H_i.

Теорема о связи равномерной сходимости и неравномерной изучаемости

Пусть $H = \bigcup_{n \in \mathbb{N}} H_n$, где каждый H_n обладает свойством равномерной сходимости. Тогда H — неравномерно изучаемый

Доказательство критерия

Необходимость:

- имеем $H = \bigcup_{n \in \mathbb{N}} H_n$, H_n — PAC-learnable
- H_n обладает равномерной сходимостью
- чтд (по теореме о связи)

Достаточность:

- имеем H — неравномерно изучаемый
- определим $H_n = \{h \in H : m_H^{NUL}(1/8, 1/7, h) \leqslant n\}$ $(H = \bigcup_{n \in \mathbb{N}} H_n)$
- $\exists h \in H_n$, т.ч. $L_D(h) = 0 \Rightarrow$ при $S \in D^n$ $\mathbb{P}[L_D(A(S)) \leqslant 1/8] > 6/7 \Rightarrow \mathrm{VCdim}(H_n) < \infty$
- чтд (по фундаментальной теореме)

PAC-изучаемость $\neq$ неравномерная изучаемость

- пусть H_n — множество полиномиальных классификаторов степени n (т.е. $h(x)\,\mathrm{sign}(p(x))$, где $p(x)$ — многочлен степени n)
- $H = \bigcup H_n$
- $\mathrm{VCdim}(H_n) = n + 1$
- $\mathrm{VCdim}(H) = \infty$

Inductive bias

- по NoFLT нужен inductive bias (априорная информация)
- пока умели только ограничить класс гипотез
- теперь попробуем «проранжировать» гипотезы

Обозначения и предположения

- пусть $H = \bigcup_{n \in \mathbb{N}} H_n$
- пусть каждый H_n обладает свойством равномерной сходимости
- введём **весовую функцию** $w : \mathbb{N} \to [0, 1]$: $\sum_i w(i) \leqslant 1$
- $\epsilon_n(m, \delta) = \min\{\epsilon \in (0, 1) : m^{UC}_{H_n}(\epsilon, \delta) \leqslant m\}$
- $\forall h \in H_n$, $|L_D(h) - L_S(h)| \leqslant \epsilon(m, \delta)$ с вероятностью не меньше, чем $1 - \delta$, если S из m элементов
- $n(h) = \min\{n : h \in H_n\}$

Structural risk minimization: bound

Теорема о верхней границе для SRM

Если выполняются предположения с предыдущего слайда, то для любого $\delta \in (0,1)$ и распределения D, с вероятностью не меньше $1-\delta$ над $S \sim D^m$ одновременно для всех $n \in \mathbb{N}$ и $h \in H_n$ выполняется:

$$|L_D(h) - L_S(h)| \leqslant \epsilon_n(m, w(n) \cdot \delta)$$

а значит, и

$$L_D(h) \leqslant L_S(h) + \min_{n:h \in H_n} \epsilon_n(m, w(n) \cdot \delta)$$

Structural risk minimization: algorithm

По теореме, выполняется: $L_D(h) \leqslant L_S(h) + \epsilon_{n(h)}(m, w(n(h)) \cdot \delta)$
SRM-алгоритм:

- **prior**: H — объединение счётного множества классов гипотез с равномерными сходимостями
- **prior**: w — весовая функция
- **вход**: $S \sim D^m$, параметр δ
- **выход**: $h \in \operatorname{argmin}_{h \in H}[L_S(h) + \epsilon_{n(h)}(m, w(n(h)) \cdot \delta)]$

Теорема о SRM и неравномерной изучаемости

Теорема о SRM и неравномерной изучаемости

Пусть $H = \bigcup_{n\in\mathbb{N}} H_n$, причём каждый H_n обладает свойством равномерной сходимости с выборочной сложностью $m^{UC}_{H_n}$. Пусть $w(n) = \dfrac{6}{n^2\pi^2}$. Тогда H является неравномерно изучаемым с помощью SRM-алгоритма и имеет выборочную сложность:

$$m^{NUL}_{H}(\epsilon, \delta, h) \leqslant m^{UC}_{H_{n(h)}}\left(\epsilon/2, \frac{6\delta}{(\pi n(h))^2}\right)$$

Замечания

- можем учить другие классы (класс всех полиномов)
- No FLT не отменяется (класс всех функций над бесконечным доменом не является объединением классов с конечной VC-размерностью)
- prior более слабый, чем в PAC $\Rightarrow$ выборочная сложность больше
- если $\mathrm{VCdim}(H_n) = n$, и $h \in H_n$ то $m_H^{NUL}(\epsilon, \delta, h) - m_{H_n}^{UC}(\epsilon/2, \delta) \leqslant 4C\dfrac{2\log(2n)}{\epsilon^2}$

SRM для счётного класса гипотез

- пусть H — счётный класс гипотез; $H = \bigcup_{n\in\mathbb{N}}\{h_n\}$
- у каждого из $H_n = \{h_n\}$ есть равномерная сходимость с
$$m^{UC}(\epsilon,\delta) = \frac{\log(2/\delta)}{2\epsilon^2} \Rightarrow \epsilon_n(m,\delta) = \sqrt{\frac{\log(2/\delta)}{2m}}$$
- SRM: $\operatorname*{argmin}_{h_n\in H}\left[L_S(h) + \sqrt{\frac{-\log(w(h)) + \log(2/\delta)}{2m}}\right]$
- prior зависит только от гипотезы

Обозначения и замечания

- зафиксируем алфавит Σ и класс гипотез H
- пусть $\sigma(h)$ — описание гипотезы h на языке Σ, т.е. вектор длины $|\sigma(h)|$, где каждый элемент из Σ
- будем называть описание **беспрефиксным**, если для любых разных h, h' выполняется, что $\sigma(h)$ не является префиксом $\sigma(h')$
- тогда (для $\Sigma = \{0,1\}$) выполняется $\sum_h \frac{1}{2^{|\sigma(h)|}} \leqslant 1$
- можем выбрать $w(h) = \frac{1}{2^{\sigma(h)}}$ и применить SRM!

Minimum description length

Теорема о MDL

Пусть H счётный или конечный класс гипотез и для него есть беспрефиксное описание над бинарным алфавитом. Тогда для любого m, $\delta > 0$ и D с вероятностью не меньше $1 - \delta$ на выборке $S \sim D^m$ выполняется:

$$\forall h \in H,\ L_D(h) \leqslant L_S(h) + \sqrt{\frac{|\sigma(h)| + \log(2/\delta)}{2m}}$$

Бритва Оккама

Бритва Оккама

Короткие объяснения *обычно* лучше длинных

- MDL — одно из применений этого принципа
- как получается, что риск зависит от выбора языка описания?

Бритва Оккама

Бритва Оккама

Короткие объяснения *обычно* лучше длинных

- MDL — одно из применений этого принципа
- как получается, что риск зависит от выбора языка описания?
- не зависит — мы выбираем язык перед тем, как смотрим на выборку (как в неравенстве Хёффдинга)

Консистентность

- если разрешить выборочной сложности зависеть от D, то получим определение **консистентного** класса и алгоритма
- консистентность — ослабление неравномерной изучаемости
- алгоритм `Memorize` является консистентным
- и в то же время «плохим учеником» из первой лекции
- может не нужно ослаблять определение?

Какой true risk у выученной гипотезы?

Какой true risk у выученной гипотезы?

- равномерная и неравномерная изучаемость даёт ответ на этот вопрос
- есть другие способы получить эту оценку (validation)

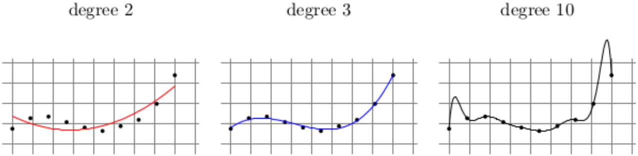
Сколько элементов должно быть в выборке, чтоб выучить лучшую гипотезу из H

Сколько элементов должно быть в выборке, чтоб выучить лучшую гипотезу из H

- PAC даёт конкретный ответ (фундаментальная теорема)
- неравномерная изучаемость и консистентность не даёт ответа на вопрос
- маленький ϵ_{est} не значит маленький ϵ_{app}!
- если PAC-алгоритм выдаёт гипотезу с большим риском, можно понять, в чём проблема (estimation error vs approximation error)

Как учить? Как выражать априорное знание

- в PAC — как только выбрали класс, сразу применяем ERM
- в неравномерной сходимости — выбрали w — применяем SRM (надо меньше априорного знания)



- в консистентных алгоритмах *иногда* даже не нужно априорное знание! (`Memorize`)

Memorize и NoFLT

- NoFLT говорит, что нельзя учиться, когда класс — все функции
- Memorize может учиться на классе всех функций
- где ошибка?

Memorize и NoFLT

- NoFLT говорит, что нельзя учиться, когда класс — все функции
- Memorize может учиться на классе всех функций
- где ошибка?
- ошибки нет: в NoFLT сначала фиксируем размер выборки

Итоги

- ввели новые модели изучаемости
- изучили SRM

Содержание

Постановка вопроса

- рассмотрели AdaBoost — можем варьировать T, чтоб управлять bias-complexity tradeoff
- как выбрать T?
- как решить, что нужен AdaBoost, а не другой алгоритм?
- надо решить задачу выбора модели (model selection)

SRM для выбора модели

- хорош, когда есть параметр, управляющий bias-complexity tradeoff
- оценка SRM зависит от эмпирического риска и «сложности» класса
- обычно, оценка SRM очень пессимистична

Валидация

- PAC-оценки на ошибку гипотезы верны для всех h и $D \Rightarrow$ часто пессимистичны
- **валидация** (validation) — проверка гипотезы на данных, не использованных для тренировки

Hold-out set

Пусть V — выборка из D, **не** использованная во время тренировки.

Hold-out set bound

Пусть h — гипотеза и функция потерь лежит в $[0;1]$. Тогда для любого $\delta \in (0,1)$ с вероятностью не меньше $1-\delta$ выполняется, что на отложенной выборке V длины m_v:

$$|L_D(h) - L_V(h)| \leqslant \sqrt{\frac{\log(2/\delta)}{2m_v}}$$

Из фундаментальной теоремы:

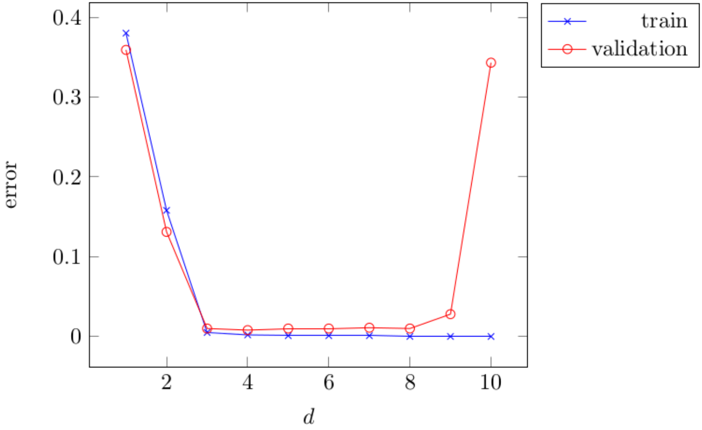
$$L_D(h) \leqslant L_S(h) + \sqrt{C\frac{d+\log(1/\delta)}{m}}$$

Использование валидации для выбора модели

- обучим r алгоритмов (или один с разными параметрами)
- каждую из r гипотез проверяем на отложенной выборке
- выбираем лучшую (ERM на конечном классе гипотез)

$$|L_D(h) - L_V(h)| \leqslant \sqrt{\frac{\log(2|H|/\delta)}{2m_v}}$$

Model selection curve



k-Fold cross validation

Алгоритм 1 k-Fold cross validation

Вход: $S = ((x_1, y_1), \ldots, (x_m, y_m))$
Вход: множество параметро Θ
Вход: алгоритм A
Вход: $k \in \mathbb{N}$

$S = \coprod_{i=1}^{k} S_i$
for $\theta \in \Theta$ **do**
  for $i = 1 \ldots k$ **do**
    $h_{i,\theta} = A(S \setminus S_i; \theta)$
  end for
  $e(\theta) = \frac{1}{k} \sum_{i=1}^{k} L_{S_i}(h_{i,\theta})$
end for
return $A(S; \mathrm{argmin}_\theta[e(\theta)])$

Train; validation; test

- выбираем минимум на валидации — оценка нечестная
- делят на три части
- на тренировочной выборке — обучают алгоритмы
- на валидационной — выбирают лучший
- на тестовой — получают оценку true risk

Содержание

Проблема

- выбрали алгоритм, класс, параметры
- обучили, на валидации выбрали лучший
- проверили на тестовой выборке, получили высокую ошибку
- что делать?

Решения

- найти больше объектов для обучения
- изменить класс гипотез:
    - увеличить его
    - уменьшить его
    - полностью изменить его
    - изменить перебираемые параметры
- изменить признаковое представление
- изменить алгоритм обучения

Большой approximation error

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

- ϵ_{app} не зависит от тренировочной выборки и алгоритма
- нет смысла увеличивать выборку, уменьшать класс, менять алгоритм
- можно поменять класс, либо увеличить его
- попробовать другое признаковое представление

Большой estimation error

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

- ϵ_{est} зависит от размера выборки
- нет смысла уменьшать выборку или увеличивать класс
- можно поменять класс полностью, либо уменьшить его
- попробовать другой алгоритм
- попробовать другое признаковое представление

Разложение ошибки с помощью валидации

$$L_D(h) = (L_D(h) - L_V(h)) + (L_V(h) - L_S(h)) + L_S(h)$$

- на синюю часть хорошая оценка
- если изумрудная большая, то «переобучение»
- если коричневая большая, то «недообучение»
- эти части не являются хорошими оценками ϵ_{est} и ϵ_{app}!

Случай большого $L_S(h)$

Пусть $L_S(h)$ большой. Запишем (h^* — лучшая гипотеза из класса):

$$L_S(h) = (L_S(h) - L_S(h^*)) + (L_S(h^*) - L_D(h^*)) + L_D(h^*)$$

- синяя скобка не больше нуля, если ERM
- изумрудная скобка хорошо оценивается
- коричневая величина и есть ϵ_{app}

Случай маленького $L_S(h)$

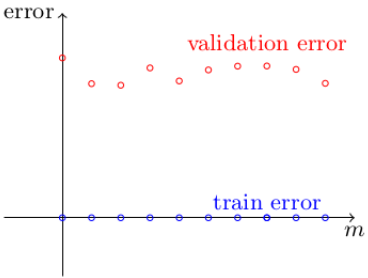
Пусть $L_S(h)$ маленький.

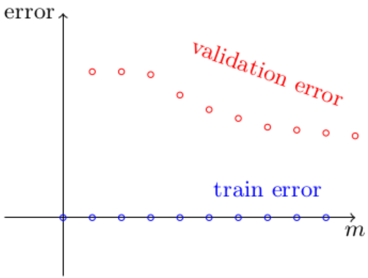
- Сценарий 1: $m < d$, ϵ_{app} большая
- Сценарий 2: $m > 2d$, $\epsilon_{\text{app}} = 0$

В обоих случаях $L_S(h_S) = 0$

Learning curves

Построить зависимость ошибок от размера выборки:





Learning curves

- если $\epsilon_{\text{app}} > 0$, то $L_S(h)$ **обычно** растёт от увеличения S
- $L_V(h_S)$ падает
- при $m \to \infty$ оба — true risk
- можно экстраполировать learning curves и найти интервал, где ϵ_{app}

Общий план

- если есть параметры, то надо начертить model-selection curve
- если $L_S(h_S)$ велико, то увеличить класс, поменять его, изменить признаковое представление
- если $L_S(h_S)$ мало, то начертить learning curves
- если ϵ_{app} велико, то добыть больше данных, уменьшить класс
- если ϵ_{app} мало, то стоит изменить класс или признаковое представление объектов

Содержание

Итоги

- рассмотрели SRM
- обсудили разные определения изучаемости
- изучили метод валидации
- составили план действий, в случае если качество гипотезы плохое

Литература

- Shai Shalev-Shwartz and Shai Ben-David — Understanding Machine Learning: From theory to algorithms (главы 8,11)