Геометрия классов в задачах статистического обучения

Никита Животовский

Московский физико-технический институт
Институт проблем передачи информации

14 марта 2015 г.

- Задан *класс функций* $\mathcal{F}$ на некотором вероятностном пространстве (Ω, μ).
- Некоторая случайная величина Y, которую необходимо приблизить с помощью $\mathcal{F}$.
- *Функция потерь* ℓ, означающая штраф за использование $f(X)$ вместо Y.
- *Обучающая выборка* $(X_i, Y_i)_{i=1}^n$, распределенная согласно произведению n совместных распределений μ и Y на исходах $(\Omega \times \mathbb{R})^n$.

Замечения:

- Для построения приближений можно использовать только обучающую выборку, информацию о классе и функции потерь.
- Функция потерь $\ell : \mathbb{R} \to \mathbb{R}_+$. Таким образом, штраф за предсказание $f(X)$ вместе Y равен $\ell(f(X) - Y)$.

Замечения:

- Для построения приближений можно использовать только обучающую выборку, информацию о классе и функции потерь.
- Функция потерь $\ell : \mathbb{R} \to \mathbb{R}_+$. Таким образом, штраф за предсказание $f(X)$ вместе Y равен $\ell(f(X) - Y)$.

Цель

По обучающей выборке построить некоторое правило $\hat{f}$, такое что $\mathbb{E}\ell(\hat{f}(X) - Y)$ оптимален.

- Случайная величина Y является фиксированной функцией $T : \Omega \to \mathbb{R}$ или эквивалентно $Y = T(X)$.
- Функция потерь квадратичная: $\ell(t) = t^2$.

Тогда цель построить такое отображение $\hat{f}$, что величина

$$\mathbb{E}\ell(f(X) - T(X)) = \mathbb{E}(f(X) - T(X))^2 = \int\limits_{\Omega} (f(t) - T(t))^2 d\mu(t)$$

как можно меньше.
Построение $\hat{f}$ производится на основании $(X_i, T(X_i))_{i=1}^n$.

- Имеет место соотношение

$$Y = g(X) + \varepsilon,$$

где ε — независимый от X гауссовский шум, а $g \in \mathcal{F}$.

- Функция потерь квадратичная: $\ell(t) = t^2$.

- Имеет место соотношение

$$Y = g(X) + \varepsilon,$$

где ε — независимый от X гауссовский шум, а $g \in \mathcal{F}$.

- Функция потерь квадратичная: $\ell(t) = t^2$.

Замечание

В этом случае Y относится уже к более сложному вероятностному пространству чем (Ω, μ).
В качестве эталона необходима функция, заданная на (Ω, μ).

Определение

Байесовское решающее правило f^*, определяемое случайной величиной $\mathbb{E}[Y|X]$, минимизирует $\mathbb{E}(f^*(X) - Y)^2$ среди случайных величин, заданных на (Ω, μ).

В примере $Y = g(X) + \varepsilon$, $g \in \mathcal{F}$ при условии независимости шума $f^* = g$.

Замечание

Однако риск f^* равен дисперсии шума. Таким образом, в этой задаче ни для какого $\hat{f}$ нельзя сделать величину $\mathbb{E}(\hat{f}(X) - Y)^2$ сколь угодно близкой к нулю.

Оптимальное решающее правило должно обладать риском, близким к риску байесовского решающего правила.

Определение

Избыточным риском правила $\hat{f}$ называется величина

$$\mathbb{E}\ell(\hat{f}(X) - Y) - \mathbb{E}\ell(f^*(X) - Y),$$

где f^* — байесовское решающее правило.

Предположим, что $Y \in \mathcal{F}$.

Определение (Valiant' 84)

Класс $\mathcal{F}$ называется PAC–обучаемым, если существует обучающий алгоритм $\psi_n : (\Omega, \mathbb{R})^n \to \mathcal{F}$, такой что с вероятность $1 - \delta$ (по отношению к обучающей выборке длины n) взяв достаточно большое n^*, зависящее от чисел ε и δ, получим, что для $n \geq n^*$

$$\mathbb{E}\ell(\hat{f}(X) - Y) - \mathbb{E}\ell(f^*(X) - Y) \leq \varepsilon,$$

где $\hat{f} = \psi_n((X_i, Y_i)_{i=1}^n)$.

Замечание

Отметим независимость от распределения данных.

No free lunch theorem

Теорема (Folklore)

- *Если про класс $\mathcal{F}$ ничего неизвестно или он слишком большой, то для любого обучающего алгоритма найдется плохое вероятностное распределение, что избыточный риск не будет стремиться к нулю.*
- *Если $f^* \notin F$ и обучающий алгоритм принимает значения в $\mathcal{F}$, то опять же избыточный риск не будет стремиться к нулю.*

Approximation–Estimation tradeoff

Обозначим риск $R(f) = \mathbb{E}\ell(f(X) - Y)$. Пусть $\hat{f}$ выбирается алгоритмом обучения, тогда

$$R(\hat{f}) - R(f^*) = \left(R(\hat{f}) - R(f^*_{\mathcal{F}})\right) + \left(R(f^*_{\mathcal{F}}) - R(f^*)\right),$$

где $f^*_{\mathcal{F}} = \arg\inf\limits_{f \in \mathcal{F}} R(f)$.

Определение

- *Выражение в первой скобке называется ошибкой оценивания (estimation error).*
- *Выражение во второй — ошибкой аппроксимации (approximation error).*

Определение (Kearns, Shapire' 91)

- *Агностический случай (Agnostic case)* — $f^* \notin \mathcal{F}$.
- *Реализуемый случай (Realizable case)* — $Y \in \mathcal{F}$.

В агностическом случае избыточным риском называют

$$R(\hat{f}) - R(f^*_{\mathcal{F}}) = \mathbb{E}\ell(\hat{f}(X) - Y) - \mathbb{E}\ell(f^*_{\mathcal{F}}(X) - Y)$$

Замечание

- *В реализуемом случае ошибка аппроксимации равна нулю.*
- *Так определенный избыточный риск совпадает с ранее введенным в реализуемом случае.*

Определение

Агностическая PAC–обучаемость класса $\mathcal{F}$ заключается в стремлении с большой вероятностью к нулю избыточного риска.

Ставится более общая задача:

Цель

Пусть $\hat{f}$ — результат обучения некоторой процедуры на $(X_i, Y_i)_{i=1}^n$. Найти наименьшую функцию $\mathcal{E}$, такую что с вероятностью не меньшей $1-\delta$ относительно обучающей выборки длины n

$$R(\hat{f}) \leq R(f^*_{\mathcal{F}}) + \mathcal{E},$$

где функция $\mathcal{E}$ может зависеть от свойств $\mathcal{F}$ и ℓ, чисел n и δ, свойств Y и так далее.

Обозначим $R_n(f) = \frac{1}{n}\sum_{i=1}^{n} \ell(f(X_i) - Y_i)$ — *эмпирический риск.*

Определение

Алгоритм обучения называется минимизатором эмпирического риска, если по обучающей выборке он выбирает

$$\hat{f} = \arg\min_{f \in \mathcal{F}} R_n(f).$$

Определение

- *Класс потерь:*

$$\ell \circ \mathcal{F} = \{(X, Y) \to \ell(f(X) - Y) : f \in \mathcal{F}\}.$$

- *Класс избыточных потерь:*

$$(\ell \circ \mathcal{F})^* = \{(X, Y) \to \ell(f(X) - Y) - \ell(f^*_{\mathcal{F}}(X) - Y) : f \in \mathcal{F}\}.$$

Пусть $\hat{f}$ минимизирует эмпирический риск, тогда

$$
\begin{aligned}
& R(\hat{f}) - R(f^*_{\mathcal{F}}) = \\
& R(\hat{f}) - R_n(\hat{f}) + R_n(f^*_{\mathcal{F}}) - R(f^*_{\mathcal{F}}) + R_n(\hat{f}) - R_n(f^*_{\mathcal{F}}) \leq \\
& R(\hat{f}) - R_n(\hat{f}) + R_n(f^*_{\mathcal{F}}) - R(f^*_{\mathcal{F}}) \leq \\
& 2 \sup_{f \in \mathcal{F}} |R(f) - R_n(f)| = \\
& 2 \sup_{g \in \ell \circ \mathcal{F}} |\mathrm{P}\, g - \mathrm{P}_n\, g|,
\end{aligned}
$$

где P и P_n означают математические ожидания соответственно по истинной и эмпирической мерам.

Замечание

- *Равномерная сходимость частот к вероятностям в классе потерь гарантирует обучаемость с помощью метода минмизиации эмпирического риска*
- *Скорость сходимости избыточного риска контролируется процессом* $\sup_{g \in \ell \circ \mathcal{F}} \left| \mathrm{P}\, g - \mathrm{P}_n\, g \right|$.

Обратимся опять к ситуации, когда $Y = T(X)$, однако не будем предполагать, что $T \in \mathcal{F}$. В этом случае требуем с вероятностью $1 - \delta$:

$$R(\hat{f}) - R(f^*_{\mathcal{F}}) = \mathbb{E}\ell(\hat{f}(X) - T(X)) - \mathbb{E}\ell(f^*_{\mathcal{F}}(X) - T(X)) \le \mathcal{E}(T, \delta),$$

Теорема (Alon, Ben-David, Cesa-Bianchi, Haussler'97 Mendelson'08)

Пусть $\hat{f}$ минимизирует эмпирический риск.

1. *Если для класса $\ell \circ \mathcal{F}$ не выполнен равномерный закон больших чисел, то*

$$\lim_{n\to\infty} \sup_{\mu} \sup_{T:\|T\|_\infty \le 1} \mathcal{E}(T,\delta) \neq 0,$$

более того, предел не равен нулю даже если $T \in \mathcal{F}$ для всех T с $\|T\|_\infty \le 1$.

2. *Если для $\ell \circ \mathcal{F}$ выполнен равномерный закон больших чисел, но не выполнена равномерная центральная предельная теорема, то $\sup_{\mu} \sup_{T:\|T\|_\infty \le 1} \mathcal{E}(T,\delta)$ сходится к нулю не быстрее чем $\frac{1}{\sqrt{n}}$*

Замечание

- *В двух приведенных ситуациях классы (классы потерь) слишком большие. Геометрические свойства $\mathcal{F}$ и $\ell \circ \mathcal{F}$ практически не влияют на $\mathcal{E}(T, \delta)$. Все контролируется лишь поведением* $\sup_{g \in \ell \circ \mathcal{F}} |\mathrm{P}\, g - \mathrm{P}_n\, g|$.
- *Выполнение равномерной центральной предельной теоремы не гарантирует порядков быстрее* $\frac{1}{\sqrt{n}}$. *В частности, существует пример, когда $\mathcal{F}$ состоит всего из двух функций, но геометрия $(\ell \circ \mathcal{F})^*$ не позволяет улучшить* $\frac{1}{\sqrt{n}}$.

Определение (Folklore)

- *Быстрые порядки (fast rates) — сходимость $\mathcal{E}$ к нулю быстрее чем $\frac{1}{\sqrt{n}}$.*
- *Медленные порядки (slow rates) — сходимость $\mathcal{E}$ со скоростью $\frac{1}{\sqrt{n}}$.*

Замечание

- *В двух приведенных ситуациях классы (классы потерь) слишком большие. Геометрические свойства $\mathcal{F}$ и $\ell \circ \mathcal{F}$ практически не влияют на $\mathcal{E}(T, \delta)$. Все контролируется лишь поведением* $\sup\limits_{g \in \ell \circ \mathcal{F}} |\mathrm{P}\, g - \mathrm{P}_n\, g|$.
- *Выполнение равномерной центральной предельной теоремы не гарантирует порядков быстрее $\frac{1}{\sqrt{n}}$.*

Откуда могут браться порядки сходимости избыточного риска быстрее $\frac{1}{\sqrt{n}}$?

Теорема (Неравенство Бернштейна)

Пусть X_i — независимые случайные величины, такие что $|X_i| \leq M$ и $\mathbb{E}X_i^2 = \sigma^2$. Тогда всех $\varepsilon > 0$

$$\mathrm{P}\left(\frac{1}{n}\sum_{i=1}^{n} X_i \geq \varepsilon\right) \leq \exp\left(-\frac{n\varepsilon^2}{2\sigma^2 + \frac{2}{3}M\varepsilon}\right)$$

Замечание

В областях с малой дисперсией хвост ведет себя как $\exp(-n\varepsilon)$.

Определение (Tsybakov' 04)

Пусть $f^* \in \mathcal{F}$. *Для пары* (ℓ, P), *где* P — *совместное распределение* Ω, Y, *выполнено условие малого шума (Low noise condition, Margin assumption) с параметрами* (β, B), *если для всех* $f \in \mathcal{F}$

$$\mathbb{E}(\ell(f(X) - Y) - \ell(f^*(X) - Y))^2 \leq B(R(f) - R(f^*))^{\beta}.$$

Условие легко понять, если взглянуть на класс избыточных потерь $(\ell \circ \mathcal{F})^*$.
Условие малого шума эквивалентно тому, что для $g \in (\ell \circ \mathcal{F})^*$:

$$\mathrm{P}\, g^2 \leq B(\mathrm{P}\, g)^{\beta}.$$

Пусть $|\mathcal{F}| = N$, $f^* \in \mathcal{F}$, ℓ ограничена единицей и выполнено условие малого шума. Тогда, объединив неравенство Буля и неравенство Бернштейна, получаем, что с вероятностью не меньшей $1 - \delta$ для любой $g \in (\ell \circ \mathcal{F})^*$:

$$\mathrm{P}\, g \le \mathrm{P}_n\, g + \sqrt{\frac{8B(\mathrm{P}\, g)^{\beta} \log(\frac{N}{\delta})}{n}} + \frac{4\log(\frac{N}{\delta})}{3n}.$$

Для минимизатора эмпирического риска $\mathrm{P}_n\, g \le 0$. Решение относительно $\mathrm{P}\, g$ дает порядок для избыточного риска:

$$R(\hat{f}) - R(f^*_{\mathcal{F}}) \le C \left(\frac{\log(\frac{N}{\delta})}{n} \right)^{\frac{1}{2-\beta}}$$

При β, пробегающих отрезок $[0, 1]$ порядки будут непрерывно меняться от $\frac{1}{\sqrt{n}}$ до $\frac{1}{n}$.

Определение (Massart' 06)

Условия малого шума Цыбакова при $\beta = 1$ называется условием малого шума Массара (Massart's low noise condition).

Пусть Y принимает значения 0 или 1, функция потерь — индикатор ошибки. Тогда $f^* = \mathbb{1}\{\eta(x) > \frac{1}{2}\}$, где $\eta(x) = \mathbb{E}[Y|X = x]$.

Утверждение

Условие малого шума Массара эквивалентно тому, что $|2\eta(x) - 1| > h$ для некоторой константы h.

Замечание

Формулировка в терминах свойств $\mathbb{E}[Y|X]$ есть и для условий Цыбакова.

Пусть в задаче бинарной классификации класс $\mathcal{F}$ имеет размерность Вапника–Червоненкиса, равную V. Обозначим $\mathcal{P}(h, \mathcal{F})$ все распределения на Ω и Y такие, что выполнено условие малого шума и $f^* \in \mathcal{F}$.

Теорема (Massart, Nedelec' 06)

Пусть $\hat{f}$ — минимизатор эмпирического риска. Тогда

$$\sup_{P \in \mathcal{P}(h,\mathcal{F})} \mathbb{E}(R(\hat{f}) - R(f^*)) \leqslant C\sqrt{\frac{V}{n}}, \text{ если } h \leq \sqrt{\frac{V}{n}}$$

и

$$\sup_{P \in \mathcal{P}(h,\mathcal{F})} \mathbb{E}(R(\hat{f}) - R(f^*)) \leqslant C\frac{V}{nh}\left(1 + \log\left(\frac{nh^2}{V}\right)\right), \text{ иначе.}$$

Теорема (Massart, Nedelec' 06)

Верна и нижняя оценка:

$$\inf_{\tilde{f}} \sup_{P \in \mathcal{P}(h,\mathcal{F})} \mathbb{E}(R(\tilde{f}) - R(f^*)) \geq \min\left(\sqrt{\frac{V}{n}}, \frac{V}{nh}\right).$$

Замечание

- *Единственное требование к классу* $f^* \in F$.
- *Условия малого шума по сути являются свойством распределений и никак не зависят от структуры классов* $\mathcal{F}, \ell \circ \mathcal{F}$ *и* $(\ell \circ \mathcal{F})^*$.
- *Условия вида конечная мощность* $\mathcal{F}$*, конечная энтропия или VC размерность необходимы лишь для гарантирования порядков* $\frac{1}{\sqrt{n}}$.

Цель

Выяснить насколько быстро стремится к нулю избыточный риск в случае, когда $f^* \notin \mathcal{F}$?

Для минимизатора эмпирического риска $\hat{f}$:

$$R(\hat{f}) - R(f^*_{\mathcal{F}}) \leq 2 \sup_{g \in \ell \circ \mathcal{F}} |\mathrm{P}\, g - \mathrm{P}_n\, g|,$$

Анализ процесса $\sup_{g \in \ell \circ \mathcal{F}} |\mathrm{P}\, g - \mathrm{P}_n\, g|$ производится с помощью симметризации.

Определение (Gine, Zinn' 84, Kolchinskii' 01)

Пусть $(\sigma_i)_{i=1}^n$ *независимые случайные величины, принимающие равновероятно значения* $+1$ *и* -1. *Радемахеровская сложность класса потерь* $(\ell \circ \mathcal{F})$:

$$\mathcal{R}(\ell \circ \mathcal{F}) = \frac{1}{n}\, \mathbb{E} \sup_{g \in \ell \circ \mathcal{F}} \left| \sum_{i=1}^{n} \sigma_i g(X_i, Y_i) \right|.$$

Лемма

$$\mathbb{E} \sup_{g \in \ell \circ \mathcal{F}} |\mathrm{P}\, g - \mathrm{P}_n\, g| \leq 2\mathcal{R}(\ell \circ \mathcal{F}),$$

а если функции $g \in \ell \circ \mathcal{F}$ *ограничены константой* C, *то*

$$\mathbb{E} \sup_{g \in \ell \circ \mathcal{F}} |\mathrm{P}\, g - \mathrm{P}_n\, g| \geq \frac{1}{2}\mathcal{R}(\ell \circ \mathcal{F}) - \frac{C}{2\sqrt{n}}.$$

Теорема (Ledoux, Talagrand' 91)

Принцип сжатия (contraction). Если $\varphi : \mathbb{R} \to \mathbb{R}$, *такая что* $\varphi(0) = 0$ *и Липшицева с константой* L, *то*

$$\mathcal{R}(\varphi \circ \mathcal{F}) \leq L\mathcal{R}(\mathcal{F}).$$

Для ограниченных классов $\mathcal{F}$ и ограниченной целевой функции Y и $\ell(t) = t^2$ можно контролировать сложность класса потерь сложностью базового класса $\mathcal{F}$.
С помощью анализа Радемахеровских средних можно легко получить основные результаты теории Вапника–Червоненкиса.

Недостатки Радемахеровского анализа:

- Используется только L_2–геометрия классов $\mathcal{F}$ или $\ell \circ \mathcal{F}$ (например, метрическая энтропия Дадли), которые характеризуют в лучшем случае лишь выполнение равномерной ЦПТ.
- Радемахеровский анализ не позволяет получить порядков быстрее чем $\frac{1}{\sqrt{n}}$.
- Радемахеровские средние не чувствительны к выпуклости классов $\mathcal{F}$ и $\ell \circ \mathcal{F}$.
- Анализ существенно опирается на концентрацию меры, которая хорошо работает лишь в задачах с ограниченными классами потерь или при сильных ограничениях, например, на $Y - f_F^*(X)$.

Возможность достижения порядков быстрее чем $\frac{1}{\sqrt{n}}$ зависит от связи математических ожиданий и дисперсий в классе избыточных потерь.

Определение (Bartlett, Mendelson' 06)

Говорят, что для класса $\mathcal{F}$ или $(\ell \circ \mathcal{F})^*$ выполнено условие Бернштейна с параметрами (β, B), если для всех $f \in \mathcal{F}$

$$\mathbb{E}(\ell(f(X) - Y) - \ell(f^*_{\mathcal{F}}(X) - Y))^2 \leq B(R(f) - R(f^*_{\mathcal{F}}))^{\beta}.$$

Замечание

По сравнению с условием малого шума мы не требуем, что $f^* \in \mathcal{F}$ и заменяем f^* на $f^*_{\mathcal{F}}$.

- Условие Бернштейна чисто геометрическое условие в отличие от условия малого шума.
- Условие очень чувствительно к структуре $\mathcal{F}$, так как удаление или добавление лишь одной функции может изменить $f^*_{\mathcal{F}}$.
- Легко проверить, что условие Бернштейна влечет единственность $f^*_{\mathcal{F}}$.
- В общем случае не является необходимым для получения порядков избыточного риска быстрее чем $\frac{1}{\sqrt{n}}$.

Рассмотрим задачу ограниченной регрессии:
$|Y| \leq 1, \sup_{f \in \mathcal{F}} |f(X)| \leq 1$ и $\ell(t) = t^2$ с замкнутым и выпуклым $\mathcal{F}$.

$$R(f) - R(f^*_{\mathcal{F}}) = \mathbb{E}\left((Y - f(X))^2 - (Y - f^*_{\mathcal{F}}(X))^2\right) = \\ -2A + \mathbb{E}\left(f(X) - f^*_{\mathcal{F}}(X)\right)^2 \geq \\ \mathbb{E}\left(f(X) - f^*_{\mathcal{F}}(X)\right)^2,$$

где $A = \mathbb{E}(Y - f^*_{\mathcal{F}}(X))(f(X) - f^*_{\mathcal{F}}(X)) \leq 0$ так как $\mathcal{F}$ — выпуклое множество.

$$\mathbb{E}\left((f(X) - Y)^2 - (f^*_{\mathcal{F}}(X) - Y)^2\right)^2 \leq 16\mathbb{E}\left(f(X) - f^*_{\mathcal{F}}(X)\right)^2.$$

Утверждение (Lee, Bartlett, Williamson' 97)

В описанном примере выполнено условие Бернштейна с параметрами $(1, 16)$.

В предыдущем примере условие Бернштейна выполнено одновременно для любой целевой функции Y, ограниченной равномерно единицей.
Рассмотрим задачу с

$$Y = 0,\ \mathcal{F} = \{f_1, f_2\}, \ell(t) = t^2,$$
$$f_1 = \mathbb{1}[0,1],\ f_2 = \mathbb{1}[-1,0],$$
$$\mathrm{P}[X = 1] = \frac{1}{2} - \frac{1}{\sqrt{n}}, \mathrm{P}[X = -1] = \frac{1}{2} + \frac{1}{\sqrt{n}}.$$

Легко убедиться, что константа $B = \sqrt{n}/2$ и условие Бернштейна не выполнено. Одновременно можно показать, что избыточный риск минимизатора эмпирического риска сходится к нулю со скоростью $\frac{1}{\sqrt{n}}$.

Причина медленных порядков сходимости предыдущего примера:

- Плохая геометрия класса $\mathcal{F}$. Для $conv(\mathcal{F})$ условие Бернштейна выполнено.
- Плохая целевая функция Y: находится слишком близко к области, где существует две проекции на класс $\mathcal{F}$.
- Помещая Y в $\mathcal{F}$, условие Бернштейна автоматически переходит в условие малого шума Цыбакова, которое в данном случае выполнено без требования выпуклости $\mathcal{F}$.

Замечание

Порядки сходимости избыточного риска к нулю существенно зависят от того, насколько расположение Y благоприятствует $\mathcal{F}$.

Замечание

Понятие быстрых порядков как сходимостей, например, $\frac{1}{n}$ и медленных порядков как сходимостей $\frac{1}{\sqrt{n}}$ упирается в тот факт, что уже необходимы сильные условия на класс, а именно возможность достигнуть хотя бы $\frac{1}{\sqrt{n}}$.

Рассматриваем задачи с $\ell(t) = t^2$.

Определение (Mendelson' 15)

Оптимистичные порядки(optimistic rate) — наилучшие порядки избыточного риска, достижимые в ситуации когда Y благоприятствует $\mathcal{F}$ следующим образом:

$$\mathbb{E}(f^*_{\mathcal{F}}(X) - Y)(f(X) - f^*_{\mathcal{F}}(X)) \geq 0.$$

Замечание

Условие $\mathbb{E}(f^*_{\mathcal{F}}(X) - Y)(f(X) - f^*_{\mathcal{F}}(X)) \geq 0$ *выполнено в двух важных ситуациях:*

- $Y \in L_2$, $\mathcal{F} \subset L_2$ *и является замкнутым и выпуклым.*
- $Y = f^*(X) + \varepsilon$, *где* ε *— независимый центрированный шум и* $f^* \in \mathcal{F}$.

В предыдущем примере оптимистичные порядки равны $\frac{1}{n}$.

Теорема (Juditsky, Rigollet, Tsybakov' 08)

Пусть $|\mathcal{F}| = M$. *Ни одна обучающая процедура, принимающая значения в классе* $\mathcal{F}$, *не может дать порядки для избыточного риска лучше чем* $\sqrt{\frac{\log(M)}{n}}$.

Цель

Есть ли способы обойти неблагоприяные Y и получать оптимистичные порядки при минимальных ограничениях Y и геометрию класса $\mathcal{F}$?

Решения:

- Рассматривать агрегационные процедуры, принимающие значения вне базового класса $\mathcal{F}$.
- Рассмотрение неточных оракульных нервенств (non-sharp), то есть вместо величины $R(f) - R(f^*_{\mathcal{F}})$ рассматривать $R(f) - CR(f^*_{\mathcal{F}})$ для $C > 1$.

Определение

Пусть $\mathcal{F}$ содержится в $\mathcal{F}^*$, нужно построить обучающий алгоритм $\psi_n : (\Omega \times \mathbb{R})^n \to \mathcal{F}^*$ такой, что с большой вероятностью

$$R(\hat{f}) - R(f^*_{\mathcal{F}}) \leq \mathcal{E}_{MSA},$$

где функция $\mathcal{E}_{MSA}$ как можно быстрее сходится к нулю с ростом n и $\hat{f} = \psi_n((X_i, Y_i)_{i=1}^n)$.

Замечание

- *Данная задача называется MS–агрегацией (Model Selection Aggregation).*
- *Обычно рассматриваются конечные словари $\mathcal{F}$.*

Ранее обсуждалось, что если $|\mathcal{F}| = M$ и $\mathcal{F} = \mathcal{F}^*$, то нельзя получить порядки быстрее $\sqrt{\frac{\log(M)}{n}}$

Теорема (Tsybakov' 04)

Для любой агрегационной процедуры $\mathcal{E}_{MSA}$ стремится к нулю не быстрее $\frac{\log(M)}{n}$.

Замечание

- *Теорема формулируется при некоторых технических ограничениях.*
- *Данная нижняя оценка соответствует нашим представлениям об оптимистичных порядках. Осталось привести примеры процедур, которые могли бы давать порядки $\frac{1}{n}$ вне зависимости от Y.*

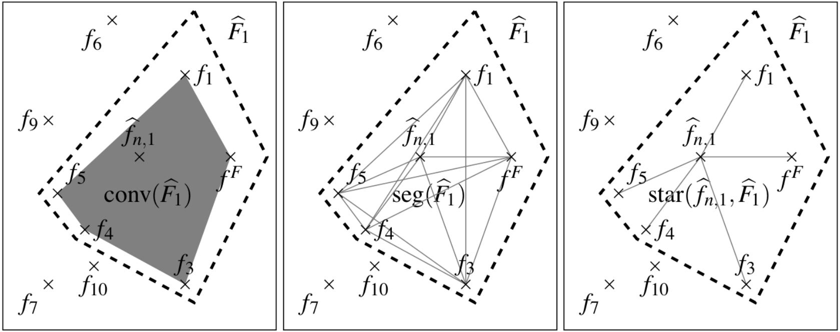
Замкнутые выпуклые классы $\mathcal{F}$ в задачах ограниченной регрессии являются классами, для которых условие Бернштейна выполнено для всех ограниченных Y.

Замечание (Lecue, Mendelson' 08)

Оказывается, что использование минимизатора эмпирического риска на $conv(\mathcal{F})$ не позволяет перейти к оптимальным порядкам. Улучшая геометрию, мы чрезмерно увеличиваем сложность.

Идея

Нужно брать выпуклую оболочку не от всего словаря, а от некоторого его подмножества, балансируя выигрыш в геометрии с увеличивающейся сложностью.



Теорема (Lecue, Mendelson' 08, Lecue' 11, Audibert' 09)

В задачах ограниченной регрессии для любого словаря $\mathcal{F}$, состоящего из M функций, и любых Y с вероятность $1 - \delta$

$$R(\hat{f}) \leq R(f^*_{\mathcal{F}}) + c\left(1 + \log\left(\frac{2}{\delta}\right)\right)\frac{\log(M)}{n}.$$

Классический способ получения оценок избыточного риска основан на анализе Радемахеровских средних, но этот подход не позволяет учесть геометрические свойства класса. Техники, основанные на локализации, позволяют получать быстрые порядки.

Определение

- $Z, Z_1, \ldots, Z_n$ — *независимые случайные величины определенные на* $(\mathcal{Z}, \mathrm{P}_{\mathcal{Z}})$.
- G — *класс функций,* $V(G)$ — *звездное замыкание:*

$$V(G) = \{\alpha g : 0 \leq \alpha \leq 1, g \in G\}$$

- *Локализованное звездное замыкание:*

$$V(G)_\lambda = \{h \in V(G) : Ph \leq \lambda\}$$

Определение

- $\| \mathrm{P} - \mathrm{P}_n \|_H = \sup_{h \in H} |(\mathrm{P} - \mathrm{P}_n)h|$.
- $\sigma(H) = \sup_{h \in H} \sqrt{\mathrm{P}\, h^2}$.
- $\|H\|_\infty = \sup_{h \in H} \|h\|_{L_\infty}$.

Теорема (Bartlett, Mendelson' 06)

Пусть G — класс действительных функций на $\mathcal{Z}$, удовлетворяющих условию $\mathrm{P}\, g^2 \le B\, \mathrm{P}\, g$ для некоторой $B > 0$. Пусть $\lambda^* > 0$ такое, что

$$\mathbb{E} \| \mathrm{P} - \mathrm{P}_n \|_{V(G)_{\lambda^*}} \le \frac{1}{8} \lambda^*.$$

Теорема (Bartlett, Mendelson' 06, Lecue' 13)

Тогда с вероятностью не меньшей $1-\delta$:

$$|\mathrm{P}\,g - \mathrm{P}_n\,g| \le \frac{1}{2}\max(\mathrm{P}\,g, \rho_n(\delta)),$$

где

$$\rho_n(\delta) = \max\left(\lambda^*, \frac{c_0(B+\|G\|_\infty)\log(\frac{4}{\delta})}{n}\right)$$

Пусть свойство $|\mathrm{P}\,g - \mathrm{P}_n\,g| \le \frac{1}{2}\max(\mathrm{P}\,g, \rho_n(\delta))$ выполнено для всех $g \in (\ell \circ \mathcal{F})^*$ с вероятностью $1-\delta$. Тогда если $\hat{f}$ — минимизирует эмпирический риск, а $\hat{g}$ — его избыточные потери, то

$$R(\hat{f}) - R(f^*_{\mathcal{F}}) = \mathrm{P}\,\hat{g} \le 2\,\mathrm{P}_n\,\hat{g} + \rho_n(\delta) \le \rho_n(\delta).$$

Остается научиться считать λ^*. Это возможно благодаря пилингу (peeling):

$$V(G)_\lambda \subset \bigcup_{i=0}^{\infty} \{\theta h : 0 \le \theta \le 2^{-i}, h \in G_{2^{i+1}\lambda}\},$$

поэтому

$$\mathbb{E} \| \mathrm{P} - \mathrm{P}_n \|_{V(G)_{\lambda^*}} \le \sum_{i=0}^{\infty} 2^{-i} \mathbb{E} \| \mathrm{P} - \mathrm{P}_n \|_{G_{2^{i+1}\lambda}}.$$

Задача сводится к оценке

$$\mathbb{E} \| \mathrm{P} - \mathrm{P}_n \|_{G_{2^{i+1}\lambda}}.$$

Способ получения быстрых порядков — рассмотрение неточных оракульных неравенств: $R(f) - CR(f^*_{\mathcal{F}})$ для $C > 1$. Оказывается, что источником быстрых порядков в данной задаче также является условие Бернштейна:

$$\mathrm{P}\, g^2 \leq B\, \mathrm{P}\, g,$$

но не для класса избыточных потерь $(\ell \circ \mathcal{F})^*$, а для класса потерь $(\ell \circ \mathcal{F})$.

Замечание

Данное условие уже не зависит от геометрии класса, а существенно опирается лишь на функцию потерь. Тривиальным образом выполнено, например, для бинарной функции потерь.

Цель

Адаптироваться к неблагоприятной целевой функции Y при условии, что класс $\mathcal{F}$ бесконечный.

Теорема (Rakhlin, Sridharan, Tsybakov' 13)

- *Разбиваем обучающую выборку длины $3n$ на 3 части.*
- *С помощью первой части выборки строим ε-сеть класса $\mathcal{F}$ по метрике* $d(f, g) = \sqrt{\frac{1}{n}\sum\limits_{i=1}^{n}(f(X_i) - g(X_i))^2}$.
- *С помощью второй части выборки строим минимизаторы эмпирического риска, на подмножествах $\mathcal{F}$ соответствующих диаграмме Вороного, порожденной ε-сетью.*

Теорема (Rakhlin, Sridharan, Tsybakov' 13)

- *С помощью третьей части выборки агрегируем построенные минимизаторы эмпирического риска любым из приведенных ранее методов для конечных словарей.*

Тогда в задаче ограниченной регрессии, ожидаемый избыточный риск построенной процедуры для специально подобранного ε обладает следующими свойствами:

Regime	Aggregation-of-leaders
Finite: $\lvert\mathcal{F}\rvert = M$	$\frac{\log M}{n}$
Parametric: $VC(\mathcal{F}) = v \le n$	$\frac{v \log(en/v)}{n}$
Nonparametric: $\mathcal{H}_2(\mathcal{F}, \epsilon) = \epsilon^{-p}$, $p \in (0,2)$	$n^{-\frac{2}{2+p}}$
$p \in [2, \infty)$	$n^{-\frac{1}{p}}$

Спасибо за внимание!

Список литературы по адресу:
nikita.zhivotovskiy@phystech.edu