Локализация оценок избыточного риска в комбинаторной теории переобучения

И. О. Толстихин
iliya.tolstikhin@gmail.com

ВЦ РАН
«Интеллектуализация обработки информации»
сентябрь 2012

Содержание

Минимизация эмпирического риска

$\mathbb{X}$ и $\mathbb{Y}$ — пространства объектов и ответов.

На $\mathbb{X} \times \mathbb{Y}$ задано вероятностное распределение P.

Обучающая выборка $\{X_i, Y_i\}_{i=1}^{\ell}$ — простая из P.

A — множество алгоритмов $a\colon \mathbb{X} \to \mathbb{Y}$.

Функция потерь $r\colon \mathbb{Y} \times \mathbb{Y} \to \mathbb{R}^+$. (ограничим $r \in [0,1]$)

Минимизация риска:

$$Pa \stackrel{\text{def}}{=} \mathsf{E} r\big(a(X), Y\big) \to \min_{a \in A}.$$

Минимизация эмпирического риска:

$$P_\ell a \stackrel{\text{def}}{=} \frac{1}{\ell} \sum_{i=1}^{\ell} r\big(a(X_i), Y_i\big) \to \min_{a \in A}. \quad \textbf{Решение} — \ \mathbf{a}^{\ell}.$$

Задача

Избыточный риск $\mathcal{E}(a) \stackrel{\text{def}}{=} Pa - \inf_{f \in A} Pf$.

Нас интересует $\mathcal{E}(a^\ell)$ — случайная величина.

Основная задача:

- Получение вероятностных неравенств вида

$$\mathsf{P}^{\otimes \ell}\{\mathcal{E}(a^\ell) \geqslant t\} \leqslant \eta(A, \ell, t),$$

  $\mathsf{P}^{\otimes \ell}$ — ℓ-я декартова степень P;
  η — неотрицательная убывающая функция от t.
- Нам важен порядок малости оценок по ℓ.
- Получение вычислимых по данным оценок $\hat{\eta}$.

Классический подход: оценки порядка $O(1/\sqrt{\ell})$

- Равномерное по классу A отклонение P_ℓ от P:

$$\mathcal{E}(a^\ell) = Pa^\ell - \inf_{a\in A} Pa = Pa^\ell - P\bar{a} =$$
$$= P_\ell a^\ell - P_\ell \bar{a} + (P - P_\ell)(a^\ell - \bar{a}) \leqslant$$
$$\leqslant \sup_{a,f\in A} |(P - P_\ell)(f - a)| \leqslant$$
$$\leqslant 2\sup_{a\in A} |(P - P_\ell)a| \stackrel{\text{def}}{=} 2\|P - P_\ell\|_A.$$

- Неравенство Буля (в случае счетного A):

$$\mathsf{P}^{\otimes\ell}\Big\{\sup_{a\in A} |(P - P_\ell)a| \geqslant t\Big\} \leqslant \sum_q \mathsf{P}^{\otimes\ell}\Big\{|(P - P_\ell)a_q| \geqslant t\Big\}.$$

- Неравенство Хефдинга:

$$\mathsf{P}^{\otimes\ell}\{|Pa - P_\ell a| \geqslant t\} \leqslant 2e^{-2\ell t^2}.$$

Основные причины завышенности классических оценок

В то же время, во многих случаях $\mathcal{E}(a^\ell) \sim o(1/\sqrt{\ell})$!

- Если A имеет конечную VC-размерность, то

$$\mathcal{E}(a^\ell) = Pa^\ell - \inf_{a\in A} Pa = Pa^\ell - P\bar{a} =$$
$$= P_\ell a^\ell - P_\ell \bar{a} + (P - P_\ell)(a^\ell - \bar{a}) \leqslant$$
$$\leqslant \sup_{a,f\in A} |(P - P_\ell)(f - a)| \leqslant$$
$$\leqslant 2 \sup_{a\in A} |(P - P_\ell)a| \stackrel{\text{def}}{=} 2\|P - P_\ell\|_A \sim O(1/\sqrt{\ell}).$$

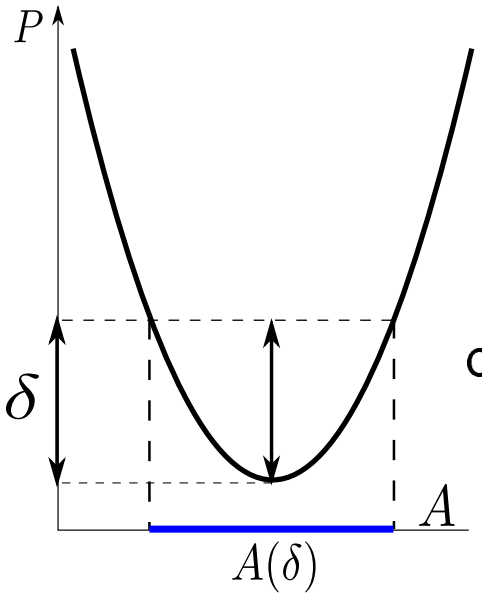
- Неравенство Буля завышего всегда, когда в классе A много схожих алгоритмов.
- Неравенство Хефдинга учитывает лишь ограниченность случайных величин и не учитывает их дисперсии.

Локализация оценок

P. Bartlett, O. Bousquet, S. Mendelson. (2005)
Local Rademacher Complexities.
P. Bartlett, S. Mendelson. (2006)
Empirical Risk Minimization.
O. Bousquet, V. Koltchinskii, D. Panchenko. (2002)
Some local measures of complexity of convex hulls and generalization bounds.
V. Koltchinskii. (2006)
Local Rademacher Complexities and Oracle Inequalities in Risk Minimization.
V. Koltchinskii, D. Panchenko. (2000)
Rademacher processes and bounding the risk of function learning.
P. Massart. (2000)
Some applications of concentration inequalities to statistics.

Идея локализации

$A(\delta) \stackrel{\text{def}}{=} \{a \in A : \mathcal{E}(a) \leqslant \delta\}$ — δ-минимальное множество.



$$\delta^\ell \stackrel{\text{def}}{=} \mathcal{E}(a^\ell) = Pa^\ell - P\bar{a} =$$

$$= P_\ell a^\ell - P_\ell \bar{a} + (P - P_\ell)(a^\ell - \bar{a}) \leqslant (P - P_\ell)(a^\ell - \bar{a}),$$

откуда с учетом $\mathcal{E}(\bar{a}) = 0$ получаем

$$\delta^\ell \leqslant \sup_{a,f \in A(\delta^\ell)} |(P - P_\ell)(a - f)|.$$

Пусть для неслучайной $U_\ell(\delta)$ с большой вероятностью $\sup_{a,f \in A(\delta)} |(P - P_\ell)(a - f)| \leqslant U_\ell(\delta)$ равномерно по δ.
Тогда с той же вероятностью $\mathcal{E}(a^\ell)$ оценивается сверху максимальным решением неравенства $\delta \leqslant U_\ell(\delta)$.

Как строить $U_\ell(\delta)$?

Неравенство Талаграна

Теорема

Пусть $Z_1, \ldots, Z_n$ — независимые случайные величины, принимающие значения в S. Для любого равномерно ограниченного константой $U > 0$ класса функций $\mathcal{F}\colon S \to \mathbb{R}$ и для всех $t > 0$ справедливо:

$$\mathsf{P}^{\otimes \ell}\left\{ n\left| \|P_n - P\|_{\mathcal{F}} - \mathsf{E}^{\otimes \ell}\|P - P_n\|_{\mathcal{F}} \right| \geqslant t \right\} \leqslant$$

$$\leqslant K \exp\left\{ -\frac{1}{K}\frac{t}{U}\log\left(1 + \frac{tU}{V}\right)\right\},$$

где $V = \mathsf{E}\sup\limits_{f\in\mathcal{F}} \sum\limits_{i=1}^{n} f^2(Z_i)$, *а* $K > 0$ — *некоторая константа.*

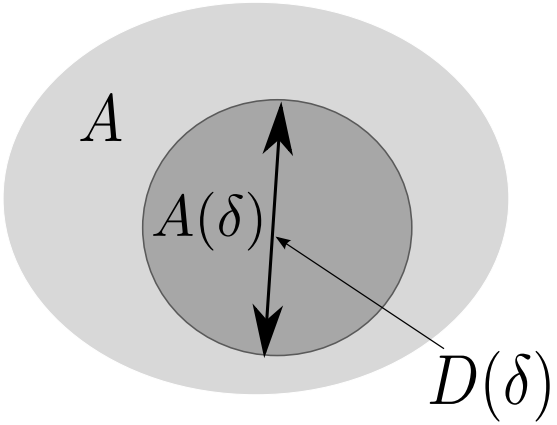
M. Talagrand. (1996) New concentration inequalities in product spaces // Inventiones Mathematicae, V, 126, pp. 505–563.

Неслучайная оценка $U_\ell(\delta)$

$$\varphi_\ell(\delta) = \mathsf{E}^{\otimes \ell} \sup_{a,f \in A(\delta)} |(P - P_\ell)(a - f)|;$$

$$D^2(\delta) = \sup_{a,f \in A(\delta)} P(a - f)^2;$$

$$U_\ell(\delta, t) = K\left(\varphi_\ell(\delta) + D(\delta)\sqrt{\frac{t}{\ell}} + \frac{t}{\ell}\right).$$



Фиксируем $\delta > 0$. Тогда для всех $t > 0$ из неравенства Талаграна [Bousquet02]:

$$\mathsf{P}^{\otimes \ell}\left\{\sup_{a,b \in A(\delta)} |(P - P_\ell)(a - b)| \geqslant U_\ell(\delta, t)\right\} \leqslant e^{-t}.$$

Локализация итерациями

Пусть $\delta_\ell^{(0)} = 1$. Тогда $A\big(\delta_\ell^{(0)}\big) = A$.

1. Используя неравенство Талаграна, получаем с большой вероятностью

$$\mathcal{E}(a^\ell) \leqslant \sup_{a,f \in A(\delta_\ell^{(0)})} |(P - P_\ell)(a - f)| \leqslant U_\ell(\delta_\ell^{(0)}, t) \stackrel{\text{def}}{=} \delta_\ell^{(1)} \sim O(1/\sqrt{\ell}).$$

2. С учетом прошлого неравенства:

$$\mathcal{E}(a^\ell) \leqslant \sup_{a,f \in A(\delta_\ell^{(1)})} |(P - P_\ell)(a - f)| \leqslant U_\ell(\delta_\ell^{(1)}, t) \stackrel{\text{def}}{=} \delta_\ell^{(2)},$$

где мы снова применили неравенство Талаграна.

И так далее...

Диаметр δ-минимального множества

$$\varphi_\ell(\delta) = \mathsf{E}^{\otimes \ell} \sup_{a,f \in A(\delta)} |(P - P_\ell)(a - f)|; \quad D^2(\delta) = \sup_{a,f \in A(\delta)} P(a-f)^2;$$

$$U_\ell(\delta, t) = K\left(\varphi_\ell(\delta) + D(\delta)\sqrt{\frac{t}{\ell}} + \frac{t}{\ell}\right).$$

Для многих задач машинного обучения $D^2(\delta) \to 0$, $\delta \to 0$:

- Задача регрессии с квадратичными потерями.
- Задача классификации $\mathbb{Y} = \{\pm 1\}$ с бинарными потерями.

В этом случае для ряда семейств A (например, VC-классов):

$$U_\ell(\delta, t) \sim o(1/\sqrt{\ell}),\ \ell \to \infty,\ \delta \to 0.$$

Одновременный учет эффектов связности и расслоения A

- Мы учли **расслоение** семейства A по значениям риска, последовательно сужая его подмножества, по которым берется супремум.
- Мы учли **схожесть** между алгоритмами семейства A с малыми значениями риска. Сделать это нам позволило неравенство Талаграна, которое учитывает диаметр этого подмножества.

Для получения оценок правильного порядка необходим одновременный учет этих эффектов.

K. V. Vorontsov. (2009) Splitting and similarity phenomena in the sets of classifiers and their effect on the probability of overfitting // PRIA, V. 19, No. 3, pp. 412–420.

Комбинаторный подход к переобучению

- Множество объектов $\mathbb{X} = \{X_1, \ldots, X_L\}$ **конечно**.
- Ответы $\{Y_1, \ldots, Y_L\}$ фиксированы.
- Все C_L^ℓ разбиений $\mathbb{X} = X^\ell \cup X^k$ на обучающую X^ℓ длиной ℓ и контрольную X^k длиной k выборки равновероятны, то есть X^ℓ вытянута из $\mathbb{X}$ **без возвращений**.

$$a = (a_i)_{i=1}^L = r\big(a(X_i), Y_i\big),\ a \in A.$$

Задача минимизации **риска**:

$$P_L a \stackrel{\text{def}}{=} \frac{1}{L}\sum_{i=1}^{L} a_i = \frac{1}{L}\sum_{i=1}^{L} r\big(a(X_i), Y_i\big) \to \min_{a \in A}.$$

К. В. Воронцов. (2011) Комбинаторная теория переобучения: результаты, приложения и открытые проблемы. // конф. ММРО-15, С. 40–43.

Комбинаторный подход к переобучению: задача

Индексы $I_\ell = \{i \in \{1,\dots,L\}\colon X_i \in X^\ell\}$ и $I_k = \{1,\dots,L\} \setminus I_\ell$.

Задача минимизации **эмпирического риска**:

$$P_\ell a \stackrel{\text{def}}{=} \frac{1}{\ell}\sum_{i\in I_\ell} a_i = \frac{1}{\ell}\sum_{i\in I_\ell} r\big(a(X_i), Y_i\big) \to \min_{a\in A}. \quad \textbf{Решение} — \ \mathbf{a^\ell}.$$

Избыточный риск $\mathcal{E}(a^\ell) = P_L a^\ell - \inf_{a\in A} P_L a$.

Задача — построение оценок

$$\mathbb{P}\{\mathcal{E}(a^\ell) \geqslant t\} \leqslant \eta(A,\ell,t), \text{ где}$$

$\mathbb{P}$ — равномерное распределение на разбиениях $\mathbb{X} = X^\ell \cup X^k$.

Проблема: неравенство Талаграна рассматривает выборку **с возвращениями**, а у нас — **без возвращений**.

Концентрационное неравенство в комбинаторном подходе

$$\text{Пусть } \sigma_A^2 = \sup_{\mathbb{X}=X^\ell \cup X^k} \frac{k}{L}\left(\frac{1}{\ell}\sup_{a\in A}\sum_{i\in I_\ell}(a_i)^2 + \frac{1}{k}\sup_{a\in A}\sum_{j\in I_k}(a_j)^2\right).$$

Рассмотрим случайную величину $Z = \sup\limits_{a\in A}|(P_L - P_\ell)a|$

и случай *счетного* класса A.

Теорема

Для любых $h > 0$ справедливо:

$$\mathbb{P}\{|Z - \mathbb{E}Z| \geqslant h\} \leqslant 2\exp\left(-\frac{\ell h^2}{16\sigma_A^2}\right).$$

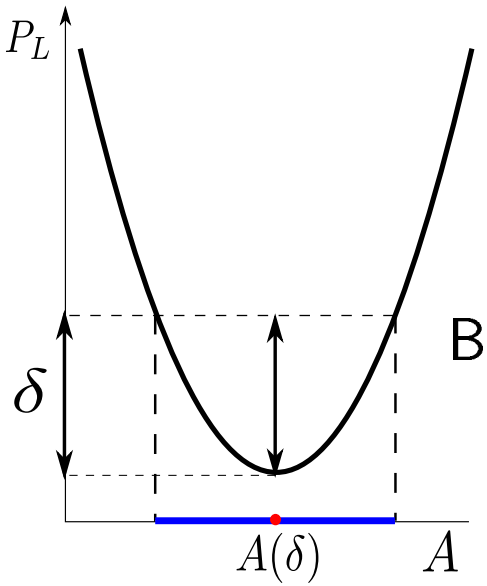
S. Bobkov. (2004) Concentration of normalized sums and a central limit theorem for noncorrelated random variables // The Annals of Probability, V. 32, N. 4, pp. 2884–2907.

Концентрационное неравенство в комбинаторном подходе

Нас будет интересовать $\sup_{a,f\in A(\delta)} |(P_L - P_\ell)(a-f)|$ и величина

$$\sigma_A^2(\delta) = \sup_{\mathbb{X}=X^\ell\cup X^k} \frac{k}{L}\left(\sup_{a,f\in A(\delta)} \sum_{i\in I_\ell} \frac{(a_i-f_i)^2}{\ell} + \sup_{a,f\in A(\delta)} \sum_{j\in I_k} \frac{(a_i-f_i)^2}{k} \right).$$

Если функция потерь бинарная и $\arg\min_{a\in A} P_L a = \bar{a}$, то:



$$A(\delta) = \left\{ a\in A\subseteq\{0,1\}^L \colon \textstyle\sum_{i=1}^L a_i \leqslant \delta + P_L\bar{a} \right\}.$$

В случае уникального решения $P_L a \to \min_{a\in A}$

$$\sigma_A^2(\delta) \to 0,\ \delta\to 0.$$

Результаты и открытые вопросы

Полученные результаты:

- Продемонстрировано, что в комбинаторном подходе возможно применение описанного метода локализации оценок.
- Получено экспоненциальное концентрационное неравенство в случае выборки без возвращений, учитывающее диаметр класса алгоритмов A.

Открытые вопросы:

- Повторение в рамках комбинаторного подхода оставшихся для получения зависящих от распределения оценок шагов и получение их вычислимых по данным аналогов.
- Экспериментальное и аналитическое сравнение оценок расслоения-связности с последними результатами SLT.