Неравенство концентрации для метода экспоненциального взвешивания

Дмитрий Островский
МФТИ ГУ

56-я научная конференция МФТИ

31 марта 2014 г.

Задача: оценить вектор $\mu \in \ell_2$, компоненты которого наблюдаются на фоне белого шума:

$$Y_k = \mu_k + \sigma\xi_k, \ \ k \in \mathbb{N},$$

где случайные величины $\xi_k \sim \mathcal{N}(0, 1)$ независимы.
Дисперсия шума σ^2 для простоты полагается известной.

Качество оценок $\hat{\mu} = \hat{\mu}(Y)$ можно измерять с помощью **квадратичного риска**

$$R(\mu, \hat{\mu}) = \mathbf{E}_\mu \|\hat{\mu} - \mu\|_2^2$$

Допустим, нам изначально задано семейство проекционных оценок вида:

$$\hat{\mu}_k^m(Y) = h_k^m Y_k,$$

где множество $h_k^m = \mathbf{1}\{k \le m\}$
Риски оценок $\hat{\mu}^m$ вычисляются очень просто:

$$R(\hat{\mu}^m, \mu) = \sigma^2 m + \sum_{k=m+1}^{\infty} \mu_k^2$$

Минимальный из этих рисков

$$r^{\mathcal{M}}(\mu) = \min_{m \in \mathbb{N}} R(\hat{\mu}^m, \mu)$$

называется **риском оракула**. Он достигается на оракульной «оценке», зависящей от неизвестного вектора μ и потому недоступной статистику.

Мы хотели бы скомбинировать из оценок $\hat{\mu}^m$ *итоговую* оценку $\overline{\mu}$ с риском, близким к риску оракула.

Достаточно естественный метод, позволяющий построить 'хорошую' оценку на основе $\hat{\mu}^m$ – их агрегирование в выпуклую комбинацию

$$\bar{\mu}^{\mathbf{w}}(Y) = \sum_{m \in \mathbb{N}} w^m \hat{\mu}^m(Y),$$

где вектор весов **w** принадлежит симплексу $\Lambda = \{w^m \geq 0, \sum_m w^m = 1\}$.

Вопрос: как правильно выбрать зависящие от наблюдений веса w^m?

[Nemirovski], [Catoni]:
Пусть у нас есть дополнительная выборка

$$Y'_k = \mu_k + \sigma \xi'_k, \quad k \in \mathbb{N}$$

с новой реализацией шума.

Можно было бы подобрать веса $w^m(Y')$, проминимизировав эмпирический риск $\bar{\mathcal{R}} = \|Y' - \bar{\mu}^{\mathbf{w}}(Y)\|^2$.

$$\|Y' - \bar{\mu}^{\mathbf{w}}(Y)\|^2 \to \min_{\mathbf{w} \in \Lambda}$$

- хотелось бы учесть нашу априорную информацию об m;
- она может быть задана в виде фиксированного вектора априорных весов $\pi \in \Lambda$.

$$\|Y' - \bar{\mu}^{\mathbf{w}}(Y)\|^2 \to \min_{\mathbf{w}\in\Lambda}$$

Пенализация

$$\|Y' - \bar{\mu}^{\mathbf{w}}(Y)\|^2 + 2\beta\sigma^2\mathcal{K}(\mathbf{w},\pi) \to \min_{\mathbf{w}\in\Lambda}$$

- штрафуем **w** за уклонение от априорных весов π;
- если π – равномерное, то $\mathcal{K}(\mathbf{w},\pi)$ превращается в $-H(\mathbf{w})$;
- параметр $\beta \geq 0$ отвечает за относительную важность априорной информации.

$$\|Y' - \bar{\mu}^{\mathbf{w}}(Y)\|^2 \to \min_{\mathbf{w}\in\Lambda}$$

Пенализация

$$\|Y' - \bar{\mu}^{\mathbf{w}}(Y)\|^2 + 2\beta\sigma^2\mathcal{K}(\mathbf{w},\pi) \to \min_{\mathbf{w}\in\Lambda}$$

- штрафуем **w** за уклонение от априорных весов π;
- если π – равномерное, то $\mathcal{K}(\mathbf{w},\pi)$ превращается в $-H(\mathbf{w})$;
- параметр $\beta \geq 0$ отвечает за относительную важность априорной информации.

Оценим сверху эмпирический риск ($\mathbf{w} \in \Lambda$ и выпуклость $\|\cdot\|^2$)

$$\|Y' - \sum_m w^m \hat{\mu}^m(Y)\|^2 \leq \sum_m w^m \|Y' - \hat{\mu}^m(Y)\|^2$$

$$\|Y' - \bar{\mu}^{\mathbf{w}}(Y)\|^2 \to \min_{\mathbf{w}\in\Lambda}$$

Пенализация

$$\|Y' - \bar{\mu}^{\mathbf{w}}(Y)\|^2 + 2\beta\sigma^2\mathcal{K}(\mathbf{w},\pi) \to \min_{\mathbf{w}\in\Lambda}$$

- штрафуем **w** за уклонение от априорных весов π;
- если π – равномерное, то $\mathcal{K}(\mathbf{w},\pi)$ превращается в $-H(\mathbf{w})$;
- параметр $\beta \geq 0$ отвечает за относительную важность априорной информации.

Оценим сверху эмпирический риск ($\mathbf{w} \in \Lambda$ и выпуклость $\|\cdot\|^2$)

$$\|Y' - \sum_m w^m \hat{\mu}^m(Y)\|^2 \leq \sum_m w^m \|Y' - \hat{\mu}^m(Y)\|^2$$

Мы приходим к оптимизационной задаче:

$$\sum_m w^m \|Y' - \hat{\mu}^m(Y)\|^2 + 2\beta\sigma^2\mathcal{K}(\mathbf{w},\pi) \to \min_{\mathbf{w}\in\Lambda}$$

$$\sum_m w^m \|Y' - \hat{\mu}^m(Y)\|^2 + 2\beta\sigma^2 \mathcal{K}(\mathbf{w}, \pi) \to \min_{\mathbf{w}\in\Lambda}$$

Простое упражнение: эта задача имеет явное решение

$$w^m \propto \pi^m \exp\left(-\frac{\|Y' - \hat{\mu}^m(Y)\|^2}{2\beta\sigma^2}\right)$$

$\|Y' - \hat{\mu}^m(Y)\|^2$ можно заменить на $\|\hat{\mu}^m(Y)\|^2 - 2\langle Y', \hat{\mu}^m(Y)\rangle$, воспользовавшись тем, что Y' не зависит от m, и перенормировав.

$$\sum_m w^m \| Y' - \hat{\mu}^m(Y)\|^2 + 2\beta\sigma^2 \mathcal{K}(\mathbf{w}, \pi) \to \min_{\mathbf{w}\in\Lambda}$$

Простое упражнение: эта задача имеет явное решение

$$w^m \propto \pi^m \exp\left(-\frac{\|\hat{\mu}^m(Y)\|^2 - 2\langle Y', \hat{\mu}^m(Y)\rangle}{2\beta\sigma^2}\right)$$

Переход к одной выборке

- используя дополнительную выборку Y', мы теряем статистическую информацию;
- проблема при переходе к одной выборке $(Y' \to Y)$ возникает в слагаемом $\langle Y, \hat{\mu}^m(Y)\rangle$.

Заметим, что

$$\langle Y', \hat{\mu}^m(Y)\rangle \approx \langle Y, \hat{\mu}^m(Y)\rangle - \sigma^2 m$$

с точностью до слагаемых с нулевым средним.

Итак, мы приходим к **экспоненциальному взвешиванию**:

$$\bar{\mu}^{\beta}(Y) = \sum_{h\in\mathcal{H}} w^m(Y)\hat{\mu}^m(Y), \quad w^m(Y) \propto \pi^m \exp\left[-\frac{r_m(Y)}{2\beta\sigma^2}\right],$$

где

$$r_m(Y) = \|\hat{\mu}^m(Y)\|^2 - 2\langle Y, \hat{\mu}^m(Y)\rangle + 2\sigma^2 m.$$

Легко проверить, что $r_m(Y)$ – несмещенная оценка риска $R(\hat{\mu}^m, \mu)$ (с точностью до постоянной, не зависящей от m).

Случай $\beta = 0$ соответствует классическому *выбору оценки с наименьшей несмещенной оценкой риска* $r_m(Y)$ *– критерию Акаике.*

Теорема (Kneip, 1994)

Для оценки $\bar{\mu}^{\beta}$ *при* $\beta = 0$ ***для всех*** $\mu \in \ell_2(1, \infty)$ *справедливо неравенство*

$$\mathbf{E}\|\bar{\mu}^{0}(Y) - \mu\|_2^2 \le r^{\mathcal{M}}(\mu) + K\sigma^2\sqrt{\frac{r^{\mathcal{M}}(\mu)}{\sigma^2}},$$

где K *– универсальная константа.*

Теорема (Kneip, 1994)

Для оценки $\bar\mu^\beta$ *при* $\beta = 0$ ***для всех*** $\mu \in \ell_2(1,\infty)$ *справедливо неравенство*

$$\mathbf{E}\|\bar\mu^0(Y) - \mu\|_2^2 \le r^{\mathcal{M}}(\mu) + K\sigma^2\sqrt{\frac{r^{\mathcal{M}}(\mu)}{\sigma^2}},$$

где K *– универсальная константа.*

Теорема (Leung & Barron, 2006)

При $\beta \ge 2$ *и* $\pi^m = 1$ ***для всех*** $\mu \in \ell_2(1,\infty)$ *выполнено*

$$\mathbf{E}\|\bar\mu^\beta(Y) - \mu\|_2^2 \le r^{\mathcal{M}}(\mu) + 2\beta\sigma^2\log(\#\ \text{оценок}),$$

Теорема (Kneip, 1994)

Для оценки $\bar{\mu}^{\beta}$ при $\beta = 0$ ***для всех*** *$\mu \in \ell_2(1, \infty)$ справедливо неравенство*

$$\mathbf{E}\|\bar{\mu}^0(Y) - \mu\|_2^2 \le r^{\mathcal{M}}(\mu) + K\sigma^2\sqrt{\frac{r^{\mathcal{M}}(\mu)}{\sigma^2}},$$

где K – универсальная константа.

Теорема (Leung & Barron, 2006)

При $\beta \ge 2$ и $\pi^m = 1$ ***для всех*** *$\mu \in \ell_2(1, \infty)$ выполнено*

$$\mathbf{E}\|\bar{\mu}^{\beta}(Y) - \mu\|_2^2 \le r^{\mathcal{M}}(\mu) + 2\beta\sigma^2 \log(\#\ \text{оценок}),$$

Теорема (Голубев, 2012)

При тех же условиях

$$\mathbf{E}\|\bar{\mu}^{\beta}(Y) - \mu\|_2^2 \le r^{\mathcal{M}}(\mu) + 2\beta\sigma^2 \log\left\{\frac{r^{\mathcal{M}}(\mu)}{\sigma^2}\left[1 + \Psi_{\beta}\left(\frac{r^{\mathcal{M}}(\mu)}{\sigma^2}\right)\right]\right\},$$

где $\Psi_{\beta}(r)$, $r \ge 1$, – ограниченная функция, $\Psi_{\beta}(r) \to 0$ при $r \to \infty$.

На первый взгляд кажется, что экспоненциальное взвешивание лучше критерия Акаике, но в действительности это не совсем так.

- Теорема Кнайпа на самом деле **концентрационная**:

$$\mathbf{P}\Big\{\|\bar{\mu}^0(Y) - \mu\|_2 \geq \sqrt{r^{\mathcal{M}}(\mu)} + x\Big\} \leq \exp\Big\{-C_1[x - C_2]_+^2\Big\},$$

  где $C_{1,2}$ – универсальные константы.

- Во второй теореме концентрация ошибки $\|\bar{\mu}^\beta(Y) - \mu\|_2^2$ вблизи риска оракула не отслеживается.

К тому же, непонятно, что происходит вблизи «абсолютного нуля», т. е. при $\beta \in (0, 2)$.

Введем избыточный риск

$$\Delta^\beta(\mu) \stackrel{\text{def}}{=} \mathbf{E}\Big[\|\bar{\mu}^\beta(Y) - \mu\|_2^2 - r^{\mathcal{M}}(\mu)\Big]_+,$$

который, в отличие от величины

$$\mathbf{E}\|\bar{\mu}^\beta(Y) - \mu\|_2^2 - r^{\mathcal{M}}(\mu),$$

уже контролирует отклонение потерь $\|\bar{\mu}^\beta(Y) - \mu\|_2^2$ агрегированной оценки от риска оракула.

Теорема

При $\beta \geq 0$, априорных весах $\pi^m = 1$ и при всех $\mu \in \ell_2(1,\infty)$ выполнено неравенство:

$$\Delta^\beta(\mu) \leq K\sigma^2\Big[r^{\mathcal{M}}(\mu) + 2\beta\sigma^2 L(\mu)\Big]^{1/2} + 2\beta\sigma^2 L(\mu),$$

$$L(\mu) = \log\left\{\frac{r^{\mathcal{M}}(\mu)}{\sigma^2}\left[1 + \Psi_\beta\left(\frac{r^{\mathcal{M}}(\mu)}{\sigma^2}\right)\right]\right\},$$

где $K > 0$ – универсальная постоянная,
$\Psi_\beta(r)$, $r \geq 1$, – ограниченная функция, $\Psi_\beta(r) \to 0$ при $r \to \infty$.

- Kneip, A. (1994). Ordered linear smoothers. *Ann. Statist.* **22** 835–866.
- Nemirovski, A. (2000). Topics in non-parametric statistics. *Lecture Notes in Math.* **1738** Springer-Verlag, Berlin.
- Catoni, O. (2004). Statistical learning theory and stochastic optimization. *Lectures Notes in Math.* **1851** Springer-Verlag, Berlin.
- Leung, G. and Barron, A. (2006). Information theory and mixing least-squares regressions. *IEEE Transactions on Information Theory* **52** 3396–3410.
- Голубев, Г.К. (2012). Экспоненциальное взвешивание и оракульные неравенства для проекционных оценок. *Проблемы передачи информации* **48** 269–280.