О ДИКИХ ТОЧКАХ*

Г. Ш. Тамасян
grigoriytamasjan@mail.ru

1 июня 2023 г.

1°. Пусть на плоскости задан набор точек $P = \{(x_1, y_1), \ldots, (x_s, y_s)\}$. Рассмотрим задачу поиска аффинной модели зависимости y_j от x_j. Как правило, эта проблема решается методом наименьших квадратов, а именно

$$\sum_{j=1}^{s} \delta_j^2 := \sum_{j=1}^{s} (ax_j + c - y_j)^2 \longrightarrow \min_{(a,c)\in\mathbb{R}\times\mathbb{R}}, \tag{1}$$

и еще реже методом наименьших модулей, т. е.

$$\sum_{j=1}^{s} |\delta_j| = \sum_{j=1}^{s} \left|ax_j + c - y_j\right| \longrightarrow \min_{(a,c)\in\mathbb{R}\times\mathbb{R}}. \tag{2}$$

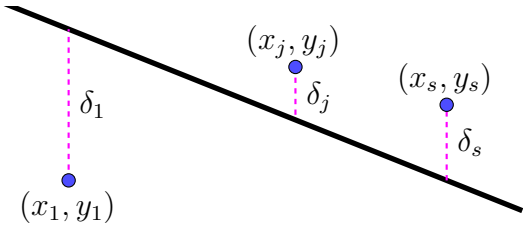


Рис. 1. Геометрическая интерпретация задач (1) и (2)

Несложно получить явное решение экстремальной задачи (1): это прямая $y = \widehat{a}x + \widehat{c}$ проходящая через центроид $(\overline{x}, \overline{y})$ набора точек P, где $\overline{x} = \frac{1}{s}\sum_{j=1}^{s} x_j$, $\overline{y} = \frac{1}{s}\sum_{j=1}^{s} y_j$,

$$\widehat{a} = \frac{\sum\limits_{j=1}^{s} x_j y_j - s\,\overline{x}\,\overline{y}}{\sum\limits_{j=1}^{s} x_j^2 - s\,\overline{x}^2}, \qquad \widehat{c} = \overline{y} - \widehat{a}\,\overline{x}. \tag{3}$$

*Семинар по оптимизации, машинному обучению и искусственному интеллекту «O&ML» http://oml.cmlaboratory.com/

Поставим вопрос об «устойчивости» точного решения $(\widehat{a}, \widehat{c})$ задачи (1) относительно изменений её данных. Рассмотрим случай, когда точки из P остаются невозмущенными, кроме одной.

Следующий пример наглядно демонстрирует чувствительность евклидовой метрики (см. задачу (1)) к возмущениям в исходных данных.

ПРИМЕР 1. Пусть P состоит из трех точек вида

$$(x_1, y_1) = (-2, -1), \quad (x_2, y_2) = (-1, 0), \quad (x_3, y_3) = (3, 0).$$

По формулам (3) получим следующее решение задачи (1): $\widehat{a} = \frac{1}{7}$, $\widehat{c} = -\frac{1}{3}$.

Если изменить ординату только второй точки, например, y_2 взять равной 2 (точка красного цвета на рис. 1), то решение «возмущенной» задачи примет вид: $\widetilde{a} = 0$, $\widetilde{c} = \frac{1}{3}$.

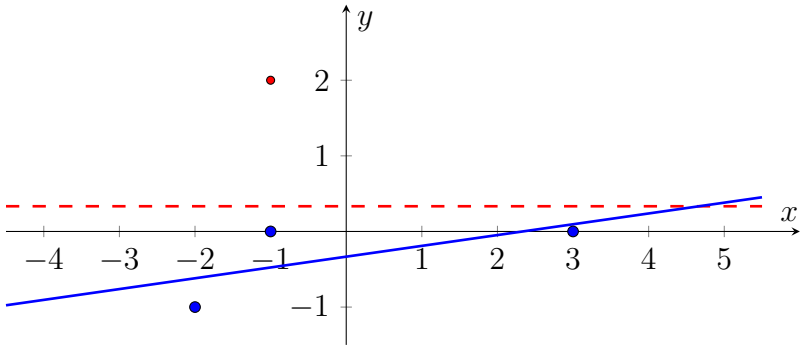
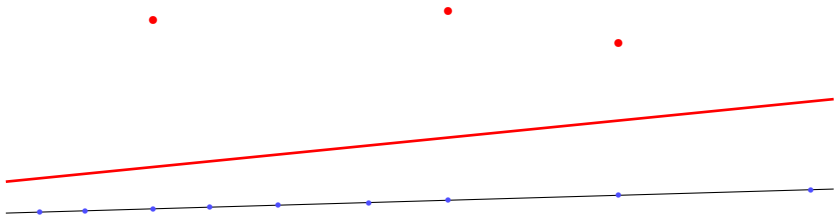


Рис. 1. $y = \frac{1}{7}x - \frac{1}{3}$ — синим цветом, $y = \frac{1}{3}$ — красный пунктир

ПРИМЕР 2. Рассмотрим пример из книги Ю.В. Линника [2, стр. 9].

x_j	0	4	10	15	21	29	36	51	68
y_j	66.7	71	76.3	80.6	85.7	92.9	99.4	113.6	125.1



$$\delta_1 \approx 0.81, \quad \delta_2 \approx -0.01, \quad \delta_3 \approx -0.09, \quad \delta_4 \approx -0.03,$$
$$\delta_5 \approx 0.09, \quad \delta_6 \approx -0.14, \quad \delta_7 \approx -0.55, \quad \delta_8 \approx -1.69, \quad \delta_9 \approx 1.61.$$

Рис. 2. Черным цветом решение исходной задачи;
красным цветом решение возмущенной задачи
при $y_3 = 576.3$, $y_7 = 599.4$, $y_8 = 613.6$

Таким образом, на примерах проиллюстрировано существенная зависимость решения от вариации во входных данных.

Далее будет показано, что решение задачи в ℓ_1-метрике (см. задачу (2)), при определенных условиях, нечувствительно к *диким точкам*, т. е. к аномальным выбросам в исходных данных.

2°. Пусть в $\mathbb{R}^2$ заданы три точки (x_j, y_j), $j \in 1:3$, такие, что $x_1 < x_2 < x_3$. Подробно исследуем задачу

$$\varphi(a, c) := \sum_{j=1}^{3} |ax_j + c - y_j| \longrightarrow \min_{(a,c)\in\mathbb{R}^2}. \tag{4}$$

УТВЕРЖДЕНИЕ 1. *Пусть $x_1 < x_2 < x_3$. Решением задачи (4) является прямая проходящая через точки (x_1, y_1) и (x_3, y_3). Оно никак не зависит от расположения точки (x_2, y_2).*

Доказательство. Покажем, что точка (x_2, y_2) лежащая между прямыми $x = x_1$ и $x = x_3$ не оказывает никакого влияния на оптимальное решение (a_*, c_*) задачи (4) (см. рис. 3).

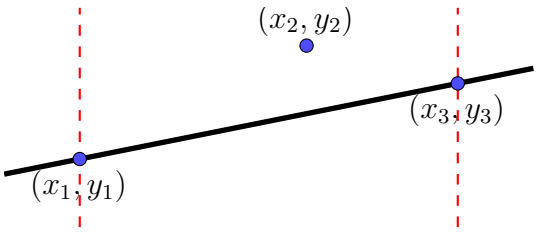


Рис. 3. Оптимальное решение $y = a_*x + c_*$

Действительно, так как целевая функция $\varphi(a, c)$ является выпуклой, а следовательно и субдифференцируемой [1], то критерием оптимальности точки (a_*, c_*) является принадлежность начало координат субдифференциальному множеству $\partial\varphi(a_*, c_*)$.

Таким образом, требуется убедиться в том, что для всех $x_2 \in (x_1, x_3)$ и любом $y_2 \in \mathbb{R}$ справедливо включение:

$$\mathbf{0} \in \partial\varphi(a_*, c_*). \tag{5}$$

Пусть a_* и c_* такие, что

$$\begin{aligned} a_*x_1 + c_* - y_1 &= 0, \\ a_*x_2 + c_* - y_2 &< 0, \\ a_*x_3 + c_* - y_3 &= 0. \end{aligned} \tag{6}$$

Случай, когда точка (x_2, y_2) лежит по другую сторону от прямой (см. рис. 3) рассматривается аналогично.

Положим $f_j(a,c) = |ax_j + c - y_j|$, $j \in 1:3$. В силу (6), имеем

$$\partial\varphi(a_*,c_*) = \sum_{j=1}^{3} \partial f_j(a_*,c_*) =$$

$$= \operatorname{conv}\left\{\begin{pmatrix}-x_1\\-1\end{pmatrix},\begin{pmatrix}x_1\\1\end{pmatrix}\right\} + \left\{\begin{pmatrix}-x_2\\-1\end{pmatrix}\right\} + \operatorname{conv}\left\{\begin{pmatrix}-x_3\\-1\end{pmatrix},\begin{pmatrix}x_3\\1\end{pmatrix}\right\}.$$

Отметим, что субдифференциальное множество $\partial\varphi(a_*,c_*)$ не зависит от y_j.

Для наглядности проверки включения (5), перенесем в левую его часть второе слагаемое из представления $\partial\varphi(a_*,c_*)$, а оставшиеся два отрезка сложим. Получим (см. рис. 4)

$$\left\{\begin{pmatrix}x_2\\1\end{pmatrix}\right\} \in \operatorname{conv}\left\{\begin{pmatrix}-x_1-x_3\\-2\end{pmatrix},\begin{pmatrix}x_3-x_1\\0\end{pmatrix},\begin{pmatrix}x_1-x_3\\0\end{pmatrix},\begin{pmatrix}x_1+x_3\\2\end{pmatrix}\right\}. \qquad (7)$$

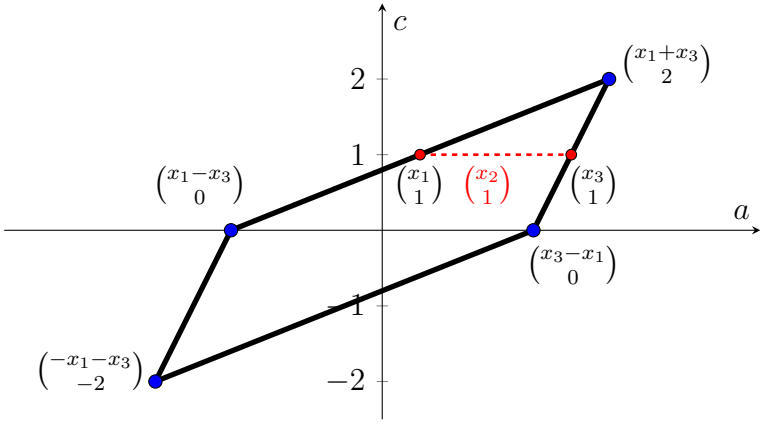


Рис. 4. Геометрическая интерпретация условия (7) при $x_1 > 0$

Несложно понять, что для всех $x_2 \in (x_1, x_3)$ и любом y_2 включение (7) выполнено, а следовательно справедливо и включение (5). □

На рис. 4 можно заметить, что критерий оптимальности (7) также выполняется, если x_2 принимает и граничные значения интервала (x_1, x_3).

УТВЕРЖДЕНИЕ 2. *Если $x_2 = x_1$, то решением задачи (4) является прямая проходящая через любую точку отрезка* $\operatorname{conv}\{(x_1,y_1),(x_1,y_2)\}$ *и точку (x_3,y_3). В случае $x_2 = x_3$ — прямая проходящая через точку (x_1,y_1) и любую точку отрезка* $\operatorname{conv}\{(x_3,y_2),(x_3,y_3)\}$.

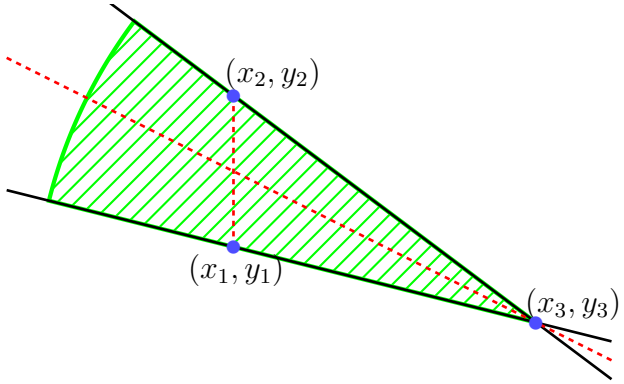


Рис. 4. Множество решений задачи (4) при $x_2 = x_1$

3°. Решение задача (4) имеет наглядное геометрическое обоснование. Пусть $x_1 < x_2 < x_3$. На рис. 5 изображены прямые проходящие через каждые две точки. Ясно, что длины пунктирных отрезков равны значениям целевой функции $\varphi(a, c)$ для соответствующей прямой $y = ax + c$. Также очевидно, что «центральный» отрезок всегда будет короче двух крайних. Таким образом, прямая проходящая через крайние точки подозрительная на оптимальность.

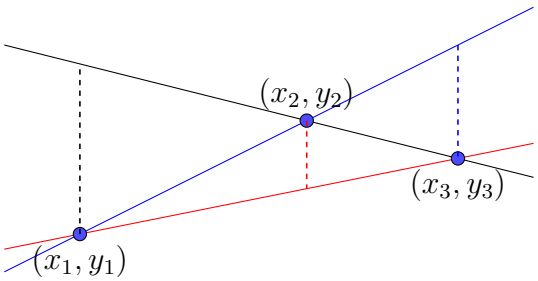


Рис. 5. Геометрическое решение

На рис. 6 изображены возможные «вариации» предполагаемого решения, на которых видно, что соответствующие им суммы невязок δ_j больше, чем до прямой проходящая через точки (x_1, y_1) и (x_3, y_3).

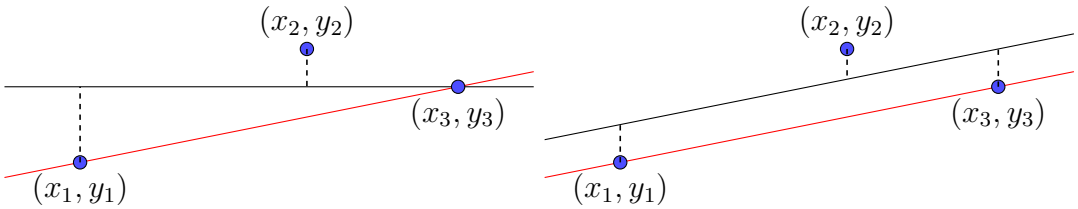


Рис. 6. Обоснование

Работа выполнена в Институте проблем машиноведения РАН при финансовой поддержке Российского научного фонда (проект № 23-41-00060).

ЛИТЕРАТУРА

1. Тамасян Г. Ш. *Элементы субдифференциального исчисления* // Семинар «O & ML». Литература. Базовые знания. (`http://oml.cmlaboratory.com/BasicKnowledge.shtml`)

2. Линник Ю.В. *Метод наименьших квадратов и основы математико-статистической теории обработки наблюдений*. М.: ФМЛ, 1958. 333 с.