

Лекція 5. РОБОТНІ МЕТОДИ ОБРОБКИ ДАНИХ

Класичні методи математичної статистики розроблені для тих випадків, коли експериментальні дані задовольняють суворим статистичним моделям і є оптимальними для цих моделей. Однак більшість з них чутлива до наявності викидів або промахів, до відхилень розподілу від гаусівського і т. п. У практичних завданнях суворі моделі нерідко порушуються і тому бажано використовувати такі методи, які слабо залежать від порушення введених припущень, проте дають досить високу точність оцінки. Такі методи, звані робастними чи стійкими [37, 38], активно розробляються в сучасній статистиці і все більш широко використовуються на практиці.

Робастні оцінки будуються таким чином, щоб їх властивості залишалися хорошими навіть у тому випадку, коли справжнє розподілення експериментальних даних відрізняється від передбачуваного. Так для багатьох наборів даних більш реалістичною є не строга модель гаусівського розподілу, а розподіл з більш "важкими" хвостами (коли ймовірності великих відхилень від середнього істотно більше, ніж при розподілі гауса). Найпростішою з таких моделей є, запропонована Т'юкі, модель засмічення (або модель грубих помилок) : розподіл вибірки представляють у вигляді суміші основного розподілу гауса $\varphi(x)$ з іншим (засмічуючим) розподілом $h(x)$:

$$p_{\varepsilon}(x) = (1 - \varepsilon)\varphi(x) + \varepsilon h(x) \quad (5.1)$$

Розподіл $h(x)$ може бути або гаусівським розподілом із значно більшою дисперсією, ніж $\varphi(x)$ (часто при дослідженні оцінок приймають відношення СКО $\sigma_h/\sigma_{\varphi} = 3$ або 10), або $h(x)$ відрізняється від гаусівського розподілу та має затягнуті хвости. Більшість робастних методів розробляється і досліджується саме на таких простих моделях, і вже таке розширення моделей дає можливість врахувати багато важливих особливостей «некласичних» ситуацій та суттєво наблизити методи оцінки до властивостей реальних даних.

Переходити до робастному оцінювання слід за розподілі вибірки з довгими хвостами, ніж в гаусівського розподілу. Досить зручно про довжину хвоста судити за величиною ексцесу $E(x)$ вибірки. Для нормального розподілу $E(x) = 0$, для розподілу з довгими хвостами $E(x) > 0$. Наприклад, за наведеними у табл. 4.4. даними, для розподілу Лапласа $E(x) = 3$, а для розподілу Стюдента $E(x) = 6/(k-4)$, де k – число ступенів свободи.

Робастні оцінки повинні відповідати двом основним вимогам:

мало поступатися ефективності оптимальним оцінкам у виконанні основний моделі, коли розподіл справді є гаусівським;

залишатися досить хорошими при відхиленнях від основної моделі, коли розподіл на відміну від гаусівського.

Для визначеності далі основні поняття та ідеї робастних методів викладаються на прикладі оцінки середнього за випадковою вибіркою $x_1, \dots, x_n$, тобто стосовно обробки даних при прямих вимірах з багаторазовими спостереженнями. Найчастіше використовується класична оцінка – середнє арифметичне результатів спостережень (2,29). Ця оцінка є найкращою при гаусівському розподілі; однак при відхиленнях від нього або наявності викидів її властивості різко погіршуються: оцінка втрачає ефективність і залежить від викидів.

Найпростіші стійкі оцінки середнього ґрунтуються на використанні впорядкованої вибірки $x'_1 < x'_2 < \dots < x'_n$. Усічені середні отримують,

$$\bar{x}(\alpha) = \frac{1}{n-2m} \sum_{i=m+1}^{n-m} x'_i, \quad (7.2)$$

відкидаючи по $m = [n\alpha]$ крайніх елементів зліва і справа впорядкованої вибірки, а потім середня залишки, що залишилися: Граничним випадком усереднених середніх при $\alpha \rightarrow 0,5$ є вибіркова медіана

$$\text{med} = \begin{cases} x'_{m+1} & \text{при } n = 2m + 1 \\ \frac{1}{2}(x'_m + x'_{m+1}) & \text{при } n = 2m \end{cases} \quad (7.3)$$

При $\alpha \rightarrow 0$ маємо середнє: $\bar{x}(0) = \bar{x}$.

До усічених середніх близькі також оцінки Віндзора [34] x_w , в яких крайні члени не відкидають, а замінюють на найближчі до них члени, що залишилися:

$$x_w(\alpha) = \frac{1}{n} \left\{ (m+1)(x'_{m+1} + x'_{n-m}) + \sum_{i=m+2}^{n-m-1} x'_i \right\} \quad (7.4)$$

Найпростішою кількісною характеристикою стійкості оцінок є точка зриву δ^* , яка визначається як максимальна частка вихідних даних, які можна довільно змінювати і при цьому не одержувати неконтрольованих похибок оцінки. Середнє арифметичне $\bar{x}$ є нестійкою оцінкою і має точку зриву $\delta^* = 0$. Для усіченого середнього $\bar{x}(\alpha)$ точка зриву дорівнює параметру усічення ($\delta^* = \alpha$), тому що більше усічення, то стійкіша оцінка. Медіана є дуже стійкою, надійною оцінкою і має максимальну точку зриву $\delta^* = 0.5$, тому її часто використовують як перше наближення в більш складних і точних алгоритмах. Проте її ефективність невелика.

Найбільш розробленим методом побудови робастних оцінок є узагальнений метод максимальної правдоподібності, в якому оцінки виходять із умови

$$\min Q(\Psi) = \sum_1^n \Psi(x_i - a) \quad (7.5)$$

де a – математичне очікування випадкової величини.

При цьому вагова функція $\psi(x)$ вибирається так, щоб при великих $|x|$ зростала повільніше, ніж квадратична $|x|^2$, але при малих $|x|$ була близька до $|x|^2$. Тому М-оцінка, що визначається умовою (7.5), слабо залежить від викидів

(промахів) і “обтяжених” хвостів розподілі, але при гауссівському розподілі вибірки близька до середнього $\bar{x}$.

Аналогічний підхід можна застосувати і в складніших завданнях обробки даних. Так, при побудові функціональних залежностей прямим узагальненням методу найменших квадратів, який є нестійким, також будуть

М-оцінки, в яких замість квадратичної використовують функції ψ з переліченими вище властивостями.

Оцінки найменших квадратів зручні тим, що виражаються у вигляді і досить прості. На жаль, М-оцінки зазвичай не виражаються у явному вигляді, навіть при оцінці середнього. У цьому випадку вони є рішеннями рівнянь

$$\sum_{i=1}^n \varphi(x_i - a) = 0 \quad (7.6)$$

де $\varphi(x) = \Psi'(x)$, на вирішення яких зазвичай використовують ітераційні методи.

Найбільш поширеними на практиці є М-оцінки Хьюбера (Ψ_H), Хампеля (Ψ_{HA}), Андрюса (Ψ_A) і Т'юкі (Ψ_T) [35, 36]; відповідні похідні вагових функцій Ψ_H , Ψ_{HA} , Ψ_A і Ψ_T (φ_H , φ_{HA} , φ_A і φ_T) наведені в табл. 7.1. Зазначимо, що ці функції містять постійні c , які мають характер параметра усічення. Параметр c пов'язаний з часткою засміченості ε розподілу вибірки і зробити вибір не завжди вдається; на практиці рекомендується використовувати значення c , наведені у табл. 7.1.

Таблиця 7.1

Оцінка	Похідні вагових функцій	Рекомендуємі значення параметрів
Хьюбера	$\varphi_H(x) = \begin{cases} -c, & x < -c; \\ x, & x \leq c; \\ c, & x > c \end{cases}$	$1 < c < 2$ $c = 1,5$
Хампеля	$\varphi_{HA}(x) = \text{sign } x \begin{cases} x , & x < a; \\ a, & a < x < b; \\ a(c - x) / (c - b), & b < x < c; \\ 0, & x > c \end{cases}$	$a = 1,7$ $b = 3,4$ $c = 8,5$
	$\left\{ \right.$	

Андрюса	$\varphi_A(x) = \begin{cases} \sin(x/c), & x \leq c\pi \\ 0, & x > c\pi \end{cases}$	$c = 2,1$
Т'юкі	$\varphi_T(x) = \begin{cases} x [1 - (x/c)^2]^2, & x \leq c; \\ 0, & x > c \end{cases}$	$\begin{matrix} 3 < c < 6 \\ c = 4.7 \\ \text{(при відомому} \\ \sigma) \end{matrix}$

В табл. 7.1. аргумент функції $\varphi(x)$ масштабують:

$$x = \frac{x_i - a_l}{S} \quad (7.7)$$

де x_i – елемент упорядкованої вибірки;

a_l – оцінка параметра зсуву, зазвичай беруть за початковий параметр зсуву медіану;

S – оцінка параметра масштабу (стала оцінка СКО).

Найбільший інтерес представляє оцінка Хьюбера, для якої функція Ψ_H в області $|x| \leq c$ є квадратичною, тобто відповідає гаусівському розподілу, а в області $|x| > c$ — лінійної, тобто. відповідає експоненційному розподілу, який має суттєво «обтяжені» хвости порівняно з гаусівським розподілом. Оцінка Хьюбера є рішенням мінімаксної задачі: максимальна асимптотична дисперсія оцінки на класі засмічених гаусівських розподілів для неї досягає мінімуму. Вона досить стійка до відхилень від розподілу гауса і має точку зриву. $\delta^* \approx 0,1$.

Для вирішення рівняння (7.6), що визначає М-оцінку, найчастіше застосовують ітераційну схему Гауса - Ньютона:

$$a_{l+1} = a_l + \frac{\sum_{i=1}^n \varphi(x_i - a_l)}{\sum_{i=1}^n \varphi'(x_i - a_l)} \quad (7.8)$$

Зокрема, при використанні функції Хьюбера Ψ_H

$$a_{l+1} = a_l + \frac{\sum_{i=k+1}^{n-k} (x_i - a_l)}{n - 2k} \quad (7.9)$$

і сума в знаменнику правої частини рівняння (7.8) дорівнюватиме кількості тих членів вибірки x_i , для яких виконано умову $|x_i - a_l| \leq c$; якщо це виконується для всіх x_i , то оцінка $a_{l+1} = \bar{x}$.

Тому переважними переважно випадків є робастні процедури, оскільки вони допускають плавний перехід від повної вибірки до суттєвого скорочення даних з урахуванням властивостей конкретної вибірки. Ця перевага стає особливо явною при переході до більш складних завдань, наприклад побудови емпіричних залежностей. Таким чином, необхідно якомога ширше використовувати робастні методи у вимірювальних задачах, для яких властивості стійкості результатів особливо важливі.

Як початкове наближення для ітераційної схеми доцільно приймати медіану вибірки: $a_0 = \text{med}$, або іншу, настільки ж просту і стійку оцінку. У багатьох практичних задачах буває достатньо виконати всього 1 ... 3 ітерації відповідно до рівняння (7.8), щоб отримати задовільні оцінки.

Інша, проста та зручна ітераційна схема для вирішення рівняння (7.6) виходить, якщо привести це рівняння до виду, аналогічного методу найменших квадратів з вагами

$$\sum_{i=1}^n \omega_{il} (x_i - a_l) = 0 \quad (7.10)$$

де множники $\omega_{il} = \varphi(x_i - a_l) / (x_i - a_l)$ аналогічні терезам у методі найменших квадратів, але обчислюються на кожному кроці ітерації. Тоді ітераційна схема має вигляд

$$a_{l+1} = \frac{\sum_{i=1}^n \omega_{il} x_i}{\sum_{i=1}^n \omega_{il}} \quad (7.11)$$

Вихідне рівняння (7.6) визначає М-оцінки при відомому значенні СКО (можна вважати $\sigma = 1$). Якщо ж СКО невідомо, то оцінки слід шукати як рішення рівняння

$$\sum_{i=1}^n \varphi[x_i - a] / S = 0, \quad (7.12)$$

де S — оцінка СКО, береться кратній медіані відхилень від медіани

$$S_1 = \text{med}\{|x_i - \text{med}\{x_i\}|\} / 0,675 \quad (7.13)$$

(де через $\text{med}\{x_i\}$ позначається медіана вибірки $\{x_i\}$) або ж за вибірковими квантилями

$$S_2 = (x'_{n-k} - x'_k) / 1,35, \quad k = [n/4]. \quad (7.14)$$

Ітераційні схеми для вирішення рівняння (7.12) мають вигляд, аналогічний (7.8) та (7.11), але із заміною $x_i - a_1$ на $(x_i - a_1)/S$.

До M -оцінок близькі за властивостями I^p - оцінки, що отримують з умови (7.5) з функцією $\Psi_p(x) = |x|^p$, де $1 \leq p < 2$. Однак вони відрізняються від M - оцінок тим, що функція Ψ_p не повинна бути близькою до квадратичної поблизу 0, і тому оцінки виявляються найбільш ефективними для розподілів, відмінних від гаусівських. Проте вони досить стійкі поблизу розподілу Гауса.

Добре відомою є оцінка найменших модулів, яка виходить із умови

$$\min \sum_1^n |x_i - \hat{x}_i(a_1, a_2, \dots)| \quad (7.15)$$

і є оптимальною при експонентному розподілі вибірки. Для середнього значення оцінка найменших модулів збігається з вибірковою медіаною; при побудові функціональних залежностей такі оцінки знаходять методами лінійного програмування.

Також є деякі інші види оцінок. Наприклад, можна відкидати крайні точки вибірки за спеціальними правилами (частка усічення а не фіксується заздалегідь, а вибирається в залежності від властивостей вибірки, як це робиться в адаптивних оцінках Хогга). Можна використовувати ранги R_i тобто. номери членів у впорядкованій вибірці, або зростаючі функції рангів та визначити оцінки з умови

$$\min Q(R) = \sum_{i=1}^n R_i(x_i' - a) \quad (7.16)$$

Крім того, можна використовувати різні адаптивні оцінки, засновані на перерахованих вище робастних оцінках, і лінійні комбінації оцінок.

Нині робастний підхід перетворився на добре розроблену область статистики. Запропоновано численні та різноманітні оцінки, які поєднують високу ефективність при основних моделях з достатньою стійкістю по відношенню до відхилень вихідних даних від моделей. Для багатьох оцінок досліджено точність (стосовно різних моделей) і розроблено апарат статистичних висновків.

Хоча з наведених ітераційних схем може здатися, що робастні методи досить трудомісткі, але насправді їх трудомісткість не дуже велика. Наприклад, у більшості випадків немає необхідності точно вирішувати рівняння (7.6), що визначає М-оцінку; достатньо виконати всього 2...3 ітерації згідно з виразами (7.8) або (7.11), щоб отримати задовільний наближення, або переконатися, що початкова оцінка практично не змінилася.

Для основних видів робастних оцінок також є оцінки похибок. Як правило, аналітичне дослідження М-оцінок ведеться стосовно «моделі засмічення», тобто передбачається, що розподіл вибірки описується сумішшю ре виду (7.1). Тоді дисперсія М-оцінки з ваговою функцією φ має вигляд:

$$D\hat{a} = \frac{\sigma^2}{n} \frac{M_\varepsilon \varphi^2}{(M_\varepsilon \varphi')^2}, \quad (7.17)$$

$$\text{де } M_\varepsilon f = \int f(u) p_\varepsilon(u) du.$$

Як зазначалося, головна мета використання робастних методів - виключити вплив грубих похибок і промахів, до яких такі чутливі класичні методи. Однак у класичній статистиці також є методи виділення промахів, тобто можна використовувати двоступінчасту процедуру:

виділити та виключити грубі похибки та промахи;

застосувати до «відредагованих» даних, що залишилися, класичну процедуру оцінювання.

Корисно зіставити наведену двоступінчасту процедуру з використанням одного з робастних методів стосовно вихідних даних. Дослідження показують, що кращою є, як правило, робастна процедура і з яких причин:

1) при використанні двоступінчастої процедури не завжди вдається чітко розмежувати етапи, особливо для складних завдань; наприклад, при побудові емпіричних залежностей грубі похибки важко виявити з урахуванням класичних оцінок;

2) навіть ретельно «відредагована» вибірка, як правило, буде мати розподіл, що відрізняється від гаусівського; тому застосування класичних методів таки залишається не цілком обґрунтованим;

3) більшість процедур виділення і виключення грубих похибок поступається ефективності робастним процедурам; крім того, вони можуть бути досить трудомісткими.

Тому переважними переважно випадків є робастні процедури, оскільки вони допускають плавний перехід від повної вибірки до суттєвого скорочення даних з урахуванням властивостей конкретної вибірки. Ця перевага стає особливо явною при переході до більш складних завдань, наприклад побудови емпіричних залежностей. Таким чином, необхідно якомога ширше використовувати робастні методи у вимірювальних задачах, для яких властивості стійкості результатів особливо важливі.

У тих випадках, коли з тих чи інших причин доцільно орієнтуватися на класичні методи, рекомендується паралельно застосовувати один із простих робастних методів; виконати 2...3 ітерації для контролю результатів та виявлення відхилень від моделі. Якщо немає істотної розбіжності між двома результатами, можна зробити звичайні статистичні висновки з урахуванням класичного підходу. В іншому випадку необхідно ретельніше проаналізувати вихідні дані і або виявити і відкинути промахи (а потім знову застосувати класичну процедуру), або орієнтуватися на один з більш точних робастних методів, для яких є оцінки похибок. Така процедура забезпечить достовірність отриманих висновків і дозволить уникнути грубих помилок.