Київський національний університет імені Тараса Шевченка
Кафедра теорії ймовірностей, статистики та актуарної математики

Р. Майборода

Непараметрична статистика

Рекомендації по виконанню індивідуальних робіт

Київ — 2017

Робота 1. Статистика зміщених вибірок

Частина 1. Оцінка функції розподілу за зміщеною вибіркою

Вибіркова процедура зветься зміщеною (biased sampling), якщо для всіх об'єктів з генеральної сукупності ймовірність потрапити до вибірки залежить від значення досліджуваної характеристики. Позначимо об'єкт літерою O, а його досліджувану характеристику — $\xi(O)$. Тоді

$$\mathsf{P}\{O \text{ потрапить до вибірки} \mid \xi(O) = t\} = cw(t),$$

де c — невідома константа, $w(t)$ — відома зміщуюча функція.

Нехай F — функція розподілу $\xi(O)$ у генеральній сукупності. У лабораторній роботі розглядається випадок, коли оцінювання F потрібно провести за двома вибірками:

(1) незміщена вибірка, до якої всі об'єкти потрапляють з однаковими ймовірностями;

(2) зміщена вибірка із заданою зміщуючою функцією w.

За незміщеною вибіркою $\xi_1, \ldots, \xi_{n_1}$ F можна оцінити, використовуючи звичайну емпіричну функцію розподілу:

$$\hat{F}_{emp}(x) = \frac{1}{n_1} \sum_{j=1}^{n_1} \mathbb{1}\{\xi_j < x\}$$

За зміщеною вибіркою $\eta_1, \ldots, \eta_{n_2}$ можна оцінити F використовуючи оцінку Горвіца-Томпсона (Horwitz-Thompson, see Shao, p. 199):

$$\hat{F}_{HT}(x) = \frac{\sum_{j=1}^{n_2} \frac{\mathbb{1}\{\eta_j < x\}}{w(\eta_j)}}{\sum_{j=1}^{n_2} \frac{1}{w(\eta_j)}}$$

За двома вибірками разом можна побудувати оцінку двома способами:

(a) Як лінійну комбінацію оцінок по зміщеній та незміщеній вибірках:

$$\hat{F}_{mix}(x) = \frac{1}{2}(\hat{F}_{emp}(x) + \hat{F}_{HT}(x))$$

(можна брати також лінійну комбінацію з ваговими коефіцієнтами не рівними $1/2$, а обраними з міркувань мінімізації дисперсії отриманої оцінки. Розробка відповідної теорії та реалізація алгоритму можуть бути хорошим додатковим завданням для тих, хто достроково виконає основну частину завдання роботи).

(б) за допомогою емпіричного методу найбільшої вірогідності, який приводить до оцінки Варді (Vardi, see Shao, p. 328):

$$\hat{F}_V(x) = \sum_{j=1}^{n_1+n_2} p_j \mathbb{1}\{\zeta_j < x\},$$

де $\zeta_1, \ldots, \zeta_{n_1+n_2}$ — об'єднання незміщеної та зміщеної вибірок,

$$p_j = \frac{W}{\lambda + n_2 w(\zeta_j)},$$

$$W = \left(\sum_{j=1}^{n_1+n_2} \frac{1}{\lambda + n_2 w(\zeta_j)} \right)^{-1},$$

Число λ знаходиться як додатній корінь рівняння

$$\sum_{j=1}^{n_1+n_2} \frac{w(\zeta_j)}{\lambda + n_2 w(\zeta_j)} = 1$$

Завдання першої частини роботи:
Для заданих функції розподілу F та зміщуючої функції w згенерувати незміщену вибірку обсягу $n_1 = 300$ та зміщену вибірку обсягу $n_2 = 300$, побудувати за ними чотири перелічені вище оцінки для F і вивести їх графіки разом із справжньою функцією F. Зробити попередні висновки щодо того, яка оцінка точніша (на око).

Частина 2. Оцінювання параметрів та характеризація якості оцінок функції розподілу Всі оцінки, описані вище, можна зобразити у вигляді

$$\hat{F}_k(x) = \sum_{j=1}^{n_1+n_2} p_j^{(k)} \mathbb{1}\{\zeta_j < x\},$$

де k — тип оцінки (емпірична функція розподілу, оцінка Горвіца-Томпсона, лінійна комбінація, оцінка Варді), $p_j^{(k)}$ — вагові коефіцієнти для оцінки k-того типу. Використовуючи ці оцінки для функції розподілу F, можна оцінювати різні функціонали від F. Наприклад, якщо потрібно оцінити

функціональний момент, $\bar{g} = \mathsf{E}\, g(\xi_j) = \int g(x)F(dx)$, то відповідна оцінка може бути навантаженим емпіричним моментом:

$$\hat{g}^{(k)} = \int g(x)\hat{F}_k(dx) = \sum_{j=1}^{n_1+n_2} p_j^{(k)} g(\zeta_j).$$

Так можна оцінювати математичне сподівання та дисперсію розподілу F.

Для оцінки квантилів (випадок неперервної, строго зростаючої F): $Q^F(\alpha) = F^{-1}(\alpha) = \sup\{t:\ F(t) < \alpha\}$ можна використовувати навантажені емпіричні квантилі:

$$\hat{Q}^{(k)}(\alpha) = \sup\{t:\ \hat{F}_k(t) < \alpha\}.$$

Так можна оцінювати медіани та інтерквартильний розмах.

Для характеризації якості оцінки $\hat{F}(x)$ для F, можна використати рівномірну відстань:

$$d_\infty(\hat{F}, F) = \sup_{t\in\mathbb{R}} |\hat{F}(t) - F(t)|,$$

L_q — відстань:

$$d_q(\hat{F}, F) = \left(\int |\hat{F}(t) - F(t)|^q F(dt) \right)^{1/q}.$$

Завдання другої частини роботи:

(а) Розробити алгоритми оцінки для заданого функціоналу (варіанти: математичного сподівання, дисперсії, медіани, інтерквартильного розмаху) на основі всіх чотирьох видів оцінки функції розподілу, отриманих у першій частині роботи. Ці алгоритми реалізувати у вигляді функцій R. Порівняти якість отриманих оцінок на модельованих вибірках, використовуючи такі характеристики якості оцінок, як зміщення, дисперсія, інтерквартильний розмах.

або

(б) На модельованих вибірках підрахувати одну з відстаней d_∞, d_1 або d_2 між оцінкою та оцінюваною функцією розподілу і порівняти середні відстані для чотирьох типів оцінок.

Конкретні умови другої частини роботи отримати у викладача.

Індивідуальні варіанти функції розподілу та зміщуючої функції:

(1) $F \sim N(1,1)$, $w(t) = 1/(1+\exp(t-1)$,

(2) $F \sim \chi^2(3)$, $w(t) = 1/(t+1)$,

(3) F — трикутний розподіл на [0,2], $w(t) = 1 - t/2$,

(4) $F \sim \mathrm{Exp}_{\lambda=1}$, $w(t) = (1+\cos(t))/2$,

(5) $F \sim \chi^2(4)$, $w(t) = 1 - 0.5/(1+x)$,

(6) F — T-Стьюдента з 6-ма ступенями вільності, $w(t) = 1/(1+\exp(t-1)$,

(7) F — бета-розподіл з параметрами $\alpha = 2$, $\beta = 3$, $w(t) = \cos(t)$.

(8) F — розподіл Лапласа з середнім 0 та інтенсивністю 1, $w(t) = 1 - 0.5 * \exp(-t^2)$,

(9) $F \sim \mathrm{Exp}_{\lambda=0.5}$, $w(t) = 1/(t+1)$,

(10) F — рівномірний розподіл на [0,1], $w(t) = 1 - t^2$.

Робота 2. Непараметричне оцінювання розподілів

У роботі три основних варіанти завдань, що відповідають різним моделям та задачам оцінювання.

Варіант (а). Оцінка функції розподілу за цензурованою вибіркою

Розглядається модель спостережень з випадковим цензуруванням з права. У цій моделі вважається, що для кожного (j-того, $j = 1, \dots, n$) спостереження існує змінна, яка досліджується Z_j, та момент цензурування Y_j. Ці змінні не спостерігаються окремо. Спостереження складаються з $X_j = \min(Z_j, Y_j)$, (відцензуровані спостереження) та $\delta_j = \mathbb{1}\{Z_j < Y_j\}$ (індикатори відсутності цензурування).

У стандартній теорії вважається, що Z_j, Y_j, $j = 1, \dots, n$ — незалежні в сукупності невід'ємні випадкові величини, Z_j має функцію розподілу F, Y_j — функцію розподілу G.

Задача полягає в оцінці функції розподілу змінної, що досліджується — F. Якщо функція розподілу цензора G — відома, то можна застосувати оцінку Горвіца-Томпсона:

$$\hat{F}^{GT}(x) = \frac{\sum_{j=1}^{n} \mathbb{1}\{X_j < x\}\delta_j/(1 - G(X_j))}{\sum_{j=1}^{n} \delta_j/(1 - G(X_j))}$$

Оцінку Каплана-Мейера можна використовувати і тоді, коли розподіл цензора невідомий. Для її підрахунку впорядкуємо спостереження у порядку зростання X_j — $(X_{[1]}, \tilde{\delta}_1)$,$(X_{[2]}, \tilde{\delta}_1)$,…,$(X_{[n]}, \tilde{\delta}_1)$. (Тобто $X_{[1]} < X_{[2]} < \dots < X_{[n]}$, $\tilde{\delta}_j$ — це δ_j, переставлені "разом з X_j").

Тоді

$$\hat{F}^{KM}(x) = 1 - \prod_{X_{[i]} \le x} \left(1 - \frac{\tilde{\delta}_i}{n - i + 1}\right)$$

(формула правильна лише для випадку, коли у вибірці немає однакових значень X_j).

Завдання роботи: згенерувати цензуровану вибірку заданого обсягу n, із заданим розподілом досліджуваної величини F та цензора G. За цією вибіркою оцінити F оцінками Горвіца-Томпсона та Каплана-Мейера. Вивести на одному графіку оцінювану функцію та її оцінки. Зробити висновки про те, яка оцінка краща (на око).

(Як додаткове завдання можна визначити середнє відхилення оцінок від оцінюваної функції в одній з відстаней, що розглядались у роботі 1).

Варіант (b). Оцінка функції розподілу за даними поточного стану

Розглядається модель спостережень даних поточного стану: досліджувана випадкова велична X_j (це час, коли для спостережуваного об'єкту відбулась певна подія A) не спостерігається, спостереження складаються з T_j і $\Delta_j = \mathbb{1}\{X_j < T_j\}$, $j = 1, \ldots, n$. Тут T_j трактують як момент спостереження, Δ_j — індикатор поточного стану: 1, якщо на момент спостереження подія вже A відбулась, 0 — якщо іще не відбулась.

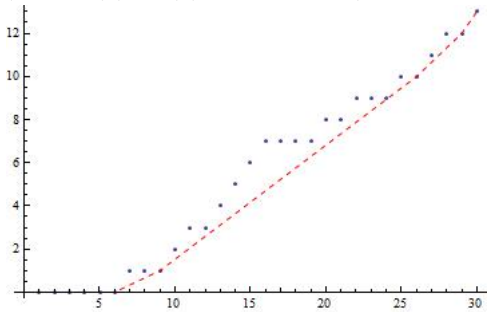
У стандартній моделі X_j, T_j — незалежні в сукупності, невід'ємні випадкові величини, функція розподілу досліджуваної характеристики X_j — F, моменту спостереження T_j — G. Функції F та G невідомі, потрібно оцінити F.

Оцінка непараметричного методу найбільшої вірогідності будується наступним чином.

Впорядковуємо пари (T_j, Δ_j) у порядку зростання T_j — отримуємо $(T_{[j]}, \tilde{\Delta}_j)$, $j = 1, \ldots, n$. Розглядаємо діаграму (i, S_i), $S_i = \sum_{j \le i} \tilde{\Delta}_j$, $i = 1, \ldots, n$ та її опуклу міноранту

$$H^*(t) = \sup\{H(t) \,:\, H(0) = 0,\ H(i) \le S_i, \forall i = 1, \ldots, n,\ H(t) \text{— опукла функція}\}$$

Позначимо w_j — похідна зліва функції $H(t)$ у точці $t = i$. (Геометрично, $(t, H(t)), t \ge 0$, це ламана, точками зламу якої є деякі з точок (i, S_i), причому множина точок, що лежать вище $(t, H(t)), t \ge 0$ є опуклою. w_j - це кутовий коефіцієнт відповідної ланки цієї ламаної):



точки — (i, S_i), ламана — $H^*(t)$.

Оцінка для F:

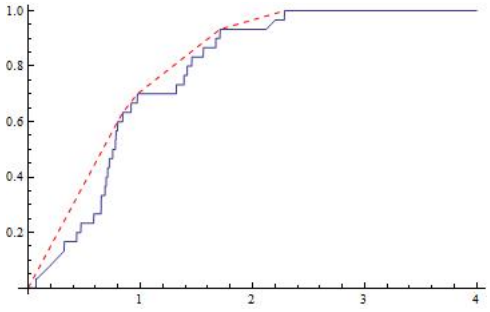
$$\hat{F}(x) = \sum_{j=0}^{n} w_j \mathbb{1}\{T_{[j]} < x \le T_{[j+1]}\},$$

тут $T_{[0]} = 0$, $T_{[n+1]} = +\infty$.

Завдання роботи: аналогічно варіанту (a).

Варіант (c). Оцінка Гренандера для щільності розподілу

Нехай $X_1,\ldots,X_n$ — незалежні, однаково розподілені спостережувані додатні випадкові величини з невідомою щільністю розподілу $f(x)$. Якщо $f(x)$ — монотонно спадна функція для додатних x, її можна оцінювати, використовуючи оцінку Гренандера (яка є оцінкою непараметричного методу найбільшої вірогідності). Для цього потрібно знайти опуклу мажоранту $\check{F}(x)$ для емпіричної функції розподілу даних $\hat{F}(x)$ на інтервалі $x \in (0,\infty)$. Похідна $\check{f}(x) = \check{F}'(x)$ від $\check{F}(x)$ по x і є оцінкою Гренандера для щільності розподілу. По суті, графік $\check{F}(x)$ це ламана, яка торкається графіка $\hat{F}(x)$ у опорних точках. Тому $\check{f}(x)$ є сталою на інтервалах між опорними точками, яка дорівнює кутовому коефіцієнту відповідної ланки ламаної.



синім — $\hat{F}(x)$, червоним — $\check{F}(x)$.

Завдання роботи: Згенерувати вибірку із заданим розподілом, написати програму для підрахунку оцінки Гренандера і намалювати графік оцінюваної щільності разом з оцінкою. Корисно для перевірки правильності оцінки вивести також графік емпіричної функції розподілу та функції $\check{F}(x)$ на одному рисунку.

Значення параметрів для індивідуальних робіт

(1) Варіант a. $F \sim \chi^2(3)$, $G \sim Exp_{\lambda=1/3}$,

(2) Варіант a. $F \sim Exp_{\lambda=1}$, $G \sim Exp_{\lambda=1/2}$,

(3)Варіант a. F — симетричний трикутний на [0,2], G — рівномірний на [0,2].

(4) Варіант a. F — логнормальний розподіл в.в. $\exp(\xi)$, де $\xi \sim N(0,1)$, $G \sim \chi^2(3)$.

(5) Варіант b. $F \sim \chi^2(3)$, $G \sim Exp_{\lambda=1/3}$,

(6) Варіант b. $F \sim Exp_{\lambda=1}$, $G \sim Exp_{\lambda=1/2}$,

(7)Варіант b. F — симетричний трикутний на [0,2], G — рівномірний на [0,2].

(8) Варіант b. F — логнормальний розподіл в.в. $\exp(\xi)$, де $\xi \sim N(0,1)$, $G \sim \chi^2(3)$.

(9)Варіант c. F — півнормальний розподіл з $\sigma = 1$.

(10)Варіант c. Щільність розподілу $f(x) = 1 - x/2$ при $x \in [0,2]$ і $f(x) \neq 0$ при $x \notin [0,1]$.

Робота 3. Вибір оптимального параметра згладжування

Спостерігається вибірка з незалежних, однаково розподілених випадкових величин $X_1, \ldots, X_n$, f — невідома щільність розподілу X_i. Задача полягає в оцінці f.

Ядерна оцінка щільності має вигляд:

$$\hat{f}(x) = \frac{1}{hn} \sum_{j=1}^{n} K\left(\frac{x - X_j}{h}\right),$$

де K — ядро, тобто щільність деякого ймовірнісного розподілу, h — параметр згладжування, який потрібно обирати, виходячи з мінімізації деякого функціоналу якості оцінки.

Типовий функціонал якості — проінтегрована середньоквадратична похибка (mean integrated squared error, MISE):

$$\mathrm{MISE}(h) = \mathsf{E} \int_{-\infty}^{\infty} (\hat{f}(x) - f(x))^2 dx.$$

Нехай виконуються наступні умови:

f — двічі неперервно диференційовна функція і $\varphi = \int_{-\infty}^{\infty} (f''(x))^2 dx < \infty$,

$d^2 = \int K^2(x)dx < \infty$, $\int xK(x)dx = 0$, $D = \int x^2 K(x)dx < \infty$.

Тоді головна частина MISE ($n \to \infty$) мінімізується при

$$h = h_{opt} = \left(\frac{d^2}{ND^2\varphi}\right)^{1/5}$$

Оскільки φ невідоме, його на практиці заміняють оцінкою за спостереженнями. Є два адаптивні підходи до побудови таких оцінок:

1. Параметричний. Вважаємо, що f близька до деякої щільності, котру можна задати параметрично, скажімо, $f(x, \vartheta)$. Знаходимо $\varphi(\vartheta) = \int_{-\infty}^{\infty} (f''(x, \vartheta))^2 dx < \infty$. Оцінюємо ϑ за спостереженнями деякою оцінкою $\hat{\vartheta}$. Підставляємо $\varphi(\hat{\vartheta})$ замість справжнього φ у формулу для h_{opt} і отримуємо оцінку $\hat{h}_{opt}$, котру і використовуємо як параметр згладжування у ядерній оцінці щільності.

Якщо вибрати на роль наближеної параметричної моделі гауссову щільність, приходимо до правила Сілвермана:

$$h_{silv} = \left(\frac{d^2 8\sqrt{\pi}}{3ND^2}\right)^{1/5} \hat{S},$$

де S — деяка оцінка середньоквадратичного відхилення нормального розподілу за даними $X_1, \ldots, X_n$. Так, при

$$\hat{S} = S_0 = \sqrt{\frac{1}{n-1}\sum_{j=1}^{n}(X_j - \bar{X})^2}$$

— отримуємо просте правило Сілвермана, а при

$$\hat{S} = \min(S_0, \mathrm{IR}/1.34)$$

— поліпшене правило Сілвермана (IR — інтерквартильний розмах).

2. Непараметричний. На першому кроці оцінюємо f деякою пілотною непараметричною оцінкою $\tilde{F}$ (це може бути ядерна оцінка з параметром згладжування, обраним за правилом Сілвермана). Оцінюємо φ за допомогою $\hat{\varphi} = \int_{-\infty}^{\infty}(\hat{f}''(x))^2 dx$. Цю оцінку підставляємо у формулу для оптимального параметра згладжування і отримуємо

$$h_{np} = \left(\frac{d^2}{ND^2\hat{\varphi}}\right)^{1/5}.$$

З цим параметром згладжування рахуємо остаточну ядерну оцінку.

Інший можливий підхід до вибору параметра згладжування полягає у використанні техніки крос-валідації. Про це можна прочитати у книжці Larry Wasserman All of Nonparametric Statistics (2007) п. 6.1.

Завдання роботи. Згенерувати вибірку з заданим розподілом і побудувати ядерну оцінку щільності використовуючи параметр згладжування

— обраний за правилом Сілвермана (простим і поліпшеним)

— обраний з використанням непараметричної оцінки φ

— обраний крос-валідацією.

Намалювати графіки отриманих оцінок разом із оцінюваною функцією щільності. Спробувати вручну підібрати параметр згладжування на око, так, щоб оцінка найточніше відповідала оцінюваній функції. Вивести також графік ядерної оцінки з теоретично оптимальним $h = h_{opt}$. Експеримент повторити з кількома вибірками з одним і тим же розподілом.

Спробувати зробити висновки по таких питаннях:

— чи є h_{opt} дійсно оптимальним значенням параметра згладжування, чи вдається знайти краще вибором на око?

— наскільки відрізняється від оптимального h обране за правилом Сілвермана? Чи доцільно його використовувати, якщо оцінювана щільність подібна до розглядуваної у Вашому варіанті?

— наскільки непараметрична адаптація поліпшує оцінку порівняно з правилом Сілвермана? Чи варто застосовувати її?

— Чи можна вважати, що метод крос-валідації дає оцінки, близькі до оптимальних?

Зауваження. Варіанти відрізняються ядрами оцінок K та щільностями f, які треба оцінити. У різних варіантах висновки щодо застосовності тих чи інших методів можуть бути різними.

Всі порівняння достатньо робити на око, але, за бажанням, можна порівнювати значення L_2 відстаней між оцінкою та оцінюваною функцією, усереднені по ~ 100 вибірках. (Це емпіричний аналог MISE).

Значення параметрів для індивідуальних завдань

(1) $f \sim N(1,1)$, K — трикутне ядро.

(2) $f \sim \chi^2(3)$, K — ядро Єпанєчнікова.

(3) f — розподіл Лапласа з мат. сподіванням 0, $\lambda = 1$, K — гауссове ядро.

(4) f — суміш розподілів $N(-1,1)$, $N(1,1)$ з ймовірністю змішування $p = 1/2$, K — ядро Єпанєчнікова.

(5) f — симетрична трикутна щільність на [-1,1], K — ядро Єпанєчнікова.

(6) f — щільність бета-розподілу з $a = 2$, $b = 4$, K — гауссове ядро.

(7) f — щільність розподілу Фішера $F(4,10)$, K — гауссове ядро.

(8) f — щільність розподілу Стьюдента $T(5)$, K — ядро Єпанєчнікова.

(9) f — півнормальна щільність з $\sigma = 1$, K — трикутне ядро.

(10) f — щільність гамма-розподілу, $\alpha = 2$, $\lambda = 1$, K — гауссове ядро.

Робота 4. Емпірично-баєсова класифікація

У роботі потрібно побудувати емпірично-баєсів класифікатор на основі навчаючої вибірки з реальних даних. При цьому використовуються два підходи: (I) класифікація у якій використовується ядерна оцінка багатовимірної щільності та (II) класифікація на основі проекції багатовимірних даних на деякий напрямок, обраний з метою мінімізації ймовірності похибки.

1. Баєсова класифікація.

Нехай спостережувані об'єкти O можуть належати одному з M класів (популяцій) $\mathcal{P}_1, \ldots, \mathcal{P}_M$. Номер популяції, якій належить об'єкт O позначимо $\mathrm{ind}(O)$. У кожного об'єкта спостерігається (можливо, багатовимірна) характеристика $\xi(O)$, яка є випадковою величиною (вектором). Простір можливих значень $\xi(O)$ позначимо $\mathcal{X}$.

Задача класифікації полягає в тому, щоб за спостережуваною характеристикою $\xi(O)$ визначити, до якої популяції належить об'єкт, тобто, вказати $\mathrm{ind}(O)$. Класифікатор це функція $g : \mathcal{X} \to \{1, \ldots, M\}$, яка можливим спостережуваним значенням x ставить у відповідність $g(x)$ — номер популяції, котрій, як ми сподіваємось, повинен належати об'єкт O з характеристикою $\xi(O) = x$.

При баєсовому підході якість класифікатора характеризують ймовірністю його помилки. Тобто, ми вважаємо, що існують ймовірності

$$\pi_k = \mathsf{P}\{\mathrm{ind}(O) = k\}$$

(апріорна ймовірність того, що об'єкт належить k-тій популяції, та розподіли спостережуваних характеристик, за умови, що об'єкт належить k-тій популяції:

$$F^{(k)}(A) = \mathsf{P}\{\xi(O) \in A \mid \mathrm{ind}(O) = k\}.$$

Виберемо міру μ, відносно якої всі розподіли $F^{(k)}$ мають щільності $f^{(k)}$:

$$F^{(k)}(A) = \int_A f^{(k)}(x)\mu(dx),$$

(у цій роботі $\mathcal{X} = \mathbb{R}^d$, μ завжди міра Лебега, тобто $f^{(k)}$ — звичайні щільності d-вимірних розподілів). Тоді ймовірність помилки класифікатора g —

$$L(g) = \mathsf{P}\{g(\xi(O) \neq \mathrm{ind}(O)\} = 1 - \int p_{g(x)} f^{g(x)}(x)\mu(dx).$$

Класифікатор з найменшою ймовірністю помилки називають баєсовим класифікатором. Легко бачити, що, якщо не накладати на функцію g додаткових обмежень, то

$$g^B(x) = \text{argmax}_{m=1,\dots,M}\, p_m f^{(m)}(x)$$

— баєсів класифікатор. Ймовірність його помилки —

$$L(g^B) = 1 - \int \max_{m=1,\dots,M} p_m f^{(m)}(x)\mu(dx)$$

Якщо $f^{(k)}$ і/або p_k невідомі, їх можна оцінити за навчаючою вибіркою, тобто за набором $\Xi_n = \{(\xi(O_1), \text{ind}(O_1)), \dots, (\xi(O_n), \text{ind}(O_n)\}$, де O_j — набір об'єктів, для яких нам відома їх справжня класифікація. Якщо оцінки підставити у формулу для баєсового класифікатора замість справжніх значень, отримуємо емпірично-баєсів класифікатор.

2. Два підходи до багатовимірної класифікації.

(I) Оцінюємо щільності розподілу $f^{(k)}$ за допомогою ядерних оцінок, наприклад:

$$\hat{f}^{(m)}(x) = \frac{1}{n_m h_1 h_2 \dots h_d} \sum_{j:\ i_j=m} \prod_{k=1}^{d} K\left(\frac{x^k - \xi_j^k}{h_k}\right),$$

де

n_m — кількість об'єктів m-того класу у навчаючій вибірці;

h_k — параметр згладжування для k-тої координати;

$x = (x^1, \dots, x^d)$, $\xi_j = \xi(O_j) = (\xi_j^1, \dots, \xi_j^d)$;

$i_j = \text{ind}(O_j)$;

K — одновимірне ядро (наприклад, Єпанечнікова).

Тепер, якщо апріорні ймовірності відомі, підставляємо оцінки у формулу для баєсового класифікатора і отримуємо:

$$\hat{g}(x) = \text{argmax}_{m=1,\dots,M}\, p_m \hat{f}^{(m)}(x)$$

— безпосередній емпірично-баєсів класифікатор.

(II) Для побудови класифікатора вибираємо деякий напрямок у $\mathbb{R}^d$ — $u = (u^1, \dots, u^d)$, $\|u\| = 1$ і робимо проекцію $\xi(O) = (\xi^1, \dots, \xi^d)$ на u: $\eta^u = \eta^u(O) = \sum_{k=1}^d u^k \xi^k$. Це — одновимірна характеристика O, за якою можна будувати класифікатори так само, як за багатовимірними, тільки

тепер оцінювати треба буде одновимірні щільності розподілу звичайними ядерними оцінками щільності $\hat{f}_u^{(m)}$, побудованими за навчаючою вибіркою $\{(\eta_1^u, \mathrm{ind}(O_1)), \ldots, (\eta_n^u, \mathrm{ind}(O_n))\}$.

Питання полягає в тому, як найкраще вибрати напрямок проектування — u? Це можна зробити, використовуючи оцінену ймовірність помилки баєсового класифікатора, побудованого за проекцією на напрямок u:

$$\hat{L}(u) = 1 - \int \max_{m=1,\ldots,M} p_m \hat{f}_u^{(m)}(x)\mu(dx).$$

Напрямок u потрібно обрати так, щоб мінімізувати цю величину.

Завдання роботи.

У файлі vino.xls містяться дані про вміст певних хімічних речовин у пробах вина з трьох виноградників. Кожен рядочок відповідає одній пробі, у стовпчику Site — номер виноградника, на якому було вироблене вино, далі у кожному стовпчику відповідна характеристика вина.

Потрібно побудувати класифікатор, який за двома заданими характеристиками (наприклад, "Alcogol"та "phenols") визначатиме, на якому винограднику вироблено дане вино. Класифікатори будуються двома способами, описаними вище. Побудувавши класифікатори потрібно підрахувати частоту їх помилок на навчаючій вибірці та зробити висновок, який з класифікаторів варто рекомендувати. Бажано також зобразити рисунки, які показують роботу класифікаторів.

Різні варіанти відрізняються різними обраними парами спостережуваних характеристик. (Для різних таких пар висновки щодо вибору класифікатора можуть бути різними).

За бажанням, студенти можуть виконувати цю роботу на власних даних, або на даних, отриманих з інтернета, якщо результати будуть цікавими.

Пари характеристик для індивідуальної роботи

(1) alcogol, phenols.
(2) ash, phenols.
(3) alcohol, magnesium.
(4) Proanthocyanins, magnesium.
(5) Flavanoids, "Alcalinity of ash".
(6) "phenols "Malic acid".
(7) NF, Flavanoids.
(8) phenols, "Hue".
(9) "Magnesium OD.

(10) "Proanthocyanins phenols.

Робота 5. Непараметрична регресія

У роботі спочатку генеруються дані, що складаються з пар чисел (X_j, Y_j), які пов'язані регресійною залежністю:

$$Y_j = g(X_j) + \varepsilon_j, j = 1, \ldots, n$$

де g — задана функція регресії, ε_j — випадкова похибка регресії. Функція регресії вважається невідомою, її треба оцінити за даними. Для оцінювання використовуються: ковзаюче середнє, ковзаюча медіана, оцінка Надарая-Ватсона, локально-лінійна оцінка та проекційні оцінки із заданими базисними функціями. Потрібно реалізувати ці оцінки у вигляді функцій R та вивести їх графіки разом з даними для порівняння. Висновки про поведінку оцінок можна робити за цими графіками "на око".

Бажано подивитись, як змінюються оцінки при зміні допоміжних параметрів — таких, як ширина вікна, параметр згладжування, вимірність простору проекції.

Добре також ввести у вибірку викид і подивитись, як відреагують на нього різні оцінки.

Ковзаюче середнє. (Sliding window, sliding mean estimate). Задаємо значення параметра h — ширина вікна. Для того, щоб оцінити $g(x)$ у точці x_0, виділяємо з усіх даних підвибірку, яка складається з тих Y_j, для яких X_j потрапляє у "вікно" ширини h навколо x_0:

$$\mathbf{Y}(x_0, h) = \{Y_j \ : \ |X_j - x_0| < \frac{h}{2}\}$$

Оцінка

$$\hat{g}_{SW}(x_0) = \text{Mean}[\mathbf{Y}(x_0, h)] = \frac{\sum_{j=1}^{n} Y_j \mathbb{1}\{|X_j - x_0| < h/2\}}{\sum_{j=1}^{n} \mathbb{1}\{|X_j - x_0| < h/2\}}.$$

Ковзаюча медіана. (Sliding median). Оцінка визначається як медіана Y_j для всіх спостережень, що потрапили до ковзаючого вікна:

$$\hat{g}_{SMed}(x_0) = \text{Median}[\mathbf{Y}(x_0, h)]$$

Якщо похибки гауссові, то ковзаюча медіана більш розкидана ніж ковзаюче середнє при тій же ширині вікна. Але, за наявності викидів, ковзаюча медіана виявляється більш стійкою до них, ніж ковзаюче середнє, яке

має тенденцію "притягатись" до викидів (так само, як оцінки Надарая-Ватсона).

Оцінка Надарая-Ватсона. (Nadaraya Watson regression). Оцінка Надарая-Ватсона з ядром K та параметром згладжування h визначається як

$$\hat{g}_{NW}(x_0) = \frac{\sum_{j=1}^{n} Y_j K((X_j - x_0)/h)}{\sum_{j=1}^{n} K((X_j - x_0)/h)}.$$

Оцінка ковзаючого вікна — це оцінка Надарая-Ватсона з прямокутним вікном $K(x) = \mathbb{1}\{|x| < 1/2\}$.

Проекційна оцінка функції регресії

Нехай $p_1,\ldots,p_m$ — набір "базисних" функцій (в принципі, це можуть бути будь-які функції, але хороші оцінки будуть, якщо справжню функцію регресії можна добре наблизити лінійною комбінацією цих функцій). Розглянемо оцінку для g вигляду

$$\hat{g}_{pr}(x) = \sum_{i=1}^{m} \hat{b}_i p_i(x)$$

де коефіцієнти $\hat{b}_i$ визначають як точку мінімуму функціоналу найменших квадратів:

$$J(b) = \sum_{j=1}^{n} \left(Y_j - \sum_{i=1}^{m} b_i p_i(X_j) \right)^2$$

Позначимо $\mathbf{X} = (p_i(X_j))_{j=1;i=1}^{n;m}$, $\mathbf{Y} = (Y_1, \ldots, Y_n)^T$, $\hat{\mathbf{b}} = (\hat{b}_1, \ldots, \hat{b}_m)^T$. Тоді

$$\hat{\mathbf{b}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}.$$

Локально-лінійна та локально-проекційна оцінка

Недолік проекційних оцінок — вони реагують на порушення гладкості оцінюваної функції "нелокально", тобто погіршуються не тільки там де є розрив, а і на тих інтервалах, де справжня функція регресії є гладенькою. Недолік оцінок Надарая-Ватсона (та оцінок, що використовують ковзаюче вікно) — неадекватна поведінка на кінцях інтервалу зміни регресора. Комбінація цих двох підходів — локально-проекційні оцінки.

Нехай $p_1,\ldots,p_m$ — набір "базисних" функцій, K — ядро, h — параметр згладжування. Для того, щоб оцінити $g(x_0)$ шукають такі коефіцієнти

$\hat{b}_i(x_0)$, які мінімізують навантажений функціонал найменших квадратів:

$$J(b) = \sum_{j=1}^{n} K((x_0 - X_j)/h) \left(Y_j - \sum_{i=1}^{m} b_i(x_0) p_i(X_j) \right)^2 .$$

Можна показати, що

$$\mathbf{b}(\hat{x}_0) = (\mathbf{X}^T \mathbf{W} \mathbf{X})^{-1} \mathbf{X}^T \mathbf{W} \mathbf{Y},$$

де $\mathbf{W}$ — діагональна матриця вагових коефіцієнтів $\mathrm{diag}(K((x_0 - X_1)/h), \ldots, K((x_0 - X_n)/h)$.

Локально-лінійна регресія відповідає вибору двох базисних функцій: $p_1(x) = 1$, $p_2(x) = x$.

Індивідуальні завдання роботи.

У завданні вказано розподіл змінної X розподіл ε та функцію регресії g. Обсяг вибірки $n = 300$.

Потрібно згенерувати дані відповідно до моделі регресії та оцінити функцію регресії, використовуючи оцінки

— у завданнях (1-5) оцінка ковзаючого середнього + ковзаюча медіана + локально-лінійна регресія з ядром Єпанєчнікова;

— у завданнях (6-10) ковзаюча медіана + оцінка Надарая-Ватсона з ядром Єпаначнікова + проекційна оцінка з базисом поліномів Лежандра.

Підбір параметра згладжування та вимірності простору проекції виконувати на око.

(Як додатове завдання можна виконати підбір параметрів методом кросс-валідації).

(1) $X \sim N(0, 2)$, ε — рівномірний на [-0.3,0.3],

$$g(x) = \begin{cases} -x^2 + 1 & \text{при } x < 0 \\ x^2 - 1 & \text{при } x \geq 0 \end{cases}$$

(2) X рівномірно розподілений на [0,3], $\varepsilon \sim N(0, 0.3)$,

$$g(x) = 3x(9 - 9x + 2x^2)$$

(3) X рівномірно розподілений на [0,3], ε рівномірно розподілений на [-1,1],

$$g(x) = -\frac{x}{7}(-242 + 805x - 742x^2 + 200x^3)$$

(4) $X \sim N(0,1)$, $\varepsilon \sim N(0.5)$,

$$g(x) = 2|x|$$

(5) $X \sim N(0,2)$, ε — рівномірний на [-0.3,0.3],

$$g(x) = \begin{cases} x^2 - 1 & \text{при } x < 0 \\ x & \text{при } x \geq 0 \end{cases}$$

(6) $X \sim N(0,2)$, ε — рівномірний на [-0.3,0.3],

$$g(x) = \begin{cases} -x^2 + 1 & \text{при } x < 0 \\ x^2 - 1 & \text{при } x \geq 0 \end{cases}$$

(7) X рівномірно розподілений на [0,3], $\varepsilon \sim N(0, 0.3)$,

$$g(x) = 3x(9 - 9x + 2x^2)$$

(8) X рівномірно розподілений на [0,3], ε рівномірно розподілений на [-1,1],

$$g(x) = -\frac{x}{7}(-242 + 805x - 742x^2 + 200x^3)$$

(9) $X \sim N(0,1)$, $\varepsilon \sim N(0.5)$,

$$g(x) = 2|x|$$

(10) $X \sim N(0,2)$, ε — рівномірний на [-0.3,0.3],

$$g(x) = \begin{cases} x^2 - 1 & \text{при } x < 0 \\ x & \text{при } x \geq 0 \end{cases}$$

Литература

[1] Боровков А.А. (1997) Математическая статистика.

[2] Гланц С.(1999) Медико-биологическая статистика.

[3] Деврой Л., Дьерфи Л.(1988) Непараметрическое оценивание плотности.

[4] Hardle, W., Muller, M., Sperlich, S., Werwatz A. (2004) Nonparametric and Semiparametric Models.

[5] Shao J. Mathematical statistics.- Springer-Verlag: New York, 1998. - 530 p.