Modelowanie i identyfikacja – laboratorium 5.
Jądrowy estymator gęstości prawdopodobieństwa

Paweł Wachel

Wymagania wstępne:

1. Wymagania wstępne z poprzednich zajęć[1].

2. Znajomość podstawowych własności estymatora dystrybuanty empirycznej (wariancja, obciążenie).

Zadania do wykonania:

1. Wykorzystując opracowane na wcześniejszych zajęciach generatory, wygenerować N-elementowy ciąg liczb losowych $\{X_1, X_2, \ldots, X_N\}$ o gęstości z rozkładu normalnego $\mathcal{N}(1,1)$.

2. Zaimplementować estymator jądrowy gęstości prawdopodobieństwa

$$\hat{f}_N(x) = \frac{1}{Nh_N} \sum_{n=1}^{N} K\left(\frac{X_n - x}{h_N}\right), \tag{1}$$

gdzie $K(\cdot)$ jest funkcją jądra (por. wykład), a h_N jest parametrem wygładzania. Wykreślić $\hat{f}_N(x)$ w funkcji x dla ustalonej wartości N (np. $N = 500$), jądra prostokątnego i kilku przykładowych wartości parametru wygładzania h_N. Przedyskutować uzyskane wyniki.

3. Wykreślić $\hat{f}_N(x)$ w funkcji x dla różnych funkcji jądra (np. dla jąder omawianych na wykładzie), ustalonej wartości N (np. $N = 500$) i ustalonego h_N. Przedyskutować uzyskane wyniki. Badania powtórzyć dla innych, samodzielnie wybranych gęstości prawdopodobieństwa (np. rozkład trójkątny z poprzednich zajęć, rozkład jednostajny, *etc.*)

4. Dla wybranej gęstości rozkładu prawdopodobieńtwa $f(x)$ wygenerować L niezależnych, N-elementowych sekwencji pomiarowych (prób) i wyznaczyć błąd empiryczny

$$Err\{\hat{f}_N\} = \frac{1}{LM} \sum_{l=1}^{L} \sum_{m=1}^{M} \left[\hat{f}_N^{[l]}(x_m) - f(x_m)\right]^2,$$

w którym $\{x_1, x_2, \ldots, x_M\}$ jest sekwencją równoodlegych punktów z pewnego odcinka $[a, b]$. Przyjąć $M = 100$ oraz $L = 10$ i wykreślić $Err\{\hat{f}_N\}$ w funkcji h_N. Przedyskutować uzyskane wyniki.

[1] Całkujemy wiedzę... przynajmniej do wakacji.

Zadania dodatkowe:

Zadanie automatycznego doboru parametru wygładzania h_N na podstawie dostępnych danych pomiarowych stanowi ważny element w obszarze zastosowań estymatorów jądrowych. Jedną z prostszych metod empirycznego doboru h_N jest tzw. technika kroswalidacji (*ang. cross-validation*).
Do wyznaczenia h_N posłużymy się błędem całkowym

$$L(h_N) = \int [\hat{f}_N(x) - f(x)]^2 dx = \tag{2}$$

$$= \int \hat{f}_N^2(x)\, dx - 2\int \hat{f}_N(x) f(x)\, dx + \int f^2(x)\, dx. \tag{3}$$

Będziemy poszukiwać wartości h_N, która minimalizuje $L(h_N)$. Zauważmy, że ostatni człon w (3) jest stałą i nie zależy od h_N. Możemy zatem wykorzystać uproszczone wyrażenie

$$J(h_N) = \int [\hat{f}_N(x)]^2 dx - 2\int \hat{f}_N(x) f(x)\, dx. \tag{4}$$

Pierwsza całka po prawej stronie jest łatwa do numerycznego oszacowania, druga natomiast zależy od *nieznanej* gęstości $f(x)$. Pamiętając, że $E\{g(X)\} = \int g(x) f(x)\, dx$ możemy ją oszacować posługując się estymatorem[2]

$$\frac{1}{N}\sum_{k=1}^{N} \hat{f}_N^{(-k)}(X_k),$$

gdzie $\hat{f}_N^{(-k)}$ jest estymatorem[3] gęstości jak w (1), lecz skonstruowanym w oparciu o obserwacje $\{X_1, X_2, \ldots, X_{k-1}, X_{k+1}, \ldots, X_N\}$ **z wyłączeniem pojedynczego pomiaru** X_k, tzn.

$$\hat{f}_N^{(-k)}(x) = \frac{1}{(N-1)h_N}\sum_{\substack{n=1 \\ n\neq k}}^{N} K\left(\frac{X_n - x}{h_N}\right).$$

W efekcie wielkość $J(h_N)$ (wzór (4)) możemy oszacować wykorzystując wzór

$$\hat{J}(h_N) = \int [\hat{f}_N(x)]^2 dx - \frac{2}{N}\sum_{k=1}^{N} \hat{f}_N^{(-k)}(X_k).$$

Jako parametr wygładzania h_N przyjmujemy wartość minimalizującą $\hat{J}(h_N)$.

1. Zaimplementować omówione powyżej podejście. Przeprowadzić symulacje dla samodzielnie wybranej gęstości $f(x)$, wykreślić $\hat{J}(h_N)$ w funkcji h_N i przedyskutować efekty działania metody na wybranym przykładzie.

[2] Dlaczego akurat takim?
[3] *ang. leave-one-out*

Literatura:

1. Jakubowski Jacek, Sztencel Rafał. Wstęp do teorii prawdopodobieństwa. Script, 2001.
2. Wasserman, Larry. All of statistics: a concise course in statistical inference. Springer Science & Business Media, 2013.
3. Wasserman, Larry. All of nonparametric statistics. Springer Science & Business Media, 2006.
4. Plucińska Agnieszka, Pluciński Edmund. Probabilistyka: rachunek prawdopodobieństwa, statystyka matematyczna, procesy stochastyczne. Wydawnictwa Naukowo-Techniczne, 2000.
5. Gajek Lesław, Kałuszka Marek. Wnioskowanie statystyczne: modele i metody. Wydawnictwa Naukowo-Techniczne, 1993.
6. Notatki z wykładu.