

Histogram

dr hab. inż. Przemysław Śliwiński

5 kwietnia 2016

1 Histogram

1. Wygenerować dwa ciągi liczb $\{X_1, \dots, X_N\}$ i $\{\xi_1, \dots, \xi_N\}$, $N = 1024$ o (standardowych) gęstościach $f_G(x)$ Gaussa i $f_C(x)$ Cauchy'ego. Przekazać je w tajemnicy osobie na prawicy.
2. Zaimplementować *histogram*, tj. estymator funkcji gęstości o postaci¹

$$\hat{f}_h(x) = \frac{1}{N} \sum_{n=1}^N \frac{1}{h} K\left(\left\lfloor \frac{x}{h} \right\rfloor - \frac{x_n}{h}\right),$$

z prostokątną funkcją jądra $K(x) = \mathbf{1}_{[0,1)}(x)$.

3. Od osoby na lewicy wziąć jej dwa ciągi i wyznaczyć histogramy podciągów tych ciągów dla $N = 128, 256, 512$ i 1024 dobierając parametr h (zwany parametrem wygładzania, ang. *bandwidth parameter*) tak, aby uzyskać najmniejszy (empiryczny) błąd średniokwadratowy

$$\text{emperror}_{G,C}(h) = \sum_{q=-Q}^Q \left[f_{G,C}(x_q) - \hat{f}_h(x_q) \right]^2, \quad (1)$$

dla wybranego (z uzasadnieniem doboru!) ciągu $\{x_q\}$, $q = -Q, \dots, Q$ (gdzie $f_{G,C}(x)$ to odpowiednio gęstość rozkładu Gaussa i Cauchy'ego).

4. Podzielić posiadane ciągi na pół i posługując się obiema częściami, jako *zbiorem uczącym* i *zbiorem testującym*, wyznaczyć h tak, aby błąd

$$\text{crosserr}(h) = \sum_{q=-Q}^Q \left[\hat{f}_h^{[1]}(x_q) - \hat{f}_h^{[2]}(x) \right]^2 \quad (2)$$

¹Wzór w poprzedniej wersji

$$\hat{f}_h(x) = \frac{1}{N} \sum_{n=1}^N \frac{1}{h} K\left(\frac{x_n - x}{h}\right)$$

był oczywiście wzorem na *estymator jądrowy* funkcji gęstości prawdopodobieństwa.

był jak najmniejszy ($\hat{f}_h^{[1]}(x)$ oraz $\hat{f}_h^{[2]}(x)$ to histogramy uzyskane, odpowiednio, z pierwszej i drugiej połowy ciągów). Uzasadnić ewentualne rozbieżności pomiędzy h dobranym przy użyciu wskaźników (1) i (2).