ՄԵՔ. ՈՒՍՈՒՑՄԱՆ ՄԱԹ. ՀԻՄՈՒՆՔՆԵՐԸ

Ա. ԴԱԼԱԼՅԱՆ
01.03.2022

5. Խտության ֆունկցիայի գնահատում

Դիտարկենք հետևյալ խնդիրը՝ մեզ տրված է $X_1, \dots, X_n$ պատահական մեծությունների մի իրականացում՝ $x_1, \dots, x_n$ s.t. $\exists \omega$ s.t. $x_1 = X_1(\omega), \dots, x_n = X_n(\omega)$։ Ենթադրում ենք, որ $X_1, \dots, X_n \overset{iid}{\sim} \mathbb{P}^*$ որտեղ $\mathbb{P}^*$-ը բաշխված է $[0,1]$-ի վրա, այնպիսին որ գոյություն ունի խտության ֆունկցիա։ Սա նշանակում է, որ

$$\exists f^* : [0,1] \to \mathbb{R}_+ \text{ s.t. } \mathbb{E}\left[h(X_i)\right] = \int_0^1 h(x) f^*(x)\, dx$$

f^* ↓ խտության ֆունկցիա
density function

$\forall h : [0,1] \to \mathbb{R}$ չափելի և բացարձակ ինտեգրելի

Մեր նպատակն է՝ օգտագործելով $X_1, \dots, X_n$-ը, կառուցել $\hat{f}_n : [0,1] \to \mathbb{R}_+$ ֆունկցիա, որի սխալանքը՝

$$L(\hat{f}_n, f^*) = \int_0^1 \left(\hat{f}_n(x) - f^*(x)\right)^2 dx$$

հնարավորինս փոքր է։ Առաջ անցնելով նշենք, որ ցույց կտանք, որ եթե f^*-ը անընդհատ դիֆերենցելի է, ապա մեկին մոտ հավանականությամբ $L(\hat{f}_n, f^*) \le \dfrac{Const}{n^{2/3}}$ ։

6. Խտության գնահատում սյունապատկերներով
(Estimation of a density by a histogramme)

Ընտրենք $K \in \mathbb{N}^*$ և նշանակենք $h = \frac{1}{K} \in (0,1)$:
Այս h պարամետրը կանվանենք պատի լայնություն (bandwidth): Կանվանենք սյունապատկեր՝ հետևյալ ֆունկցիան

$$\hat{f}_n(x) = \frac{1}{h} \sum_{j=1}^{K} \hat{p}_j \cdot \mathbb{1}_{C_j}(x)$$

որտեղ $C_j = [(j-1)h \,;\, jh)$ և $\hat{p}_j = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_{C_j}(X_i)$:

<u>Դիտարկում</u> Քանի որ $|C_j| = h = \frac{1}{K}$ and $\sum_{j=1}^{K} \mathbb{1}_{C_j}(X_i) = 1$,

$$\int_0^1 \hat{f}_n(x)\,dx = \frac{1}{h} \sum_{j=1}^{K} \hat{p}_j \int_0^1 \mathbb{1}_{C_j}(x)\,dx$$

$$= \frac{1}{h} \sum_{j=1}^{K} \hat{p}_j |C_j| = \sum_{j=1}^{K} \hat{p}_j = \frac{1}{n} \sum_{i=1}^{n} \underbrace{\sum_{j=1}^{K} \mathbb{1}_{C_j}(X_i)}_{=1}$$

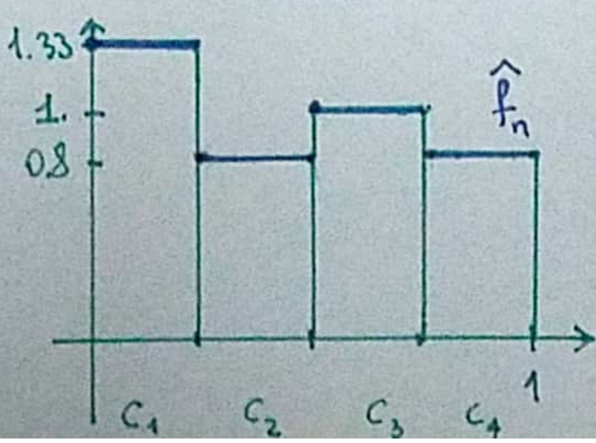
$$= 1.$$

Հետևաբար, $\hat{f}_n$-ը խտության ֆունկցիա է:

Մեր նպատակն է ցույց տալ, որ h-ի «ճիշտ» ընտրության դեպքում $\hat{f}_n$ սյունապատկերը մոտ է f^* անհայտ ֆունկցիային:

<u>Օրինակ</u> $n = 15 \,;\, K = 4$

$(x_1, \dots, x_{15}) = (0.1; 0.9; 0.8; 0.15;$
$0.7; 0.2; 0.05; 0.1; 0.3; 0.4;$
$0.25; 0.85; 0.6; 0.55; 0.6)$



7. Ռիսկի սահմանափակում (bounds on the risk)

Սկսենք հեշտագույն պարզ արդյունքից:

Լեմմա 1. Կամայական $j \in \{1, \dots, K\}$-ի համար, $\hat{p}_j \cdot n$ պատահական մեծությունն ունի բինոմային բաշխում (n, p_j^*) պարամետրերով, որտեղ $p_j^* = \mathbb{P}(X_1 \in C_j)$:
Հետևաբար

$$\mathbb{E}[\hat{p}_j] = p_j^* \quad և \quad \mathbb{E}\left[(\hat{p}_j - p_j^*)^2\right] = \frac{p_j^*(1-p_j^*)}{n} \leq \frac{1}{4n}:$$

Ապացույց Դիտարկենք $Z_i = \mathbb{1}(X_i \in C_j) = \mathbb{1}_{C_j}(X_i)$:
Քանի որ $X_1, \dots, X_n$-ը անկախ են, նույնը ճիշտ է նաև $Z_1, \dots, Z_n$-ի համար: Ավելին, բոլոր Z_i-երն ընդունում են միայն 0 և 1 արժեքները: Հետևաբար Z_i-ն բեռնուլիի բաշխում ունի: Այդ բաշխման պարամետրի արժեքն է

$$\mathbb{P}(Z_i = 1) = \mathbb{P}(X_i \in C_j) = \mathbb{E}[\mathbb{1}_{C_j}(X_i)] = \int_0^1 \mathbb{1}_{C_j}(x) f^*(x) dx$$
$$= \int_{C_j} f^*(x) dx = p_j^*:$$

Ուրեմն $X_1, \dots, X_n \overset{iid}{\sim} \mathbb{P}^*$-ից հետևում է, որ $Z_1, \dots, Z_n \overset{iid}{\sim} \mathcal{B}(p_j^*)$: Հետևաբար, քանի որ $n\hat{p}_j = \sum_{i=1}^n Z_i \sim \mathcal{B}(n, p_j^*)$

$$\mathbb{E}[\hat{p}_j] = \frac{1}{n} \sum_{i=1}^n \mathbb{E}[Z_i] = \frac{1}{n} \sum_{i=1}^n p_j^* = p_j^*$$

$$Var(\hat{p}_j) = \frac{1}{n^2} Var\left(\sum_{i=1}^n Z_i\right) = \frac{1}{n^2} \sum_{i=1}^n Var(Z_i) = \frac{n p_j^*(1-p_j^*)}{n^2}$$

Թեորեմ 4 Կամայական $K\in\mathbb{N}^*$-ի համար

$$\mathbb{E}\left[L(\hat{f}_n, f^*)\right] = \sum_{j=1}^{K}\left\{\int_{C_j}\left(f^*(x)-\frac{p_j^*}{h}\right)^2 dx + \frac{p_j^*(1-p_j^*)}{nh}\right\} \quad (*)$$

$$\leqslant \sum_{j=1}^{K}\int_{C_j}\left(f^*(x)-\frac{p_j^*}{h}\right)^2 dx + \frac{1}{nh} \quad (**)$$

Ապացույց Քանի որ $[0,1]=\bigcup_{j=1}^{K} C_j$,

$$\int_0^1 \left(f^*(x)-\hat{f}_n(x)\right)^2 dx = \sum_{j=1}^{K}\int_{C_j}\left(f^*(x)-\hat{f}_n(x)\right)^2 dx$$

$$= \sum_{j=1}^{K}\int_{C_j}\left(f^*(x)-\frac{1}{h}\hat{p}_j\right)^2 dx :$$

Հետևաբար, $\mathbb{E}\left[L(\hat{f}_n, f^*)\right] = \sum_{j=1}^{K}\int_{C_j}\mathbb{E}\left[\left(f^*(x)-\frac{1}{h}\hat{p}_j\right)^2\right]dx :$

Կիրառելով միջ.-քառակուսային կորստի 5 հատկությունը ներկայացման ձևով՝

$$\mathbb{E}\left[\left(f^*(x)-\hat{p}_j/h\right)^2\right] = \mathbb{E}\left[\left(f^*(x)-p_j^*/h\right)^2\right] + \mathbb{E}\left[\left(p_j^*/h-\hat{p}_j/h\right)^2\right]$$

$$= \left(f^*(x)-p_j^*/h\right)^2 + \frac{1}{h^2}\mathbb{E}\left[\left(p_j^*-\hat{p}_j\right)^2\right]$$

$$= \left(f^*(x)-p_j^*/h\right)^2 + \frac{p_j^*(1-p_j^*)}{nh^2} :$$

Ինտեգրելով այս հավասարության 2 կողմերը C_j-ի վրա, ստանում ենք $(*)$-ը (քանի որ $\int_{C_j} 1\,dx = h$): Իսկ $(*)$-ից $(**)$-ը ստանալու համար՝ բավական է նկատել, որ

$$\sum_{j=1}^{K} p_j^*(1-p_j^*) \leqslant \sum_{j=1}^{K} p_j^* = 1 :$$

Ապացուցված արդյունքները պարզեցնելու համար, ենթադրենք, որ f^* խտության ֆունկցիան պատկանում է Հյոլդերի (Hölder) դասին՝

$$f^* \in \mathcal{H}_{\beta,L} = \left\{ f:[0,1] \to \mathbb{R} : |f(x)-f(y)| \leq L|x-y|^\beta \ \forall x,y \right\}$$

β-պարամետրը, որը պատկանում $(0,1]$-ին է, կոչվում է ողորկության աստիճան (degree of smoothness):

Լեմմա 2 Եթե $f^* \in \mathcal{H}_{\beta,L}$, ապա $\forall x \in C_j$,

$$\left| f^*(x) - \frac{p_j^*}{h} \right| \leq L \cdot h^\beta$$

Ապացույց Նշենք, որ $\int_{C_j} 1\,dy = h$ և $p_j^* = \int_{C_j} f^*(y)\,dy$:

Հետևաբար,

$$\left| f^*(x) - \frac{1}{h} p_j^* \right| = \left| \int_{C_j} \frac{f^*(x)}{h} dy - \frac{1}{h} \int_{C_j} f^*(y)\,dy \right|$$

$$\leq \frac{1}{h} \int_{C_j} |f^*(x) - f^*(y)|\,dy$$

$$\leq \frac{1}{h} \int_{C_j} L|x-y|^\beta dy \leq \frac{1}{h} \int_{C_j} L h^\beta dy \leq L h^\beta : \blacksquare$$

Վարժություն Ցույց տալ, որ կարելի է ստանալ հետևյալ ավելի ճշգրիտ գնահատականը $\left|f^*(x) - \frac{1}{h} p_j^*\right| \leq L \cdot (h/2)^\beta$: