Oulun yliopisto
Matemaattisten tieteiden laitos
Funktioiden estimointi
1. harjoitus, 24. 1. 2007

1. Tässä ja seuraavassa tehtävässä tarkastelemme histogrammin käyttöä tiheysfunktion estimoinnissa.

Reaaliakselilla on määriteltynä tasavälinen (hila-)pisteistö t_k, $k \in \mathbb{Z}$, jossa $t_k < t_{k+1}, \forall k$ ja jossa perättäisten pisteiden etäisyys toisistaan on vakio h, $t_{k+1} - t_k = h, \forall k$. Näin määritellään reaaliakselin ositus väleihin $B_k = [t_k, t_{k+1}[$. Lisäksi on havaittu satunnaisotos $X_1, \ldots, X_n$ jakaumasta, jonka tiheysfunktio on f. Määrittelemme, että satunnaismuuttuja N_k saa arvokseen niiden otospisteiden X_i lukumäärän, jotka sattuvat välille B_k,

$$N_k = \sum_{i=1}^{n} 1_{B_k}(X_i), \qquad k \in \mathbb{Z}\,.$$

(Tässä $1_{B_k}(X_i)$ saa arvon yksi, jos $X_i \in B_k$, ja muussa tapauksessa arvon nolla.)

Tiheysfunktion histogrammiestimaattori pisteessä $x \in \mathbb{R}$ määritellään nyt kaavalla

$$\hat{f}(x) = \frac{N_k}{nh}, \quad \text{kun } x \in B_k.$$

(Tästä erotuksena tavanomaisessa frekvenssihistogrammissa piirretään yksinkertaisesti arvo N_k kullekin reaaliakselin välille B_k.)

(a) Näytä, että $\hat{f}$:n integraali reaaliakselin yli on yksi.

(b) Näytä, että kun $x \in B_k$, niin

$$\mathbb{E}\hat{f}(x) = \frac{p_k}{h}, \quad \text{missä} \quad p_k = \int_{B_k} f(u)\,du.$$

2. Jatkamme tiheysfunktion histogrammiestimaattorin $\hat{f}(x)$ tarkastelua edellisen tehtävän oletuksin ja merkinnöin. Määrittelemme

$$\begin{aligned}
\mathrm{Bias}[\hat{f}(x)] &= \mathbb{E}\hat{f}(x) - f(x) \\
\mathrm{Var}[\hat{f}(x)] &= \mathbb{E}[\hat{f}(x) - \mathbb{E}\hat{f}(x)]^2 \quad \left(= \mathbb{E}\left\{[\hat{f}(x) - \mathbb{E}(\hat{f}(x))]^2\right\}\right) \\
\mathrm{MSE}[\hat{f}(x)] &= \mathbb{E}[\hat{f}(x) - f(x)]^2.
\end{aligned}$$

(a) Laske $\mathrm{Var}[\hat{f}(x)]$, kun $x \in B_k$. (Vihje: N_k on binomijakautunut.)

(b) Näytä, että kun $x \in B_k$, niin

$$\mathrm{MSE}[\hat{f}(x)] = \left(\frac{p_k}{h} - f(x)\right)^2 + \frac{p_k(1-p_k)}{nh^2}.$$

(Vihje: osoita ensin, että $\mathrm{MSE}[\hat{f}(x)] = (\mathrm{Bias}[\hat{f}(x)])^2 + \mathrm{Var}[\hat{f}(x)]$.)

3. Olkoon satunnaismuuttujilla X ja Y jatkuva (yhteis-)jakauma, jonka tiheysfunktio on $f(x,y)$. Tällöin X:n (reuna-)tiheysfunktio on

$$f_X(x) = \int_{-\infty}^{\infty} f(x,y)\,dy,$$

ja Y:n ehdollinen tiheys ehdolla $X = x$ on

$$f_{Y|X}(y \mid x) = f(x,y)/f_X(x).$$

(Oletetaan yksinkertaisuuden vuoksi, että $f_X(x) > 0$ kaikilla x.) Tässä tapauksessa Y:n ehdollinen odotusarvo ehdolla $X = x$ eli (Y:n) *regressiofunktio* (X:n suhteen) saadaan kaavasta

$$m(x) = \int_{-\infty}^{\infty} y f_{Y|X}(y \mid x)\,dy.$$

Olkoon $g : \mathbb{R} \to \mathbb{R}$ mielivaltainen funktio. Osoita, että

$$\mathbb{E}[Y - g(X)]^2 = \mathbb{E}[Y - m(X)]^2 + \mathbb{E}[m(X) - g(X)]^2.$$

(Vihje: Kirjoita $Y - g(X) = Y - m(X) + m(X) - g(X)$, ja huomaa, että ristitermin odotusarvo, $2\mathbb{E}[(Y - m(X))(m(X) - g(X))]$, häviää regressiofunktion määritelmän perusteella.)

4. Johda periodogrammille lauseke

$$\hat{f}(\lambda) = \frac{1}{2\pi n}\left|\sum_{t=1}^{n} e^{-i\lambda t} X_t\right|^2.$$

5. Tarkastelemme hahmontunnistusta kahden luokan tapauksessa, kun

$$f_1(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}, \quad f_2(x) = \frac{2}{\sqrt{2\pi}} e^{-2x^2}, \quad x \in \mathbb{R}.$$

Oletamme luokille yhtä suuret prioritodennäköisyydet. Määrittelemme luokittimen $g : \mathbb{R} \to \{1, 2\}$ kaavalla

$$g(x) = \begin{cases} 1, & \text{jos } f_1(x) > f_2(x), \\ 2, & \text{muuten.} \end{cases}$$

Määrää päätösalueet $g^{-1}(1)$ ja $g^{-1}(2)$ sekä laske luokitteluvirhe $\mathbb{P}(g(X) \neq J)$.