Oulun yliopisto
Matemaattisten tieteiden laitos
Funktioiden estimointi
6. harjoitus, 28. 2. 2007

1. Osoita oikeaksi kaavat

(a) $\int \min(p, q) = 1 - \frac{1}{2} \int |p - q|$

(b) $\int |p - q| \le 2H_2(p, q)$.

Tässä p ja q ovat $\mathbb{R}^n$:n tiheysfunktiota, kaikki integraalit otetaan koko $\mathbb{R}^n$:n yli ja $H_2(p, q)$ on tiheyksien p ja q välinen Hellinger-etäisyys, eli

$$H_2(p, q) = \left[\int (\sqrt{p} - \sqrt{q})^2 \right]^{1/2}.$$

2. Todista luentojen lemma 3.15. (Vihe: funktiot $p(x_1, \ldots, x_n) = \prod_{i=1}^n f(x_i)$ ja $q(x_1, \ldots, x_n) = \prod_{i=1}^n g(x_i)$ ovat $\mathbb{R}^n$:n tiheysfunktioita, joihin voit soveltaa edellisen tehtävän tuloksia.)

3. Lauseen 3.16 todistamiseksi tarvitaan funktiot f_0 ja g, jotka toteuttavat luentojen ss. 73-74 annetut ehdot (i)–(vi). Konstruoi tällaiset funktiot.

4. Huomautus: Tässä ja seuraavassa tehtävässä kehitetään tiheysfunktion histogrammiestimaattorin asymptoottista teoriaa. Laskujen lopputulos tekee uskottavaksi sen tosiasian, että toisin kuin ydinestimaattori, *histogrammiestimaattori ei saavuta minimax-optimaalista konvergenssivauhtia* tiheysfunktion estimoinnissa.

Palautetaan aluksi mieliin ensimmäisissä harjoituksissa johdetut tulokset. Tiheysfunktion histogrammiestimaattorin harha

$$\text{Bias}[\hat{f}(x)] = \frac{p_k}{h} - f(x), \qquad \text{kun } x \in B_k,$$

ja varianssi

$$\text{Var}[\hat{f}(x)] = \frac{p_k(1 - p_k)}{nh^2}, \qquad \text{kun } x \in B_k.$$

Tässä B_k on väli $[t_k, t_{k+1}[$ (kun $k \in \mathbb{Z}$), välin B_k pituus on h, $p_k = \int_{B_k} f(u)\, du$ ja n on otoskoko.

Oletamme (yksinkertaisuuden vuoksi), että f on kahdesti jatkuvasti derivoituva funktio, joka on nolla välin $[-1, 1]$ ulkopuolella.

(a) Johda tulos

$$\int \text{Var}[\hat{f}(x)]\, dx = \frac{1}{nh} - \frac{1}{n} R(f) + \frac{1}{n} \varepsilon_1(h),$$

missä $\varepsilon_1(h) \to 0$, kun $h \to 0+$ ja $R(f) = \int f^2$. (Tässä integraalit voidaan ottaa yhtä hyvin joko välin $[-1, 1]$ tai koko reaaliakselin yli.)

(b) Johda tulos

$$\int \mathrm{Bias}^2[\hat{f}(x)]\,dx = \frac{1}{12}h^2 R(f') + h^2 \varepsilon_2(h),$$

missä $\varepsilon_2(h) \to 0$, kun $h \to 0+$.

Vihjeitä (a)- ja (b)-kohtia varten: Jos g on jatkuva funktio, joka on nolla välin $[-1, 1]$ ulkopuolella, ja pisteet $\xi_k \in B_k$ kaikilla k, niin

$$\sum_{k \in \mathbb{Z}} g(\xi_k)h = \int_{-1}^{1} g(t)\,dt + \varepsilon(h),$$

missä $\varepsilon(h) \to 0$, kun $h \to 0+$. Tämä johtuu siitä, että välin $[-1, 1]$ ositukseen liittyvä Riemannin summa suppenee kohti välin yli otettua integraalia osituksen tihentyessä. (Varoitus: Tilanne olisi monimutkaisempi, ellei g häviäisi kompaktin välin ulkopuolella.)

Seuraava versio integraalilaskun väliarvolauseesta saattaa osoittautua hyödylliseksi. Jos g_1 ja g_2 ovat välillä $[a, b]$ jatkuvia funktioita, joista $g_2 \geq 0$ välillä $[a, b]$, niin on olemassa ξ siten, että $a < \xi < b$ ja

$$\int_a^b g_1(t)g_2(t)\,dt = g_1(\xi) \int_a^b g_2(t)\,dt.$$

5. Olkoon nyt histogrammin välinpituus h_n otoskoon n funktio, ja olkoon $\hat{f}_n$ otoskoolla n ja välin pituudella h_n saatava histogrammiestimaattori. Oletamme, että $h_n \to 0$, kun $n \to \infty$.

(a) Johda kaava

$$\mathrm{MISE}[\hat{f}_n] = \mathrm{AMISE}[\hat{f}_n] + o(\frac{1}{nh_n} + h_n^2),$$

jossa histogrammiestimaattorin asymptoottinen MISE

$$\mathrm{AMISE}[\hat{f}_n] = \frac{1}{nh_n} + \frac{1}{12}h_n^2 R(f').$$

(Käytä hyväksesi edellisen tehtävän tuloksia.)

(b) Määrää optimaalinen välin pituus h_n^* siten, että $\mathrm{AMISE}[\hat{f}_n]$ minimoituu.

(c) Osoita, että kun käytetään optimaalista välin pituutta h_n^*, niin asymptoottinen MISE on kertaluokkaa $O(n^{-2/3})$.