

Oulun yliopisto
Matemaattisten tieteiden laitos
Tilastollinen hahmontunnistus
12. harjoitus, viikko 15, 2010

1. Luennolla johdettiin luokan j posterioritodennäköisyydelle $P(j|x)$ estimaattori

$$\hat{P}(j|x) = \frac{\sum_{i=1}^{n_j} K_h(x - X_{ji})}{\sum_{l=1}^n K_h(x - X_l)},$$

missä $X_{j1}, \dots, X_{jn_j}$ ovat opetusvektorit luokasta j . Osoita, että $\hat{P}(j|x) = \bar{a}_j(x)$, missä $a_j = \bar{a}_j(x)$ minimoi lausekkeen

$$\sum_{i=1}^n (a_j - Y_{ji})^2 K_h(x - X_i),$$

ja $Y_{ji} = 1_{\{J_i=j\}}$.

2. Tarkastellaan Bayesin luokitinta g^* , kun todelliset luokkatiheysfunktiot f_1 ja f_2 tunnetaan, ja $P_1 = P_2$. Osoita, että Bayesin virhetodennäköisyydelle $e(g^*)$ pätee kaava

$$e(g^*) = \frac{1}{2} - \frac{1}{4} \int_{\mathbb{R}^d} |f_1 - f_2|.$$

Miten tämän perusteella tulkitset Bayesin virhetodennäköisyyden riippuvan luokkatiheysfunktioden L^1 -etäisyydestä $\int_{\mathbb{R}^d} |f_1 - f_2|$? Vihje kaavan johtamiseksi: kirjoita

$$e(g^*) = \frac{1}{2} e_1(g^*) + \frac{1}{2} e_2(g^*) = \frac{1}{2} \int_{\{f_2 > f_1\}} f_1 + \frac{1}{2} \int_{\{f_1 \geq f_2\}} f_2$$

(vrt. luku 3.2) ja kirjoita sitten $\int_{\mathbb{R}^d} |f_1 - f_2|$ tässä kaavassa esiintyvien integraalien avulla. Tässä on merkitty lyhyesti $\{f_2 > f_1\} = \{x \mid f_2(x) > f_1(x)\}$ ja $\{f_1 \geq f_2\} = \{x \mid f_1(x) \geq f_2(x)\}$.

3. Johda $\mathbb{R}^d$:n yksikköpallon tilavuudelle C_d kaava

$$C_d = \frac{\pi^{d/2}}{\Gamma(\frac{1}{2}d + 1)},$$

jossa

$$\Gamma(t) = \int_0^\infty u^{t-1} e^{-u} du.$$

Menettele seuraavasti:

1. Olkoon

$$V(r) = m(\{x \in \mathbb{R}^d : \|x\|^2 \leq r^2\}) = \int_{\mathbb{R}^d} 1_{[0, r^2]}(\|x\|^2) dx$$

d -ulotteisen r -säteisen pallon tilavuus. Totea suorittamalla muuttujanvaihto integraalissa, että

$$V(r) = r^d V(1), \quad r > 0.$$

2. Koska $\int_a^\infty e^{-u} du = e^{-a}$, on (integroimisjärjestyksen vaihtamista koskevan Fubinin lauseen nojalla)

$$\begin{aligned} \int_{\mathbb{R}^d} e^{-\|x\|^2} dx &= \int_{\mathbb{R}^d} \left(\int_{\|x\|^2}^\infty e^{-u} du \right) dx \\ &= \int_{\mathbb{R}^d} \left(\int_0^\infty e^{-u} 1_{[\|x\|^2, \infty[}(u) du \right) dx \\ &= \int_0^\infty e^{-u} \left(\int_{\mathbb{R}^d} 1_{[\|x\|^2, \infty[}(u) dx \right) du. \end{aligned}$$

Vasemmalla olevan integraalin osat laskea ja oikealla olevan voit lausua $V(1)$:n ja Γ -funktion avulla.

4. Olkoot f ja g tiheysfunktioita. Osoita seuraava identiteetti (joka tunnetaan nimellä Scheffén lause)

$$2 \sup_B \left| \int_B f - \int_B g \right| = \int |f - g|.$$

Tässä supremum otetaan kaikkien (Borelin) joukkojen yli. Vihje: seuraavien tulosten todistamisesta on hyötyä

$$2 \int_B (f - g) \leq 2 \int_{\{f > g\}} (f - g) = \int |f - g|.$$

5. Kokeillaan tiheysfunktion ydineistointia **snowfall**-aineistolla, jonka saat käyttöösi kurssin kotisivun kautta. Aineisto sisältää tuumissa mitattuna vuosittaisen lumentulon v. 1910–1972 Buffalon kaupungissa New Yorkin osavaltiossa. Toteuta ydineistimaattori haluamassasi ohjelmointiympäristössä. Etsi sitten sellainen silotusparametrin h arvo, jolla estimaatissa näkyy selkeästi kolme moodia (eli lokaalia maksimia), kun ytimenä käytetään Gaussin ydintä. Seuraavaksi etsi sellainen h , jolla estimaatissa näkyy ainoastaan yksi moodi. Koeta toistaa edelliset kohdat käyttämällä ns. tasaista ydintä $K = \frac{1}{2} 1_{[-1, 1]}$.