

Oulun yliopisto
 Matemaattisten tieteiden laitos
 Tilastollinen hahmontunnistus
 8. harjoitus, viikko 10, 2010

Seuraavissa kolmessa tehtävässä tarkastellaan luokittelua diskreetin piirrevektorin tapauksessa.

1. Olkoon X satunnaisvektori, joka voi saada ainoastaan äärellisen monta arvoa $x_1, \dots, x_m \in \mathbb{R}^d$. Kuten ennenkin, piirrevektorin X oikean luokan kertoo satunnaismuuttuja J , joka voi saada arvot $1, \dots, c$ ja luokan j prioritodennäköisyyttä merkitään P_j :llä. Luokan j posterioritodennäköisyys on Bayesin kaavan nojalla

$$P(j|x_k) \equiv P(J=j|X=x_k) = \frac{P_j P(X=x_k|J=j)}{P(X=x_k)}, \quad k=1, \dots, m$$

ja Bayesin luokitin määritellään kuten ennenkin,

$$g^*(x_k) = \arg \max_{j=1, \dots, c} P(j|x_k).$$

Todista, että g^* minimoi luokitteluvirheen.

2. Olkoon $c=2$ ja $X = [X^{(1)}, \dots, X^{(d)}]^T$ d -ulotteinen binäärinen satunnaisvektori, eli kukin komponentti $X^{(i)}$ voi saada vain arvot 0 tai 1. Olkoon

$$P(X^{(i)}=1|J=1) = p_i, \quad P(X^{(i)}=1|J=2) = q_i,$$

$i=1, \dots, d$. Oletamme, että kummalekin luokalle komponentit $X^{(i)}$ ovat riippumattomia, jolloin mitatulle piirrevektorille $x = [x^{(1)}, \dots, x^{(d)}]^T$ pätee

$$P(X=x|J=1) = \prod_{i=1}^d p_i^{x^{(i)}} (1-p_i)^{1-x^{(i)}}, \quad P(X=x|J=2) = \prod_{i=1}^d q_i^{x^{(i)}} (1-q_i)^{1-x^{(i)}}.$$

Osoita, että Bayesin luokitin on silloin lineaarinen,

$$a^T x + \alpha \stackrel{1}{\underset{2}{>}} 0,$$

missä

$$a = [a^{(1)}, \dots, a^{(d)}], \quad a^{(i)} = \log \left[\frac{p_i(1-q_i)}{q_i(1-p_i)} \right], \quad i=1, \dots, d$$

ja

$$\alpha = \sum_{i=1}^d \log \left[\frac{1-p_i}{1-q_i} \right] + \log \frac{P_1}{P_2}.$$

3. Samaa laitetta valmistetaan kahdella eri tuotantolinjalla ja valmis laite koostuu kolmesta osasta, jotka voivat toisistaan riippumatta olla toimivia tai viallisia. Tiedetään, että linjalta 1 valmistuneessa laitteessa osa i on toimiva todennäköisyydellä p_i ja että vastaavasti linjalta 2 valmistuneessa laitteessa osa i on toimiva todennäköisyydellä q_i ($i = 1, 2, 3$). Oletetaan, että tarkastettava valmis laite voi yhtä hyvin olla peräisin kummasta tahansa tuotantolinjasta ja että $p_1 = p_2 = 0.8$, $p_3 = 0.7$ ja $q_1 = q_2 = q_3 = 0.5$. Konstruoi Bayesin luokittimen, joka valmiin laitteen kolmen osan toimivuuden perusteella yrittää määrittää kummalta linjalta laite on peräisin. Miten luokitellaan laite, jonka osista vain ensimmäinen toimii?

4. Olkoon kahden luokan tapauksessa luokkien ehdolliset odotusarvot ja kovarianssimatriisit

$$\mu_1 = \begin{bmatrix} 0 \\ 0 \end{bmatrix}, \quad \Sigma_1 = \begin{bmatrix} 2 & 1 \\ 1 & 2 \end{bmatrix}$$

$$\mu_2 = \begin{bmatrix} 0 \\ 2 \end{bmatrix}, \quad \Sigma_2 = \begin{bmatrix} 2 & -1 \\ -1 & 2 \end{bmatrix}$$

Laske tässä tapauksessa Fisherin lineaarisen luokittelijan päätöspinnan normaalivektori a ja hahmottele luokittelijan päätösalueet, kun kynnysparametri $\alpha = -1$.

5. Tämän harjoituksen voit tehdä jollain matemaattisella/tilastollisella ohjelmistolla (MATLAB, R, ...).

Ohjelmoi luentojen kaavassa (4.1) määritelty kvadraattinen luokittelija.

Tarkastellaan kahden normaalijakautuneen luokan $N(\mu_1, \Sigma_1)$, $N(\mu_2, \Sigma_2)$ tilannetta, kun $P_1 = P_2 = 0.5$ ja luokkien odotusarvot ja kovarianssimatriisit ovat samat kuin tehtävässä 4. Estimoi luokittimesi (Bayesin) virhetodennäköisyys generoimalla suuri määrä vektoreita kummastakin luokasta, estimoimalla niiden avulla virhetodennäköisyys kummallekin luokalle erikseen (väärin luokiteltujen vektoreiden suhteelliset osuudet) ja yhdistämällä sitten tulokset prioritodennäköisyyksillä painottaen (vrt. luentojen luku 3.2).