Matematika B4
XII. gyakorlat

2005. május 4.

1. Általános regresszió

Adott (X, Y) kétdimenziós valószínűségi változó az együttes eloszlásával (folytonos esetben sűrűségfüggvényével). X-et megfigyelve Y-t szeretnénk közelíteni egy $k(X)$ alakú tippelő függvénnyel. A közelítés azt jelenti, hogy az elkövetett négyzetes hiba – $(Y - k(X))^2$ – átlagát szeretnénk minimalizálni. Pontosabban azt az k függvényt keressük, amire

$$\mathbf{E}\left((Y - k(X))^2\right)$$

minimális. A tétel azt mondja, hogy

$$k(x) = \mathbf{E}(Y \mid X = x).$$

Ha pedig az elkövetett abszolút hibát, azaz $\mathbf{E}(|Y - k(X)|)$ -t szeretnénk minimalizálni, akkor a lehető legjobb tippelés az Y mediánja, feltéve, ha már tudjuk, hogy mit vett fel az X, azaz éppen a feltételes medián függvény. A feltételes sűrűségfüggvény

$$f_{2|1}(x) = \frac{f(x, y)}{\int_{-\infty}^{\infty} f(x, y)dy}$$

amelynek az első esetben a várható értékét kell kiszámolni, a második esetben $\frac{1}{2}$-edel egyenlővé tenni és belőle y-t mint x függvényét kifejezni.

2. Lineáris regresszió

A gyakorlatban ritkábban tudjuk az együttes sűrűségfüggvényt becsülni, könnyebb a kovarianciát, várható értéket és a szórást. Itt jön be a **lineáris regresszió**. Y-t közelítjük X-szel egy egyenes segítségével:

$$y - \mathbf{E}(Y) = \frac{\mathrm{cov}(X, Y)}{\mathbf{D}^2(X)}(x - \mathbf{E}(X)).$$

Ilyenkor az $\mathbf{E}\left((Y - k(X))^2\right)$ kifejezésben olyan k függvényeket keresünk, amelyek linárisak.
Ez egy kicsit másképp:

$$\frac{y - \mathbf{E}(Y)}{\mathbf{D}(Y)} = R(X, Y)\frac{x - \mathbf{E}(X)}{\mathbf{D}(X)}.$$

ahol $R(X, Y) = \frac{\mathrm{cov}(X,Y)}{\mathbf{D}(X)\mathbf{D}(Y)}$ az X és a Y korrelációs hányadosa. Ha $|R(X, Y)| = 1$, akkor a két valószínűségi változó közt lineáris függés van, ha 0, akkor nincs függés.
A kovariancia definíciója:

$$cov(X, Y) = \mathbf{E}(XY) - \mathbf{E}(X)\mathbf{E}(Y)$$

A kovariancia mátrixban az i. oszlop j. sorában az i. és a j. valószínűségi változó kovarianciája áll, vagyis ez egy szimmetrikus mátrix, melynek főátlójában pedig épp a szórásnégyzetek helyezkednek el, aza két valváltozóra ez így néz ki:

$$\begin{pmatrix} \mathbf{D}^2(X) & cov(X,Y) \\ cov(X,Y) & \mathbf{D}^2(Y) \end{pmatrix}$$

Feladatok:

1. A Duna holnaputáni budapesti vízállását akarjuk becsülni a mai bécsi vízállásból. Bár a két vízállás közt szoros kapcsolat van, azért pontosan nem lehet megmondani a vízállást, mindkettőt egy-egy valószínűségi változó írja le.Tegyük fel, hogy mindkét vízállást egy 0 és 1 közti számmal tudunk jellemezni, melynek legyen az együttes eloszlásfüggvénye $f(x,y) = \frac{6}{5}(x + (y-1)^2)$ ahol $0 < x < 1$ és $0 < y < 1$.

   a) Határozzuk meg a budapeati vízállást a bécsi ismeretében, azaz. mi a feltételes sűrűségfüggvény?

   b) Mi annak a valószínűsége, hogy budapesten alacsonynak nevezhető (azaz 0 és $1/2$ közé esik) a vízállás, ha Bécsben x volt? (Mennyi ez $x = 1/3$-ra?)

   c) ha már ismerjük a bécsi vízállást, mire tippelünk a budapestire, ha a lehető legkisebb négyzetes hibát akarjuk elkövetni?

   d) és ha az abszolút hibát akarjuk minimalizálni?

2. (Az előadáson hallott eloszlás regressziója) $X = RND1$, $Y = RND1 * RND2$ múlt órán láttuk, hogy az együttes sűrűségfüggvény $f(x,y) = 1/x$, ha $0 < y < x < 1$ háromszögön vagyunk. Mi a regressziós görbe, ha az a bszolút hibát, illetve ha a négyzetes hibát szeretnénk minimalizálni?

3. Egy kétdimenziós valószínűségi változó sűrűségfüggvénye $\frac{1}{6}xy$ $(0 < x < 2, x < y < 2x)$. Milyen $k(y)$ függvénnyel érdemes a második koordinátából az elsőt tippelni, ha az a célunk, hogy a tippelésnél elkövetett hiba négyzetének átlagos értéke sok kísérlet esetén minél kisebb legyen,

   (a) ha feltesszük, hogy $k(y)$ lineáris,

   (b) ha $k(y)$ tetszőleges valós lehet?

4. Ugyanaz a problema, mint az elozo feladatban, de most a tippelo fuggvenyunk csak $c\sqrt{y}$ alalku lehet?
   Segitseg: itt a
   $m(c) = \mathbf{E}((X - c\sqrt{Y})^2) = \mathbf{E}(X^2) + c^2\mathbf{E}(Y) - 2c\mathbf{E}(X\sqrt{Y}) = \mathbf{E}(X^2) + \mathbf{E}(Y)(c^2 - \frac{\mathbf{E}(X\sqrt{Y})}{\mathbf{E}(Y)})^2 - \frac{\mathbf{E}(X\sqrt{Y})^2}{\mathbf{E}(Y)}$
   fuggvenyt kell minimalizalni, ahol c valtozhat.

5. X és Y együttes sűrűségfüggvénye $h(x,y) = 60xy^2$, ha $0 \le x \le 1$, $0 \le y \le 1 - x$. Határozzuk meg a kovarianciájukat!
   Tegyük fel, hogy a második koordinátát tudjuk megfigyelni és az elsőt ezen megfigyelt adattól függően becsüljük az $x = \frac{2}{3}(1 - y)$ képlet alapján. Van-e ennél jobb módszer, ha négyzetes eltérés hibáját akarjuk minimalizálni?

6. Az egységkörön választunk egyenletes eloszlás szerint egy (X, Y) pontot. Az X koordinata ismeretében hogyan kozelítené $|Y|$-t, feltéve, hogy a hiba abszolútértéknégyzetét szertné minimalizálni?

7. Az (X, Y) kovarianciamátrixa $\left(\begin{smallmatrix} 8 & 4 \\ 4 & 2 \end{smallmatrix}\right)$ Van-e linearis kapcsolat X és Y között?

8. Statisztikai adatok alapján annak a valószínűsége, hogy ikerszületéskor mindkét gyerek fiú, 0.32, annak a valószínűssge, hogy mindkét gyermek lány, 0.28. Annak a valószínűsége, hogy az első iker fiú és a masodik lány ugyanannyi, mint fordítva. Jelölje X illetve Y az első, illetve a masodik gyerek nemét, legyen a felvett ertékük fiú esten 1, lány esetén 0. Számítsuk ki az X és a Y korrelációs együtthatóját! Hogyan tippelnénk Y ismeretében X-re lineáris függvénnyel, ha a tippeles átlagos hibáját akarjuk minimalizálni?

9. Legyen (X, Y) egyenletes eloszlású a $(0,0)$, $(1,0)$, $(0,2)$ pontok által meghatározott háromszögön. Számítsuk ki Y-nak X-ra vonatkozó regressziós függvényét!

10. Legyenek X és Y két véges szórasú valószínűségi változó. Legyen $A = X + Y$, $B = X - Y$ Bizonyítsa be, hogyha tudjuk, hogy B-nek A-ra vonatkozó regressziós egyenese konstans, akkor a X es Y szórasa egyenlő!

11. Többpártrendszer eseten az egyes pártokra leadott szavazatok százalékos aránya valószínűségi változó. A Zöldek az összes szavazatok X, a Demokraták az összes szavazatok Y hányadát kapják, együttes eloszlásuk $h(x,y) = 24xy$, ha $0 < x, 0 < y, x + y < 1$.
    Ha a Demokraták az összes szavazatok 40%-át kaptak, mire tippelünk, mennyit kaptak a zöldek?

12. Magyarországon a 18 év feletti férfiak testmagasságának átlagos érteke 178 cm, szórása 10 cm. nőknel ugyanezek az adatok:166 cm, és 8 cm. Focimeccseken a drukkerek 10%-a nő, a többiek férfiak. Mindkét nem testmagasságának eloszlását normalis eloszlásúnak véve:

    a.) Mi annak a valószínűsége, hogy egy 170 cm-nel alacsonyabb szurkoló nő?

    b.) Adja meg x függvényében annak a valószínűségét, hogy egy x cm magas drukker férfi!

    c.) Hogyan tippeljunk a szurkolók testmagasságából a nemükre, ha a celunk az, hogy a lehető legnagyobb valószínűséggel helyesen tippeljünk?