Matematika B4
XI. gyakorlat

2006. április 27.

1. Általános regresszió

Adott egy (X, Y) kétdimenziós valószínűségi változó együttes eloszlása (például folytonos esetben sűrűségfüggvényével). X-et megfigyelve Y-t szeretnénk közelíteni egy $k(X)$ alakú tippelő függvénnyel. Az elkövetett $(Y - k(X))^2$ négyzetes hiba átlagát szeretnénk minimalizálni, vagyis azt a k függvényt keressük, amire

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

minimális. Az órán tanult tétel kimondja, hogy ebben az esetben a megoldás a feltételes várható érték:

$$k(x) = E(Y \mid X = x)$$

Ha pedig az elkövetett abszolút hiba várható értékét, azaz $E(|Y - k(X)|)$ -t szeretnénk minimalizálni, akkor a lehető legjobb tippelés az Y feltételes mediánja.

Az első esetben a feltételes sűrűségfüggvényből a feltételes várható értékét kell kiszámolni:

$$k(x) = E(Y \mid X = x) = \int_{-\infty}^{\infty} y f_{2|1}(y|x) dy$$

A második esetben először a feltételes eloszlásfüggvényt számoljuk ki:

$$F_{2|1}(y|x) = \int_{-\infty}^{y} f_{2|1}(v|x) dv$$

utána ezt $\frac{1}{2}$-del tesszük egyenlővé:

$$F_{2|1}(y|x) = \frac{1}{2}$$

és ebből az egyenletből y-t x függvényeként ki kell fejezni, így jutunk az $y = k(x)$ függvényhez..

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 budapesti vízállás eloszlását 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. $X = RND1$, $Y = RND1 * RND2$ eloszás esetén láttuk, hogy az együttes sűrűségfüggvény $f(x, y) = 1/x$, ha $0 < y < x < 1$ háromszögön vagyunk. Mi a regressziós görbe, ha az abszolút hibát, illetve ha a négyzetes hibát szeretnénk minimalizálni?

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

4. 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?

2. Lineáris regresszió

A gyakorlatban sokszor előnyben részesítjük a lineáris függvényeket, és keressük azt a lineáris függvényt, amivel tippelve X-ből Y-ra, a hiba négyzetének a várható éréke a legkisebb. Ennek az optimális lineáris függvénynek a neve *regressziós egyenes*, az egyenlete:

$$y = \frac{\mathrm{cov}(X, Y)}{\sigma^2(X)}(x - E(X)) + E(Y),$$

ahol $cov(X, Y)$ az X és Y közötti kovariancia. Az egyenletet kicsit másképp felírva, az alábbi alakot kapjuk:

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

ahol $R(X, Y) = \frac{\mathrm{cov}(X,Y)}{\sigma(X)\sigma(Y)}$ az X és az Y közötti korrelációs együttható.

Ha $|R(X, Y)| = 1$, akkor a két valószínűségi változó között determinisztikus lineáris függés van. Ha $R(X, Y) = 0$, akkor a két valószínűségi változót egymástól korrelálatlannak nevezzük.

A kovariancia definíciója:

$$cov(X, Y) = E((X - E(X))(Y - E(Y)))$$

Kibontva a zárójeleket:

$$cov(X, Y) = E(XY) - E(X)E(Y)$$

Az (X, Y) kovariancia mátrixa egy szimmetrikus mátrix, melynek főátlójában a szórásnégyzetek helyezkednek el, a mellékátlóban pedig az X és Y közötti kovariancia. A kovariancia mátrix tehát így néz ki:

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

Feladatok:

5. 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?

6.* Ugyanaz a probléma, mint az előző feladatban, de most a tippelő függvényünk csak $c\sqrt{y}$ alakú lehet?
Segitseg: itt a
$m(c) = E((X - c\sqrt{Y})^2) = E(X^2) + c^2E(Y) - 2cE(X\sqrt{Y}) = E(X^2) + E(Y)(c^2 - \frac{E(X\sqrt{Y})}{E(Y)})^2 - \frac{E(X\sqrt{Y})^2}{E(Y)}$
függvényt kell minimalizálni, ahol c változhat.

7. X és Y együttes sűrűségfüggvénye $h(x,y) = 60xy^2$, ha $0 \leq x \leq 1$, $0 \leq y \leq 1 - x$. 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?

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

9. 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 értékük 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?

10. 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!

11. 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 értékük 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?

12.* 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 és Y szórasa egyenlő!

13. 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?