10. Kalkyyliä pistetodennäköisyysfunktioille

Tarkastellaan diskreettejä sm:ia (sm = satunnaismuuttuja), X, Y, K, joista kukin voi saada vain äärellisen määrän eri arvoja; arvojoukot S_X, S_Y, S_K äärellisiä joukkoja. Oletetaan epäoleellisten vaikeuksien välttämiseksi, että

$$P(X=x, Y=y, K=k) > 0 \quad \forall\, x \in S_X, y \in S_Y, k \in S_K$$

Tällöin kaikki alla käytetyt ehdolliset tn:t ovat hyvin määriteltyjä. Jatkossa jätetään ehdot $x \in S_X$, $y \in S_Y$, $k \in S_K$ kirjoittamatta.

Marginalisointi, eli miten siirrytään yhteisjakaumasta reunajakaumaan:

$$P(X=x, Y=y) = \sum_k P(X=x, Y=y, K=k)$$

(todennäköisyyden additiivisuus!)

$$P(X=x) = \sum_x P(X=x, Y=y)$$

$$= \sum_x \sum_k P(X=x, Y=y, K=k)$$

jne.

Yhteis-ptnf:stä (ptnf = pistetodennäköisyysfunktio) saadaan reuna-ptnf summaamalla siitä tarpeeton muuttuja tai tarpeettomat muuttujat pois.

Ehdollinen ptnf ≡ ehdollinen tn. Esim.

$$P(X=x \mid Y=y) = \frac{P(X=x, Y=y)}{P(Y=y)}$$

Kertolaskukaava (eli -sääntö)

$$P(X=x, Y=y) = P(X=x)\, P(Y=y \mid X=x)$$
$$= P(Y=y)\, P(X=x \mid Y=y)$$

(on sama asia kuin ehdollisen ptnf:n määritelmä.)

<u>Riippumattomuus</u>: $X \perp\!\!\!\perp Y$, jos

$$P(X=x, Y=y) = P(X=x) P(Y=y) \quad \forall x, y$$

Jos $X \perp\!\!\!\perp Y$, niin $P(Y=y) = P(Y=y \mid X=x)$ kaikilla x, y, ja $P(X=x) = P(X=x \mid Y=y)$ kaikilla x, y. Tämä seuraa riippumattomuuden määritelmästä sekä kertolaskukaavasta.

Marginalisointi onnistuu samalla tavalla myös ehdolliselle yhteis-ptnf:lle:

$$P(X=x \mid K=k) = \sum_y P(X=x, Y=y \mid K=k)$$

(ehdollinen tn $B \mapsto P(B \mid K=k)$ on additiivinen!)

<u>Ehdollinen riippumattomuus</u>: $X \perp\!\!\!\perp Y \mid K$, eli X ja Y ovat riippumattomia ehdolla K, jos ne ovat riippumattomia ehdollisessa yhteisjakaumassa

$$P(X=x, Y=y \mid K=k) \quad \forall k$$

ts. jos

$$P(X=x, Y=y \mid K=k) = P(X=x \mid K=k)\, P(Y=y \mid K=k)$$
$$\forall x, y, k$$

Vaihtoehtoinen luonnehdinta ehdolliselle riippumattomuudelle (HT):

$X \perp\!\!\!\perp Y \mid K$, jos

$$P(X = x \mid Y = y, K = k) = P(X = x \mid K = k)$$

tai

$$P(Y = y \mid X = x, K = k) = P(Y = y \mid K = k)$$

$\forall x, y, k$

Huomautus: nämä kaikki kaavat pätevät myös, jos X onkin (diskreetti) satunnaisvektori, esim.

$$X = \underline{X} = (X_1, X_2, \ldots, X_n)$$

tai Y on (diskreetti) satunnaisvektori tai K on (diskreetti) satunnaisvektori.

Lisäksi nämä kaavat yleistyvät myös jatkuvien jakaumien tapauksiin, jolloin summien sijasta pitää laskea integraaleja. Palaamme jatkuvaan tapaukseen myöhemmin.

11. Ennustaminen: mikä pallo seuraavaksi?

[AS: jakso 1.5]

Pallot kulhossa: olemme havainneet

$\underline{X} = \underline{x}$ eli $X_1 = x_1$, $X_2 = x_2, \dots, X_n = x_n$

Millä todennäköisyydellä seuraava nosto tuottaa valkoisen pallon, eli $X_{n+1} = 1$?

Tähän kysymykseen on hankala antaa vastausta frekventistisen paradigman puitteissa; voitaisiin yrittää ennustaa sijoittamalla parametrin K järkevä frekventistinen estimaatti kaavaan K/N ts. ennuste olisi $\hat{K}/N$. Tällöin täysin sivuutetaan se ongelma, että K:n estimointiin liittyy epävarmuutta.

Bayesläisessä paradigmassa ennustaminen on periaatteessa suoraviivaista: pitää tarkastella <u>ennuste</u>- eli <u>prediktiivistä</u> todennäköisyyttä

$$P(X_{n+1} = 1 \mid \underline{X} = \underline{x})$$

Tämä tapa ottaa huomioon parametrin estimointiin liittyvän epävarmuuden!

Miten ennustetodennäköisyys lasketaan? Yksinkertaisin tapa on hyödyntää ehdollista riippumattomuutta. Kun $K=k$, on

$$P(\underline{X}=\underline{x},\ X_{n+1}=x_{n+1} \mid K=k)$$
$$=\left(\frac{k}{N}\right)^{x_1}\left(1-\frac{k}{N}\right)^{1-x_1}\cdots\left(\frac{k}{N}\right)^{x_n}\left(1-\frac{k}{N}\right)^{1-x_n}\left(\frac{k}{N}\right)^{x_{n+1}}\left(1-\frac{k}{N}\right)^{1-x_{n+1}}$$
$$=P(\underline{X}=\underline{x} \mid K=k)\,P(X_{n+1}=x_{n+1} \mid K=k)$$

Siis $\underline{X} \perp\!\!\!\perp X_{n+1} \mid K$.

Nyt

$$P(X_{n+1}=x_{n+1} \mid \underline{X}=\underline{x})$$
$$=\sum_k P(X_{n+1}=x_{n+1},\ K=k \mid \underline{X}=\underline{x}) \quad \text{(marginalisointi)}$$
$$=\sum_k P(K=k \mid \underline{X}=\underline{x})\,P(X_{n+1}=x_{n+1} \mid \underline{X}=\underline{x},\ K=k)$$
(kertolaskukaava)
$$=\sum_k P(K=k \mid \underline{X}=\underline{x})\,P(X_{n+1}=x_{n+1} \mid K=k)$$
$(\underline{X} \perp\!\!\!\perp X_{n+1} \mid K)$

Tässä siis summataan posterioritn $P(K=k \mid \underline{X}=\underline{x})$ kertaa

$$P(X_{n+1}=x_{n+1} \mid K=k)=\begin{cases}\frac{k}{N}, & \text{jos } x_{n+1}=1\\ 1-\frac{k}{N}, & \text{jos } x_{n+1}=0\end{cases}$$

Ennustejakauma on tietenkin ptnf, ell

$$\underbrace{P(X_{n+1}=0 \mid \underline{X}=\underline{x})}_{\geq 0}+\underbrace{P(X_{n+1}=1 \mid \underline{X}=\underline{x})}_{\geq 0}=1.$$