8. Todennäköisyyslaskentaa

Emmen bayesläisen päättelyn käsittelyä kertaamme ehdolliseen todennäköisyyteen liittyvää kalkyyliä. [AS, jakso 1.4; TN I, jaksot 1.7 ja 1.10]
[TN I ≡ P. Tuominen: Todennäköisyyslaskenta I]

Olkoot A ja B tapahtumia. Oletetaan, että $P(B) > 0$ ja että tiedämme, että B on sattunut (mutta emme tiedä mitään muuta). Miten tällöin pitää määritellä A:n tn? Vastaus: pitää laskea A:n todennäköisyys ehdolla B, eli

$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

[TN I, määr. 1.7.1]

Olkoot $P(A) > 0$ ja $P(B) > 0$. Ehdollisen tn:n määritelmästä saadaan (todennäköisyyksien) kertolaskukaava eli ketjusääntö:

$$P(A \cap B) = P(B)\,P(A|B) = P(A)\,P(B|A)$$

Tästä voidaan ratkaista toinen ehdollisista todennäköisyyksistä, $P(B|A)$, jos $P(A|B)$, $P(A)$ ja $P(B)$ tunnetaan:

$$P(B|A) = \frac{P(B)\,P(A|B)}{P(A)}$$

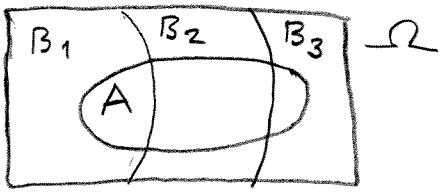
Olkoon $B_1, B_2, \ldots, B_M$ jokin perusjoukon Ω ("varman tapahtuman") ositus ts.

- joukot B_i ovat pistevieraita
- niiden yhdiste $\bigcup_{j=1}^{M} B_j = \Omega$

Oletetaan, että tunnetaan

- tn:t $P(B_k)$, $k=1,\ldots,M$
- ehdolliset tn:t $P(A \mid B_k)$, $k=1,\ldots,M$

Havaitaan, että tapahtuma A sattuu.
Miten lasketaan tämän tiedon valossa tapahtumien B_k todennäköisyydet $P(B_k|A)$?



Ratkaisu:

$$P(B_k \mid A) = \frac{P(B_k \cap A)}{P(A)} = \frac{P(B_k)\,P(A \mid B_k)}{P(A)}$$

Miten lasketaan $P(A)$? Ratkaisu:
A voidaan osittaa pistevieraisiin paloihin $A \cap B_i$, $i=1,\ldots,M$, joten (todennäköisyyden additiivisuus:)

$$\begin{aligned} P(A) &= P(A \cap B_1) + \cdots + P(A \cap B_M) \\ &= P(B_1)\,P(A \mid B_1) + \cdots + P(B_M)\,P(A \mid B_M) \end{aligned}$$

Tämä on <u>kokonaistodennäköisyyden kaava.</u>

Nyt saadaan <u>Bayesin kaava</u>:

$$P(B_k \mid A) = \frac{P(B_k)\,P(A \mid B_k)}{\sum_{i=1}^{M} P(B_i)\,P(A \mid B_i)}$$

Laskemista helpottava huomio:
Bayesin kaava voidaan esittää verrannollisuustuloksena

$$P(B_k|A) \propto P(B_k)\, P(A|B_k) =: q_k$$

Tämä tarkoittaa sitä, että

$$P(B_k|A) = c\, q_k = c\, P(B_k) P(A|B_k)$$

jossa $c = 1/P(A)$ on muuttujasta k riippumaton vakio. Tämä vakio voidaan määrittää siitä ehdosta, että

$$1 = \Sigma_k P(B_k|A) = c \Sigma q_k$$

$$\Rightarrow \quad \frac{1}{c} = \Sigma q_i = \Sigma P(B_i) P(A|B_i) = P(A)$$

Miksi sitten on ilmeistä, että $\Sigma_k P(B_k|A) = 1$?
Vastaus 1: asian näkee Bayesin kaavasta!
Vastaus 2: ehdollinen tn $B \mapsto P(B|A)$ on todennäköisyysmitta ja (B_k) on Ω:n ositus, Tn:n additiivisuuden nojalla

$$1 = P(\Omega|A) = \Sigma_k P(B_k|A).$$

Bayes läisessä tilastotieteessä
- $P(B_k)$, $k = 1, \dots, M$ on priorijakauma
  [lat. a priori ≡ ennen (havaintoa)]
- A vastaa havaintoa
- $k \mapsto P(A|B_k)$ on uskottavuusfunktio.
- $P(B_k|A)$, $k = 1, \dots, M$ on posteriorijakauma
  [lat. a posteriori = (havainnon) jälkeen]

Resepti (Priorista ja uskottavuusfunktiosta posteriorin, diskreetti tapaus)

Ainekset: Priorijakauma ja uskottavuusfunktio

Ohjeet: 1) Laske luvut q_k kertomalla

priori × uskottavuus, eli

$$q_k = P(B_k) \cdot P(A \mid B_k)$$

2) Laske summa $s = \sum q_k$.

3) Posteriorijakauma on

$$P(B_k \mid A) = q_k / s.$$

Esimerkkejä: kts. AS esim. 1.1 ja 1.2

Esim 1.1 : $n = 2$

$P(B_1) = 0.001$ $\quad P(B_2) = 0,999$

$P(A \mid B_1) = 0,997$ $\quad P(A \mid B_2) = 0,015$

$q_1 = 0,001 \times 0,997$ $\quad q_2 = 0,999 \times 0,015$

$s = q_1 + q_2 = 0,015982$

$$P(B_1 \mid A) = \frac{q_1}{s} \approx 0,062 \qquad P(B_2 \mid A) = \frac{q_2}{s} \approx 0,938$$

Suurin osa työstä kuluu siihen, että tilanteen selostuksesta ymmärretään, mikä on priori ja mikä uskottavuusfunktio. Varsinainen laskeminen on helppoa.

[vrt. AS: jakso 1.3]

9. Bayesläinen päättely: parametri satunnaismuuttujana

Frekventistisessä päättelyssä parametri on tuntematon, mutta kiinteä suure (kiinteä ≡ deterministinen ≡ ei-satunnainen). Tämän takia frekventistisessä tilastotieteessä parametriavaruudessa ei ole määriteltynä todennäköisyysjakaumaa, ja kaikkei todennäköisyyksiä koskevat lausumat [esim. luottamusvälin luottamustaso] pitää ajatella niin, että aineisto tulkitaan niissä satunnaiseksi. Jotta saataisiin konkreettinen tulkinta, voidaan ajatella, että alkuperäistä koetta toistetaan vastaavissa oloissa hyvin monta kertaa, ja sitten lasketaan kyseisen tapahtuman suhteellinen frekvenssi toistoissa (=> frekventistinen päättely).

Bayesläisessä päättelyssä toimitaan toisin:

- parametri ajatellaan satunnaismuuttujaksi
- havaintoja vastaavalle satunnaisvektorille $\underline{X}$ formuloidaan sen ehdollinen todennäköisyysjakauma ehdolla parametri
- kun havainnot $\underline{x}$ on saatu, satunnaisvektorille $\underline{X}$ kiinnitetään arvo $\underline{X} = \underline{x}$
- päättelyn tulos on parametrin posteriorijakauma, joka lasketaan Bayesin kaavalle:

$$\text{posteriori} \propto \text{priori} \times \text{uskottavuus}$$

Tarkastellaan ensin diskreettiä tapausta, jossa satunnaisen parametrin K arvo kuuluu johonkin äärellisen joukkoon S_K, ja havaintovektorin

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

arvo kuuluu johonkin äärelliseen joukkoon $S_{\underline{X}} \subset \mathbb{R}^n$.

Ptmf (pistetodennäköisyysfunktio)
$k \mapsto P(K=k)$ on priorijakauma
Ehdollinen ptmf

$$P(\underline{X} = \underline{x} \mid K=k) \qquad \underline{x} \in S_{\underline{X}}, \quad k \in S_K$$

kuvaa aineiston jakaumaan ehdolla $K=k$.

Yhdessä nämä kaksi funktiota spesifioivat parametrin ja havaintovektorin $\underline{X}$ yhteisjakauman, sillä (kertolaskukaava!)

$$P(K=k, \underline{X} = \underline{x}) = P(K=k)\ P(\underline{X}=\underline{x} \mid K=k)$$

Nämä kaksi funktiota määrittelevät bayesläisen tilastollisen mallin; malli M^* [AS].

Kun havaintovektorin $\underline{X}$ arvo $\underline{x}$ havaitaan, mitä sen jälkeen kaikki tieto parametrin K jakaumasta sisältyy sen posteriorijakaumaan

$$k \mapsto P(K=k \mid \underline{X}=\underline{x})$$

$$\propto \underbrace{P(K=k)}_{\text{priori}}\ \underbrace{P(\underline{X}=\underline{x} \mid K=k)}_{\text{uskottavuusfunktio}}$$

Esimerkki: mustat ja valkoiset pallot kulhossa:
N palloa, joista K (parametri) valkoista ja $N-K$ mustaa. Poimitaan palloja satunnaisesti palauttaen.

$$X_i = \begin{cases} 1, & \text{jos } i\text{:s nostettu pallo on valkoinen} \\ 0, & \text{jos } i\text{:s nostettu pallo on musta} \end{cases}$$

$$\begin{aligned} P(\underline{X} = \underline{x} \mid K = k) &= P(X_1 = x_1, \dots, X_n = x_n \mid K = k) \\ &= \left(\frac{k}{N}\right)^{x_1} \left(1 - \frac{k}{N}\right)^{1-x_1} \cdots \left(\frac{k}{N}\right)^{x_n} \left(1 - \frac{k}{N}\right)^{1-x_n} \\ &= \left(\frac{k}{N}\right)^{T(\underline{x})} \left(1 - \frac{k}{N}\right)^{n - T(\underline{x})} \end{aligned}$$

jossa $T(\underline{x}) = \sum_{i=1}^{n} x_i$ = nostettujen valkoisten lkm

$n - T(\underline{x})$ = nostettujen mustien lkm

$$\frac{k}{N} = P(\text{"nostetaan valkoinen"} \mid K = k)$$

$$1 - \frac{k}{N} = P(\text{"nostetaan musta"} \mid K = k)$$

Huomautus: satunnaismuuttujat $X_1, \dots, X_n$ ovat ehdollisesti riippumattomia ehdolla $K = k$, eli ne ovat riippumattomia niiden ehdollisessa yhteisjakaumassa, sillä

$$P(\underline{X} = \underline{x} \mid K = k) = \prod_{i=1}^{n} P(X_i = x_i \mid K = k)$$

$$P(X_i = x_i \mid K = k) = \left(\frac{k}{N}\right)^{x_i} \left(1 - \frac{k}{N}\right)^{1-x_i}$$

Ne eivät (yleensä) ole marginaalisesti riippumattomia eli riippumattomia (reuna-) yhteisjakaumassa

$$P(\underline{X} = \underline{x}) = \sum_{k} P(K = k, \underline{X} = \underline{x}) = \sum_{k} P(K = k) P(\underline{X} = \underline{x} \mid K = k)$$

joka ei yleensä faktoroidu reunajakaumiensa tuloksi.

Kun havaitaan, että $\underline{X} = \underline{x}$, sen jälkeen lasketaan posterionijakauma

$$P(K=k \mid \underline{X}=\underline{x}) = \frac{P(K=k)\,P(\underline{X}=\underline{x} \mid K=k)}{\sum_i P(K=i)\,P(\underline{X}=\underline{x} \mid K=i)}$$

$$\propto P(K=k)\,P(\underline{X}=\underline{x} \mid K=k)$$

"Pallot kulhossa" -esimerkissä riittää tuntea nostettujen valkoisten pallojen lkm $T(\underline{x})$, sillä tämä tunnusluku kertoo havainnoista kaiken tarvittavan (se on ns. tyhjentävä tunnusluku): uskottavuusfunktion $P(\underline{X}=\underline{x} \mid K=k)$ arvot voidaan laskea heti, kun $T(\underline{x})$ tiedetään.

Mitä parametrin satunnaisuus tarkoittaa?

Yksi mahdollisuus: K on oikeasti poimittu jostakin jakaumasta, ja sitten kulho on täytetty.

Toinen mahdollisuus: pidän valkoisten pallojen lukumäärää satunnaismuuttujana, koska en tiedä mikä arvo sillä on. Tässä todennäköisyyden <u>subjektiivisessa tulkinnassa</u> priorijakauma kuvaa kvantitatiivisesti subjektini (esim. minun tai sinun) epävarmuuden parametrin arvosta. Havainnon teon jälkeen tämän epävarmuuden ilmaisee posterionijakauma, joka saadaan lasketuna Bayesin kaavalla.

Bayesläisessä tilastotieteessä käytetään yleensä tätä subjektiivista tulkintaa.