J Kohonen

6A Bayes-päättely II

Tehtäviin kannattaa valmistautua tutustumalla luentoihin 5A ja 5B.

Tuntitehtävät

6A1 (Ennustejakauma) Eräs kolikko tuottaa kruunia (merkitään 1:llä) tuntemattomalla tn:llä Θ, ja klaavoja (0) tn:llä $1 - \Theta$, joka heitolla riippumattomasti. Θ:n priori on tasajakauma välillä $[0, 1]$.

(a) Jos parametrilla on tietty arvo $\Theta = \theta$, mikä on silloin todennäköisyys, että 10:llä peräkkäisellä heitolla kaikki tulokset ovat kruunia?

(b) Priorijakauman perusteella (ts. ennen yhtään havaintoa), mikä on todennäköisyys että ensimmäiset 10 heittotulosta ovat kruunia? Ohje: Jos tulosjonoa merkitään vektorilla $\vec{x}$, niin osituskaavan perusteella

$$f(\vec{x}) = \int_0^1 f_{\vec{X}|\Theta}(\vec{x} \,|\, \theta) \; f_\Theta(\theta) \; d\theta.$$

Toinen integraalin sisällä olevista suureista on jo laskettu a-kohdassa ja toinen on priori. Integraali pitää vielä laskea.

(c) Kolikkoa on heitetty 20 kertaa ja kaikki tulokset olivat kruunia. Mikä on nyt parametrin Θ posteriorijakauma? (Voit joko laskea posterioritiheyden, tai päätellä jakauman nimen ja parametrit luentojen perusteella.)

(d) C-kohdan havaintojen perusteella, mikä on todennäköisyys sille, että *seuraavat* 10 heittotulosta ovat kaikki kruunia? Ohje: Käytä jälleen osituskaavaa, mutta tällä kertaa käytä Θ:n posteriorijakaumaa priorin sijasta (ks. luento 5B).

(e) Vertaa b- ja d-kohtien numeerisia tuloksia seuraavaan asetelmaan: *reiluksi tiedettyä* kolikkoa ($\theta = 0.5$) heitetään 10 kertaa ja kysytään todennäköisyyttä saada 10 kruunaa. Koeta selittää näiden kolmen luvun suuruusjärjestys arkijärjellä.

Ratkaisu.

(a) θ^{10}.

(b) Olkoon $\vec{X}$ heittotuloksia kuvaava satunnaisvektori ja $\vec{1}$ vektori, jossa on kymmenen ykköstä.

$$\begin{aligned} f_{\vec{X}}(\vec{1}) &= \int_0^1 f_{\vec{X}|\Theta}(\vec{1} \,|\, \theta) f_\Theta(\theta) d\theta \\ &= \int_0^1 \theta^{10} \cdot 1 \; d\theta \\ &= \frac{1}{11} \approx 0.091. \end{aligned}$$

(c) Olkoon $\vec{X}$ ensimmäisiä 20 tulosta kuvaava satunnaisvektori ja $\vec{1}$ vektori, jossa on 20 ykköstä. Tällöin Θ:n normalisoimaton posterioritiheys on priori kertaa uskottavuus eli $1 \cdot \theta^{20}$, kun $0 \leq \theta \leq 1$, ja kun se normalisoidaan, saadaan posteriori

$$f_{\Theta|\vec{X}}(\theta \,|\, \vec{1}) = 21\theta^{20}.$$

Toisin sanoen Θ:n posteriorijakauma on $\mathrm{Beta}(21, 1)$.

(d) Olkoot ensimmäiset 20 heittoa $\vec{X}$ ja seuraavat 10 heittoa $\vec{Y}$. Molemmissa tapauksissa merkitsemme “kaikki ykkösiä”-tulosta vektorilla $\vec{1}$. Osituskaavan nojalla

$$\begin{aligned} f_{\vec{Y}|\vec{X}}(\vec{1} \,|\, \vec{1}) &= \int_0^1 f_{\vec{Y}|\Theta}(\vec{1} \,|\, \theta) f_{\Theta|\vec{X}}(\theta \,|\, \vec{1}) \, d\theta \\ &= \int_0^1 \theta^{10} \cdot 21\theta^{20} \, d\theta \\ &= 21 \int_0^1 \theta^{30} \, d\theta \\ &= \frac{21}{31} \approx 0.677. \end{aligned}$$

(e) Reiluksi tiedetyllä kolikolla kymmenen kruunan tn on $(\frac{1}{2})^{10} = \frac{1}{1024} \approx 0.001$.

Selitys arkijärjellä:

- Reiluksi tiedetyllä kolikolla on aika epätodennäköistä saada pelkkiä kruunia 10 kertaa.
- B-kohdassa priorijakauman mukaan oli kohtuullisen mahdollista, että kolikko on paljonkin kruunapainoinen, esim. oli 10%:n todennäköisyys sille, että $\Theta \geq 0.90$. Siten 10 kruunan saaminen ei olisi kovinkaan yllättävää.
- D-kohdassa on jo saatu havainnoista lisätukea sille, että kolikko olisi kruunapainoinen (Θ suuri), joten on luonnollista että 10 kruunan tn on nyt entistäkin suurempi.

6A2 (Eksponenttijakauman Bayes-päättely) Jos jatkuvalla satunnaismuuttujalla Λ on tiheys

$$f(\lambda) = \begin{cases} c\ \lambda^{\alpha-1}\ e^{-\beta\lambda} & \text{kun } \lambda > 0, \\ 0 & \text{muuten,} \end{cases}$$

sanomme että Λ:lla on **gammajakauma** parametrein $\alpha > 0$ ja $\beta > 0$. Tätä merkitään $\Lambda \sim \mathrm{Gam}(\alpha, \beta)$. (Jos $\alpha = 1$, saadaan ennestään tuttu eksponenttijakauma.) Tiheydessä c on normalisointivakio, joka saa aikaan että $\int_0^\infty f(\lambda) = 1$. Normalisointivakion arvo voidaan (jos α on kokonaisluku) laskea kaavalla $c = \beta^\alpha/(\alpha-1)!$, missä ! tarkoittaa kertomaa. Lisäksi tiedetään, että jos $\Lambda \sim \mathrm{Gam}(\alpha, \beta)$, niin $\mathrm{E}(\Lambda) = \alpha/\beta$.

(a) Eräitä alkeishiukkasia hajoaa satunnaisin välein riippumattomasti tuntemattomalla taajuusparametrilla Λ (hajoamista sekunnissa). Parametrille oletetaan priorijakauma $\Lambda \sim \mathrm{Gam}(2, 10)$. Kirjoita priorin tiheysfunktio. Laske Λ:n odotusarvo. (Yllä annetuilla kaavoilla selviät ilman integroimista.)

(b) Olkoon X_i $(i = 1, 2, \ldots)$ väliaika i:nnen ja $(i+1)$:nnen hajoamisen välillä. Jos taajuusparametrilla on arvo $\Lambda = \lambda$, niin X_i:llä on eksponenttijakauma taajuusparametrilla λ, tiheydellä

$$f_{X_i|\Lambda}(x_i \mid \lambda) = \lambda e^{-\lambda x_i},$$

kun $x_i > 0$. Laske Λ-parametrin normalisoimaton posterioritiheys

$$f_\Lambda(\lambda)\ f_{\vec{X}|\Lambda}(\vec{x} \mid \lambda),$$

kun on havaittu kolme hajoamisten väliaikaa (sekunteina) $\vec{x} = (x_1, x_2, x_3) = (3.0, 12.2, 16.1)$. Ohje: Koska kolme väliaikaa ovat riippumattomat, $f(\vec{x} \mid \lambda) = f(x_1 \mid \lambda)\ f(x_2 \mid \lambda)\ f(x_3 \mid \lambda)$. Sievennä lauseke mahdollisimman yksinkertaiseksi.

(c) Tarkastele edellä laskettua normalisoimatonta posterioria. Etsi posteriorimoodi eli posterioritiheyden maksimikohta. (Vihje: Logaritmista ja derivoimisesta voi olla apua. Normalisointivakion arvoa ei tarvitse ratkaista.)

(d) Ratkaisemasi posterioritiheyden pitäisi itse asiassa olla erään gammajakauman tiheysfunktio (älä välitä tässä normalisointivakiosta). Mikä gammajakauma on kyseessä? Vihje: Katso λ:n ja e:n eksponentteja.

(e) Nyt kun tunnet Λ:n posteriorijakauman nimen ja parametrit, laske sen odotusarvo (taaskin helposti edellä annetulla kaavalla).

(f) (Vapaaehtoinen lisätehtävä, edellyttää tietokonetta.) $\mathrm{Gam}(\alpha, \beta)$-jakauman q-kvantiilin voi laskea R-komennolla `qgamma(q, alpha, beta)`. Etsi Λ-parametrille posteriorijakauman mukainen 95% uskottavuusväli.

Ratkaisu.

(a) Sijoittamalla annettuun kaavaan $\alpha = 2$ ja $\beta = 10$ saadaan tiheys

$$f(\lambda) = 100\,\lambda\,e^{-10\lambda}$$

kun $\lambda > 0$. Annetun kaavan mukaisesti jakauman $\mathrm{Gam}(2, 10)$ odotusarvo on $2/10 = 0.2$. Jos λ:n arvo on lähellä odotusarvoa, niin arvioimme silloin hiukkasia hajoavan osapuilleen taajuudella 0.2 hajoamista sekunnissa, eli yksi hajoamimen 5 sekunnissa.

(b) Priorin ja uskottavuuden tulo on

$$\begin{aligned}
&100\,\lambda\,e^{-10\lambda}\ \lambda e^{-\lambda x_1}\ \lambda e^{-\lambda x_2}\ \lambda e^{-\lambda x_3} \\
=&100\,\lambda^4\,e^{-\lambda\,(10+x_1+x_2+x_3)} \\
=&100\,\lambda^4\,e^{-\lambda\,(10+3.0+12.2+16.1)} \\
=&100\,\lambda^4\,e^{-\lambda\,41.3},
\end{aligned}$$

kun $\lambda > 0$.

(c) Normalisoimattoman posteriorin logaritmi on

$$\log(100) + 4\log(\lambda) - 41.3\lambda,$$

sen derivaatta on

$$4/\lambda - 41.3$$

ja derivaatan ainoa nollakohta on

$$\lambda = 4/41.3 \approx \mathbf{0.097}.$$

Tämä on siis λ-parametrin posteriorimoodi.

(d) Tiheysfunktio on muotoa

$$c\,\lambda^4\,e^{-41.3\lambda}.$$

Vihjeen mukaisesti, ja vertaamalla gammajakauman tiheysfunktioon, huomataan, että λ:n eksponentti on $\alpha - 1 = 4$, joten $\alpha = 5$. Vastaavasti e:n eksponentti on $-\beta\lambda = -41.3\lambda$, joten $\beta = 41.3$. Kyseessä on siis jakauma $\mathrm{Gam}(5, 41.3)$.

(e) Käyttäen annettua odotusarvon kaavaa, posteriorijakauman $\mathrm{Gam}(5, 41.3)$ odotusarvo on $5/41.3 \approx \mathbf{0.121}$.

(f) Lasketaan välin päätepisteet R-komennoilla as `qgamma(0.025, 5, 41.3)` ja `qgamma(0.975, 5, 41.3)`. Uskottavuusväliksi saadaan $[0.039, 0.248]$.

Kolmen havainnon jälkeen olemme 95% varmoja siitä, että hiukkasten hajoamistaajuus on välillä $[0.039, 0.248]$ (hajoamista sekunnissa). Vrt. harjoitustehtävään 5A4.

Kotitehtävät

6A3 (DNA-malli) Ihmisen DNA:ta voidaan pitää merkkijonona, jossa on $3 \cdot 10^9$ kirjainta, kukin aakkostosta A, C, G, T. Merkkijonon pituudesta vain noin 1.5% on *eksoneja* (proteiinia koodaavia alueita eli suoria rakennusohjeita proteiinien rakentamiseen aminohappoja ketjuttamalla). Eksonien ulkopuolisilla DNA:n osilla on muita tehtäviä.

Erään tutkimuksen mukaan eksoneissa kirjaimia esiintyy osuuksin $(0.16, 0.30, 0.32, 0.22)$, ja muualla DNA:ssa osuuksin $(0.25, 0.25, 0.25, 0.25)$. Käytämme yksinkertaista stokastista mallia, jonka mukaan kukin kirjain on satunnainen em. todennäköisyyksin, riippuen vain siitä kummanlaisessa alueessa merkkijonoa ollaan.

Tutkimme merkkijonossa erästä satunnaisesti valittua paikkaa i. Sitä, kuuluuko kyseinen paikka eksoniin vai ei, merkitsemme tuntemattomalla parametrilla Θ ($\Theta = 1$ jos eksonissa, $\Theta = 0$ jos ei). Tutkimme paikan i ympäriltä 100 peräkkäistä kirjainta, ja havaitsemme siinä erään jonon AACTG…TGA, jossa kirjaimia ACGT on lukumäärät 14, 30, 36 ja 20. Oletamme yksinkertaisuuden vuoksi, että koko tutkimamme 100-merkkinen jono on joko kokonaan eksonia tai kokonaan eksonien ulkopuolella.

(a) Mikä on parametrin Θ priorijakauma, kun käytetään vain tietoa, että paikka DNA-jonossa valittiin satunnaisesti, mutta ei katsota jonon sisältöä?

(b) Jos $\Theta = 0$, mikä on todennäköisyys havaita juuri se 100 kirjaimen jono, jonka havaitsimme? Vihje. Ajattele luennon 5B järjestettyjä jonoja, niin et tarvitse multinomikertoimia. Laske jonon todennäköisyys ainakin kolmella merkitsevällä numerolla. Älä hämmenny siitä, että jonon tn on aika pieni.

(c) Jos $\Theta = 1$, mikä on todennäköisyys havaita juuri se 100 kirjaimen jono, jonka havaitsimme?

(d) Laske Θ:n posteriorijakauma ja selitä sanallisesti, mitä se tarkoittaa. (Huomaa, että Θ on diskreetti parametri, sen mahdolliset arvot ovat 0 ja 1.)

Tämä on erittäin yksinkertaistettu malli DNA:n rakenteesta, ja luvut ovat osittain keksittyjä. Tämäntapaisilla malleilla on kuitenkin todella etsitty DNA:sta eksonikohtia.

Arviointiohje. 0.5 p per kohta.

Ratkaisu.

(a) $P(\Theta = 0) = 0.985$ and $P(\Theta = 1) = 0.015$. Koska eksonien osuus DNA-jonon pituudesta on 1.5%, tämä on myös tn, että satunnaisesti valittu kohta kuuluu eksoniin.

(b) Olkoon $\vec{X} = (X_1, X_2, \ldots, X_{100})$, missä $X_j \in \{A, C, G, T\}$. Huomaa, että kirjaimet A, C, G, T ovat X_i:n mahdolliset *arvot*. (Jos tuntuu luontevammalta, voit ajatella niitä kokonaislukuina $1, 2, 3, 4$.)

Todennäköisyys havaita juuri tämä jono $\vec{x}$ eksonien ulkopuolella on

$$f_{\vec{X}|\Theta}(\vec{x} \,|\, 0) = 0.25^{14} \cdot 0.25^{30} \cdot 0.25^{36} \cdot 0.25^{20} = 0.25^{100} \approx 6.22 \cdot 10^{-61}.$$

(c) Todennäköisyys havaita juuri tämä jono $\vec{x}$ eksonialueella on

$$f_{\vec{X}\,|\,\Theta}(\vec{x}\,|\,1) = 0.16^{14} \cdot 0.30^{30} \cdot 0.32^{36} \cdot 0.22^{20} \approx 1.60 \cdot 10^{-58}.$$

(d) Koska Θ:lla on vain kaksi mahdollista arvoa, päättely on kätevää esittää taulukkona.

θ	priori $f(\theta)$	uskottavuus $f(\vec{x}\,\|\,\theta)$	tulo	posteriori $f(\theta\,\|\,\vec{x})$
0	0.985	$6.22 \cdot 10^{-61}$	$6.13 \cdot 10^{-61}$	0.203
1	0.015	$1.60 \cdot 10^{-58}$	$2.41 \cdot 10^{-60}$	0.797
$\sum$	1.000		$3.02 \cdot 10^{-60}$	1.000

Posteriorijakauma kertoo, että paikka, jota tarkastelemme, kuuluu noin 80% todennäköisyydellä eksoniin.

6A4 (Kolikon ravistelu) Professori Abel on rakentanut kolikonheittokoneen. Aluksi kolikko asetetaan juomalasin pohjalle klaavapuoli ylöspäin (0). Kone ravistaa lasia vähän aikaa, minkä jälkeen tutkitaan onko kolikossa klaava (0) vai kruuna (1) ylöspäin. Ravistelua toistetaan ja näin saadaan jono satunnaislukuja $X_1, X_2, X_3, \ldots$, missä $X_i \in \{0, 1\}$ on kolikon asento i:n ravistelun jälkeen. Alkutila $X_0 = 0$ on tunnettu.

Ravistelu tehtiin 50 kertaa ja näin saatiin jono

$$\vec{x} = (x_1, x_2, \ldots, x_{50}) = (01111110000000011111110011111100111110000000000000)$$

(Merkintä tarkoittaa 50:n numeron jonoa, vaikka siinä ei ole pilkkuja välissä.)

Fysikaalisen ymmärryksemme mukaan arvioimme, että jokaisessa ravistelussa kolikko *kääntyy* eräällä tn:llä θ ja *ei käänny* tn:llä $1-\theta$, riippumatta asennosta ennen ravistelua ja aiemmista tapahtumista. Merkitsemme $K_i = 1$ jos kolikko kääntyy i:nnessä ravistelussa ja $K_i = 0$ jos ei käänny.

(a) Katso havaittua jonoa $\vec{x}$. Vaikuttaako se mielestäsi reilun kolikon heittotulosten jonolta (jossa kukin heittotulos on aiemmista riippumaton)?

(b) Huomaamme, että kokeen aikana kolikko kääntyi 8 kertaa. Pidetään kääntymistodennäköisyyttä tuntemattomana parametrina Θ, jolla on tasapriori välillä $[0, 1]$. Määritä Θ:n posteriorijakauma ja jokin haluamasi piste-estimaatti sille (esimerkiksi moodi tai odotusarvo).

Vihje. Ajattele kääntymistä kuvaavien indikaattorien K_i jonoa. Mikä jakauma niillä on? Millainen havainto saatiin? Muista, että binaarimallin parametrin posteriori on eräs beetajakauma. Voit lisäksi käyttää tietoa, että jakauman Beta(a, b) odotusarvo on $a/(a + b)$ (vrt. luento 5B).

(c) Määritä Θ:lle 95% uskottavuusväli, ts. sellainen väli, että Θ on kyseisellä välillä 95% todennäköisyydellä (havainnoista lasketun posteriorijakauman perusteella). Vihje: Kannattaa käyttää tietokonetta. Esim. R-komento `qbeta(q, a, b)` antaa Beta(a, b)-jakauman q-kvantiilin.

(d) Professori Abel huomaa, että kone tuotti 23 kertaa kruunan ja 27 kertaa klaavan. Tämän perusteella hän pitää konettaan menestyksenä ja väittää kokeen osoittavan, että kone tuottaa kruunia ja klaavoja yhtä todennäköisesti, täysin satunnaisesti ja toisistaan riippumatta. Pohdi mitä mieltä olet Abelin järkeilystä.

(e) Oletetaan, että hyvin pitkän koesarjan jälkeen olemme päätyneet siihen, että $\theta = 0.17$ hyvin suurella tarkkuudella. Laske todennäköisyys sille, että kun kolikkoa ravistetaan kymmenen kertaa, niin se päätyy samaan asentoon kuin ennen ravisteluja. Ilmoita tulos neljällä desimaalilla. Vihje: Mieti kääntymisten lukumäärää ravistelusarjassa. Mitä tapahtuu, jos lukumäärä on parillinen?

Arviointiohje. 0.4 pistettä per kohta, pyöristetään ylöspäin (ts. kolme kohtaa riittää 2 pisteeseen). B-kohdassa riittää, että löydettiin beetajakauma suunnilleen oikein parametrein, ja joku piste-estimaatti. D-kohdassa pisteet saa mielekkäästä pohdinnasta, esim. sen toteamisesta, että peräkkäiset tulokset eivät vaikuta riippumattomilta. E-kohdassa pisteet oikeasta numeerisesta tuloksesta.

Ratkaisu.

(a) Ei vaikuta. Kolikko vaikuttaa pysyvän usein pitkän aikaa samassa asennossa. (Jos heitettäisiin reilua kolikkoa riippumattomasti, niin joka heitolla pitäisi olla 50% todennäköisyys saada eri tulos kuin edellisellä heitolla. Nyt kolikko ei näyttäisi kääntyvän näin usein.)

(b) Θ:n priori on tasajakauma eli $f(\theta) = 1$.

Kääntymiset noudattavat binaarimallia, sillä joka ravistelukerralla on sama tn θ kolikon kääntymiselle (ja $1-\theta$ kääntymättä jäämiselle), aiemmista tapahtumista riippumattomasti. Siten uskottavuus 8 kääntymisen ja 42 kääntymättä jäämisen jonolle on $\theta^8(1-\theta)^{42}$. Normalisoimaton posteriori on $1 \cdot \theta^8(1-\theta)^{42}$. Voisimme normalisoida sen integroimalla, mutta vihjeen mukaisesti käytämme tietoa, että kyseessä on jakauma $\mathrm{Beta}(9, 43)$.

Eräät piste-estimaatit ovat

- posteriorin odotusarvo $= \frac{9}{9+43} \approx 0.173$, käyttäen vihjettä beetajakauman odotusarvosta.
- posteriorin moodi $= \frac{8}{50} = 0.160$, jonka voi laskea etsimällä posteriorin logaritmin derivaatan nollakohdan.

(c) R-komennoilla `qbeta(0.025, 9, 43)` ja `qbeta(0.975, 9, 43)` saamme halutunlaisen välin päätepisteet. Väli on $[0.084, 0.286]$.

(d) On totta, että jokainen koneen tuottama tulos on *satunnainen* (eräästä jakaumasta), mutta tulokset ovat vahvasti toisistaan *riippuvia*. Edellisessä kohdassa saimme estimoitua, että kääntymistodennäköisyyttä kuvaava parametri Θ on luultavasti välillä $[0.084, 0.286]$, eli paljon alle puolikkaan. — Pitkän päälle kone kyllä tuottaa noin puolet kruunia ja puolet klaavoja, mutta ne esiintyvät pitkinä kruunajonoina ja klaavajonoina.

(e) Jotta kolikko on samassa asennossa kuin ennen ravisteluja, sen on pitänyt kääntyä parillinen määrä kertoja (0, 2, 4, 6, 8 tai 10). Koska kääntymiset ovat kussakin ravistelussa riippumattomia, kääntymisten määrä on binomijakautunut parametrein $n = 10$ ja $\theta = 0.17$. Yhteenlaskukaavan perusteella tn, että kääntymisten määrä on parillinen, on

$$b(0) + b(2) + b(4) + b(6) + b(8) + b(10) \approx 0.5078,$$

missä $b(k) = \binom{10}{k} 0.17^k 0.83^{10-k}$ on em. binomijakauman tiheys pisteessä k. Helppo tapa tämän laskemiseen on R-komento

```
sum(dbinom(c(0,2,4,6,8,10), 10, 0.17))
```

Professori Abel oli osittain oikeassa tässä mielessä: Koneella saadaan aikaan kohtuullisen riippumattomia ravistelutuloksia, kunhan tarkastellaan tuloksia riittävän monen ravistelun välein. Esim. 10 ravistelun välein tarkastelemalla saadaan jono $X_{10}, X_{20}, X_{30}, \ldots$, jossa seuraava tulos on lähes riippumaton edellisestä, sillä kolikko kääntyy 10 ravistelussa suurin piirtein todennäköisyydellä 1/2.

Tällainen riippuvan havaintojonon *ohentaminen* on osa moderneja Markov Chain Monte Carlo -menetelmiä (MCMC), joilla pystytään tutkimaan varsin monimutkaisiakin posteriorijakaumia. Nämä ovat tämän johdantokurssin ulkopuolella.