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.

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

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

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

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

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, $f(\vec{x} \,|\, \lambda) = f(x_1 \,|\, \lambda)\; f(x_2 \,|\, \lambda)\; f(x_3 \,|\, \lambda)$. 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.

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.

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

$$\vec{x} = (x_1, x_2, \ldots, x_{50}) = (01111110000000011111110011111001111100000000000000)$$

(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?