Bayes tilastotiede kotitentti, kevät 2011

Ohjeet: Kaksi hyvin laskettua tehtävää riittää kurssin suorittamiseen hyvällä arvosanalla. Voit käyttää kirjallisuutta, ja tarvittaessa pyytää neuvoa opettajalta. Kotitentti palautetaan ennen juhannusta.

1. Oletetaan että olet päättynyt tuntemattomaan maan suurkaupunkiin, jonka kokoa et tiedä. Ensimmäiseksi näet kahta bussia, jonka numerot ovat 87 ja 144. Oletamme että bussilinjat numeroidaan kokonaisluvuin ykkösestä lähtien ja kaikilla bussilinjoilla on sama todennäköisyys tulla havaituksi. Mitä voit päätellä kaupungin bussilinjojen määrästä havaintojen perusteella ? Aseta priori bussilinjoille määrälle, kokeile yhdellä aidolla priorilla ja yhdelllä epäaidolla priorilla. Laske posteriori jakauma, sen odotusarvon ja maksimi.

   Tarvittaessa voit approksimoida summia integraaleilla:

$$\sum_{n=1}^{\infty} f(n) \sim \int_0^{\infty} f(x)dx$$

2. Vertaillaan liinearinen ja kvadraattinen regressiomallit:

$$\begin{aligned} \mathcal{M}_1 : &\quad Y_i = a + bX_i + \sigma\varepsilon_i \\ \mathcal{M}_2 : &\quad Y_i = \alpha + \beta X_i + \gamma X_i^2 + \eta\varepsilon_i \end{aligned}$$

   jossa $\varepsilon_i \sim \mathcal{N}(0,1)$ on standardi-gaussinen.

   Data on $(X_i, Y_i, i = 1, \ldots, N)$ jossa X_i (metri/sekunnissa) on auton nopeus jarrutushetkellä ja Y_i (metri) on jarrutusmatka, eli montako metria auto vielä kulkee jarrutuksen alkuhetkestä ennen kun se pysähtyy kokonaan.

   - Aseta informatiiviset priorit parametreille (a, b, σ), mallissa $\mathcal{M}_1$ ja parametreille $(\alpha, \beta\gamma, \eta)$ mallissa $\mathcal{M}_2$.
     Jos olet itse ajanut tai ollut auton kyydissä, käytä omaa käytännon kokemusta auton jarrutuksesta.Jos haluat voit käyttää myös tietoa perusfyysikasta.
     Tämä valinta ei perustu Y dataan, älä vielä katso Y dataa ennen kun olet valinnut prioreja !
     Käytä gaussinen-käänteisgamma konjugatti-prioreja.
   - Piirra R:lla lineaarinen ja qvadrattinen regressiokäyrät

     $X = (4, 4, 7, 7, 8, 9, 10, 10, 10, 11, 11, 12, 12, 12, 12, 13, 13, 13, 13, 14, 14, 14, 14, 15, 15, 15,$
     $16, 16, 17, 17, 17, 18, 18, 18, 18, 19, 19, 19, 20, 20, 20, 20, 20, 22, 23, 24, 24, 24, 24, 25)$
     $Y = (2, 10, 4, 22, 16, 10, 18, 26, 34, 17, 28, 14, 20, 24, 28, 26, 34, 34, 56, 26, 36, 60, 80, 20, 26,$
     $54, 32, 40, 32, 40, 50, 42, 56, 76, 84, 36, 46, 68, 32, 48, 52, 56, 64, 66, 54, 70, 92, 93, 120, 85)$

     (poimi data liitetiedostosta joka löytyy kurssin Koppa-sivulta)

- Laske parametrien posteriorit molemmissa malleissa.
- Laske posteriorit myös winbugsilla ja vertaile tuloksia.
- Laske myös posteriorit epäinformatiivisilla prioreilla

$$\pi(\beta) = \pi(\alpha) = \pi(\gamma) = \pi(b) = \pi(a) = 1, \quad \pi(\eta^2) = \frac{1}{\eta^2}, \pi(\sigma^2) = \frac{1}{\sigma^2}$$

- Bayesin kerroin on

$$\frac{p(Y|X,\mathcal{M}_2)}{p(Y|X,\mathcal{M}_1)} = \frac{\int_0^\infty \int_{-\infty}^\infty \int_{-\infty}^\infty \int_{-\infty}^\infty p(Y|\alpha,\beta,\gamma,\eta^2,\mathcal{M}_1)\pi(\alpha|\eta^2)\pi(\beta\eta^2)\pi(\gamma|\eta^2)\pi(\eta^2)d\alpha d\beta d\gamma d\eta^2}{\int_0^\infty \int_{-\infty}^\infty \int_{-\infty}^\infty p(Y|a,b,\sigma^2,\mathcal{M}_1)\pi(a|\sigma^2)\pi(b|\sigma^2)\pi(\sigma^2)dbdad\sigma^2}$$

  Bayesin kerroin riippuu priorin valinnasta. Muistetaan että kun verrataan eri malleja, Bayesin kertoimissa sallitaan vain aitoja priorijakaumia, paitsi silloin kun sama parametri esintyy kaikissa malleissa. Silloin sen parametri priori voi olla myös epäaito. Esimerkiksi residuaalien varianssi parametreilla σ^2, η^2, voisi olla sama epäaito priori kaikissa malleissa.

  i) Laske Bayesin kerroin Normaali approksimaatiolla posteriorin-maximin estimaattorin ympärillä. (käytä omia informatiivisia prioreja)

  ii) Laske sama Bayesin kerroin Winbugsilla. Voit muokata winbugs-esimerkki

  Pine: Bayes factors using pseudo priors, sivu 38 http://www.mrc-bsu.cam.ac.uk/bugs/winbugs/Vol3.pdf

  iii) Eksaktisti: integroimalla parametreja konjugaatti-priorin suhteen analyyttisesti, tämä onnistuu kun tiedät konjugaatti jakauma perheiden normalisointi vaikioita.

  iv) Jos mallin indikaattorin I: priori on $\pi(\mathcal{M}_1) = \pi(\mathcal{M}_2) = 1/2$, laske quadrattisen mallin posteriori todennäköisyys $P(I = 2|Y, X)$.

  v) Kokeile miten Bayes kerroin muuttuu kun käytetään aitoja mutta paljon epäinformatiivisempia prioreja.

3. (Montako komponentteja sekoituksessa ?)

   Olkoon $\{\mathcal{M}_m\colon m \in \mathbb{N}\}$ gaussinen sekoitus malli, jossa datalle X on tiheys

$$p(x|\theta_m, \mathcal{M}_m) = \sum_{\ell=1}^{m} \omega_\ell \phi\left(\frac{x - x_\ell}{\eta_l}\right)$$

   jossa

$$\phi(x) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right)$$

   on standardi gaussinen tiheys.

   $\theta_m = (\omega_\ell, x_\ell, \eta_\ell : \ell = 1, \ldots, m)$

   on mallin $\mathcal{M}_m$ parametri.

Tässä $\omega_k \in [0,1]$ ja $(\omega_1 + \cdots + \omega_m) = 1$, siis ω on todennäköisyysvektori. Sen konjugaatti jakauma multinomiaaliselle datalle on Dirichlet jakauma jolla on tiheys.

$$p(\omega_1, \ldots, \omega_{m-1}, \omega_m) \propto \left\{ \prod_{k=1}^{m} \omega_k^{\alpha_k - 1} \mathbf{1}(\omega_k \in [0,1]) \right\} \mathbf{1}(\omega_1 + \cdots + \omega_m)$$

Mallin identifiotuvuuden takia voidaan olettaa $x_1 \leq x_2 \leq \cdots \leq x_m$.

Olkoon $I \in \{0, 1, 2, \ldots, \}$ mallin indikaattori, jossa komponenttien määrä ei ole rajoitettu.

Voidaan asettaa epäaito priori mallin indikaattorille $\pi(I) = 1$.

Mallien vertalilussa jokaisen mallin parametrien (θ_m) priori $\pi(\theta_m | \mathcal{M}_m)$ pitää olla aito todennäköisyysjakauma, paitsi silloin kun sama parametri esintyy kaikissa malleissa.

Olkoon otoskoko $N = 100$, data on

$X = c(-0.0153, 0.0305, -0.0120, 0.0856, 0.1742, 0.0465, 0.1982, 0.1541, 0.3241, 0.0961, 0.0981, 0.5720$
$0.6387, 0.8265, 0.8685, 1.0255, 1.1858, 1.1593, 1.2203, 1.3839, 1.5694, 1.3769, 1.4005, 1.5517, 1.4777, 1.$
$1.6918, 1.6941, 1.5925, 1.7422, 1.9580, 2.2188, 2.2998, 2.6073, 2.7795, 2.7310, 3.0587, 3.3305, 3.5364, 3.$
$3.6299, 3.8725, 3.7669, 3.8347, 3.5253, 3.5071, 3.3509, 3.3257, 3.2347, 3.0698, 3.3924, 2.3479, 2.1231, 1.$
$1.4268, 1.3024, 1.0590, 0.8167, 0.9264, 0.5895, 0.8094, 0.8233, 0.7721, 0.6843, 0.5710, 0.6625, 0.5485, 0.$
$0.4660, 0.3813, 0.5022, 0.4332, 0.4025, 0.3460, 0.5640, 0.4776, 0.3380, 0.3575, 0.4999, 0.5250, 0.4405, 0.$
$0.2838, 0.1174, 0.0647, 0.1204, 0.1818, 0.1004, 0.0915, 0.1725, 0.0785, -0.0797, 0.1072)$

(poimi data liite-tiedostosta kurssin Koppan sivulta).

Laske Bayesin kerroin winbugsilla (eksaktiisti olisi liian iso lasku niin suurella datamäärällä, miksi?)

$$\frac{p(X|\mathcal{M}_m)}{p(X|\mathcal{M}_1)}$$

jossa $\mathcal{M}_1$ on malli jossa on 1 komponentti. Kokeile kun $m = 2, 3, 4$.

Koska on kyse kahden mallien välisestä Bayesin kertoimestä, muita malleja ei tartvitse otta huomioon laskussa. Siis voit olettaa että a priori, mallien priori todennäköisyydet ovat

$$\pi(\mathcal{M}_1) = \pi(\mathcal{M}_m) = 1/2$$

4. Tähän kysymykseen ei ole pakko vastata eikä vastauksesi vaikuta kurssin suorittamiseen:

   Kerro lyhyesti rehellinen mielipiteesi Bayes-teoriasta,

- Onko sinusta tullut valistunut Bayeslainen tilastotietelija, joka hyväksyy että kaikki sinulle tuntemattomat suureet ovat samalla tavalla satunnaisia ja todennäköisyysjakauma kuvaa sinun informatiotilaa tuntemattomasta.
- Vai sitten ilmiöiden todennäköisyys määräytyy yksikäsitteisesti luonnon lakien seurauksena, ja Bayes teoria on vain käyttökelpoinen menetelmä, joka teknisestä syystä vaati priori jakauman asettaminen myös tuntemattomille parametreille, jotka kuitenkin eivät ole "oikeasti satunnaisia".

(Kyllä/Ei vastaus myös riittää).