Bayes-päättely, 6. harjoitukset (1.–2.3.2017)

1. Lineaarinen regressio. Tarkastellaan aineistoa[1] `diabetes.csv`, joka sisältää 442 diabetespotilaan iän, sukupuolen, painoindeksin, verenpaineen ja mittaukset kuudesta eri veren (ehkä kolesteroli?)-arvosta (`S1-S6`) ja selitettävän muuttujan `y`, joka kuvaa taudin etenemistä vuosi mittausten tekemisen jälkeen. Tarkoituksena on siis selittää, mitkä riskitekijät vaikuttavat diabeteksen pahenemiseen. Käytä tämän ja seuraavien tehtävien analyyseihin tiedostoa `diabetes_sd.csv`, joka sisältää saman aineiston, mutta normalisoituna siten, että jokaisen muuttujan keskiarvo on 0, ja kaikkien selittävien muuttujien euklidinen normi on 1.

(a) Suorita tavallinen lineaarinen regressio `R`:n `lm`-funktiolla, siten että selität muuttujaa `y` kaikilla aineiston muilla muuttujilla. Mitkä selittäjistä ovat tilastollisesti merkitseviä merkitsevyystasolla 0.05?

(b) Suorita vastaava bayesiläinen lineaarinen regressio stanilla. Malli on tarkemmin

$$Y_i \,|\, \alpha, \boldsymbol{\beta}, \sigma \sim N(\alpha + \boldsymbol{\beta}\mathbf{x}_i, \sigma^2) \quad \text{kaikille } i = 1, \ldots, N,$$

tai matriisimuodossa

$$\mathbf{Y} \,|\, \alpha, \boldsymbol{\beta}, \sigma \sim N_n(\alpha \mathbf{1}_n + \boldsymbol{\beta}\mathbf{X}, \sigma^2 \mathbf{I}),$$

missä

$$\mathbf{X} = \begin{bmatrix} x_{11} & \ldots & x_{1K} \\ \vdots & & \vdots \\ x_{N1} & \ldots & x_{NK} \end{bmatrix}$$

on matriisi, joka sisältää $K = 10$ selittävän muuttujan arvot $N = 442$ havainnolle,

$$\boldsymbol{\beta} = \begin{bmatrix} \beta_1 \\ \vdots \\ \beta_K \end{bmatrix}$$

on vektori joka sisältää regressiokerrointen arvot, α on regression vakiotermi, ja σ virhetermien keskihajonta. Voit käyttää epäoleellista prioria

$$p(\alpha, \boldsymbol{\beta}, \sigma) \propto 1, \quad \sigma > 0,$$

jolloin prioria ei tarvitse määrittää stanille erikseen (mutta muista rajoittaa σ positiiviseksi!).

Jos määrität selittävät muuttujat matriisina: `matrix[N,K] x`; ja selitettävän muuttujan (ja regressiokertoimet) vektorina: `vector[N] y`; voit määrittää mallin kätevästi matriisimuodossa: `y ~ normal(x * beta + alpha, sigma);`

2. Jatkoa edelliseen tehtävään.

(a) Piirrä kuva regressiokerrointen $\beta_1, \ldots, \beta_K$ mediaaneista ja 50% ja 95% uskottavuusväleistä (Tämän pitäisi onnistua automaattisesti `plot(stan_fit)`-komennolla. Piirrettäviä parametreja voi säätää `par`-argumentilla, ja ulompaa Uskottavuusväliä `outer_level`-argumentilla), missä `stan_fit` on `stan`-funktion palauttama olio. Uskottavuusvälit (credible interval), eli "Bayesiläiset luottamusvälit" ovat Bayes-päättelyn vastine frekventistisille luottamusväleille. Tulkinnallisesti ne ovat huomattavasti miellyttävämpiä, sillä nyt voit sanoa, että parametrin todellinen arvo on 95% todennäköisyydellä sen 95% uskottavuusvälillä!

[1] Tätä on käytetty esimerkkiaineistona artikkeleissa Least Angle Regression [1] ja Bayesian Lasso [2]

(b) Piirrä vastaava kuva, johon merkitset (a)-kohdassa estimoimoimasi regressiokerrointen $\beta_1, \ldots, \beta_K$ suurimman uskottavuuden estimaatit $\hat{\beta}_1, \ldots, \hat{\beta}_K$, ja regressiokerrointen luottamustasojen $\alpha = 0.5$ ja $\alpha = 0.05$ (frekventistiset) luottamusvälit. Luottamustason α frekventistinen luottamusväli regressiokertoimelle β_k ovat muotoa

$$\left(\hat{\beta}_k - t_{N-K-1}(\alpha/2) \cdot \mathrm{se}(\hat{\beta}_k),\ \hat{\beta}_k + t_{N-K-1}(\alpha/2) \cdot \mathrm{se}(\hat{\beta}_k)\right),$$

missä $t_\nu(u)$ on t:n jakauman vapausasteella ν yläkvantiili, eli piste, jonka oikealle puolelle jää osuus u jakauman todenäköisyysmassasta,[2] ja $\mathrm{se}(\hat{\beta}_j)$ on estimaattorin β_j keskivirhe[3]. Lisäksi voit hyödyntää `segments`-funktiota.

Vertaa kuvia. Millä jakaumatuloksilla selittäisit havaintojasi?

3. Jatkoa edellisille tehtäville. Tehdään samalle aineistolle Lasso, eli L1-regularisoitu lineaarinen regressio. Tämä on toteutettu valmiiksi `R`:n `glmnet`-kirjastossa.

(a) Lue "Quick start"-osio osoitteesta `https://web.stanford.edu/~hastie/glmnet/glmnet_alpha.html`, ja tee Lasso diabetes-aineistolle siten, että selität jälleen `y`:tä kaikilla muilla muuttujilla. (Vihje: Tämän pitäisi onnistua suoraan komennolla `glmnet(x, y)`, missä `y` on selitettävä muuttuja, ja `x` matriisi, joka sisältää selittäjät.) Piirrä lasso-polut (tämän pitäisi onnistua suoraan komennolla `plot(lasso_fit)`, missä `lasso_fit` on `glmnet`:in palauttama olio), eli regressiokerrointen arvot niiden yhteenlaskettujen pituuksien funktiona. Katso saatko samanlaiset lasso-polut, kuin artikkelin [1] kuvan 1 vasemmanpuolimmaisessa paneelissa!

(b) Etsi 'rankaisu-parametrin' λ optimaalinen arvo ristiinvalidoinnin avulla (Vihje: voit käyttää valmista funktiota `cv.glmnet`. Tarkemmat ohjeet löytyvät yllä linkitetystä tutoriaalista). Mitkä ovat regressiokerrointen arvot tällä λ:n arvolla?

4. Jatkoa edellisille tehtäville. Tehtävän 2 Bayesiläinen lineaarinen regressio, mutta tällä kertaa epäinformatiivisen priorin sijasta oletetaankin, että osa regressiokertoimista on lähellä nollaa, eli asetetaan regressiokertoimille informatiivinen Laplace-priori, joka kutistaa regressiokertoimia kohti nollaa Lasso-tyylisesti. Tehdään täysi Bayesiläinen malli antamalla myös parametrin λ neliölle Gamma-priori. Malli on siis muuten sama kuin tehtävässä 2, mutta

$$\begin{aligned} \beta_k \,|\, \sigma, \lambda &\sim \mathrm{Laplace}(0, \sigma/\lambda) \quad \text{kaikille } k = 1, \ldots, K, \\ \lambda^2 &\sim \mathrm{Gamma}(1, 1.78). \end{aligned}$$

Keskihajonnalle voit olettaa edelleen epäinformatiivisen priorin $p(\sigma) \propto 1, \sigma > 0$.

(a) Sovita malli stanilla (Vihje Laplace-jakauman saat komennolla `double_exponential(0, sigma / lambda)`). Piirrä histogrammi parametrin λ posteriorijakaumasta. Vertaa tulosta tehtävän 3 ristiinvalidoinnilla saatuun parametrin λ arvoon.

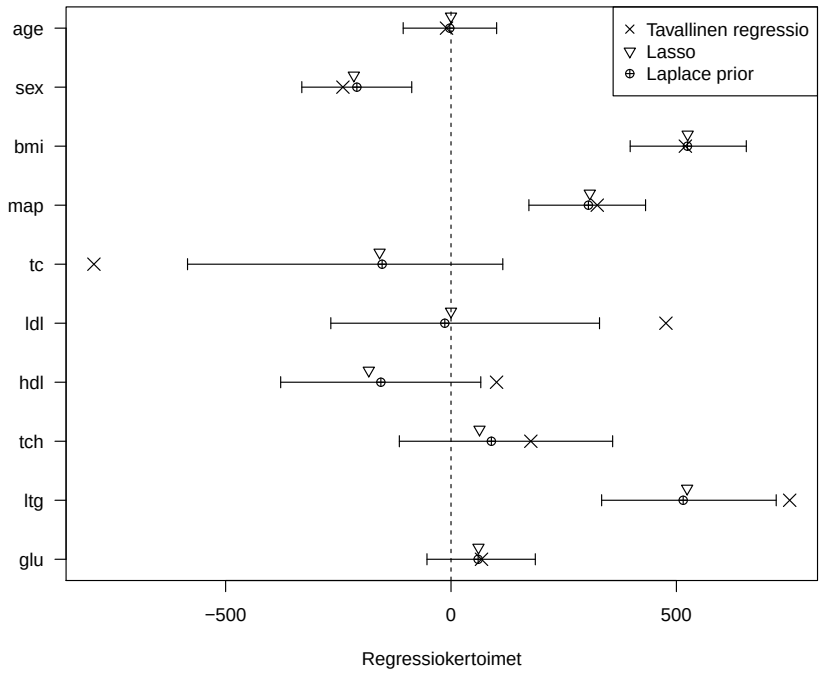
(b) Koetetaan replikoida artikkelin [1] kuvaa 2! Piirrä parametrien $\beta_1, \ldots, \beta_{10}$ posteriorijakaumien mediaanit ja 95% luottamusvälit. Lisää kuvaan tehtävässä 1 laskemasi suurimman uskottavuuden estimaatit $\hat{\boldsymbol{\beta}}_{\mathrm{MLE}}$, ja tehtävässä 3 laskemasi lasso-estimaatit $\hat{\boldsymbol{\beta}}_{\mathrm{Lasso}}$ ristiinvalidoimalla valitsemallasi parametrin arvolla λ. Lopputulos voisi muistuttaa kuvaa 1.

(c) Vertaa tehtävän 1 tuloksiin. Mille parametreille 95%:n uskottavuusväli ei sisällä nollaa?

[2] saadaan `R`-ssä funktiolla `qt(u, nu, lower = FALSE)`.

[3] Löytyvät esimerkiksi komennon `summary(lm_fit)$coefficients`, missä `lm_fit` on `lm`-funktion palauttama olio, palauttamasta matriisista.

Kuva 1: Regressiokerrointen $\boldsymbol{\beta}$ 95% mediaanit ja uskottavuusvälit. Lisäksi myös suurimman uskottavuuden estimaatit $\hat{\boldsymbol{\beta}}_{\mathrm{MLE}}$ ja Lasso-estimaatit $\hat{\boldsymbol{\beta}}_{\mathrm{Lasso}}$.



Viitteet

[1] Bradley Efron, Trevor Hastie, Iain Johnstone, Robert Tibshirani, et al. Least angle regression. *The Annals of statistics*, 32(2):407–499, 2004.

[2] Trevor Park and George Casella. The bayesian lasso. *Journal of the American Statistical Association*, 103(482):681–686, 2008.