Lab 2 Bayesian Learning

Hugo Morvan

2024-04-16

Linear and polynomial regression

The dataset Linkoping2022.xlsx contains daily average temperatures (in degree Celcius) in Linköping over the course of the year 2022. Use the function read_xlsx(), which is included in the R package readxl (install.packages("readxl")), to import the dataset in R. The response variable is temp and the covariate time that you need to create yourself is defined by:

$$time = \frac{the\ number\ of\ days\ since\ the\ beginning\ of\ the\ year}{365}$$

A Bayesian analysis of the following quadratic regression model is to be performed:

$$temp = \beta_0 + \beta_1 * time + \beta_2 * time^2 + \epsilon, \epsilon \overset{\text{iid}}{\sim} N(0, \sigma^2).$$

a)

Use the conjugate prior for the linear regression model. The prior hyperparameters μ_0, Ω_0, ν_0 and σ^2 shall be set to sensible values. Start with $\mu_0 = (0, 100, -100)^T$, $\Omega_0 = 0.01 * I_3$, $\nu_0 = 1$ and $\sigma^2 = 1$. Check if this prior agrees with your prior opinions by simulating draws from the joint prior of all parameters and for every draw compute the regression curve. This gives a collection of regression curves; one for each draw from the prior. Does the collection of curves look reasonable? If not, change the prior hyperparameters until the collection of prior regression curves agrees with your prior beliefs about the regression curve. [Hint: R package mvtnorm can be used and your $Inv - \chi^2$ simulator of random draws from Lab 1.]

b)

Write a function that simulate draws from the joint posterior distribution of β_0, β_1,β_2 and σ^2.

i) Plot a histogram for each marginal posterior of the parameters.

ii) Make a scatter plot of the temperature data and overlay a curve for the posterior median of the regression function $f(time) = E[temp|time] = \beta_0 + \beta_1 * time + \beta_2 * time^2$, i.e. the median of f (time) is computed for every value of time. In addition, overlay curves for the 90% equal tail posterior probability intervals of f (time), i.e. the 5 and 95 posterior percentiles of f (time) is computed for every value of time. Does the posterior probability intervals contain most of the data points? Should they?

c)

It is of interest to locate the time with the highest expected temperature (i.e. the time where f (time) is maximal). Let's call this value $\tilde{x}$. Use the simulated draws in (b) to simulate from the posterior distribution of $\tilde{x}$. [Hint: the regression curve is a quadratic polynomial. Given each posterior draw of β_0, β_1 and β_2 , you can find a simple formula for $\tilde{x}$.]

d)

Say now that you want to estimate a polynomial regression of order 10, but you suspect that higher order terms may not be needed, and you worry about overfitting the data. Suggest a suitable prior that mitigates this potential problem. You do not need to compute the posterior. Just write down your prior. [Hint: the task is to specify μ_0 and Ω_0 in a suitable way.]

Posterior approximation for classification with logistic regression

The dataset WomenAtWork.dat contains n = 132 observations on the following eight variables related to women (see table in lab compendium).

a)

Consider the logistic regression model:

$$Pr(y = 1|x, \beta) = \frac{exp(x^T\beta)}{1 + \exp(x^T\beta)},$$

where y equals 1 if the woman works and 0 if she does not. x is a 7-dimensional vector containing the seven features (including a 1 to model the intercept). The goal is to approximate the posterior distribution of the parameter vector β with a multivariate normal distribution

$$\beta|y, x \sim N(\tilde{\beta}, J_y^{-1}(\tilde{\beta})),$$

where $\tilde{\beta}$ is the posterior mode and $J_y^{-1} = -\frac{complete}{later}$ is the negative of the oberved Hessian evaluated at the posterior mode. Nota that $\frac{complete}{later}$ is a 7*7 matrix with second derivatives on the diagonal and cross derivatives $inserteq$ on the off-diagonal. You can compute this derivative by hand, but we will let the computer do it numerically for you. Calculate both $\tilde{\beta}$ and $J(\tilde{\beta})$ by using the optim function in R. [Hint: You may use code snippets from my demo of logistic regression in Lecture 6.] Use the prior $beta \sim N(0, \tau^2 I)$, where $\tau = 2$.

Present the numerical values of $\tilde{\beta}$ and $J_y^{-1}(\tilde{\beta})$ for the WomenAtWork data. Compute an approximate 95% equal tail posterior probability interval for the regression coefficient to the variable NSmallChild. Would you say that this feature is of importance for the probability that a woman works? [Hint: You can verify that your estimation results are reasonable by comparing the posterior means to the maximum likelihood estimates, given by: r glmModel<- glm(Work ~ 0 + ., data = WomenAtWork, family = binomial).]

b)

Use your normal approximation to the posterior from (a). Write a function that simulate draws from the posterior predictive distribution of $Pr(y = 0|x)$, where the values of x corresponds to a 40-year-old woman, with two children (4 and 7 years old), 11 years of education, 7 years of experience, and a husband with an income of 18. Plot the posterior predictive distribution of $Pr(y = 0|x)$ for this woman. [Hints: The R package mvtnorm will be useful. Remember that $Pr(y = 0|x)$ can be calculated for each posterior draw of β .]

c)

Now, consider 13 women which all have the same features as the woman in (b). Rewrite your function and plot the posterior predictive distribution for the number of women, out of these 13, that are not working. [Hint: Simulate from the binomial distribution, which is the distribution for a sum of Bernoulli random variables.]