Computer Lab 2
732A54 - Bayesian Learning

Jooyoung Lee - joole336
Zuxiang Li - zuxli371

April 14, 2020

1. Linear and polynomial regression The dataset TempLinkoping.txt contains daily temperatures (in Celcius degrees) at Malmslätt, Linköping over the course of the year 2016 (366 days since 2016 was a leap year). The response variable is temp and the covariate is

$$time = \frac{the\ number\ of\ days\ since\ beginning\ of\ year}{366}$$

The task is to perform a Bayesian analysis of a quadratic regression

$$temp = \beta_0 + \beta_1 \cdot time + \beta_2 \cdot time^2 + \epsilon, \epsilon\ N(0, \sigma^2)$$

(a) Determining the prior distribution of the model parameters. Use the conjugate prior for the linear regression model. Your task is to set the prior hyperparameters $\mu_0, \Omega_0, v_0\ and\ \sigma^2$ to sensible values. Start with $\mu_0 = (-10, 100, -100)^T$, $\Omega_0 = 0.01 \cdot I_3, v_0 = 4\ and\ \sigma^2 = 1$. 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. Do 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: the R package mvtnorm will be handy. And use your $Inv - \chi^2$ simulator from Lab 1.]

(b) Write a program that simulates from the joint posterior distribution of $\beta_0, \beta_1, \beta_2 and \sigma^2$ Plot the marginal posteriors for each parameter as a histogram. Also produce another figure with a scatter plot of the temperature data and overlay a curve for the posterior median of the regression function $f(time) = \beta_0 + \beta_1 \cdot time + \beta_2 \cdot time^2$, computed for every value of time. Also overlay curves for the lower 2.5% and upper 97.5% posterior credible interval for $f(time)$. That is, compute the 95% equal tail posterior probability intervals for every value of time and then connect the lower and upper limits of the interval by curves. Does the interval bands contain most of the data points? Should they?

(c) It is of interest to locate the time with the highest expected temperature (that is, the time where $f(time)$ is maximal). Let's call this value $\tilde{x}$ Use the simulations in b) to simulate from the posterior distribution of $\tilde{x}$ [Hint: the regression curve is a quadratic. You can find a simple formula for ~x given $\beta_0, \beta_1\ and\ \beta_2$

(d) Say now that you want to estimate a polynomial model of order 7 but you suspect that higher order terms may not be needed, and you worry about overfitting. 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 $\mu_0\ and\ \Omega_0$ in a smart way.]

2. Posterior approximation for classification with logistic regression

The dataset WomenWork.dat contains n = 200 observations (i.e. women) on the following nine variables:

(a) Consider the logistic regression

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

where y is the binary variable with y = 1 if the woman works and y = 0 if she does not. x is a 8-dimensional vector containing the eight features (including a one for the constant term that models the intercept). Fit the logistic regression using maximum likelihood estimation by the command:glmModel

<-glm(Work ~ 0 + ., data = WomenWork, family = binomial). Note how I added a zero in the model formula so that R doesn't add an extra intercept (we already have an intercept term from the Constant feature). Note also that a dot (.) in the model formula means to add all other variables in the dataset as features. family = binomial tells R that we want to fit a logistic regression.

(b) Now the fun begins. Our goal is to approximate the posterior distribution of the 8-dim parameter vector β with a multivariate normal distribution

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

where $\tilde{\beta}$ is the posterior mode and $J(\tilde{\beta}) = -\frac{\partial^2 \ln p(\beta|y)}{\partial\beta\partial\beta^T}|_{\beta=\tilde{\beta}}$ is the observed Hessian evaluated at the posterior mode. Note that $\frac{\partial^2 \ln p(\beta|y)}{\partial\beta\partial\beta^T}$ is an 8x8 matrix with second derivatives on the diagonal and cross-derivatives $\frac{\partial^2 \ln p(\beta|y)}{\partial\beta\partial\beta^T}$ on the offdiagonal. It is actually not hard to compute this derivative by hand, but don't worry, we will let the computer do it numerically for you. Now, both $\tilde{\beta}$ *and* $J(\tilde{\beta})$ are computed by the optim function in R. See my code https://github.com/mattiasvillani/BayesLearnCourse/raw/master/Code/MainOptimizeSpam. zip here I have coded everything up for the spam prediction example (it also does probit regression, but that is not needed here). I want you to implement you own version of this. You can use my code as a template, but I want you to write your own file so that you understand every line of your code. Don't just copy my code. Use the prior $\beta \ N(0, \tau^2 I), with \ \tau = 10$. Your report should include your code as well as numerical values for $\tilde{\beta}$ *and* $J_y^{-1}(\tilde{\beta})$ for the WomenWork data. Compute an approximate 95% credible interval for the variable NSmallChild. Would you say that this feature is an important determinant of the probability that a women works?

(c) Write a function that simulates from the predictive distribution of the response variable in a logistic regression. Use your normal approximation from 2(b). Use that function to simulate and plot the predictive distribution for the Work variable for a 40 year old woman, with two children (3 and 9 years old), 8 years of education, 10 years of experience. and a husband with an income of 10. [Hint: the R package mvtnorm will again be handy. And remember my discussion on how Bayesian prediction can be done by simulation.]

Appendix A : Code Question 1

Appendix B : Code Question 2