Bertil Wegmann
Div of Statistics and Machine Learning
Dept of Computer and Information Science
Linköping University

Bayesian Learning

Computer Lab 2

You are recommended to use R for solving the labs.
You work and submit your labs in pairs, but both of you should contribute equally and understand all parts of your solutions.
`It is not allowed to share exact solutions` with other student pairs.
The submitted lab reports will be verified through OURIGINAL and indications of plagiarism will be investigated by the Disciplinary Board.
Submit your solutions via LISAM, no later than May 5 at 23:59.

Please note the following about the format of the submitted lab report:
1. The lab report should include all solutions and plots to the stated problems with necessary comments.
2. Submit the lab report with your code attached to the solution of each sub-problem (1a), 1b),...) in **one** PDF document.
3. Submit a separate file containing all code.

---

1. **Linear and polynomial regression**
   The dataset `temp_linkoping.csv` contains daily average temperatures (in degree Celcius) in Linköping between July 2023 and June 2024. Import the dataset in R. The response variable is *temp* and the covariate *time*, which is defined as

$$time = \frac{\text{the number of days since the beginning of the observation period}}{365}.$$

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

$$temp = \beta_0 + \beta_1 \cdot time + \beta_2 \cdot time^2 + \varepsilon, \; \varepsilon \overset{iid}{\sim} N(0, \sigma^2).$$

   (a) Use the conjugate prior for the linear regression model. The prior hyperparameters μ_0, Ω_0, ν_0 and σ_0^2 shall be set to sensible values. Start with $\mu_0 = (0, -100, 100)^T$, $\Omega_0 = 0.01 \cdot I_3$, $\nu_0 = 1$ and $\sigma_0^2 = 1$. 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 $Scaled - Inv{-}\chi^2$ 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 $f(time) = \mathrm{E}[temp|time] = \beta_0 + \beta_1 \cdot time + \beta_2 \cdot time^2$, 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 lowest expected temperature (i.e. the *time* where $f(time)$ is minimal). Let's call this value $\tilde{x}$. Use the simulated draws in (b) to simulate from the **posterior distribution of** $\tilde{x}$. You can solve this analytical **or** numerical. [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 and motivate your choice. Repeat task (a) for the selected prior and the higher order polynomial to see if it behaves as expected. [Hint: the task is to specify μ_0 and Ω_0 in a suitable way.]

2. **Posterior approximation for classification with logistic regression**

The dataset `Disease.csv` contains $n = 313$ observations on the following six variables related to a certain disease:

Variable	Data type	Meaning	Role
Class_of_diagnosis	Binary	Disease or not?	Response y
Gender	Binary	Woman (1) or Male (0)	Feature
Age	Counts	Age	Feature
Duration_of_symptoms	Numeric	Duration of symptoms in days	Feature
Dyspnoea	Binary	1 if person has laboured breathing	Feature
White_blood	Counts	N white blood cells per microliter	Feature

(a) Consider the logistic regression model:

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

where y equals 1 if the person has a disease and 0 if not. $\mathbf{x}$ is a 6-dimensional vector containing five features and a column of 1's to include an intercept in the model. The values of the variables Age, Duration_of_symptoms, and White_blood in $\mathbf{x}$ need to be standardized to mean 0 and variance 1. For each of these variables x_j, calculate the standardized value for each observation i by using the formula

$$\frac{x_{ij} - \bar{x_j}}{s_{x_j}},$$

where $\bar{x_j}$ and s_{x_j} are the sample mean and standard deviation of x_j, respectively. The goal is to approximate the posterior distribution of the parameter vector β with a multivariate normal distribution

$$\beta|\mathbf{y}, \mathbf{x} \sim N\left(\tilde{\beta}, J_{\mathbf{y}}^{-1}(\tilde{\beta})\right),$$

where $\tilde{\beta}$ is the posterior mode and $J(\tilde{\beta}) = -\frac{\partial^2 \ln p(\beta|\mathbf{y})}{\partial\beta\partial\beta^T}|_{\beta=\tilde{\beta}}$ is the negative of the observed Hessian evaluated at the posterior mode. Note that $\frac{\partial^2 \ln p(\beta|\mathbf{y})}{\partial\beta\partial\beta^T}$ is a 6×6 matrix with second derivatives on the diagonal and cross-derivatives $\frac{\partial^2 \ln p(\beta|\mathbf{y})}{\partial\beta_i\partial\beta_j}$ 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 $\beta \sim \mathcal{N}(0, \tau^2 I)$, where $\tau = 2$.
Present the numerical values of $\tilde{\beta}$ and $J_{\mathbf{y}}^{-1}(\tilde{\beta})$ for the `Disease` data. Compute an approximate 95% equal tail posterior probability interval for the regression coefficient to the variable Age. Would you say that this feature is of importance for the probability that a person has the disease?
[Hint: You can verify that your estimation results are reasonable by comparing the posterior means to the maximum likelihood estimates, given by: `glmModel <- glm(Class_of_diagnosis ~ 0 + ., data = Disease, family = binomial)`.]

(b) Use your normal approximation to the posterior from (a). Write a function that simulate draws from the posterior predictive distribution of $\Pr(y = 1|\mathbf{x})$, where the certain diagnosis method was used and where the values of $\mathbf{x}$ corresponds to a 38-year-old woman, with 10 days of symptoms, no laboured breathing and 11000 white blood cells per microliter. Note that the corresponding standardized values need to be calculated for the variables Age, Duration_of_symptoms, and White_blood in $\mathbf{x}$ by using the formula in (a). Plot the posterior predictive distribution of $\Pr(y = 1|\mathbf{x})$ for this person.
[Hints: The R package `mvtnorm` will be useful. Remember that $\Pr(y = 1|\mathbf{x})$ can be calculated for each posterior draw of β.]

GOOD LUCK!
BEST, BERTIL