THE UNIVERSITY OF NEW SOUTH WALES

MATH 3871/5960 Bayesian Inference and Computation
Assignment

Assignment

This assignment covers material until (and including) Lecture 6 and Lab/Tute 6. It is worth 20% of the final course grade. Total marks available = 45.

Please refer to the following instructions. Failure to do so will result in loss of marks.

- Assignment to be submitted via moodle by 15 November 11:59 PM Sydney time.
- You must use the template .R script to produce code for this assignment. **Your assignment will not be marked unless you use the template .R script**. All necessary files for the assignment can be found in the Assessments Hub section.
- Submit your assignment via the submission portal in Moodle. You will need to submit two files, as detailed below. Use the appropriate tabs in moodle to submit each of the files.
  1. A report as a .pdf or .docx file containing plots and any text-based responses to questions and
  2. The template .R script filled in with your code. Ensure that you have specified your zID in the filename of the .R file.
- Print, sign and attach the plagiarism statement to your report.
- Refer to course outline for grading of late submissions

Plagiarism Statement

I declare that this assessment item is my own work, except where acknowledged, and has not been submitted for academic credit elsewhere. I acknowledge that the assessor of this item may, for the purpose of assessing this item reproduce this assessment item and provide a copy to another member of UNSW; and/or communicate a copy of this assessment item to a plagiarism checking service (which may then retain a copy of the assessment item on its database for the purpose of future plagiarism checking).

I certify that I have read and understood UNSW Rules in respect of Student Academic Misconduct.

Name (print clearly):

Student Number:

Signature:

Date:

Bayesian analysis: Vinho Verde

In lecture 5, you saw an example of Bayesian linear regression. In this assessment, you will be performing Bayesian *logistic* regression on a real dataset. This dataset is related to red wine variants of the Portuguese "Vinho Verde" wine, described in the publication Cortez, P., Cerdeira, A., Almeida, F., Matos, T., & Reis, J. (2009). Modeling wine preferences by data mining from physicochemical properties

The dataset you will be working with is available for download on Moodle in the Assessments Hub section. The input variables (or covariates) of interest for this assignment are as follows:

1. `fixed acidity`
2. `volatile acidity`
3. `citric acid`
4. `residual sugar`
5. `chlorides`
6. `free sulfur dioxide`
7. `total sulfur dioxide`
8. `density`
9. `pH`
10. `sulphates`
11. `alcohol`

The output or response variable (based on sensory data) is `quality`, measured as a score between 0 and 10.

Background on logistic regression

In linear regression, the assumed model relating the expectation of the response variables and covariates is given by

$$\mathbb{E}[Y_i|\mathbf{x}_i] = \mathbf{x}_i\boldsymbol{\beta}, \quad i = 1, 2, \ldots n$$

where

- i denotes the index of the (response, covariate) pair, i.e. $(y_i, \mathbf{x}_i)$ denotes the ith pair of values in the dataset and n denotes the total number of observations.
- $\mathbf{x}_i$ is the i-th row of the design matrix $\mathbf{X}$ of covariates, therefore its dimension is $1 \times (k+1)$ where k is the number of different covariates of interest. The matrix $\mathbf{X}$ is of size $n \times (k+1)$ where n is the number of observations.
- $\boldsymbol{\beta}$ is a column vector of parameters of dimension $(k+1) \times 1$

As the name suggests, *generalised linear models* (GLMs) are a generalisation of linear regression models for when the response variable Y_i conditioned on $\mathbf{x}_i$ is distributed according to a member from the exponential family. In generalised linear models, the covariates are related to the expected value of the response variable via an invertible transformation $g(\cdot)$, i.e.

$$\mathbb{E}[Y_i|\mathbf{x}_i] = g^{-1}(\mathbf{x}_i\boldsymbol{\beta}), \quad i = 1, 2, \ldots n \tag{1}$$

$$\eta_i := g(\mathbb{E}[Y_i]) = \mathbf{x}_i\boldsymbol{\beta} \tag{2}$$

where

- η_i is the linear predictor, i.e. the transformed version of the expected response variable that is linearly related to the covariates.
- $g(\cdot)$ is called the link function.

Notice that, in the case of linear regression (as seen in Lecture 5), $g(\cdot)$ is the identity function and the response variable Y_i conditioned on $\mathbf{x}_i$ is normally distributed. In this assignment, we will always assume that $Y_i,\ i = 1, 2, \ldots, n$ are independent and identically distributed.

One of the most popular GLMs is the logistic regression model. In logistic regression, the response variable is binary:

$$Y_i \in \{0, 1\}, \quad i = 1, 2, \ldots n$$

The values 0 and 1 are usually interpreted as labels for two possible classes (e.g. 'success' and 'failure'). In this case, logistic regression can then be used for binary classification and the response variables are Bernoulli distributed:

$$Y_i = \begin{cases} 1 & \Pr(y_i = 1) = p_i \\ 0 & \Pr(y_i = 0) = (1 - p_i) \end{cases}$$

such that $\mathbb{E}[Y_i] = p_i \in [0, 1]$. The link function g in logistic regression is the logit function, given by

$$g(p_i) = \log\left(\frac{p_i}{1 - p_i}\right), \quad p_i \in (0, 1)$$

taking exponential of both sides and re-arranging using (2) yields

$$p_i = \frac{1}{1+\exp(-\mathbf{x}_i\boldsymbol{\beta})} = \frac{\exp(\mathbf{x}_i\boldsymbol{\beta})}{1+\exp(\mathbf{x}_i\boldsymbol{\beta})}$$

Frequentist analysis: For maximum likelihood estimation of the coefficients $\boldsymbol{\beta}$, it is not possible to analytically optimise the model to obtain the MLE. It is necessary to use an iterative algorithm (e.g. iterative weighted least squares) to approximate the solution of the optimisation problem, as is done in the `glm` function in `R`, see `https://stat.ethz.ch/R-manual/R-devel/library/stats/html/glm.html`.

It holds that for large n, the MLE, $\hat{\boldsymbol{\beta}}$ is approximately normally distributed,

$$\hat{\boldsymbol{\beta}} \sim \mathcal{N}(\boldsymbol{\beta}^*, (X^TWX)^{-1})$$

where $\boldsymbol{\beta}^*$ denotes the true parameter and W is a matrix depending on β arising from the second derivative of the likelihood function (the form of this matrix is not important for this assignment.)

Bayesian analysis: As is standard in Bayesian inference, one needs to choose a prior distribution for $\boldsymbol{\beta}$ and obtain an expression for the posterior

$$\pi(\boldsymbol{\beta}|\mathbf{y},\mathbf{x}) \propto L(\mathbf{y}|\boldsymbol{\beta},\mathbf{x})\pi(\boldsymbol{\beta})$$

In Part II of this assignment, you will obtain a form for the likelihood function $L(\mathbf{y}|\boldsymbol{\beta},\mathbf{x})$.

Questions

In this assignment, you will perform Frequentist and Bayesian logistic regression. In Bayesian logistic regression, the goal is to draw samples from $\pi(\boldsymbol{\beta}|\mathbf{y}, \mathbf{x})$, the posterior distribution on the regression coefficients using prior knowledge.

PART 1: Frequentist analysis

1. **[2 marks]** Read the data into `R` into a dataframe called `wine`. Check if there are missing values (NA) and, in case there are, remove them.

2. **[2 marks]** Since we are implementing logistic regression, we require a response variable which takes value of either 0 or 1. Suppose we consider a 'good' wine one with quality above 6.5 (included). Create a variable named `wine$good` with the appropriate 0 or 1 label for the response variables.

3. **[4 points]** Run a frequentist analysis on the logistic model to obtain the maximum likelihood estimates of the coefficient $\boldsymbol{\beta}$ using the `glm()` function in `R`, specifying the `family` as binomial. Output the MLE into a numeric variable called `mleest`.

PART II: Bayesian analysis

1. **[4 marks]** Show that the log-likelihood function for the logistic regression model with n iid data pairs is given by

$$\log L(\mathbf{y}|\boldsymbol{\beta}, \mathbf{x}) \;= \sum_{i=1}^{n} y_i \mathbf{x}_i \boldsymbol{\beta} - \log[1 + \exp(\mathbf{x}_i \boldsymbol{\beta})]. \tag{3}$$

2. **[4 marks]** Consider the following prior distribution on the coefficients $\boldsymbol{\beta}$,

$$\pi(\boldsymbol{\beta}) = \mathcal{N}_{k+1}(\boldsymbol{\alpha}, \boldsymbol{\Omega}).$$

   with $\boldsymbol{\alpha} = \mathbf{0}$ (i.e. a $k+1$ dimensional vector of zeros) and $\boldsymbol{\Omega} = 100 I_{k+1}$ where I_{k+1} is the $k+1 \times k+1$ identity matrix. This is a weakly informative prior. Write a function named `lpost.LR` that takes as input the following three variables:

   - `beta`, the value of the regression coefficients
   - `X`, the value of the covariates
   - `y`, the response variable

   and outputs the value of the log posterior for the given prior.

3. **[10 marks]** Write a function named `mhmcmc` that implements Metropolis-Hastings MCMC to sample from the posterior distribution of the regression coefficients. Within the function, set the proposal distribution to be the random walk proposal, i.e.

$$q(\boldsymbol{\beta}|\boldsymbol{\beta}_t) = \mathcal{N}_{k+1}(\boldsymbol{\beta}_t, \Sigma)$$

where Σ is the covariance matrix to be specified as a variable input to the `mhmcmc` function. You must use `mvrnorm` for random sampling from a multivariate normal. Ensure the function produces the following outputs (with the same variable name)

- `beta_star`, the proposed value at each iteration
- `beta_mat`, the value of the chain at each iteration
- `accprob`, the acceptance probability at each iteration
- `acc`, the total number of accepted samples.

4. **[5 marks]** Run the `mhmcmc` algorithm with 10^5 iterations. Initialise the algorithm at the MLE obtained from the `glm` function and use the Σ provided in the `R` template for the proposal covariance. Provide the following in your report:

   (a) Trace plots of all the coefficients.

   (b) The proportion of accepted moves as a variable named `mh4acc`

   (c) **In no more than 100 words** comment on the suitability of both the proposal distribution and the initialisation of the MCMC chain.

5. **[6 marks]** One again, run the `mhmcmc` algorithm with 10^5 iterations and the same proposal distribution as in Question 4, but this time with 4 different sets of initialisations for each of the coefficients. Plot the chains for all coefficients (overlay the 4 chains on the same plot) and comment on convergence.

6. **[8 marks]** Tune the metropolis hastings proposal distribution to improve the acceptance rate so that it is in the range of 10% to 30%. You may consider only 1 chain. *Hint: Consider setting the covariance proportional to the covariance of the MLE.* In your report, provide:

   (a) Trace plots of all coefficients

   (b) Plots of the approximate posterior marginals (you may plot as histograms) and indicate the MLE calculated in Part I as a vertical line. *Hint: consider what practices you should adopt when extracting iid samples from an MCMC chain.*

   (c) **In no more than 100 words** explain in what ways the results are superior and what features of the proposal make it better suited to this problem than the proposal distribution in Questions 4 & 5.