MAST90104: A First Course in Statistical Learning

Assignment 4, 2022

Due: 11:59 pm Sunday 23 October. Please submit a scanned or other electronic copy of your work via the Learning Management System

This assignment is worth 5% of your total mark. The total point is 30. Include your R commands and output in your answer (or add them as an appendix). Late submissions will have their mark deducted.

1. (10 pts) The data *winequality-white.csv* includes data from the paper *Modeling wine preferences by data mining from physicochemical properties* by P. Cortez, A. Cerdeira, F. Almeida, T. Matos and J. Reis (2009). The data consists of 4898 observations of white variants of the Portuguese "Vinho Verd" wine. The variables are:

   - fixed acidity
   - volatile acidity
   - citric acid
   - residual sugar
   - chlorides
   - free sulfur dioxide
   - total sulfur dioxide
   - density
   - pH
   - sulphates
   - alcohol : percent alcohol content of the wine
   - quality : Output variable (score between 0 and 10)

   The quality score in this data set ranges from 3 to 9, but we will recode the levels as *bad*, *average* and *good*:

   ```
   wine$quality = factor(wine$quality)
   levels(wine$quality ) <- c("bad","bad","average","average","good" , "good","good")
   ```

   (a) Fit a multinomial model to predict the wine quality by category, considering all available predictors. Refine the model using stepwise selection.

   (b) Repeat the analysis with an ordinal model. Comment on any differences.

   (c) We have a new observation with these attributes:

   ```
   > newobs
     fixed.acidity volatile.acidity citric.acid residual.sugar chlorides
              7.5              0.5         0.3           2.25      0.09
     free.sulfur.dioxide total.sulfur.dioxide  density   pH   sulphates alcohol
                    14                   36    0.997   3.3        0.63      9.5
   ```

   Report the probability that this wine variant is a bad wine, according to the multinomial and ordinal models.

   (d) Under the ordinal model, what is the odds ratio of being classified as "average" or "bad" of a wine variant with residual sugar level 3 compared to a variant with residual sugar level 5, given that the other attributes are the same?

2. (10 pts) Let x be a random variable, $x \in (0,1)$ with pdf

$$f(x) = bx^n(1-x)^n.$$

Easy to see that x follows a Beta distribution with normalizing constant b.

(a) Let the envelope $h(x)$ be $U(0,1)$. Construct a rejection sampling algorithm to generate from the above distribution with $n = 5$.

(b) Implement your algorithm in R and generate $10,000$ samples. Compare the sample pdf curve with the actual curve of the corresponding Beta distribution. You should choose a seed for the random number generator and specify that in your R code.

3. (10 pts) Consider the simple linear regression model

$$\begin{aligned} y_i &= \beta_0 + \beta_1 x_i + \epsilon_i, \quad i = 1, \ldots, n \\ \epsilon_i &\sim N(0, \sigma^2). \end{aligned}$$

We assume the following priors for the parameters:

$$\beta_0 \sim N(0, \sigma_\beta^2), \quad \beta_1 \sim N(0, \sigma_\beta^2), \quad \sigma^2 \sim IG(a, b).$$

The Inverse Gamma $IG(a,b)$ distribution has pdf

$$f(x) = \frac{b^a}{\Gamma(a)} x^{-a-1} \exp\left(-\frac{b}{x}\right)$$

You can generate values from $IG(a,b)$ distribution by using the function `rinvgamma` in the package `invgamma`, setting *shape* and *rate* arguments to a and b.

(a) Derive the conditional posterior distribution of β_0, β_1 and σ^2

(b) Consider the following data

Test cracking (x)	Actual cracking (y)
2.0	1.9
3.0	2.7
4.0	4.2
5.0	4.8
6.0	4.8
7.0	5.1

Based on your answer in part (a), implement a Gibbs sampler to sample from the joint posterior distribution of $(\beta_0, \beta_1, \sigma^2)$ given this data. For the priors, take $\sigma_\beta^2 = 10; a = 2; b = 1$. Run your algorithm for $200,000$ iterations where the first $50,000$ iterations are burn-in. Report the posterior median of the 3 parameters.

You should choose a seed for the random number generator and specify that in your R code.