MAST90104: Introduction to Statistical Learning

Assignment 4

Please submit a scanned or other electronic .pdf of your work, **named XabYcd.pdf where XabYcd is your name,** via the Learning Management System - see this link for instructions.

The .pdf must have in **one file**:

- handwritten or typed answers to the questions
- handwritten or typed R code used to produce your answers
- graphics required to answer the questions

If you have more than one file submitted, *only the last LMS .pdf file with your name on it will be marked.*

1. College juniors (undergraduates) at US universities were asked if they were "unlikely", "somewhat likely", or "very likely" to apply to graduate school. Data on whether the parents have graduate education (`pared`), whether the undergraduate institution is public (`public`), and current GPA (`gpa`), were also collected.

```
load("ologit.Rdata")
head(dat)

##             apply pared public  gpa
## 1     very likely     0      0 3.26
## 2 somewhat likely     1      0 3.21
## 3        unlikely     1      1 3.94
## 4 somewhat likely     0      0 2.81
## 5 somewhat likely     0      0 2.53
## 6        unlikely     0      1 2.59

ftable(xtabs(~ public + apply + pared, data = dat))

##                        pared   0   1
## public apply
## 0      unlikely              175  14
##        somewhat likely        98  26
##        very likely            20  10
## 1      unlikely               25   6
##        somewhat likely        12   4
##        very likely             7   3
```

   (a) Fit a multinomial model to predict the probabilities for graduate application by category, "unlikely", "somewhat likely", or "very likely".

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

   (c) Under the ordinal model, what is the fitted probability that a student whose parents have graduate degrees and who went to a public university, with a gpa of 3.0, is very likely to undertake graduate study?

2. The $t(3)$ distribution has pdf $p(x) = \frac{2}{\sqrt{3}\pi}\left(1+\frac{x^2}{3}\right)^{-2}$, $-\infty < x < \infty$, and the Cauchy$(\sqrt{3}, 0)$ distribution has pdf $g(y) = \frac{1}{\sqrt{3}\pi}\left(1+\frac{y^2}{3}\right)^{-1}$, $-\infty < y < \infty$.

(a) Suppose $U \overset{d}{=} U(0,1)$, and $Y = \sqrt{3}\tan(\pi(U - \frac{1}{2}))$. Show that $Y \overset{d}{=} \text{Cauchy}(\sqrt{3}, 0)$.

(b) Construct an Acceptance-Rejection (A-R) sampling algorithm (or a mixture of A-R and transformation algorithm) for generating random numbers from $t(3)$ by using the result of (a).

(c) Write an R function to implement (b). Then use it to generate a sample of 1000 numbers from $t(3)$, and compare the sample pdf curve with the actual $t(3)$ curve.

3. The Dirichlet distribution is a multivariate generalisation of the beta distribution. It takes values in $\{\mathbf{x} = (x_1, \ldots, x_d) : x_i \in [0,1], \sum_i x_i = 1\}$, and, for $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_d) \in \mathbb{R}_+^d$, has density

$$f(\mathbf{x}) = \frac{1}{B(\boldsymbol{\alpha})} \prod_{i=1}^{d} x_i^{\alpha_i - 1}, \text{ where } B(\boldsymbol{\alpha}) = \frac{\prod_i \Gamma(\alpha_i)}{\Gamma\left(\sum_i \alpha_i\right)}.$$

(a) Prove that if $\mathbf{X} = (X_1, \ldots, X_d)$ has a Dirichlet distribution with parameter $\boldsymbol{\alpha}$ (we write $\mathbf{X} \sim \text{Dir}(\boldsymbol{\alpha})$), then $\mathbb{E}X_i = \alpha_i / \sum_i \alpha_i$. Hint: $\int \cdots \int f(\mathbf{x}) dx_1 \cdots dx_d = 1$.

(b) Show that the Dirichlet distribution is the conjugate prior for the multinomial. That is, if $\mathbf{p} \sim \text{Dir}(\boldsymbol{\alpha})$ and $\mathbf{X}|\mathbf{p} \sim \text{multinomial}(n, \mathbf{p})$, then $\mathbf{p}|\{\mathbf{X} = \mathbf{x}\} \sim \text{Dir}(\boldsymbol{\beta})$, where $\boldsymbol{\beta}$ depends on $\boldsymbol{\alpha}$ and $\mathbf{x}$.

(c) In 2003 Briggs, Ades and Price reported on a trial for the treatment of asthma. Patients received one of two treatments (seretide or fluticasone), and their status was monitored from week to week. Possible states were

STW Successfully treated week

UTW Unsuccessfully treated week

HEX Hospital managed exacerbation

PEX Primary-care managed exacerbation

TF Treatment failure (treatment ceased and patient removed from the trial)

For patients on seretide, the number of transitions from one state to another were

	To					
	STW	UTW	HEX	PEX	TF	
From						Total
STW	210	60	0	1	1	272
UTW	88	641	0	4	13	746
HEX	0	0	0	0	0	0
PEX	1	0	0	0	1	2
TF	0	0	0	0	81	81

The rows of this table can be considered as observations from independent multinomial random variables. Using Dir(1,...,1) priors, give Bayesian estimates (posterior means) for

$$p_{ij} = \mathbb{P}(\text{ state changes from } i \text{ to } j \text{ })$$

for $i = \text{STW}, \ldots, \text{PEX}$ and $j = \text{STW}, \ldots, \text{TF}$.

What prior would be appropriate for the transitions from state TF?

4. There is a famous data set from RA Fisher's 1936 paper on *The use of multiple measurements in taxonomic problems* which consists of 50 samples from each of three species of Iris (Iris setosa, Iris virginica and Iris versicolor). Four features were measured from each sample: the length and the width of the sepals and petals, in centimetres.

The data is available in R in the data set `iris`.

Perform a principal components and cluster analysis on the four variables using the R commands `prcomp` and `kmeans`. Use two separate methods to determine the number of clusters.

Use the `ggplot2` and `ggfortify` libraries to do plots of the first two principal components that indicate where the directions of the original variables and the positions of the cluster means.

Comment on your results.