SIT743 Bayesian Learning and Graphical Models
Assignment-1
Total Marks = 120, Weighting - 25%
Due date: 26 April 2020 by 11.30 PM

INSTRUCTIONS:

- For this assignment, you need to submit the following **THREE** files.

  1. **A written document** (*A **single pdf only***) covering all of the items described in the questions. All answers to the questions must be written in this document, i.e, **not** in the other files (code files) that you will be submitting. ***All the relevant results (outputs, figures) obtained by executing your R code must be included in this document.***
     *For questions that involve mathematical formulas, you may write the answers manually (hand written answers), **scan it to pdf** and **combine with your answer document**. Submit a combined **single pdf** of your answer document.*

  2. A **separate** ".R" file or '.txt' file containing your code (R-code script) that you implemented to produce the results. Name the file as "*name-StudentID-Ass1-Code.R*" (where `*name*' is replaced with your name - you can use your surname or first name, and *StudentID* with your student ID).

  3. **A data file** named "*name-StudentID-LzMyData.txt*" (where `*name*' is replaced with your name - you can use your surname or first name, and *StudentID* with your student ID).

- All the documents and files should be submitted (uploaded) via *SIT 743 Clouddeakin Assignment Dropbox* by the due date and time.
- **Zip files are NOT accepted**. All three files should be uploaded **separately** to the CloudDeakin.
- E-mail or manual submissions are **NOT** allowed. Photos of the document are **NOT** allowed.
- The questions Q2 and Q3 **do not** require any R programming.

==========

Some of the questions in this assignment require you to use the "**Lizard Island**" dataset. This dataset is given as a CSV file, named "**LZIsData.csv**". You can download this from the Assignment folder in CloudDeakin. Below is the description of this dataset.

Lizard Island dataset:
This dataset gives the weather measurements collected at ***Lizard Island,*** which is an island in the Great Barrier Reef (North Queensland, Australia)**.**
[http://weather.aims.gov.au/#/station/1166].

The data gives 10 minutes sample measurements collected over a 1 month period between May 2019 and June 2019.

The variables include the following (4 variables; in the same order of columns appear in the file **LZIsData.csv**):

Air Temperature: Air temperature in degrees Celsius.

Humidity: Humidity in percentage.

Wind Speed: Maximum Wind speed in kilometre per hour

Air Pressure: pressure measurements expressed in units of Hectopascals

Q1) [19 Marks]:

- Download the data file "**LZIsData.csv**" and save it to your R working directory.
- Assign the data to a matrix, e.g. using

```
the.data <- as.matrix(read.csv("LZIsData.csv", header = FALSE, sep = ","))
```

- Generate a sample of 1500 data using the following:

```
my.data <- the.data [sample(1: 4464,1500),c(1:4)]
```

Save "`my.data`" to a text file titled "*name-StudentID-LzMyData.txt*" using the following R code (**NOTE: you 'must' upload this data text file and the R code along with your submission. If not, ZERO marks will be given for this whole question**).

```
write.table(my.data,"name-StudentID-LzMyData.txt")
```

Use the sampled data ("my.data") to answer the following questions.

1.1) Draw histograms for 'Air temperature' and 'Air Pressure" values, and comment on them. [2 Marks]

1.2) Draw a parallel Box plot using the two variables; 'Air Temperature' and the 'Wind Speed'.
Find five number summaries of these two variables.
Use both five number summaries and the Boxplots to compare and comment on them. [5 Marks]

1.3) Which summary statistics would you choose to summarize the center and spread for the 'Humidity' data? Why (support your answer with proper plot/s)? Find those summary statistics for the "Humidity" data.
[4 Marks]

1.4) Draw a scatterplot of ''Air Temperature' (as x) and 'Humidity' (as y) for the ***first 1000 data vectors selected from the "my.data"*** (name the axes).
Fit a linear regression model to the above two variables and plot the (regression) line on the same scatter plot.
Write down the linear regression equation.
Compute the *correlation coefficient* and the *coefficient of Determination*.
Explain what these results reveal. [8 Marks]

Q2) [21 Marks]

2.1) The table shows results of a survey conducted about the favorite sports, in different states over some period in 2020.

		State		
		New south Wales (N)	**Victoria (V)**	**Queensland (Q)**
Sports	**Footy (F)**	1000	2000	1300
	Basketball (B)	1500	500	500
	Cricket (C)	1400	1000	800

Suppose we select a person at random,

a) What is the probability that the person is from Victoria (V)? [1 mark]

b) What is the probability that the person likes cricket (C) and from New South Wales (N)? [1 Mark]

c) What is the probability that the person likes Footy (F) given that he/she is from Queensland (Q)? [2 Marks]

d) What is the probability that the person, who likes Basketball (B) is from Victoria (V)? [2 Marks]

e) What is the probability that the person is from Victoria (V) or likes cricket (C)? [2 Marks]

f) Find the marginal distribution of sports. [3 marks]

g) Are sports and state mutually exclusive? Explain [2 Marks]

h) Are sports and state independent? Explain [3 marks]

2.2) The weather in Victoria can be summarised as follows
If it rains one day there is a 75% chance it will rain the following day. If it is sunny one day there is a 30% chance it will be sunny the following day. Assume that the prior probability it rained yesterday is 0.6, what is the probability that it was sunny yesterday given that it is rainy today? **[5 Marks]**

Q3) [5 Marks]

3.1) State two differences between frequentist way and the Bayesian way of estimating a parameter [2 marks]

3.2) Why conjugate priors are useful in Bayesian statistics? [1 mark]

3.3) Give two examples of Conjugate pairs (i.e., give two pairs of distributions that can be used for prior and likelihood) [2 marks]

Q4) Frequentist and Bayesian estimations [31 Marks]

An Artificial Intelligence solutions provider, BigSecAI Ltd. houses several computing servers to perform computationally intensive processing, such as deep learning, on sensitive (secure) data for customers, including government agencies. In order to provide reliable service, BigSecAI wants to improve their monitoring and maintenance activities of their computer servers. As part of their planning, the BigSecAI wants to model the lifetime pattern of their servers. BigSecAI assumes that the length of time $\boldsymbol{x_i}$ (in years) a computer server i lasts follows a form of exponential distribution with an unknown parameter θ, as shown below. Here, the quantity $\left(\frac{1}{\theta}\right)$ represents *on average, how long a certain server last*.

$$\boldsymbol{x_i} \sim \boldsymbol{Exp(\theta)}$$

$$\boldsymbol{Exp(\theta)} = \boldsymbol{p(x_i|\theta)} = \boldsymbol{\theta e^{-(x_i\theta)}}$$

Assume that there are N servers used, and each of their lifetime are independently and identically distributed (iid).

4.1) BigSecAI first decided to use a ***frequentist approach*** to arrive at an estimate for θ. Answer the following questions.

a) Show that the joint distribution of lifetime of N servers can be given by the below equation (**show the steps clearly**).

$$\boldsymbol{p(X|\theta)} = \boldsymbol{\theta^N\, e^{-(\theta S)}}, \quad \text{where } \boldsymbol{S} = \sum_{i=1}^{N} \boldsymbol{x_i}$$

[3 marks]

b) Find a simplified expression for the log-likelihood function $\boldsymbol{L(\theta)} = \mathbf{ln}\,(\boldsymbol{p(X|\theta)})$ **[3** marks**]**

c) Show that the Maximum likelihood Estimate ($\hat{\theta}$) of the parameter θ is given by:

$$\widehat{\boldsymbol{\theta}} = \frac{\mathbf{1}}{\overline{\boldsymbol{X}}} \ , \quad where \ \overline{\boldsymbol{X}} = \frac{\mathbf{1}}{\boldsymbol{N}} \sum_{\boldsymbol{i=1}}^{\boldsymbol{N}} \boldsymbol{x_i}$$

[4 Marks]

d) Suppose that the lifetimes of six of their servers are {2, 7, 6, 10, 8, 3}, what is the Maximum likelihood Estimate $\hat{\theta}$ (MLE) of parameter θ given this data? [2 Marks]

e) Hence, on the average, how long would 7 servers last if they are used one after another? [2 Marks]

f) **W**hat is the probability that a server lasts between six and twelve years?

*Hint: Use cumulative distribution function (**cdf**) of exponential distribution. The **cdf** of the exponential distribution is given by* $F(x) = 1 - e^{-(x\theta)}$.

[4 marks]

4.2) BigSecAI has now consulted an overseas computer hardware vendor, ***HardwareExpert,*** which has more experience working with large servers, and obtained some prior information about the lifetime of servers of similar capacity and processing capabilities. The ***HardwareExpert*** mentioned that their θ value follows a pattern that can be described using a form of Gamma distribution, ***Gamma (a,b)***, where a and b are the hyper-parameters of the Gamma distribution, with $a = 0.1$ and $b = 0.1$.

$$\mathbf{Gamma}\ (\boldsymbol{a, b}) \ = \boldsymbol{K} \ \boldsymbol{b^a \theta^{(a-1)} e^{-b\theta}}\ \mathbf{,}$$ Where K is a constant.

a) BigSecAI has decided to use this prior information from ***HardwareExpert*** for their estimation. If it uses the Gamma distribution prior, Gamma **(a,b)**, obtain **an expression** for the **posterior distribution** (show all the steps)**.**
Show that the posterior distribution is also a Gamma distribution, Gamma **(a', b')**, with different hyper-parameters a' and b'. Express a' and b' **in terms of *a*, *b*, *N* and *S***. [5 Marks]

b) Use the values for a and b hyper-parameters suggested by the ***HardwareExpert,*** *and* the server lifetimes that has been observed from 6 servers: {2, 7, 6, 10, 8, 3}, to find the value of a' and b'. What is the posterior mean estimate of θ **?** [4 Marks]

c) Write a R program and plot the obtained likelihood distribution, the prior distribution and the posterior distribution on the same graph. Use different colors to show the distributions on the plot. [4 Marks]

Q5) Bayesian inference for Gaussians (unknown mean and known variance) [15 marks]

A metal factory in Queensland produces iron bars. They are quality tested by measuring the amount of sag they undergo under a standard load. A random sample of n iron bars shows an average sag measurement of 5cm. Assume that the sag measurements are normally distributed with unknown mean θ and known standard deviation 0.25 cm. Suppose your prior distribution for θ is normal with mean 4 cm and standard deviation of 2 cm.

a) Find the posterior distribution for θ in terms of n. (Do not derive the formulae) [3 Marks]

b) For n=20, find the mean and the standard deviation of the posterior distribution. Comment on the posterior variance [3 Marks]

c) For n=100, find the mean and the standard deviation of the posterior distribution. Compare with the results obtained for n=20 in the above question Q5(b) and comment. [3 Marks]

d) Assume that the **prior** distribution is **changed**, and now the prior is distributed as a triangle defined over the range 4 to 6, as shown below:

$$P(\theta) = \begin{cases} \frac{4}{3}\theta - \frac{16}{3} & for \;\; 4 \leq \theta \leq 4.75 \\ -\frac{4}{5}\theta + \frac{24}{5} & for \;\; 4.75 < \theta \leq 6 \end{cases}$$

Write a R program to implement this triangular prior, and compute the posterior distribution considering $n = 1$. Using R program find the posterior mean estimate of θ. Sketch, on a single coordinate axes, the prior, likelihood and the posterior distributions obtained. [6 Marks]

(**Hint**. Use 'Bolstad' package in R to perform this.
library(Bolstad)
#https://cran.r-project.org/web/packages/Bolstad/Bolstad.pdf)

Q6) Clustering: [11 marks]

6.1) ***K-Means clustering:*** Use the data file "IOCdata2020.txt" provided in CloudDeakin for this question. Load the file "IOCdata2020.txt" using the following:

```
zz<-read.table("IOCdata2020.txt")

zz<-as.matrix(zz)
```

a) Draw a scatter plot of the data. [1 mark].

b) State the number of classes/clusters that can be found in the data (by visual examination of the scatter plot) [1 marks].

c) Use the above number of classes as the k value and perform the k-means clustering on that data. Show the results using a scatterplot (show the different clusters with different colours). Comment on the clusters obtained. [2 Marks]

d) Vary the number of clusters (k value) from 2 to 20 in increments of 1 and perform the k-means clustering for the above data. Record the *total within sum of squares (TOTWSS)* value for each k, and plot a graph of TOTWSS verses k. Explain how you can use this graph to find the correct number of classes/clusters in the data. [3 marks]

6.2) ***Spectral Clustering:*** Use the same dataset (`zz`) and run a spectral clustering (use the number of clusters/centers as 3) on it. Show the results on a scatter plot (with colour coding). Compare these clusters with the clusters obtained using the k-means above and comment on the results. [4 Marks]

Q7) [18 Marks]

For this question you will be using "**HeronIslandWaterTemp**" dataset. This dataset is given as a CSV file, named "**HeronIslandWaterTemp.csv**". You can download this dataset from the Assignment folder in CloudDeakin.

This dataset contains only one variable, namely **"Water Temperature" (WT)**.

Use the following R code to load the whole data for WT variable

```
WTempdata <- as.matrix(read.csv("HeronIslandWaterTemp.csv", header =
                                TRUE, sep = ","))
]
```

7.1) Provide a time series plot of the WT data (use the index as the time (x-axis)) using R code. [1 Marks]

7.2) Plot the histogram for WT data. Comment on the shape. How many ***modes*** can be observed in the data? [2 Marks]

7.3) Fit a **single Gaussian** model $\mathcal{N}(\mu, \sigma^2)$ to the distribution of the data, where μ is the **mean** and σ is the **standard deviation** of the Gaussian distribution.

Find the maximum likelihood estimate (MLE) of the parameters, i.e., the **mean** μ and the **standard deviation** (σ).

Plot the obtained (single Gaussian) density distribution along with the histogram on the same graph.
[3 Marks]

7.4) Fit a **mixture of Gaussians** model to the distribution of the data using **the number of Gaussians equal to the number of modes** found in the data (in Q7.2 above) . Write the R code to perform this. Provide the **mixing coefficients**, **mean and standard deviation for each of the Gaussians** found. [4 Marks]

7.5) Plot these Gaussians on top of the histogram plot. Include a plot of the combined density distribution as well (use different colors for the density plots in the same graph). [3 Marks]

7.6) Provide a plot of the **log likelihood values** obtained over the iterations and comment on them. [2 Marks]

7.7) Comment on the distribution models obtained in Q7.3 and Q7.4. Which one is better? [1 Marks]

7.8) What is the main problem that you might come across when performing a maximum likelihood estimation using mixture of Gaussians? How can you resolve that problem in practice? [2 Marks]