SIT743 Multivariate and Categorical Data Analysis
Assignment-2

Total Marks = 100, Weighting - 20%

Due date: 23rd September 2018 by 11.30 PM

For this assignment, you need to submit the following two 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 should be included in this document.*
2. A separate ".R" file or '.txt' file containing your R code (R-code script) that you implemented to produce the results. Name the file as "*name-StudentID-Ass2-Code.R*" (where `*name*' is replaced with your name - you can use your surname or first name).

All the files should be submitted (uploaded) via *SIT 743 Clouddeakin Assignment Dropbox* by the due date and time.

- E-mail or manual submissions are **NOT** allowed.
- Zip files are **NOT** allowed
- **Photos or other formats (other than pdf) of the document** are **NOT** allowed.

==

Assignment tasks

Q1) [Marks 4 +4+4+5+4+2+2+3=28]

For this question you will be using the "AIMSNingalooReefAirPressure.csv" dataset. This dataset gives the air pressure measurements collected at Ningaloo reef in Western Australia over a one year period between August 2017 and August 2018.

You can download this dataset from the Assignment folder in CloudDeakin. You can use the following R code to load the data:

```
AIMSDataAirPres<-
as.matrix(read.csv("AIMSNingalooReefAirPressure.csv", header = TRUE,
sep = ",", quote = "\"", dec = ".", fill = TRUE, comment.char = ""))
```

1.1) Provide a time series plot of the data (use the index as the time (x-axis)). Use the following R code to plot it:

```
plot(AIMSDataAirPres)
```

Provide the **five point summary**, **mean** and the **standard deviation** of the air pressure data.

1.2) Plot the histogram of the air pressure data. Comment on the shape. How many ***modes*** can be observed in the data?

1.3) 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). You can use the following code to perform the fitting:

```
library(MASS)
fit1<-fitdistr(AIMSDataAirPres,"normal")
```

Plot the obtained density distribution.

1.4) 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 Q1.2 above) . Write the R code to perform this. Provide the **mixing coefficients**, **mean and standard deviation for each of the Gaussians** found.

1.5) 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).

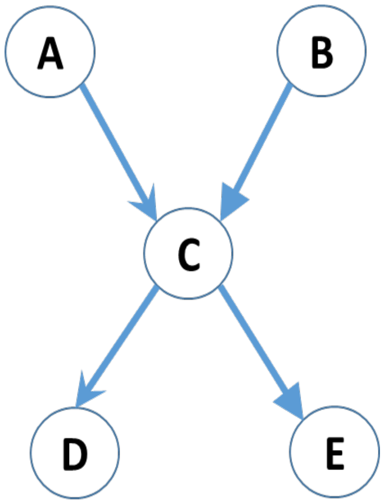
1.6) Provide a plot of the **log likelihood values** obtained over the iterations and comment on them.

1.7) Comment on the distribution models obtained in Q1.3 and Q1.4. Which one is better?

1.8) 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?

Q2) [Marks 2+3+3+2=10]

Consider the following Bayes network shown below.



The nodes represent the following variables:

A ∈{winter, spring, summer, autumn}

B ∈{high river flow, low river flow}

C ∈ {Bass, Barramundi, Cod}

D ∈ {light, medium, dark}

E ∈ {wide, thin}

2.1) Write down the joint distribution $P(A, B, \mathrm{C}, \mathrm{D}, \mathrm{E})$ for the above network.

2.2) How many parameters are required to fully specify the distribution according to the above network?

2.3) How many parameters are required if there are no independencies among the variables is assumed. Compare with the result of above question Q2.2.

2.4) Write down the equation (only) to compute P(A=summer | D=dark, E=wide)

Q3) [Marks 4+5+3 = 12]

A belief network models the relation between the variables ***oil; inf; eh; bp; rt*** which stand for the price of oil, inflation rate, economy health, British Petroleum Stock price, and retailer stock price. Each variable takes the states *low; high*, except for ***bp*** and ***rt*** which have states *low; high; normal*. The belief network model for these variables has tables as shown below

$P(eh = low) = 0.2$	
$p(bp = low \mid oil = low) = 0.8$	$p(bp = normal \mid oil = low) = 0.15$
$p(bp = low \mid oil = high) = 0.1$	$p(bp = normal \mid oil = high) = 0.4$
$p(oil = low \mid eh = low) = 0.9$	$p(oil = low \mid eh = high) = 0.05$
$p(rt = low \mid inf = low, eh = low) = 0.6$	$p(rt = low \mid inf = low, eh = high) = 0.1$
$p(rt = low \mid inf = high, eh = low) = 0.2$	$p(rt = low \mid inf = high, eh = high) = 0.05$
$p(rt = normal \mid inf = low, eh = low) = 0.3$	$p(rt = normal \mid inf = low, eh = high) = 0.2$
$p(rt = normal \mid inf = high, eh = low) = 0.2$	$p(rt = normal \mid inf = high, eh = high) = 0.1$
$p(inf = low \mid oil = low, eh = low) = 0.9$	$p(inf = low \mid oil = low, eh = high) = 0.1$
$p(inf = low \mid oil = high, eh = low) = 0.2$	$p(inf = low \mid oil = high, eh = high) = 0.02$

3.1) Draw the belief network for this distribution

3.2) Use the below libraries in R to create this belief network in R along with the probability values as shown in the above table.

You may use the following **libraries** for this:

```
library("gRain")
source("https://bioconductor.org/biocLite.R")
biocLite("RBGL")
library(RBGL)
library(gRbase)
library(gRain)
biocLite("Rgraphviz")
```

```
#define the appropriate network and use the "compileCPT()"function to Compile list of conditional probability tables and create the network.
```

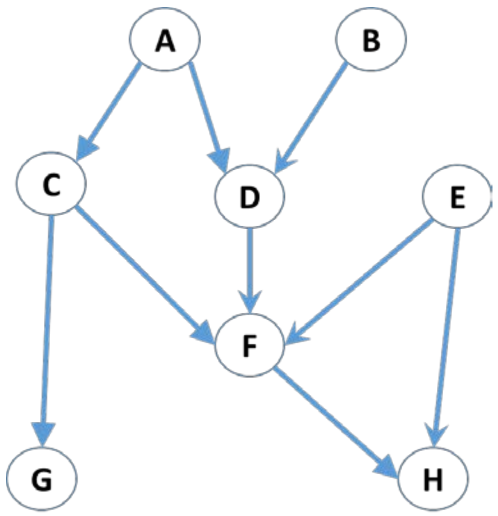
Show the probability tables obtained in the R output and verify with the above table.

3.3) Use R program to compute the following:

Given that the **BP stock price** is *high* and the **retailer stock pric**e is *normal*, what is the probability that **inflation** is *high*? `

Q4) **[Marks : (2+2+2+2+2+3+3) + 5=21]**

Consider the Bayesian network shown in the diagram below.



4.1) Use the *d-separation* method to find out whether each of the following statements is correct or not and **mention the reason for it.**

a) $C \perp G \mid \emptyset$ (C is marginally independent of G)

b) $C \perp H \mid E$ (C is conditionally independent of H given E)

c) $G \perp \mathrm{E} \mid \mathrm{D}$

d) $C \perp \mathrm{H} \mid \mathrm{F}$

e) $\mathrm{B} \perp \mathrm{G} \mid F$

f) $B \perp G \mid \{\mathrm{D}, \mathrm{C}, \mathrm{E}\}$

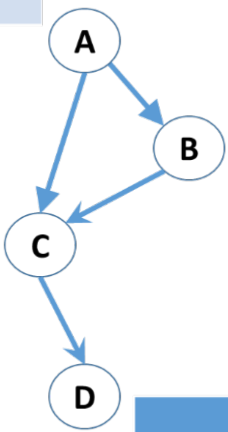
g) $A \perp H \mid \{D, F\}$

4.2) Write a **R-Program** to produce this Bayesian network and **perform the *d-sepration* tests** for all of the above cases mentioned in Q4.1 (a) to (g).

Q5) [Marks: 8+2=10]

Consider four **binary** variables A, B, C, D. The Directed Acyclic Graph (DAG) shown below describe the relationship between these variables along with their conditional probability tables (CPT).

A=0	A=1
α	$1-\alpha$



	B=0	B=1
A=0	0.2	0.8
A=1	β	$1-\beta$

		C	
		0	1
A, B	00	**0.2**	**0.8**
	01	**0.4**	**0.6**
	10	**0.7**	**0.3**
	11	**0.9**	**0.1**

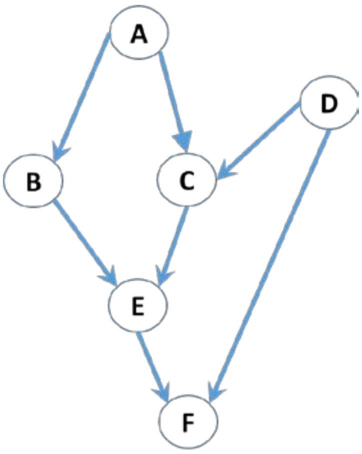
	D=0	D=1
C=0	γ	$1-\gamma$
C=1	0.3	0.7

5.1) Find an expression for $P(D = 1|A = 0)$ and show that it only depends on the γ values.

5.2) Find the value of $P(D = 1|A = 0)$ when $\gamma = 0.1$.

Q6) [Marks : 3 + 10 =13]

Consider the Bayesian network shown below.



6.1) Write down the expression for computing $P(\mathrm{F} \mid \mathrm{A} = 1)$.

6.2) Show the step by step process to perform *variable elimination* to compute $P(\mathrm{F} \mid \mathrm{A} = 1)$. Use the following variable ordering for elimination process: B, C, D, E

Q7) Examples of Bayesian applications [6 Marks]

An example of a real world application of Bayesian methods is described in the following article, which describes a scenario for locating a missing plane (AF447) in the ocean.

- http://apps.npr.org/documents/document.html?id=1096813-af447-final-report-to-bea-jan-2011-2

Do a research (using journal or conference papers/publications) and describe **two other real world applications** of any Bayesian methods/Bayesian nets. Briefly describe what the application is about, and what techniques are used. **Provide references** for each of the applications/papers. Description **should not exceed 300 words** (**for both the applications together, including references**).