SIT743 Multivariate and Categorical Data Analysis
Assignment-2
Total Marks = 100, Weighting - 40%
Due date: 25th September 2019 by 11.30 PM

INSTRUCTIONS:

- 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 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-Ass2-Code.R*" (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 two files should be uploaded **separately** to the CloudDeakin.
- E-mail or manual submissions are **NOT** allowed. Photos of the document are **NOT** allowed.

==========

Assignment tasks

Q1) [32 Marks]

A survey has been conducted in Melbourne to study the *travel mode choice (M)* behavior of people. The list of factors that influence the travel mode choice, along with their possible values, is provided below. A Bayesian network that has been created based on the survey results is shown below, which represents the relationship between these various factors (variables).

J (Occupation) ∈{Student, Employee, Individual, Others}

A (Age) ∈{<18, 18-35, 36-55, >55}

S (Salary – monthly in dollars) ∈ {<2000, 2000-6000, 6000-10000, >10000}

V (owning a private car) ∈ {Yes, No}

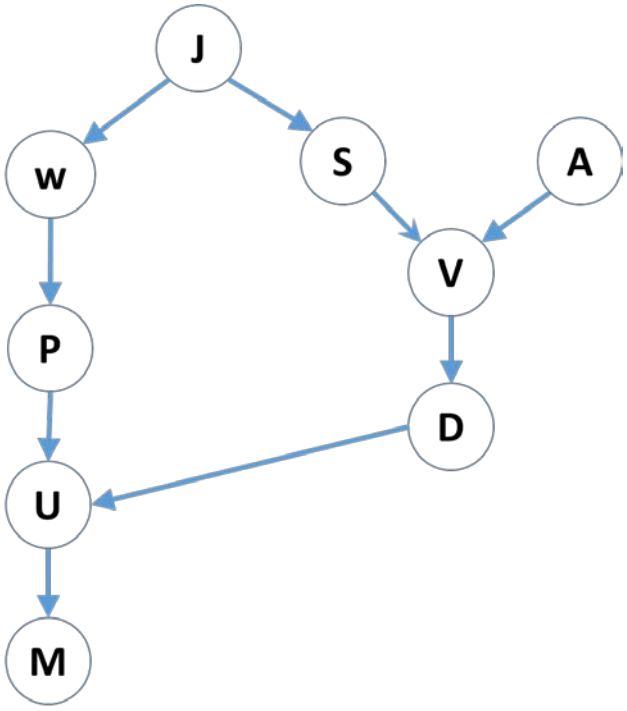
W (Trip purpose) ∈ {Commute to work, other}

D (Trip distance in km) ∈ {<1, 1-3, 3-6, >6}

P (Trip time period) ∈ {Peak hour, Off-peak hour}

U (Trip duration in mins) ∈ {<30, 30-60, >60}

M (Travel mode choice) ∈ {Walking, Bicycle, Public transport, Car}



1.1) Write down the joint distribution $P(J, \mathrm{W}, \mathrm{S}, \mathrm{A}, \mathrm{P}, \mathrm{V}, \mathrm{D}, \mathrm{U}, \mathrm{M})$ for the above network.

1.2) Find the minimum number of parameters required to fully specify the distribution according to the above network.

1.3) How many parameters are required, at a minimum, if there are **no independencies among the variables is assumed?** Compare with the result of the above question (Q1.2) and comment.

1.4) ***d-separation*** method can be used to find two sets of independent or conditionally independent variables in a Bayesian network. For **each of the statements** given below from (a) to (c), perform the following:

- List **all** the possible paths from the first (set of) node/s to the second (set of) node/s.
- State if each of those paths is *blocking* or *non-blocking* **with reasons**.

- Hence, mention if the statement is **true** or **false**.

a) $W \perp \mathrm{V} \mid \emptyset$ (W is marginally independent of V)

b) $A \perp \mathrm{M} \mid \{\mathrm{D}, \mathrm{W}\}$ (A is conditionally independent of M given {D, W})

c) $\{A, W\} \perp \mathrm{D} \mid \mathrm{V}$

1.5) Write a R-Program to produce the above Bayesian network, and perform the d-separation tests for all of the above cases mentioned in Q1.4 (a) to (c). Show the **plot of the network** you obtained and the **output (of d-separation test)** from your program.

1.6) Show the step by step process to perform **variable elimination** to compute $\boldsymbol{P}(\mathbf{M} \mid \mathbf{S} = \mathbf{3000} - \mathbf{6000},\ \mathbf{V} = \boldsymbol{Yes},\ \boldsymbol{P} = \boldsymbol{Peak\ hour},\ \boldsymbol{D} = < \mathbf{1})$. Use the following variable ordering for the elimination process: **J, W, A, U**.

[Marks 2+5+3+11+3+8 = 32]

Q2) [16 Marks] **Implementing a Bayesian network in R and performing inference**

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* respectively. Each variable takes different states as given below.

$\boldsymbol{eh}$ $(economic\ health) \in \{low, high\}$

$\boldsymbol{oil}$ $(price\ of\ oil) \in \{low, high\}$

$\boldsymbol{inf}$ $(inflation\ rate) \in \{low, high\}$

$\boldsymbol{bp}$ $(British\ Petrol\ stock\ price) \in \{low, lower\ middle\ (LM), upper\ middle\ (UM), high\}$

$\boldsymbol{rt}$ $(retailer\ stock\ price) \in \{low, high\}$

The belief network that models these variables has (probability) tables as shown below.

$P(eh = low) = 0.3$	
$p(oil = low \mid eh = low) = 0.3$	$p(oil = high \mid eh = high) = 0.4$
$p(bp = low \mid oil = low) = 0.2$	$p(bp = UM \mid oil = low) = 0.3$
$p(bp = high \mid oil = low) = 0.05$	$p(bp = LM \mid oil = high) = 0.5$
$p(bp = UM \mid oil = high) = 0.2$	$p(bp = high \mid oil = high) = 0.1$
$p(inf = low \mid oil = low, eh = low) = 0.1$	$p(inf = high \mid oil = high, eh = low) = 0.4$
$p(inf = low \mid oil = low, eh = high) = 0.5$	$p(inf = high \mid oil = high, eh = high) = 0.6$
$p(rt = low \mid inf = low, eh = low) = 0.2$	$p(rt = high \mid inf = low, eh = high) = 0.3$
$p(rt = low \mid inf = high, eh = low) = 0.6$	$p(rt = high \mid inf = high, eh = high) = 0.8$

2.1) 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:

```
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.
```

a) Show the obtained **belief network** for this distribution
b) Show the probability tables **obtained from the R output,** (and verify with the above table).

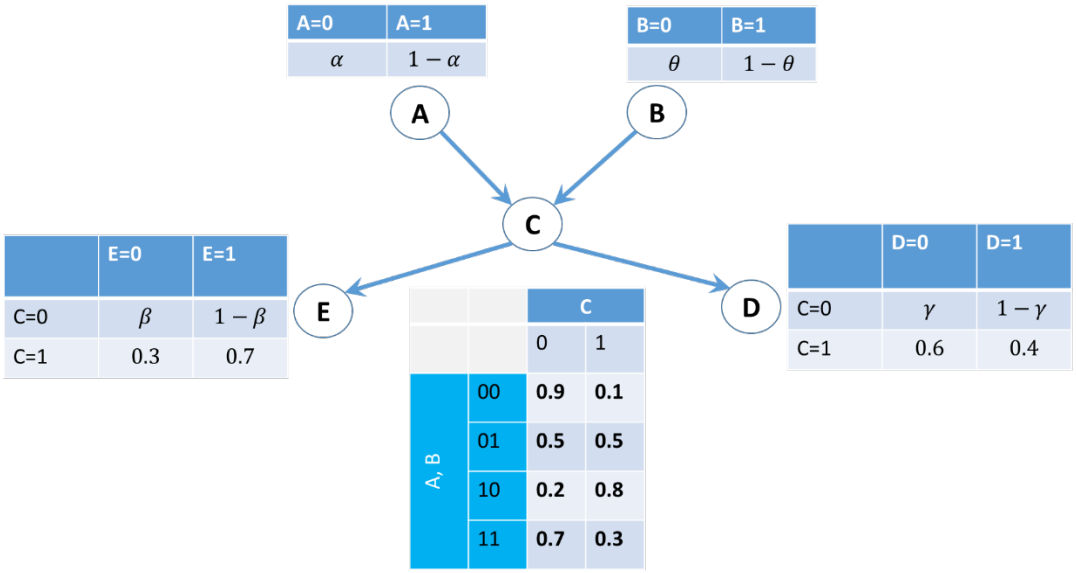
2.2) Use R program to compute the following probabilities:

a) Given that the **Oil price** is *low* and the **retailer stock price** is *low*, what is the **most possible state** of the **British petroleum stock price**?
b) Given that the **inflation rate** is *low*, what is the probability that **retailer stock price** is *high*?
c) Find the marginal distribution of ***price of oil.***
d) Find the joint distribution of **inflation,** and **British petroleum stock price.**

[Marks: (3+5) + (2+2+2+2) = 16]

Q3) [16 Marks]

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



A=0	A=1
α	$1-\alpha$

B=0	B=1
θ	$1-\theta$

	E=0	E=1
C=0	β	$1-\beta$
C=1	0.3	0.7

	D=0	D=1
C=0	γ	$1-\gamma$
C=1	0.6	0.4

		C	
		0	1
A, B	00	**0.9**	**0.1**
	01	**0.5**	**0.5**
	10	**0.2**	**0.8**
	11	**0.7**	**0.3**

3.1) In the above network, state why A is independent of B, i.e., A⊥B.

3.2) Hence, find an expression (in a simplified form) for $\boldsymbol{P(E=1|A=1,B=0)}$ in terms of β.

3.3) The table shown below provides 20 simulated data obtained for the above Bayesian network. Use this data to find the maximum likelihood estimates of α, β, γ and θ.

	A	B	C	D	E
1	1	1	1	0	0
2	1	0	1	1	1
3	1	1	0	1	0
4	1	1	0	1	1
5	0	1	1	1	1
6	1	1	0	0	0
7	0	1	1	0	1
8	1	1	0	1	0
9	1	1	0	1	1
10	1	1	0	1	1
11	1	0	1	0	1
12	1	1	0	0	0
13	1	0	1	0	1
14	1	1	0	1	1
15	0	1	0	1	1
16	1	1	0	1	1
17	1	0	1	1	1
18	1	1	0	1	1
19	1	1	0	1	0
20	1	1	1	1	1

3.4) Find the value of $\boldsymbol{P(E = 1|A = 1, B = 0)}$ using the values obtained for β from the above question Q3.3.

[Marks 2+ 8 + 4 + 2 = 16]

Q4) Bayesian Structure Learning [30 Marks]

For this question, you will be using a dataset, called ***"Child"***, which contains 20 variables. This dataset provides information about diagnosing congenital heart disease in a new born "blue baby". The csv file (**"CHILD10k.csv"**) containing the dataset can be downloaded from CloudDeakin.

Use the following R code to load the ***Child*** dataset:

```
ChildData <- read.csv(file="CHILD10k.csv", header=TRUE, sep=",")
```

The ***true network structure*** of this dataset can be viewed (plot) using the following R code.

```
library(bnlearn)
#create and plot the network structure.
modelstring = paste0("[BirthAsphyxia][Disease|BirthAsphyxia][LVH|Disease][DuctFlow|Disease]",
            "[CardiacMixing|Disease][LungParench|Disease][LungFlow|Disease][Sick|Disease]",
            "[HypDistrib|DuctFlow:CardiacMixing][HypoxiaInO2|CardiacMixing:LungParench]",
            "[CO2|LungParench][ChestXray|LungParench:LungFlow][Grunting|LungParench:Sick]",
            "[LVHreport|LVH][Age|Disease:Sick][LowerBodyO2|HypDistrib:HypoxiaInO2]",
            "[RUQO2|HypoxiaInO2][CO2Report|CO2][XrayReport|ChestXray][GruntingReport|Grunting]")

dag = model2network(modelstring)
par(mfrow = c(1,1))
#source("https://bioconductor.org/biocLite.R")
#biocLite("Rgraphviz")
graphviz.plot(dag)
```

Use R programming, as appropriate, to answers the following questions.

4.1) Use the *Child* dataset to learn Bayesian network structures using **hill-climbing (hc) algorithm**, utilizing two different scoring methods, namely **Bayesian Information Criterion score (BIC score)** and the **Bayesian Dirichlet equivalent (Bde score), for each of the** following **sample sizes** of the data**:**

a) **100 (first 100 data)**
b) **500 (first 500 data)**
c) **1000 (first 1000 data)**
d) **5000 (first 5000 data)**

For each of the above cases,

- provide the scores obtained for BIC and BDe,
- Plot the network structure obtained for the BIC and BDe scores.

4.2) Based on the results obtained for the above question (Q 4.1), discuss how the BIC score compare with BDe score for different sample sizes in terms of **structure** and **score** of the learned network.

4.3)

a) Find the Bayesian network structures utilising the **full dataset, and using both BIC and Bde scores.** Show the scores and the obtained networks.

b) **C**ompare the networks obtained above (in Q4.3.a) for each BIC and Bde scoring methods with the ***true network structure*** and **comment**. Use the "compare()" function and "graphviz.compare()" function available in the "bnlearn" R package to perform the comparisons.

c) Use the network obtained using the **BIC score** in the above question (Q4.3.a) to find the **maximum likelihood estimate of the parameters of the network**, i.e, find the conditional probability distribution table entries (CPD table values). Show the obtained CPD table entries.

d) Use the above learned network obtained (in Q4.3.c) to find the probability of : **P(Disease ="*Lung*" | CO2 = "*High*", LungParench ="*Abnormal*")**

[Marks (4*4) + 3 + (3+3+3+2) = 30]

Q5) Bayesian usage examples [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 networks. Briefly describe on your own words ***what the application is about***, and ***the details of the techniques used***. Provide **references** for each of the applications/papers. Description **should NOT exceed 400 words** (**for both the applications together, including references**).

[Marks (3+3) = 6]