SIT743 Bayesian Learning and Graphical Models
Assignment-2
Total Marks = 100, Weighting - 40%
Due date: 24th May 2020 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) [31 Marks]

Weather conditions influence the production of good quality coffee in a region. A list of factors that influence the coffee cultivation, along with their possible values, and a Bayesian network that represents the relationship between these factors (variables) are given below.

M (Maximum Temperature) ∈ { < 20, 20-30, 30-40, > 40 }

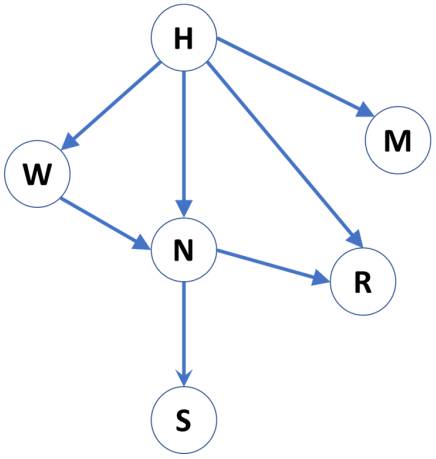
N (Minimum Temperature) ∈ { < 0, 0-10, 10-20, > 20}

W (Wind speed) ∈ {Low, Medium, High}

H (Relative humidity) ∈ {< 50, 50-60, > 60}

R (Precipitation) ∈ {Low, High}

S (Solar radiation) ∈ {Low, Medium, High}



1.1) Write down the joint distribution $P(H, W, N, M, R, S)$ 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)

a) Write down a joint probability density function if there are **no independence among the variables is assumed.**
b) How many parameters are required, at a minimum, if there are **no independencies among the variables is assumed?**
c) 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) $M \perp \mathrm{S} \mid \emptyset$ (M is marginally independent of S)

b) $W \perp \mathrm{R} \mid \{\mathrm{N}, \mathrm{H}\}$ (W is conditionally independent of R given {N, H})

c) $\{R, S\} \perp \mathrm{W} \mid \mathrm{H}$

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)

a) Show the step by step process to perform **variable elimination** to compute $\boldsymbol{P(W \mid S = Low,\ R = Low)}$. Use the following variable ordering for the elimination process:
N, H, M.

b) What is the treewidth of the network, given the above elimination ordering?

[Marks 2+4+5+10+3+7 = 31]

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

A belief network models the relation between the variables ***A, B, C, D and E,*** which represents the *season, river flow rate, fish species, color* and *size* respectively. Each variable takes different states as given below.

$\boldsymbol{A}\ (season) \in \{wet, dry\}$

$\boldsymbol{B}\ (river\ flow\ rate) \in \{low, high\}$

$\boldsymbol{C}\ (fish\ species) \in \{Bass, Cod\}$

$\boldsymbol{D}\ (colour) \in \{light, medium, dark\}$

$\boldsymbol{E}\ (size) \in \{wide, thin\}$

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

$P(A = wet) = 0.3$	$P(B = high) = 0.2$
$p(C = bass \mid A = wet, B = high) = 0.4$	$p(C = bass \mid A = dry, B = high) = 0.5$
$p(C = bass \mid A = wet, B = low) = 0.6$	$p(C = bass \mid A = dry, B = low) = 0.3$
$p(D = light \mid C = bass) = 0.2$	$p(D = medium \mid C = bass) = 0.4$
$p(D = light \mid C = cod) = 0.5$	$p(D = medium \mid C = cod) = 0.3$
$p(E = wide \mid C = bass) = 0.6$	$p(E = wide \mid C = cod) = 0.4$

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:

```
#https://www.bioconductor.org/install/

#BiocManager::install(c("gRain", "RBGL", "gRbase"))

#BiocManager::install(c("Rgraphviz"))

library("Rgraphviz")

library(RBGL)

library(gRbase)

library(gRain)

#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).

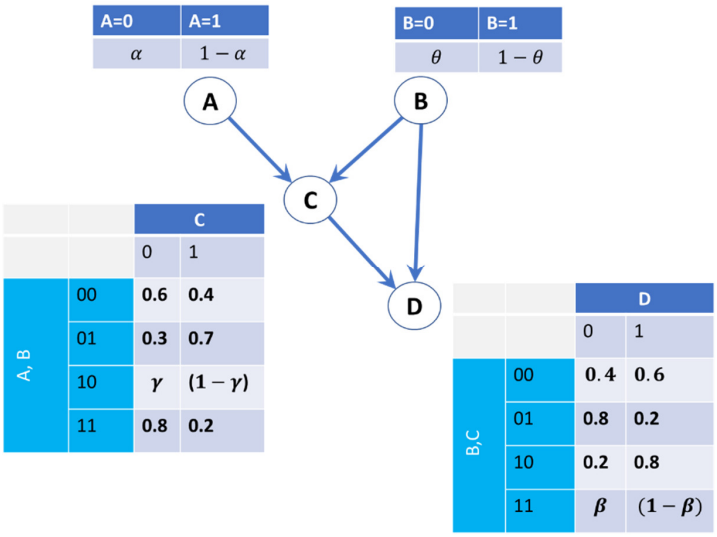
2.2) Use R program to compute the following probabilities:

a) Given that the **river flow rate** is *low*, what is the probability that **size** is *thin*?
b) Given that the **colour** is *dark* and the **season** is *dry*, what is the probability that the **fish species** is **Cod**?
c) Find the joint distribution of **colour** and **fish species.**
d) Find the marginal distribution of ***fish species*.**

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

Q3) [15 Marks]

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



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

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

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

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

[Marks 3+ 7 + 4 + 1 = 15]

Q4) Bayesian Structure Learning [27 Marks]

For this question, you will be using a dataset, called ***"hailfinder"*** available from the 'bnlearn' R package. which contains 56 variables. This has meteorological data.

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

```
library (bnlearn)
# load the data.
data(hailfinder)
summary(hailfinder)
```

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("[N07muVerMo][SubjVertMo][QGVertMotion][SatContMoist][RaoContMoist]",
                     "[VISCloudCov][IRCloudCover][AMInstabMt][WndHodograph][MorningBound][LoLevMoistAd][Date]",
                     "[MorningCIN][LIfr12ZDENSd][AMDewptCalPl][LatestCIN][LLIW]",
                     "[CombVerMo|N07muVerMo:SubjVertMo:QGVertMotion][CombMoisture|SatContMoist:RaoContMoist]",
                     "[CombClouds|VISCloudCov:IRCloudCover][Scenario|Date][CurPropConv|LatestCIN:LLIW]",
                     "[AreaMesoALS|CombVerMo][ScenRelAMCIN|Scenario][ScenRelAMIns|Scenario][ScenRel34|Scenario]",
                     "[ScnRelPlFcst|Scenario][Dewpoints|Scenario][LowLLapse|Scenario][MeanRH|Scenario]",
                     "[MidLLapse|Scenario][MvmtFeatures|Scenario][RHRatio|Scenario][SfcWndShfDis|Scenario]",
                     "[SynForcng|Scenario][TempDis|Scenario][WindAloft|Scenario][WindFieldMt|Scenario]",
                     "[WindFieldPln|Scenario][AreaMoDryAir|AreaMesoALS:CombMoisture]",
                     "[AMCINInScen|ScenRelAMCIN:MorningCIN][AMInsWliScen|ScenRelAMIns:LIfr12ZDENSd:AMDewptCalPl]",
                     "[CldShadeOth|AreaMesoALS:AreaMoDryAir:CombClouds][InsInMt|CldShadeOth:AMInstabMt]",
                     "[OutflowFrMt|InsInMt:WndHodograph][CldShadeConv|InsInMt:WndHodograph][MountainFcst|InsInMt]",
                     "[Boundaries|WndHodograph:OutflowFrMt:MorningBound][N34StarFcst|ScenRel34:PlainsFcst]",
                     "[CompPlFcst|AreaMesoALS:CldShadeOth:Boundaries:CldShadeConv][CapChange|CompPlFcst]",
                     "[InsChange|CompPlFcst:LoLevMoistAd][CapInScen|CapChange:AMCINInScen]",
                     "[InsSclInScen|InsChange:AMInsWliScen][R5Fcst|MountainFcst:N34StarFcst]",
                     "[PlainsFcst|CapInScen:InsSclInScen:CurPropConv:ScnRelPlFcst]")

dag = model2network(modelstring)
par(mfrow = c(1,1))
#BiocManager::install(c("Rgraphviz"))
graphviz.plot(dag)
```

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

4.1) Use the *hailfinder* 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) **1000 (first 1000 data)**
c) **10000 (first 10000 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 these comparisons and comment.

c) Fit the data to the network obtained using the **BIC score** in the above question (Q4.3.a) in order to compute the conditional probability distribution table entries (CPD table values). Show the obtained CPD table entries for the variable "**CombClouds**".

d) Use the above learned network obtained (in Q4.3.c) to find the probability of : $P(CombClouds = ''Cloudy'' \mid MeanRH = ''VeryMoist'',\ IRCloudCover = ''Cloudy'')$

[Marks (3*4) + 3 + (4+3+3+2) = 27]