THE UNIVERSITY OF SYDNEY

COMP3308 Assignment 2

Due: Friday 5pm, week 13 (3rd June)

This assignment is worth 15% of your final assessment

Late submissions policy

A penalty of -1 mark will apply for each day late and the assignment will not be accepted if it is submitted more than 7 days after the due date. The cut-off time is 5pm.

Programming languages

The only allowed languages for this assignment are Python, Java, C, C++, and Matlab.

Submission and pair work

Working in pairs is strongly encouraged. Both students will receive the same mark.

Electronic version (report, plagiarism coversheet, and code) via eLearning. All files should be zipped together in a single file. The zip file should be named 0123456.zip, where 0123456 is your SID. In case of a pair submission, put both SIDs separated by an underscore: 0123456_0789123.zip. Only one of the two students needs to make a submission.

The electronic submissions has to contain the signed plagiarism cover sheet either scanned or as a picture. It can be added separately to the zip file or embedded in the report.

To assist with marking, your zip file must have the following structure:

```
0123456.zip
  |- src/
  |    +- main2.py, main3.py, Main.java, or main.m
  |    +- Makefile (required only if you are using C or C++)
  |    +- ... any additional source/program/data files
  +- report.pdf
```

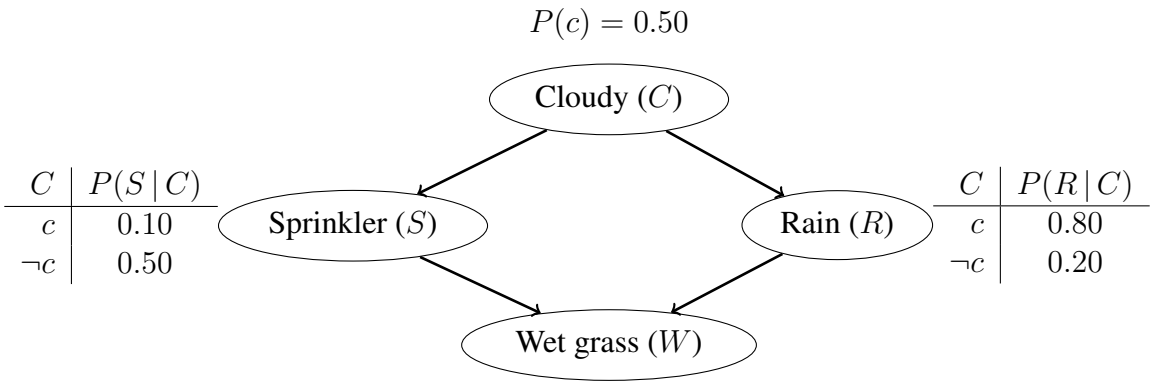
- if you are using **`C`**, **`C++`** you must provide a **`Makefile`** that compiles your program into an executable named **`main`**;
- if you are using **`Python 2`**, your main file must be called **`main2.py`**;
- if you are using **`Python 3`**, your main file must be called **`main3.py`**;

Task description

The goal of this assignment is to:

1. Build a Bayesian network for diagnostic problems and verify independence statements
2. Implement an inference algorithm on a Bayesian network with seven nodes and compute posterior probabilities.

Task 1 Bayesian network construction



C	$P(S \mid C)$
c	0.10
$\neg c$	0.50

C	$P(R \mid C)$
c	0.80
$\neg c$	0.20

S	R	$P(w \mid S, R)$
s	r	0.99
s	$\neg r$	0.90
$\neg s$	r	0.90
$\neg s$	$\neg r$	0.00

Given the above Bayesian network, create a program that estimates $P(c \mid s, w)$ using likelihood weighting. Your program does not need to be general, i.e. it can be specific to this particular query.

Your program should read two positive integers m, n from standard input (the integers will be separated by a space). It should use m as the number of estimates and repeat this calculation n times.

Your program should then output the mean and variance of the posterior estimate to six decimal places to standard output. The two values should be separated by a space.

Sample input:	**Sample output:**
`100 10`	`0.167064 0.003569`
`1000 100`	`0.167309 0.000743`
`1000 1000`	`0.017544 0.000082`

Note that the if you want to verify if your implementation is correct only the sample output for $m = 1000, n = 1000$ should be considered as for the other settings the values can vary greatly.

Task 2

Use JavaBayes (available at http://www.cs.cmu.edu/~javabayes/) to construct a small Bayes net modelling the relationship between metastatic cancer, brain tumor, increased total serum calcium, coma and severe headaches. Metastatic cancer is a possible cause of a brain tumor and is also an explanation for increased total serum calcium. In turn, either of these could explain a patient falling into a coma. Severe headache is also possibly associated with a brain tumor.

The prior probability of metastatic cancer $P(m)$ is 0.20.

The conditional probability of increased total serum calcium $P(I \mid M)$ is:

$$P(i \mid m) = 0.80$$
$$P(i \mid \neg m) = 0.20$$

The conditional probability of brain tumor $P(B \mid M)$ is:

$$P(b \mid m) = 0.20$$
$$P(b \mid \neg m) = 0.05$$

The conditional probability of coma $P(C \mid I, B)$ is:

$$P(c \mid i, b) = 0.80$$
$$P(c \mid \neg i, b) = 0.80$$
$$P(c \mid i, \neg b) = 0.80$$
$$P(c \mid \neg i, \neg b) = 0.05$$

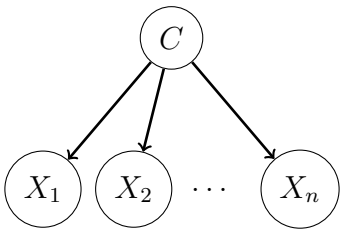
The conditional probability of severe headache $P(S \mid B)$:

$$P(s \mid b) = 0.80$$
$$P(s \mid \neg b) = 0.60$$

1. Construct and show the equivalent graphical model.
2. What is the prior probability of coma P(C)?
3. What is the probability of metastatic cancer given the patient has severe headaches and has not fallen into coma?
4. What is the Markov blanket of coma?
5. Are increased total serum calcium and brain tumor independent given coma? Explain.
6. What is the probability of fallen into coma given the patient has metastatic cancer?

Task 3

The naive Bayes classifier can be represented by the Bayes net below:



1. Show that the naive Bayes factorisation of the equation below follows the independence assumptions encoded by the graph.

$$P(C, X_1, X_2, \ldots, X_n) = P(C) \prod_{i=1}^{n} P(X_i \,|\, C)$$

2. Show that if all variables are binary valued, then:

$$\log \frac{P(c \,|\, X_1, X_2, \ldots, X_n)}{P(\neg c \,|\, X_1, X_2, \ldots, X_n)}$$

   is a linear function of the value of the input variables, that is, it can be written as:

$$\sum_{i=1}^{n} \alpha_i X_i + \alpha_0$$

   where ($X_i = 0$ if $X = \neg x$ and 1 otherwise).

Task 4

Create a Bayes net for a problem of your choice. The network should have at least 5 nodes. Some of the nodes can be binary variables but some should have at least three discrete states. As a suggestion, you can think of diagnosis problems where symptoms are given and you want to infer the underlying causes (e.g. Microsoft Clippy, car breakdowns, medical diagnosis).

1. List your random variables, explain the meaning of each, and specify the conditional probability distributions of the network (including prior probabilities for the variables without parents).
2. Explain the method used to construct such a network.

Task 5 Report

Write a report (similar to a research paper) describing your analysis and findings. It should include the following sections

1. Aim – briefly state the aim of your study (e.g. inference in Bayes nets) and write a paragraph on why the problem is important.
2. Methods – briefly describe Bayes nets, and two inference algorithms: variable elimination, and likelihood weighting.
3. Results and discussion – present the solutions for the previous questions. Present the accuracy results for likelihood weighting compared to exact inference using JavaBayes for different values of m and n. Also note the mean and standard deviation for the posterior probability. Discuss.
4. Conclusions – summarize your main findings and, if possible, suggest future work.
5. Reflection – what was the most important thing you learned from this assignment? (1-2 paragraphs).
6. Include detailed instructions on how to run your code.

Marking sheet

Student(s):

CRITERIA	YOUR MARK	COMMENTS
(10 marks) Report (0.5 marks) Introduction – What is the aim of the study – Why is this study (the problem) important (2 marks) Methods – Bayes net is well described – Inference methods are well described (6 marks) Questions and Results – (2 marks, 0.33 per item) Task 2 – (1 mark, 0.5 per item) Task 3 – (1 mark, 0.5 per item) Task 4 – (2 marks) Task 5, the analysis is well presented. (0.5 mark) Conclusions and future work – Meaningful conclusions based on the results – Meaningful future work suggested (0.5 mark) Reflection – Meaningful and relevant personal reflection (0.5 mark) English and presentation – Academic style, grammatical sentences, no spelling mistakes – Good structure and layout; consistent formatting		
(4 marks) Code – Code runs and computes the correct posterior probability using likelihood weighting.		
(1 mark) At the discretion of the marker – For impressing the marker, excelling expectation, i.e. a general model, other inference algorithms etc.		
Penalties: – 2 mark maximum for badly written code or code that is not well documented and difficult to read. – 0.5 marks for not including instructions on how to run your code – Penalty for late submission: -1 mark for each day late		
Total (out of 15):		