COMP3308 Artificial Intelligence
Semester 1, 2014

Assignment 2. Bayesian networks

Submission deadline: Friday, week 13, 5pm (6 June).

Late policy submission: 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.

This assignment can be completed individually or in pairs. Working in pairs is encouraged. Both students will receive the same mark.

Submission instructions: You need to submit 2 versions:

1. Electronic version (report + 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 submit.
2. Printed version (report + code + plagiarism cover sheet) in the locker COMP3308/3608 (School of IT building, level 1, in the undergraduate labs wing, close to the room where the AI tutorials are held). You can submit the printed version on Monday (9 June) by 5pm.

Programming language: You can write the program in a language of your choice (e.g. Python, Java, C, C++, Matlab) but we need to be able to test your code on the University machines. You need to include instructions on how to run your code.

Weight: This assignment is worth 15 marks = 15% of your final mark.

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.

The questions below should be answered in the report as described in 5.

1. 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 0.2. The conditional probability of increased total serum calcium P(I | M) is: P(i | m) = 0.8 and P(i | ¬m) = 0.2. The conditional probability of brain tumor P(B | M) is: P(b | m) = 0.2 and P(b | ¬m) = 0.05. The conditional probability of coma P(C | I, B) is: P(c | i, b) = 0.8, P(c | ¬i, b) = 0.8, P(c | i, ¬b) = 0.8 and P(c | ¬i, ¬b) = 0.05. The conditional probability of severe headache P(S | B) is P(s | b) = 0.8 and P(s | ¬b) = 0.6.

a) Construct and show the equivalent graphical model.
b) What is the prior probability of coma P(C)?
c) What is the probability of metastatic cancer given the patient has severe headaches and has not fallen into coma?
d) What is the Markov blanket of coma?
e) Are increased total serum calcium and brain tumor independent given coma? Explain.
f) What is the probability of fallen into coma given the patient has metastatic cancer?

2. The naïve Bayes classifier can be represented by the Bayes net below:

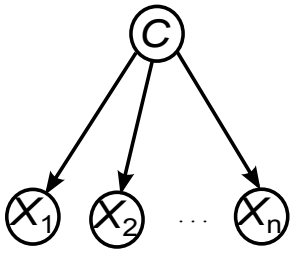

Figure 1: Naïve Bayes classifier

a. Show that the naïve Bayes factorization of the equation below follows the independence assumptions encoded by the graph,

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

b. Show that if all variables are binary-valued, then, $\log \frac{P(C = c|X_1, X_2, \ldots, X_n)}{P(C = \neg c|X_1, X_2, \ldots, X_n)}$ 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 $(X_i = 0 \text{ if } X = \neg x \text{ and } 1 \text{ otherwise})$.

3. 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).
   a. 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).
   b. Explain the method used to construct such a network.

4. Given the Cloudy-Rain-Sprinkler-WetGrass network in Figure 2, write a program that estimates the probability of P(cloudy | sprinkler, wetgrass) using likelihood weighting. Your program does not need to be general, i.e. it can be specific to this query. The program should take as the input the number of samples used to construct the estimate (N). Run the program M=1000 times for N=10,

100, 1000 and 5000 and compute a mean and variance of the posterior probability estimate for each value of N.

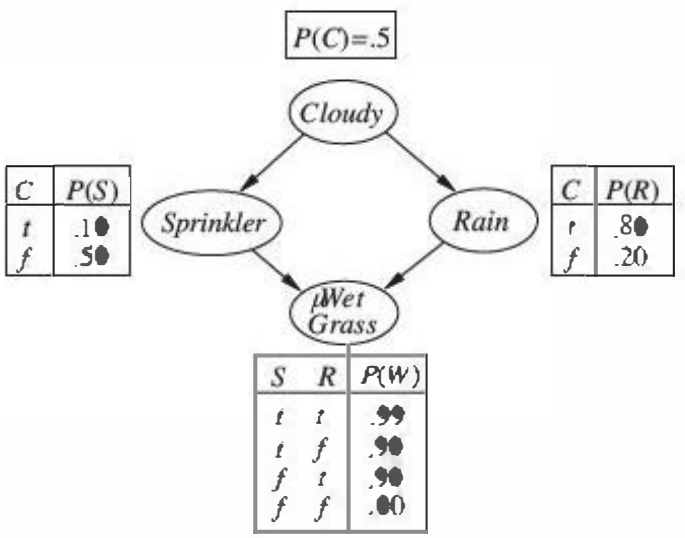

Figure 2. Cloudy-Rain-Sprinkler-WetGrass network.

5. Write a report (similar to a research paper) describing your analysis and findings. It should include the following sections
    a. Aim – briefly state the aim of your study (e.g. inference in Bayes nets.) and write a paragraph why the problem is important.
    b. Methods – briefly describe Bayes nets, and two inference algorithms: variable elimination, and likelihood weighting.
    c. Results and discussion – Present the solutions for questions 1, 2, 3 and 4 above. Present the accuracy results for likelihood weighting compared to exact inference using JavaBayes for each value of N in question 4. Also show the mean and standard deviation for the posterior probability. Discuss.
    d. Conclusions – summarize your main findings and, if possible, suggest future work.
    e. Reflection – what was the most important thing you learned from this assignment? [1-2 paragraphs]
    f. Include detailed instructions on how to run your code.

COMP3308 Assignment 1 – Marking Sheet
Marked out of 15

Student(s):

	Your mark	Comments
1. [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] Question 1 – [1 mark, 0.5 per item] Question 2 – [1 mark, 0.5 per item] Question 3 – [2 marks] Question 4, 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		
2. [4 marks] Code Code runs and computes the correct posterior probability using likelihood weighting.		
3. [1 mark] At the discretion of the marker - for impressing the marker, excelling expectation.		
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):		