FIT5047: Fundamentals of AI – Assignment 2

Due date: Sunday, 1 Nov 2020 (23:55PM Melbourne time).

Evaluation: 100 marks = 20%.

Penalty: 10 marks for every hour of late submission.

Submission: You have to submit your report (in PDF format) via Moodle.

Uploading your report in Moodle: exact instructions sent in week 9 via Moodle.

Report: You should submit a file called "FIT5047_StudentId_2020S2_Ass2.pdf" to the relevant place on Moodle. Please, note that StudentID refers to your Student ID number. Then, as an example, if your StudentID is 12345678, then you should submit the PDF file FIT5047_12345678_2020S2_Ass2.pdf

Note 1: Please follow the University policies and regulations regarding Academic Integrity. Details can be found online from the following hyperlink: click!.

Note 2: You will be required to present your work during your tutorial in week 12 individually. Please be prepared since it is a compulsory component of this assessment. The exact time of your interview will be determined by your lecturer and tutor at a later stage.

Q1: Probability I

{1+1+1+1+3+3+4=14 marks}

Consider a model with the following random variables:

Alarm, Fire, Tampering, Smoke, Evacuation, Report.

These variables can have Boolean values and are used to represent the following situations:

- Alarm: the fire alarm in your flat sounds;
- Fire: there is a fire in your flat;
- Tampering: the smoke detector in your flat was tampered with;
- Smoke: there is smoke in your flat;
- Evacuation: your flat's building was evacuated;
- Report: a local newspaper writes a report on your flat's building evacuation.

We know that a fire increases the probability of the alarm sounding, and that if someone tampered with your smoke detector then your alarm is more likely to sound when there is no fire and less likely to sound when there is a fire. We also know that: the probability that there is a fire is 0.01, whereas the probability that your smoke alarmed is not tampered with is 0.98; the probability that there is smoke when there is a fire is 0.95, while the probability that there is smoke when there is no fire is 0.05; when there is a fire alarm, there is a 0.9 probability that there will be an evacuation, and there is never an evacuation when there is no fire alarm; if there is an evacuation, there is a 0.7 probability that the local newspaper will write a report on it, and if there is no evacuation there is a 0.9 probability that the local newspaper will not report it. We also know the following conditional probabilities about your probabilistic model:

	P(Alarm=T | Tampering , Fire)	P(Alarm=F | Tampering , Fire)
Tampering=T , Fire=T	0.50	0.50
Tampering=T , Fire=F	0.85	0.15
Tampering=F , Fire=T	0.99	0.01
Tampering=F , Fire=F	0.00	1.00

Based on the information above, answer the following questions:

(i) What is the marginal probability that your smoke detector has been tampered with?

(ii) What is the marginal probability that either there is or there is not a news report?

(iii) Let us assume that you have observed that there is smoke in your flat. What is the posterior probability that there will be a news report?

(iv) Let us assume that you have observed that there was no fire, and that there was a news report about your flat. What is the posterior probability that your smoke detector has been tampered with?

(v) Let us assume that you have observed that there is no smoke in your flat. What is the posterior probability that your smoke detector has been tampered with?

(vi) Let us assume that you have observed that there has been a news report about your flat, and there is no smoke in your flat. What is the posterior probability that your smoke detector has been tampered with?

(vii) Let us assume that you have observed that there was no fire, that there was a news report about your flat, and that there is smoke in your flat. What is the posterior probability that your smoke detector has been tampered with?

Q2: Probability II

{3 + 3 + 1.5 +1.5 + 1.5 + 1.5 + 2 = 14 marks}

Creutzfeldt-Jakob is a rare disease: about 1 in 1,000 people get it. Doctors have found that 90% of the people with Creutzfeldt-Jakob disease are ham eaters, while 1% of people who do not have Creutzfeldt-Jakob disease eat ham. These are fictitious figures. Let CJ represent the variable corresponding to having Creutzfeldt-Jakob disease. Let HE represent the variable corresponding to being a ham eater. Based on this information, answer the following questions (your answer should be typed and not handwritten).

a What are the following probabilities? (That is, calculate their value.)

- Pr(+CJ)
- Pr(+HE | +CJ)
- Pr(-HE | -CJ)

b What is the probability that a ham eater will have Creutzfeldt-Jakob disease? Show the formula before you provide any values.

c Should you stop eating ham? Why or why not?
(No marks are given for an absent or incorrect explanation).

Now, consider a probabilistic model for determining a patient's risk of heart disease. The model consists of the following five random variables:

- H: Heart Disease in {present, absent}
- B: Blood Pressure in {high, low}
- E: Exercise in {frequent, infrequent}
- G: Gender in {male, female}
- S: Smoking in {smoker, non-smoker}

The model admits the following factorisation of the joint distribution for the five variables:

$$P(H, B, E, G, S) = P(H \mid B , E , G , S) * P(B \mid E , G) * P(E) * P(G) * P(S)$$

(a) If nothing is known about the patient, is Smoking (S) independent of Exercise (E) and why?

(b) Suppose you are told that the patient has high Blood Pressure (B=high). Is Smoking (S) independent of Exercise (E)?

(c) And if you are told that the patient has Heart Disease (H) are S and E independent?

(d) You know from your medical training that frequent exercise lowers the risk of heart disease while smoking increases it (irrespective of gender). If you are told that a particular patient with the disease was NOT a smoker, what can you infer about their level of exercise?

Q3: Reasoning under Uncertainty (Bayesian reasoning)

{6+6=12 marks}

(a) Suppose we have one red, one blue, and one yellow box. In the red box we have 3 apples and 5 oranges, in the blue box we have 4 apples and 4 orange, and in the yellow box we have 3 apples and 1 orange. Now suppose we randomly selected one of the boxes and picked a fruit. If the fruit that is picked is an apple, what is the probability that it was picked from the yellow box?

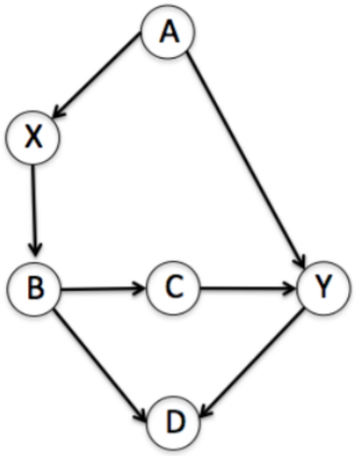
Note that the chances of picking the red, blue, and yellow boxes are 50%, 30%, and 20% respectively and the selection chance for any of the pieces from a box is equal for all the pieces in that box. Please, show the full procedure in your report.

(b) Suppose there are three coins in a bag, a fair coin and two not fair coins: a two-head coin and a two-tail coin. At random, a coin is taken out of the bag and tossed. The result is heads. What is the probability that the coin that was taken out of the bag was the fair coin? What is the probability of having tossed the fair coin if the result had been tails?

Q4: Reasoning under Uncertainty (D-Separation)

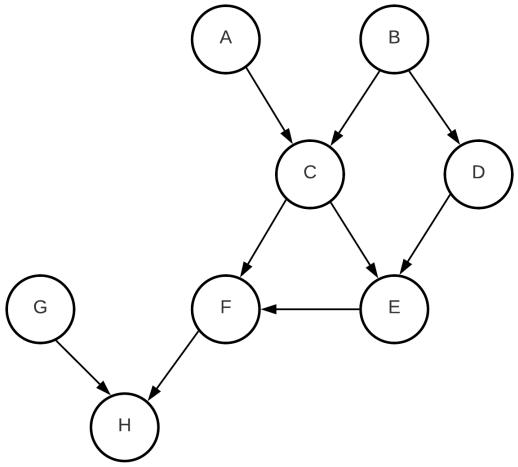
{4+8=12 marks}

(a) Consider the following graph.



Find all the sets of nodes that d-separate X and Y, that is, list *all possible* subsets of A, B, C and D that, if they were known, would cause X to be independent of Y.

(b) Based on the figure below, determine whether the following claims are True or False. (Justify your answer.)



(b.a) $B \perp\!\!\!\perp G \mid A$

(b.b) $C \perp\!\!\!\perp D \mid F$

(b.c) $C \perp\!\!\!\perp D \mid A$

(b.d) $H \perp\!\!\!\perp B \mid C, F$

Q5: Reasoning under Uncertainty (Decision networks)

{20 marks}

A fictional company (let us call it Curie-Franklin-Meitner-Mirza-Noether-Rubin, or CFMMNR) has branched out from its other work on building and launching fictional telescopes (where it bought out the fictional FWO operation) and visiting web sites, and is now apparently involved in mining. We are pleased to say that this company, CFMMNR, is responsible and cares about the environment and the welfare of its workers and employees. CFMMNR has several options for a mining project. It starts out with $13,000.

- There is a probability of $50\% = 0.5 = 1/2$ that a precious material A is underground and that it would be worth $20,000.
- There is a probability of $50\% = 0.5 = 1/2$ that nothing useful is underground and that it would be worth $0.

CFMMNR is facing a number of choices:

- Option 1 is to do nothing and keep the money they started with.
- Option 2 is to spend $10,000 and mine for A.

CFMMNR has the option of spending $1,000 on a consultant. When the consultant (call her Rosalind) says that material A is present, she has a probability of 0.7 of being correct and a probability of 0.3 of being incorrect. When the consultant says that mineral A is not present, then she has a probability of 1/30 of being incorrect and a probability of 29/30 of being correct. After doing the mathematics, it is concluded that there would be a probability of 0.7 that she would say that material A is present. Thus, CFMMNR has a third option.

- Option 3 is to spend $1,000 to hire the consultant to make a decision (to follow through with mining or doing nothing further) based on the consultant's recommendation.

Using Netica, set up a decision network for a choice between Option 1, Option 2 and Option 3.

Note: In your answer, give any prior probabilities, any conditional probability tables (CPTs), any utility functions, explain your working and show your choice out of Option 1, Option 2 and Option 3. State how much money you expect CFMMNR to have.

Q6: Machine Learning (Decision Trees)

{1+2+2+1+6=12 marks}

The "tic-tac-toe.arff" dataset is available in Moodle. Each example in this dataset represents a different game of tic-tac-toe (http://en.wikipedia.org/wiki/Tic-tac-toe) where the player writing crosses ("x") has the first move. Only those games that do not end in a draw are included, with the positive class being the case where the first player wins and the negative class the case where the first player loses. The features encode the status of the game at the end, so each square either contains a cross "x", a nought "o" or a blank "b".

(a) Train the Decision Tree Learner: J48
How many leaf nodes are there in the tree learnt?

(b) In terms of overall accuracy, how accurate is the decision tree model overall?

(c) How often does it predict a win (positive) when the true result was a loss (negative), and vice versa?

(d) What would the learnt decision tree predict for the following game?

x	x	x
o	o	x
	o	

(e) Calculate the information gain provided by the first split in the tree.
(Show your calculations.)

Q7: Machine Learning (Naive Bayes)

{2+4.5+2+3.5+2.5+1.5=16 marks}

Ingrid wants to schedule a day for playing tennis next week. She collects data, shown in the table below, about some past days and their suitability for playing tennis based on the weather and wind condition. Your answer should be typed and not handwritten.

Weather (WE)	Wind (WI)	Tennis (T)
rainy (ra)	weak (wk)	no (N)
cloudy (cl)	strong (st)	no (N)
cloudy (cl)	weak (wk)	yes (Y)
sunny (su)	strong (st)	yes (Y)

(a) Given the formula for Entropy:

$$H(X) = -\sum_{i=1}^{n} \Pr(x_i) \log_2 \Pr(x_i)$$

(i) What is the entropy of the random variable Tennis: H(Tennis)?
(Spell out the formula for this calculation before you provide any values.)

(ii) Calculate the information gain if the attribute Weather is used as the root of a Decision Tree that is used to determine if Ingrid can play tennis.
(Spell out the formula for this calculation before you provide any values.)

(iii) Assuming that Weather and Wind are the only two variables, do you think that splitting on Weather is a good idea? Why or why not?
(No marks will be given for an absent or incorrect explanation.)

(b) Given the Naive Bayes formula:

$$\Pr(C \mid v_{i1} \dots v_{in}) = \alpha \prod_{k=1}^{n} \Pr(v_{ik} \mid C_i = c) \Pr(C_i = c)$$

(i) Given the four instances in the above table, use Maximum Likelihood Estimation to estimate the probabilities of the parameters required for this formula in order to determine if Ingrid can play tennis. Use the given variable names and values to indicate which parameter you are calculating.

(ii) Using these estimates, find the predicted class of a new datapoint with Weather=sunny and Wind=weak. (Spell out the formula for this calculation before giving any values.)

(iii) Are your results reasonable? Why or why not? And if not, how can we improve them from estimation? (No marks will be given for an absent or incorrect explanation.)