Probabilistic Modelling and Reasoning
Assignment 1

Instructor: Prof Chris Williams

Published: Monday, October 10th, 2011
Due: Monday, October 31st, 2011 by 4pm

Remember that plagiarism is a university offence. Please read the policy at `http://www.ed.ac.uk/schools-departments/academic-services/students/postgraduate-taught/discipline/plagiarism` .

Marking Breakdown

A Results/answer correct plus extra achievement at understanding or analysis of results. Clear explanations, evidence of creative or deeper thought will contribute to a higher grade.

B Results/answer correct or nearly correct and well explained.

C Results/answer in right direction but significant errors.

D Some evidence that the student has gained some understanding, but not answered the questions properly.

E Serious error or slack work.

Mechanics

You should submit this assignment **manually** to the ITO office in Appleton Tower by the deadline. Handwritten paper submissions are acceptable if the handwriting is neat and legible.

The policy stated in the School of Informatics MSc Degree Guide is that normally you will not be allowed to submit coursework late. For exceptions to this, e.g. in case of serious medical illness or serious personal problems, see `http://www.inf.ed.ac.uk/student-services/teaching-organisation/for-taught-students/coursework-and-projects/late-coursework-submission` .

Inference in a Belief Network

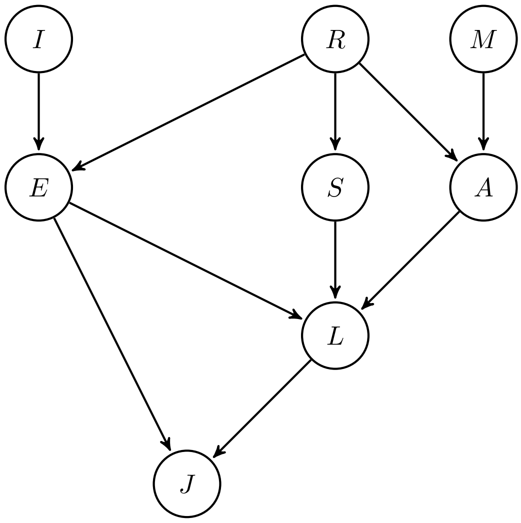

Figure 1: Motor Racing Belief Network.

You are in charge of a new racing team and you wish to join [**J**] the National Open-wheel Car Championship in 2013. The board of the championship usually accepts new cars conditioned on the fact that candidate teams are able to lap [**L**] a nearby circuit in less than 2 minutes.

There is a set of technical regulations [**R**] that consists of design requirements each car needs to meet. It refers to domains such as driver safety [**S**], engine CO_2 emissions [**E**] or aerodynamics [**A**]. The regulations for the 2013 competition will be announced in early 2012, together with the month [**M**] of the test, which may be May or August.

Unsurprisingly, sponsors interfere with your team's decisions. You know that a high level of interference [**I**] commonly translates into you having to design a low emissions car. Nevertheless, this comes at the cost of performance. You also believe that you will have to spend a lot of time improving the aerodynamics of the car. Consequently, you expect to do a better job if the race is scheduled later.

Previous years' experience also tells you that even if you are not fast enough, you might be allowed to join the championship based on an outstandingly low emissions engine.

Your probability model is shown in Figure 1, with the CPTs given below.

A javabayes file of this network (called `racing`) is available from the PMR webpage.

$$
\begin{array}{rl}
P(I = high) = 0.5 & P(R = high) = 0.8 \\
P(M = may) = 0.3 & \\
P(E = low|I = high, R = high) = 0.9 & P(A = good|M = may, R = high) = 0.2 \\
P(E = low|I = low, R = high) = 0.7 & P(A = good|M = may, R = low) = 0.3 \\
P(E = low|I = high, R = low) = 0.5 & P(A = good|M = aug, R = high) = 0.5 \\
P(E = low|I = low, R = low) = 0.1 & P(A = good|M = aug, R = low) = 0.8 \\
P(S = high|R = high) = 0.9 & P(S = high|R = low) = 0.7 \\
P(L = yes|E = high, S = low, A = good) = 0.9 & P(L = yes|E = low, S = low, A = good) = 0.5 \\
P(L = yes|E = high, S = low, A = bad) = 0.7 & P(L = yes|E = low, S = low, A = bad) = 0.3 \\
P(L = yes|E = high, S = high, A = good) = 0.8 & P(L = yes|E = low, S = high, A = good) = 0.4 \\
P(L = yes|E = high, S = high, A = bad) = 0.5 & P(L = yes|E = low, S = high, A = bad) = 0.1 \\
P(J = yes|L = yes, E = low) = 0.99 & P(J = yes|L = yes, E = high) = 0.9 \\
P(J = yes|L = no, E = low) = 0.05 & P(J = yes|L = no, E = high) = 0
\end{array}
$$

(a) Write down the joint distribution defined by the belief network.

(b) Suppose you now have a car with good aerodynamics but high CO_2 emissions. Using simple probability rules, *efficiently* compute the probability that sponsor interference was high. In other words, calculate $P(I = high|A = good, E = high)$. Show your working. One way to tackle this question is to use the elimination algorithm. Do not construct a junction tree at this stage.

(c) You want to compute your chance of lapping the circuit in less than 2 minutes with a car that has good aerodynamics and when there is low sponsor interference. One of the possible junction trees for the network is given by Figure 2. Use the assignment of potentials:

$$
\begin{aligned}
\Psi(I, E, R) &= P(I)P(E|I, R) \\
\Psi(E, R, S, A) &= P(R)P(S|R) \\
\Psi(R, A, M) &= P(M)P(A|M, R) \\
\Psi(E, S, A, L) &= P(L|E, S, A) \\
\Psi(J, E, L) &= P(J|E, L)
\end{aligned}
$$

Formally, compute $P(L = yes|I = low, A = good)$ by hand, this time using the junction tree algorithm. Show your working. HINT: the handout "Worked example of inference in a junction tree" (see week 3) may be helpful here.

(d) Give an elimination order which leads to the junction tree given in Figure 2 and demonstrate that this ordering does indeed produce the clique structure. Note that there may be more than one ordering giving rise to the same cliques.

(e) Suppose that after observing $J = yes$ and calculating $P(I = high|J = yes)$, you had the option of observing one extra variable in order to obtain a more accurate posterior distribution for I. Are there any variables whose observation would lead to no extra information, and would therefore be poor choices? In other words, is I conditionally independent of any other node given J? Show your reasoning.

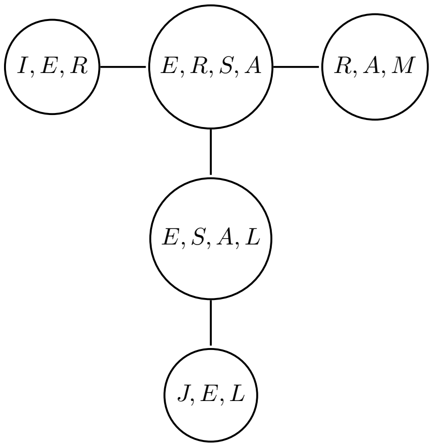

Figure 2: One possible junction tree for the Motor Racing belief network.

(f) You clearly want to know as much as possible about J. A friend can tell you in advance either the month of the test or how big the sponsor interference will be. Which of these two variables would you rather observe?

The general issue here is that of active variable selection. In our case we can compute $P(J|X)$, where X denotes either I or M. Of course we do not know in advance what value X will take on. However, we can compute the conditional entropy

$$H(J|X) = -\sum_{J,X} P(J, X) \log_2 P(J|X),$$

which is a measure (in bits) of the uncertainty in J after observing X. Higher values of entropy denote greater uncertainty.

Compute $H(J|I)$ and $H(J|M)$ working to 4 decimal places, and comment on which variable you would select. Show your working. You may use Java Bayes to help with your computations if that is helpful.

(g) One simple way to approximate inference in a belief network is via drawing samples form the joint distribution, e.g. using ancestral sampling. For example to compute $P(R = low|A = bad, E = low)$ we sample from the network N times and then set:

$$P(R = low|A = bad, E = low) = \frac{count(R = low, A = bad, E = low)}{count(A = bad, E = low)}$$

Say we draw $N = 100$ samples from the joint distribution. Comment on how accurate you expect the resulting estimate of $P(R|A = bad, E = low)$ to be. Hint: you will need to compute some probabilities to answer this question, using e.g. JavaBayes.

(h) We know that the JTA allows up to compute posterior marginals *within a clique*. How could $P(I = high, M = may|L = no)$ be efficiently computed using the JT given in part (c)? Note that I and M do not appear in the same clique. You are **not** required to compute the answer numerically.