Probabilistic Modelling and Reasoning: Assignment
School of Informatics, University of Edinburgh

Instructor: Dr. Amos Storkey
Assignment TA: Andrew Stewart.
Please email amos+pmrass@inf.ed.ac.uk with any problems regarding the assignment.

Published 16 Feb 2015. Due in 4pm Fri 13 March 2015

Remember that plagiarism is a university offence. Please read the policy at http://www.inf.ed.ac.uk/teaching/plagiarism.html

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/F/G serious error or slack work.

0 Mechanics

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

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

The files mentioned below in the assignment are available in a zip file from the course website.
Imagine a clinic set up to help Ebola patients, in an area affected by the virus. The belief network below describes the current dynamic by which people suffer symptoms, might arrive at the clinic and might see a specialist. There is some chance that each person has an Ebola infection (E=yes), shows symptoms, e.g. (Bleeding, B = Yes), (Fever, F = Yes), and Visits the Clinic (V=yes). Bleeding increases the risk of Complications (C=yes). The person may also be referred for Specialist help (S=Yes) if a specialist is available.

\[
\begin{align*}
P(E = yes) &= 0.01 \\
P(F = yes|E = yes) &= 0.6 \\
P(F = yes|E = no) &= 0.1 \\
P(B = yes|E = yes) &= 0.8 \\
P(B = yes|E = no) &= 0.05 \\
P(V = yes|F = yes, B = yes) &= 0.8 \\
P(V = yes|F = yes, B = no) &= 0.5 \\
P(V = yes|F = no, B = yes) &= 0.7 \\
P(V = yes|F = no, B = no) &= 0 \\
P(C = yes|B = yes) &= 0.75 \\
P(C = yes|B = no) &= 0.10 \\
P(S = yes|V = yes) &= 0.6 \\
P(S = yes|V = no) &= 0.0
\end{align*}
\]

(a) Write down the form of the joint distribution of this belief network, in terms of conditional probabilities such as \(P(V|E)\).

(b) You see a patient being assigned to see a Specialist (\(S = Yes\)). Calculate \(P(E = no|S = yes)\), the probability of this patient being clear of Ebola infection. Show clear working of the steps to calculate this. Double check your result using the BRML Toolbox. The BRML can be downloaded from the following link:

http://www0.cs.ucl.ac.uk/staff/d.barber/brml

(c) Convert this into an undirected graphical model (or Markov network), and draw the network. Capture as many conditionally independence relationships as possible. Use of the Barber BRML Toolbox is encouraged for plotting.

(d) Draw a minimal undirected factor graph corresponding to the original directed belief network. Write down a particular choice for the factorisation. Comment on the equivalence or otherwise of the conditional independence relationships encoded by this factor graph, compared with the undirected graph in the previous section.

(e) In a health awareness campaign, people are encouraged to visit the clinic if they have a fever, and this increases the rate of visits by people who have a fever, irrespective of the
state any other variables. What conditional probabilities in the network are changed by this move, and in what direction? Describe any effect this has on the proportion of people with complications that visit the clinic. State clearly what conditional probabilities you are considering to draw the conclusions you do.

(f) Suppose someone does see a specialist but does not have a fever. What conditional independence relationship exists in the distribution that could not be discovered through the graph alone?

2 ICA [35 %]

Independent Component Analysis (ICA) represents data \( v \) using a linear transformation to variables \( h \). ICA tries to make \( h \) maximally-independent. ICA is often used to separate parts of audio recordings. We can write the transformation that ICA makes in the form

\[
v = Ah, \quad h = Bv
\]

(1)

where \( A \) is the square mixing matrix, and \( B = A^{-1} \) is the unmixing matrix.

Two sound files can be found at [http://www.inf.ed.ac.uk/teaching/courses/pmr/pmrassignment.html](http://www.inf.ed.ac.uk/teaching/courses/pmr/pmrassignment.html). They can be loaded using the Matlab function:

```matlab
[sound1 sr] = wavread('sound_mic1.wav');
sound2 sr] = wavread('sound_mic2.wav');
```

(a) Load this data into Matlab or Octave and inspect the data. Plot the waveforms, compare scatter and boxplot, use sound(X) to listen to the wav data, and describe the observations you make regarding the sound files. Make any comment about the form the sources may take.

(b) Using the BRML Toolbox, ICA can be run as follows:

```matlab
import brml.*
Xs = [sound1 sound2];
Xs = Xs';

opts.maxits=500; opts.tol=10e-6; opts.whiten=0; opts.plotprogress=1;
[A,B,Xout] = ica(Xs,opts);

sound(Xout(1,:))
```

Run ICA on the sound data using this code, and state the learnt mixing matrix \( A \). Examine the learned independent components. Compare the ICA latent values (each of the 'rows' of the matrix outputted) to the original data. Explain the differences between the original sound vectors and the ICA latent values.
(c) If ICA were to be used to cleanly separate three overlapping voices in a recording, what else would be needed?

(d) The log likelihood of the data under the model can be calculated as:

$$\log p(v) = N \log |B| + 2N (\log \beta - \log 2\Gamma(1/\beta)) - \sum_{n=1}^{N} \sum_{i=1}^{2} |B_i \cdot v_n|^{\beta}$$

where $B_i$ is the $i$th row of $B$, and $\Gamma(.)$ is the $\Gamma$ function.

Taking $\beta = 1$, write a Matlab function to calculate the log likelihood of the given wave files under your current estimated ICA model.

3 Car Driver Identification [25 %]

Car insurers currently charge based on the car and driver. Using trip telemetry data, there may be an opportunity to charge more appropriately, given a driver’s behaviour.

Look at the details of the driver telematics Kaggle competition (see [http://www.kaggle.com/c/axa-driver-telematics-analysis](http://www.kaggle.com/c/axa-driver-telematics-analysis)). These webpages describe a problem to identify driver characteristics as an unsupervised classification problem. The task is to develop an algorithmic signature of driving type, such that a telemetric fingerprint can identify when a trip was driven by a given driver.

Carefully read all the public information about the challenge. There is no need to look at the data[1] or accept the conditions of the challenge (though you may choose to do so). Please write a maximum of one page to describe how you might choose to tackle the problem described here, and justify the decisions and recommendations you make. You could, as part of your answer, explain why you choose not to do things in other ways. Your answer should be informed by what you have learnt from PMR. You may discuss this (with acknowledgement), but the write-up should be your own work and understanding.

---

[1] In any real data driven modelling scenario, you would be mad not to look at the data, but we do not want to build an assignment on the basis of an external company’s terms and conditions that you might not wish to accept.