Homework 5: Bayesian Graphical Models

Instructions:

1. You may discuss this assignment with other students in the class, but you must submit your own answers to the questions below.
2. Include an honor pledge with your submission.
3. Submit online.
4. This homework is worth 50 points with the possibility of 15 points extra credit, and the point totals for each question are shown in parentheses.

Assignment:

1. (10) Suppose we have a model for lung cancer based on the following joint probability: $p(S, R, C) = p(C|S, R)p(S)p(R)$, where S = smoking; R = Radon exposure; and C = lung cancer.
   (a) Is $R \perp\!\!\!\perp S \,|\, C$? Show why or why not.
   (b) How do we expand this model to account for miners who have both a higher likelihood of smoking and a higher likelihood of radon exposure?

2. (20) A graphical model for a pulmonary clinic is shown in Figure 1 [1], where the nodes are associated with the following variables:

   X_1 Visited Southeast Asia
   X_2 Smoker
   X_3 Turberculosis
   X_4 Lung cancer
   X_5 Bronchitis
   X_6 Either tuberculosis or lung cancer
   X_7 Positive X-ray image
   X_8 Dyspnea (Shortness of breath)

   For clinic patients, we have the base rate probabilities shown below:

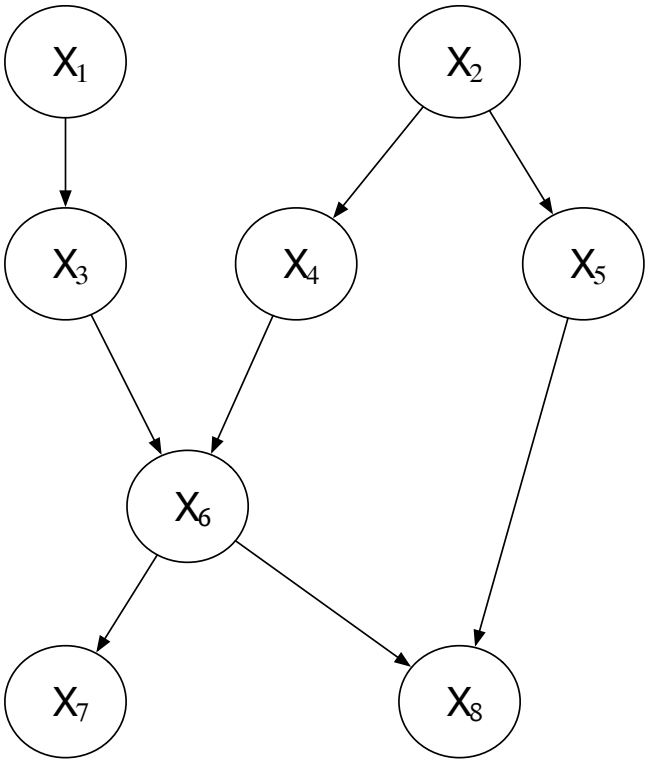

Figure 1: Pulmonary Clinic Graphical Model

$$
\begin{aligned}
p(X_1 = T) &= 0.01 \\
p(X_2 = T) &= 0.3 \\
p(X_3 = T|X_1 = T) &= 0.05 \\
p(X_3 = T|X_1 = F) &= 0.01 \\
p(X_4 = T|X_2 = T) &= 0.25 \\
p(X_4 = T|X_2 = F) &= 0.01 \\
p(X_5 = T|X_2 = T) &= 0.6 \\
p(X_5 = T|X_2 = F) &= 0.2 \\
p(X_7 = T|X_6 = T) &= 0.98 \\
p(X_7 = T|X_6 = F) &= 0.05 \\
p(X_8 = T|X_5 = T, X_6 = T) &= 0.9 \\
p(X_8 = T|X_5 = T, X_6 = F) &= 0.7 \\
p(X_8 = T|X_5 = F, X_6 = T) &= 0.8 \\
p(X_8 = T|X_5 = F, X_6 = F) &= 0.1
\end{aligned}
$$

Also $p(X_6 = T|X_3, X_4) = 0$ if both $X_3 = X_4 = F$ and $= 1$ otherwise.

(a) Find $p(X_8), p(X_8 = T|X_2 = T)$ and $p(X_8 = T|X_2 = F)$.

(b) Compare a noncausal interpretation of $p(X_8 = T|X_5 = F)$ with a causal interpretation, where the physician asks the question, "If I could treat the patient's bronchitis, how would that change their chances of having shortness of breath?"

3. (20) You are tracking the performance of a set of companies with the idea that you might possibly buy stock in them. You decide to automate this process using HMM and you implement your first version for one company. This company has three states that are hidden from investors: (1) in trouble; (2) static; and (3) major growth potential. You have estimated the transition probabilities between states as follows:

$$
\begin{bmatrix}
.6 & .3 & .1 \\
.4 & .4 & .2 \\
.1 & .4 & .5
\end{bmatrix}
$$

You have a text analysis system to process the quarterly reports and assess their sentiment into one of three categories: (1) Fine, (2) Good, and (3) Very good. Your estimates for the probabilities of these sentiments given the state of the company are shown in the following matrix (the sentiments are in the rows and the states of the company are in the columns):

$$\begin{bmatrix} .45 & .4 & .15 \\ .3 & .4 & .3 \\ .2 & .5 & .3 \end{bmatrix}$$

You have three quarterly reports with the assessments: Fine, Good, Very Good.

(a) In order to decide whether to invest, find the most likely current state.

(b) To get a sense of management performance, find the most likely path to the current state.

4. (15 Extra Credit) Read the paper by Blei, et al., [2] describing Latent Dirichlet Allocation, then answer the following questions:

   (a) (10) How do the authors use variational inference to approximate the desired posterior distribution?

   (b) (5) How do the authors use the EM algorithm for parameter estimation?

References

[1] Lauritzen, S.L. and D.L. Spiegelhalter. "Local computations with probabilities on graphical structures and their applications to expert systems." *Journal of the Royal Statistical Society B*, 50(2) 157–224, 1988.

[2] Blei, David M., Andrew Y. Ng, and Michael I. Jordan. "Latent dirichlet allocation." *Journal of machine learning research* Jan (2003): 993–1022.