Exercise Sheet #9

Due: 2019-12-19 16:00

Fred Hamprecht
Manuel Haußmann

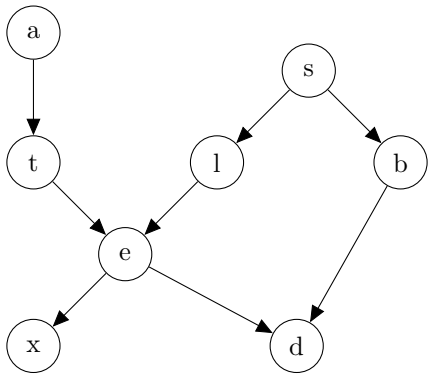
General Regulations.

- You should hand in your solutions in groups of at least two people (recommended are three).
- The theoretical exercises can be either handwritten notes (scanned), or typeset using LaTeX.
- Practical exercises should be implemented in python and submitted as jupyter notebooks (`.ipynb`). Always provide the (commented) code as well as the output, and don't forget to explain/interpret the latter.
- Submit all your files in a single `.zip` archive to `mlhd1920@gmail.com` using the following standardized format: The subject line should consist of the full names of all team members as well as the exercise, and the title of the zip archive the last names. I.e. assuming your group consists of Ada Lovelace, Geoffrey Hinton and Michael Jordan, this means

  Subject: [EX09] Michael Jordan, Geoffrey Hinton, Ada Lovelace
  Zip Archive: `ex09-jordan-hinton-lovelace.zip`

1 A Medical Probabilistic Graphical Model (20 pt +5 pt)

The setup of this exercise is the diagnosis of lung diseases in a clinical environment. Consider the (hypothetical) probabilistic graphical model as specified in Figure 1.



Variable	Meaning
a	visit to Asia?
s	smoking?
t	tuberculosis?
l	lung cancer?
b	bronchitis?
e	tuberculosis or lung cancer
x	positive X-ray?
d	dyspnea (shortness of breath)

Figure 1: A probabilistic graphical model for the diagnosis of lung diseases.

Assume that we have the following conditional probabilities given by prior knowledge

$$
\begin{array}{llll}
p(a) & = 0.01 & p(s) & = 0.5 \\
p(t|a) & = 0.05 & p(t|\bar{a}) & = 0.01 \\
p(l|s) & = 0.1 & p(l|\bar{s}) & = 0.01 \\
p(b|s) & = 0.6 & p(b|\bar{s}) & = 0.3 \\
p(e|t,l) & = 1 & p(e|t,\bar{l}) & = 1 \\
p(e|\bar{t},l) & = 1 & p(e|\bar{t},\bar{l}) & = 0 \\
p(x|e) & = 0.98 & p(x|\bar{e}) & = 0.05 \\
p(d|e,b) & = 0.9 & p(d|e,\bar{b}) & = 0.7 \\
p(d|\bar{e},b) & = 0.8 & p(d|\bar{e},\bar{b}) & = 0.1
\end{array}
$$

Here we used the shorthand of $p(v)$ to refer to the probability that a variable is true and $p(\bar{v})$ to mean the probability that variable v is wrong (e.g. $p(a)$ gives the probability of a visit to Asia, while $p(\bar{a})$ gives the probability of not having visited Asia.)

i) Having specified a graphical model, we can directly read from the graph certain dependency/independency structures. State whether the following claims are true and false, and explain why not if they aren't.

- tuberculosis $\perp$ smoking | shortness of breath
- lung cancer $\perp$ bronchitis | smoking
- visit to Asia $\perp$ smoking | lung cancer
- visit to Asia $\perp$ smoking | lung cancer, shortness of breath

ii) Given the probabilities we have multiple approaches to compute the different marginal/conditional distributions we are interested in.

- The first is via variable elimination as discussed in the lecture. Use this approach to compute the following distributions analytically:

  What is the probability of a patient showing up with dyspnea ($p(d)$)? How do these probabilities change if we know that he/she is a smoker ($p(d|s)$) or does not smoke ($p(d|\bar{s})$)?
- Another approach is to use simulations. A sample from this graphical model consists of eight binary variables. Assuming that we have a set of N samples, computing marginals then simply consists of simply counting how many samples have the desired values. Similarly, computing conditional probabilities consists of counting the number of samples given some constraints. Implement a function that gives you a sample from this joint distribution. Compute $N = 100000$ samples and compare your numerical approximations to your analytical solutions for the three probabilities above.

iii) Using your simulator give numerical estimates for the probabilities in the following scenario:

According to our model, what is the marginal probability of a patient having lung cancer ($p(l)$)? A patient arrives complaining about shortness of breath. How does that change your estimate for that patient $p(l|d)$ having lung cancer? You decide to take some x-rays, which come back positive. What is the new $p(l|x, d)$? While waiting for the results you discovered that your patient just had a nice vacation in Asia and is a chain smoker. How do each of these new pieces of information to change your lung cancer estimate (i.e. what are your results for $p(l|x, d, a), p(l|x, d, s), p(l|x, d, s, a)$)? Also, discuss how many samples you get to estimate each of the probabilities. As you increase the number of conditions, do you observe any pattern and if so is it problematic or irrelevant?

iv) *(technical +5pt)* Compute the analytical distributions using variable elimination for the distributions discussed in *iii)*. How do they compare?

2 HMM: Robot localization in a 1D world (10 pt)

The task of this exercise is robot localization following a similar setup to the one presented in the lecture. Our robot lives in a circular world consisting of 10 discrete locations (i.e. if the robot is in the tenth location and moves to the right it comes out at the first). The robot has access to a map of this world which shows that all of the locations consist of squares except for three, which are circular. The map is given as

$$[X, X, X, O, X, O, X, O, X, X]. \tag{1}$$

At each time step the robot tries to move either to the right or to the left with equal probability. This movement succeeds in 90% of the cases, while 5% of the time it does not move at all and the last 5% it

moves a step in the opposite direction. After moving, its sensors try to read whether the current location contains a square or a circle. This sensor reading is correct 95% of the time.

Implement the algorithm to update the beliefs of the robot as to where it is given a sequence of actions $\mathbf{a}_T$ and sensor readings $\mathbf{r}_T$ and visualize how the beliefs $\alpha(x_t)$ as to where the robot is change from an initial belief that assigns equal probability ($\alpha(x_t = \text{location i}) = 1/10$) as we get new information with each time step. See `hmm-data.npz` for the data.