Department of Mathematics and Computer Science
University of Southern Denmark, Odense

June 17, 2011
Marco Chiarandini

DM825 - Introduction to Machine Learning

Sheet 8, Spring 2011 [pdf format]

Exercise 1 Bayesian Networks – Inference.
Figure shows a graphical model with conditional probabilities tables about whether or not you will panic at an exam based on whether or not the course was boring ("B"), which was the key factor you used to decide whether or not to attend lectures ("A") and revise doing the exercises after each lecture ("R").
You should use the model to make exact *inference* and answer the following queries:

- what is the probability that you will panic or not before the exam given that you attended the lectures and revised after each lecture?

  From the CPT, 0.

- what is the probability that you will panic or not before the exam?

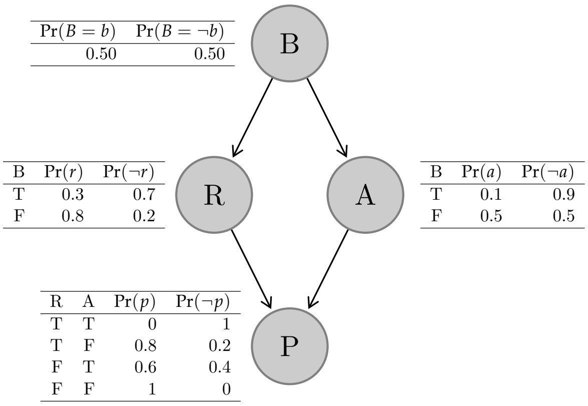
$$\begin{aligned}
\Pr(p) &= \sum_{b,r,a} \Pr(b,r,a,p) \\
&= \sum_{b,r,a} \Pr(b)\Pr(r|b)\Pr(a|b)\Pr(p|r,a) \\
&= \sum_{b} \Pr(b) \sum_{r,a} \Pr(r|b)\Pr(a|b)\Pr(p|r,a) \\
&= 0.5 \cdot (0.3 \cdot 0.1 \cdot 0 + 0.3 \cdot 0.9 \cdot 0.8 + 0.7 \cdot 0.1 \cdot 0.6 + 0.7 \cdot 0.9 \cdot 1) \\
&\quad +0.5 \cdot (0.8 \cdot 0.5 \cdot 0 + \cdot 0.8 \cdot 0.5 \cdot 0.8 + 0.2 \cdot 0.5 \cdot 0.6 + 0.2 \cdot 0.5 \cdot 1) \\
&= 0.684
\end{aligned}$$

- Your teacher saw you panicking at the exam and he wants to work out from the model the reason for that. Was it because you did not come to the lecture or because you did not revise?

$$\begin{aligned}
\Pr(r|p) &= \frac{\Pr(p|r)\Pr(r)}{\Pr(p)} \\
&= \frac{\sum_{b,a} \Pr(b,a,r,p)}{\Pr(p)} \\
&= \frac{0.5 \cdot (0.3 \cdot 0.1 \cdot 0 + 0.3 \cdot 0.9 \cdot 0.8) + 0.5 \cdot (0.8 \cdot 0.5 \cdot 0 + 0.8 \cdot 0.5 \cdot 0 + 0.8 \cdot 0.5 \cdot 0.8)}{\Pr(p)} \\
&= \frac{0.268}{0.684} = 0.3918
\end{aligned}$$

$$\begin{aligned}\Pr(r|a) &= \frac{\Pr(p|a)\Pr(a)}{\Pr(p)} \\ &= \frac{0.144}{0.684} = 0.2105\end{aligned}$$

Repeat the inference in the last two queries by means of stochastic inference implementing in R the Prior-Sample, Rejection-Sampling, Likelihood-weighting and Markov Chain Monte Carlo procedures.



$\Pr(B=b)$	$\Pr(B=\neg b)$
0.50	0.50

B	$\Pr(r)$	$\Pr(\neg r)$
T	0.3	0.7
F	0.8	0.2

B	$\Pr(a)$	$\Pr(\neg a)$
T	0.1	0.9
F	0.5	0.5

R	A	$\Pr(p)$	$\Pr(\neg p)$
T	T	0	1
T	F	0.8	0.2
F	T	0.6	0.4
F	F	1	0

Figure 1: The graphical model of exercise 1. Lower-case letter indicate the outcome that the upper-case letter can take.

Exercise 2 Directed Graphical Models Consider the graph in Figure left.

- Write down the standard factorization for the given graph.

  Solutions The standard factorization for any directed graphical model can be written as $p(x) = \Pi_{v \in V} p(x_v | x_{pa(v)})$, where $x_{pa(v)}$ are the nodes parent of x_v. Here, this yields

  $$p(x) = p(x_1)p(x_2)p(x_3|x_{10})p(x_4|x_2,x_6,x_7)p(x_5|x_9)p(x_6|x_1,x_2)p(x_7)p(x_8)p(x_9|x_3,x_7,x_8)p(x_{10}|x_3).$$

- For what pairs (i, j) does the statement X_i is independent of X_j hold? (Don't assume any conditioning in this part.)

  The goal is to find all pairs (i, j) such that Xi and Xj are independent. We can achieve this by computing from each node its reachability, that is, the nodes that

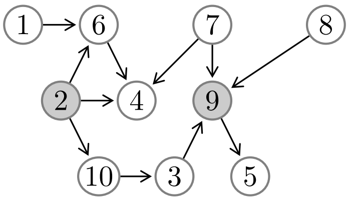


Figure 2: A directed graph.

are reachable by a path that does not have head-to-head subcomponents. From node 1 we can get to nodes 6 and 4. From node 2 we can reach nodes 6, 4, 10, 3, 9, and 5. From nodes 3 and 10 we can reach the sames nodes as node 2. From node 4 we can reach every node but node 8. From node 5 we can reach every node but node 1. From node 6 we can reach any node but nodes 7 and 8. From node 7 we can reach node 9, 4, and 5. Node 8 can only reach nodes 9 and 5. Node 9 can't reach node 1. Finally, node 10 can't reach nodes 1, 7, and 8. Thus (1, 2), (1, 3), (1, 5), (1, 7), (1, 8), (1, 9), (1, 10), (2, 7), (2, 8), (3, 7), (3, 8), (4, 8), (6, 7), (6, 8), (7, 8), (7, 10), and (8, 10) are all independent pairs. In all there are 17 distinct pairs.

- Suppose that we condition on $\{X_2, X_9\}$, shown shaded in the graph. What is the largest set A for which the statement X_1 is conditionally independent of X_A given $\{X_2, X_9\}$ holds?

  Solution We say that $\mathbf{X} \perp \mathbf{Y} \mid \mathbf{Z}$ if $\mathbf{X}$ and $\mathbf{Y}$ are d-separated given $\mathbf{Z}$ in the digraph, that is, if there is no *active* path between any node $X \in \mathbf{X}$ to $Y \in \mathbf{Y}$ given $Z \in \mathbf{Z}$. In class we defined the four conditions for a path to be *active*.

  Checking d-separation implies checking all paths from a vertex to another. This maybe exponential. The following is a linear time algorithm for d-separation. We begin by traversing the graph bottom up, from the leaves to the roots, marking all nodes that are in Z or that have descendants in Z. Intuitively, these nodes will serve to identify a head-to-head structure, ie., $X \to Z \leftarrow Y$. In the second phase, we traverse breadth-first from X to Y, stopping the traversal along a path when we get to a blocked node. A node is blocked if: (a) it is in the "middle" node of a structure $X \to Z \leftarrow Y$ and unmarked in phase I, or (b) is not such a node and is in Z. If our breadth-first search gets from X to Y, then there is a path between them through Z.

  Conditioned on $\{X_2, X_9\}$ there is no active path to nodes 3, 10, 7, 8 and 5. Hence, $A = \{3, 5, 7, 8, 10\}$. Note that nodes 2 and 9 are not elements of the set A because we are conditioning on them.

- What is the largest set B for which X_8 is conditionally independent of X_B given $\{X_2, X_9\}$ holds?

  Solutions Conditioned on $\{X_2, X_9\}$ starting at node 8 we cannot reach with an active path nodes 1, 5, and 6. Therefore, $B = \{1, 5, 6\}$.

- Suppose that I wanted to draw a sample from the marginal distribution $p(x_5) = \Pr[X_5 = x_5]$. (Don't assume that X_2 and X_9 are observed.) Describe an efficient algorithm to do so without actually computing the marginal.

  Solution

  We wish to generate a sample of x_5 from the marginal distribution $p(x_5)$. We can achieve this by sampling from the join distribution $p(\vec{x})$ and then marginalizing. For example, given a joint distribution $\Pr[x1, x2]$, one can generate a sample from the marginal distribution of x_1 by sampling from the joint distribution and discarding x_2. To see this, let A be the event that the sample $\bar{x}_1$ lies in some set F. Therefore, $\Pr[A] = \Pr[x_1 \in F \cup x_2 \in \mathbb{R}] = \Pr[x_1 \in F]$. Thus, $\bar{x}_1$ and x_1 have the same distribution.

  Hence we can avoid unnecessary computations applying the following algorithm:

  - calculate the Topological order of the graph

– sample using the factorization and the topological sorting until you sample x_5.

Hence, we can first generate a sample of x_2, x_7, and x_8. Then, using factorization, we can generate a sample of from the distribution $p(x_{10}|x_2)$, followed by a sample from the distribution $p(x_3|x_{10})$ by using the sample obtained of x_{10}. Next, we can generate a sample of x_9 from the distribution $p(x_9|x_7, x_8, x_2)$. Finally, we can obtain a sample for x_5 by sampling from the distribution $p(x_5|x_9)$. We can immediately see that generating a sample of x_5 did not require actually sampling from x_1, x_6, or x_4 because conditioned on nodes 2, 7, and 8, x_5 is independent of nodes 1, 6, and 4. Thus, what we've done is generating a sample from the joint distribution $p(x_2, x_3, x_{10}, x_7, x_8, x_9, x_5) = p(x_2)p(x_7)p(x_8)p(x_{10}|x_2)p(x_3|x_{10})p(x_9|x_3, x_8, x_7)p(x_5|x_9)$.