**T.61.5140 Machine Learning: Advanced Probablistic Methods**
Hollmén, Raiko (Spring 2008)
Problem session, 8th of February, 2008
http://www.cis.hut.fi/Opinnot/T-61.5140/

1. Consider a model for five binary variables $x_i$. (a) What is the number of parameters needed to represent the distribution $P(x_1, x_2, x_3, x_4, x_5)$ if no assumptions are made? (b) How about if the model in Figure 1 is assumed? (c) And how about, if additionally, the Markov assumption $(P(x_{i+1} \mid x_i) = P(x_i \mid x_{i-1}), i = 2, 3, 4)$ is made? Note: Sometimes only free parameters are counted, that is, for every sum of probabilities that equals 1, one parameter is determined by the others and is thus not counted. We can ignore this here and count all the parameters.

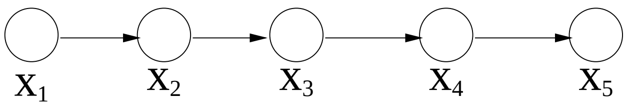


Figure 1: Problems 1 and 2.

Solution:

(a) When no assumptions are made about $P(x_1, x_2, x_3, x_4, x_5)$, it can be represented as a 5-dimensional table where each variable $x_i$ can get two values, 0 and 1. The probability of each combination of the values is given as a number in the table and the sum of these numbers is 1. The size of the table is $2^5 = 32$.

(b) Representation of $P(x_1)$ takes 2 parameters. Each additional representation $P(x_{i+1} \mid x_i)$ takes $2^2 = 4$ parameters. The answer is thus 18.

(c) Representation of $P(x_1)$ takes 2 parameters as before, but in this case, the rest are handled with a shared table for each $P(x_{i+1} \mid x_i)$. The number of parameters is 6.

2. For the Bayesian network in Figure 1, solve $P(x_3 \mid x_1, x_5)$.

Solution 1:

First we use the Bayes theorem to get $x_5$ and $x_3$ in the correct order.

$$P(x_3 \mid x_1, x_5) = \frac{P(x_5 \mid x_1, x_3)P(x_3 \mid x_1)}{P(x_5 \mid x_1)} \tag{1}$$

Then we forget about the normalization constant $P(x_5 \mid x_1)$ for a while and concentrate on the numerator, let us call it $A(x_3)$:

$$A(x_3) = P(x_5 \mid x_1, x_3)P(x_3 \mid x_1) \tag{2}$$

Some of the variables conditioned on can be dropped out, for instance $P(x_5 \mid x_1, x_3) = P(x_5 \mid x_3)$ can be shown using d-separation ($x_5$ and $x_1$ are independent given $x_3$). Then we need marginalization principle to include $x_2$ and $x_4$:

$$\begin{align}
A(x_3) &= P(x_5 \mid x_3)P(x_3 \mid x_1) \tag{3}\\
&= \left[\sum_{x_4} P(x_5 \mid x_4, x_3)P(x_4 \mid x_3)\right]\left[\sum_{x_2} P(x_3 \mid x_2, x_1)P(x_2 \mid x_1)\right] \tag{4}\\
&= \left[\sum_{x_4} P(x_5 \mid x_4)P(x_4 \mid x_3)\right]\left[\sum_{x_2} P(x_3 \mid x_2)P(x_2 \mid x_1)\right] \tag{5}
\end{align}$$

Now we have written the equation with probabilities of type $P(x_{i+1} \mid x_i)$ of which the model consists. Finally, we reinclude the normalization constant:

$$P(x_3 \mid x_1, x_5) = \frac{A(x_3)}{A(0) + A(1)} \tag{6}$$

Note how $A(x_3)$ is a product of two messages, one coming in the direction of the edges and the other coming against them.
Solution 2:

This solution is simpler but it does not highlight the message passing and its implementation would be heavier (exponential in the number of unobserved variables).

$$\begin{align}
P(x_3 \mid x_1, x_5) &= \frac{P(x_1, x_3, x_5)}{P(x_1, x_5)} \tag{7}\\
&= \frac{\sum_{x_2,x_4} P(x_1, x_2, x_3, x_4, x_5)}{\sum_{x_2,x_3,x_4} P(x_1, x_2, x_3, x_4, x_5)} \tag{8}\\
&= \frac{\sum_{x_2,x_4} P(x_5 \mid x_4)P(x_4 \mid x_3)P(x_3 \mid x_2)P(x_2 \mid x_1)P(x_1)}{\sum_{x_2,x_3,x_4} P(x_5 \mid x_4)P(x_4 \mid x_3)P(x_3 \mid x_2)P(x_2 \mid x_1)P(x_1)} \tag{9}\\
&= \frac{\sum_{x_2,x_4} P(x_5 \mid x_4)P(x_4 \mid x_3)P(x_3 \mid x_2)P(x_2 \mid x_1)}{\sum_{x_2,x_3,x_4} P(x_5 \mid x_4)P(x_4 \mid x_3)P(x_3 \mid x_2)P(x_2 \mid x_1)} \tag{10}
\end{align}$$

3. Show that the property of there being no directed cycles in a directed graph follows from the statement that there exists an ordered numbering of the nodes such that for each node there are no links going to a lower-numbered node.

Solution:

Consider a directed graph in which the nodes of the graph are numbered such that are no edges going from a node to a lower numbered node. If there exists a directed cycle in the graph then the subset of nodes belonging to this directed cycle must also satisfy the same numbering property. If we traverse the cycle in the direction of the edges the node numbers cannot be monotonically increasing since we must end up back at the starting node. It follows that the cycle cannot be a directed cycle.

4. Using the d-separation criterion, show that the conditional distribution for a node $x_i$ in a directed graph, conditioned on all of the nodes in the Markov blanket, is independent of the remaining variables in the graph.

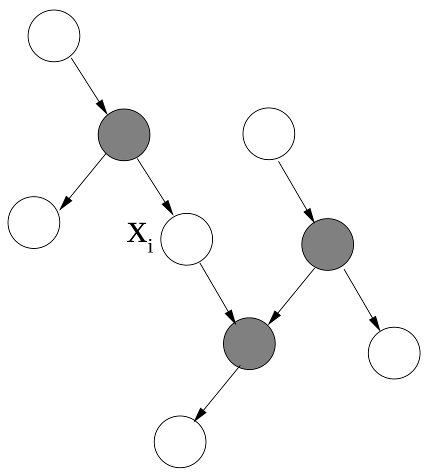


Figure 2: Problem 4. The Markov blanket of a variable includes its parents, children, and co-parents. To solve problem 4, one needs to show that all paths from $x_i$ to nodes outside of its Markov blanket are blocked.

Solution:

Consider Figure 2. In order to apply the d-separation criterion we need to consider all possible paths from the central node $x_i$ to all possible nodes external to the Markov blanket. There are three possible categories of such paths. First, consider paths via the parent nodes. Since the link from the parent node to the node $x_i$ has its tail connected to the parent node, it follows that for any such path the parent node must be either tail-to-tail or head-to-tail with respect to the path. Thus the observation of the parent node will block any such path. Second consider paths via one of the child nodes of node $x_i$ which do not pass directly through any of the co-parents. By definition such paths must pass to a child of the child node and hence will be head-to-tail with respect to the child node and so will be blocked. The third and final category of path passes via a child node of $x_i$ and then a co-parent node. This path will be head-to-head with respect to the observed child node and hence will not be blocked by the observed child node. However, this path will either tail-totail or head-to-tail with respect to the co-parent node and hence observation of the co-parent will block this path. We therefore see that all possible paths leaving node $x_i$ will be blocked and so the distribution of $x_i$, conditioned on the variables in the Markov blanket, will be independent of all of the remaining variables in the graph.