cis32-ai — lecture # 24 — wed-3-may-2006

today's topics:

- probabilistic models

terminology

- First, a little graph theory terminology.
- A *directed graph* is a set of variables and a set of directed arcs between them.
- A directed graph is *acyclic* if it is not possible to start at a node, follow the arcs in the direction they point, and end up back at the starting node.
- We will only talk about directed acyclic graphs.
- We won't worry about whether it should be "acyclic directed graph".

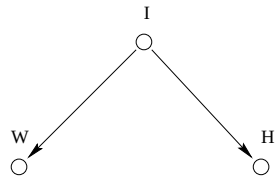
- A is the *parent* of B if there is a directed arc from A to B.
- B is the *child* of A if A is the parent of B.
- Any node with no parents is known as a *root* of the graph.
- Any node with no children is known as a *leaf* of the graph.
- The parents of node A and the parents of those parents, and the parents of those parents, and so on, are the *ancestors* of A.
- If A is an ancestor of B, then B is a *descendent* of A.

- We say that there is a *link* between two nodes A and B if A is a parent of B or B is a parent of A.
- We say that two nodes A and C in a graph have a *path* between them if it is possible to start at A and follow a series of links through the graph to reach C.
- Note that in defining a path we ignore the direction of the arrows.
- In other words we are considering the underlying undirected graph.

- A graph is said to be *singly-connected* if it includes no pairs of nodes with more than one path between them.
- A graph which is not singly-connected is *multiply-connected*.
- A singly-connected graph with one root is called a *tree*.
- A singly-connected graph with several roots is called a *polytree*.
- A polytree is sometimes called a *forest*.

Causal reasoning

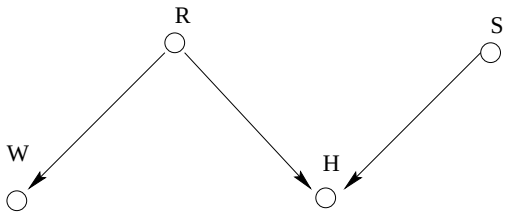
- We start by considering some examples of this notion of "has no effect".
- Consider representing the information that I (icy roads) makes it more likely that Watson will crash W and that Holmes will crash H.
- This information can be captured in the graph:



- Now, if we learn that Watson has crashed, this makes us believe it is more likely that the roads are icy.
- Thus learning something about W allows us to conclude something about I.
- This in turn makes us believe that it is more likely that Holmes has crashed.
- Thus the new information about I can be propagated to learn something about H.
- However, if we learn that the roads are not icy, then the news about Watson has no effect on our beliefs about Holmes.
- Thus knowing the value of I for certain prevents information about W affecting H.

- In other words our notion of "has an effect" depends upon what we know, and changes as we learn new things.
- When nothing is known about I, H is *dependent* on W.
- Once we know the value of I with certainty, then H is *independent* of W.
- This phenomenon is known as *conditional independence*.
- As we shall see this notion of independence is exactly statistical independence, which is why the factorisation given above works.
- Before we can formalise this idea, we need some further examples of exactly what we are formalising.

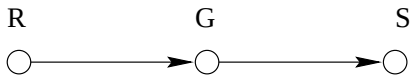
- Holmes knows that there are two causes of his grass being wet, H, either the sprinkler has been on, S, or it has been raining, R.
- If it has been raining, then Watson's grass will be wet as well, W.
- This example may be captured by the graph:



- When Holmes sees that his grass is wet, he is more inclined to believe both that it was raining and that the sprinkler was on.

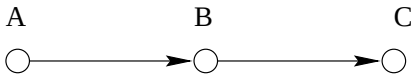
- Thus observing H makes both S and R more likely.
- However, when Holmes sees that Watson's grass is also wet, his belief that it was raining increases.
- Thus observing W provides more evidence for R.
- Because rain seems to be a very likely explanation for the wet grass, Holmes now has little belief that the sprinkler was on.
- Thus the increase in belief in R leads to a decrease in S.
- We say that the increasing probability of rain *explains away* the sprinkler.
- Knowledge about W makes S dependent on R.

- Our third example is a variation on a theme of the second.
- Consider that the grass being wet G is a cause of Holmes' shoes being wet S as he walks to his car.
- A subset of the relevant information is thus:



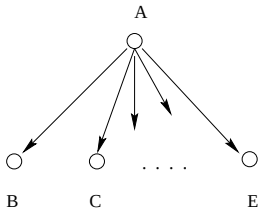
- In a similar way to previous examples, if Holmes finds his shoes are wet, then his belief in the grass being wet increases as does his belief that it was raining.
- Thus knowing S increases belief in G and R.

- However if Holmes is told by his wife that the grass is wet before he leaves the house, then realising his shoes are wet will not change his belief about the likelihood of rain.
- Thus once we know the value of G, R and S are independent.
- This gives us some standard patterns of connection and their associated dependencies.
- In a serial connection such as:



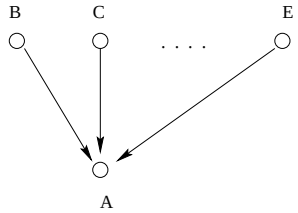
A and C are independent once B is known, and we say that A and C are *d-separated* given B.

- In the following generalisation of the icy roads example:



  $B, C, \ldots E$ are d-separated given A.
- We refer to this connection as *diverging*
- Thus in both the serial and diverging case, learning some thing leads to d-separation.

- The final case is the converging connection which generalises the sprinker example:



- Here the parents of A are independent *unless* something is known about A which does not come from the parents.
- In other words, if there is direct evidence about A, or evidence from the children of A, then the parents of A become dependent upon one another.

- Unlike in the other cases here the change is from independence to dependence.
- We note in passing that there are two types of evidence.
- Evidence which fixes the exact state of a variable (which value it takes) is called *hard* evidence.
- Other evidence is *soft* evidence.
- Hard evidence is required to induce conditional independence in the serial and diverging cases.
- Soft evidence is sufficient to create dependence in the converging case.
- We are now ready to formally define probabilistic networks.
- For that, however, we will need another lecture...

Bayesian networks

- A Bayesian network is defined by:

  Definition 1.1 *A* Bayesian network *is a directed acyclic graph where each variable is a random variable with a finite set of mutually exclusive states. For every variable A there is a probability function $\Pr(A \mid B_1, \ldots, B_n)$ where $B_1, \ldots, B_n$ are the parents of A.*

  Note that for root nodes the probability function is just $\Pr(A)$.
- Bayesian networks are also known as probabilistic causal networks.

- We also have:

  Definition 1.2 *Given random variables A, B and C, A and C are* conditionally independent given B, *if:*

  $$\Pr(A \mid B) = \Pr(A \mid B, C)$$

  If B is empty then A and C are independent in the sense we are used to.

- Applying Bayes' rule this means that:

$$\begin{aligned}\Pr(C \mid B, A) &= \frac{\Pr(A \mid C, B)\Pr(C \mid B)}{\Pr(A \mid B)} \\ &= \frac{\Pr(A \mid B)\Pr(C \mid B)}{\Pr(A \mid B)} \\ &= \Pr(C \mid B)\end{aligned}$$

- Finally, we have:

  Definition 1.3 *Two variables in a causal network are* d-separated *if, for all paths between A and B there is an intermediate variable V such that:*

  1. *The connection is serial or diverging and there is hard evidence for V; or*
  2. *The connection is converging and there is no evidence for either V or any of its descendants.*

- Now the neat thing about Bayesian networks is that:

  Theorem 1.1 *If A and B are d-separated in a Bayesian network G, and evidence e is entered, then:*

  $$\Pr(A \mid B, e) = \Pr(A \mid e)$$

- This is neat because it means that it is easy to spot conditional independencies from the graphical representation of the Bayesian network.
- It also makes it easy to show that:

  Theorem 1.2 *Given a graph which includes two d-separated nodes A and B, then changes in the probability of A have no impact on the change in probability of B.*

  and this, in turn makes it easy to devise algorithms for computing the probabilities of variables as evidence is obtained.

- Another consequence of Theorem **??** is that the joint probability over all the variables in a Bayesian network is just the product of all the conditional probabilities in the Bayesian network.

- In other words

  Theorem 1.3 *If G is a Bayesian network which includes only the set of variables $U = \{A_1, \ldots, A_m\}$, then the joint probability distribution over U is:*

  $$\Pr(U) = \prod_i \Pr(A_i \mid pa(A_i))$$

  where $pa(A_i)$ is the set of parents of A_i.

- This, of course, is exactly the same factorisation as we saw in the last lecture.
- What this means from the computational side is that provided we start at the root nodes, for which we have unconditional probabilities, we can conceive of a message passing algorithm to compute $\Pr(U)$.