

# Belief Networks

Chris Williams

School of Informatics, University of Edinburgh

September 2011

# Overview

- ▶ Independence
- ▶ Conditional Independence
- ▶ Belief networks
- ▶ Constructing belief networks
- ▶ Inference in belief networks
- ▶ Learning in belief networks
- ▶ Readings: e.g. Bishop §8.1 (not 8.1.1 nor 8.1.4), §8.2, Russell and Norvig, §15.1, §15.2, §15.5, Jordan handout §2.1 (details of Bayes ball algorithm not examinable)

# Independence

- ▶ Let  $\mathbf{X}$  and  $\mathbf{Y}$  be two disjoint subsets of variables. Then  $\mathbf{X}$  is said to be *independent* of  $\mathbf{Y}$  if and only if

$$P(\mathbf{X}|\mathbf{Y}) = P(\mathbf{X})$$

for all possible values  $\mathbf{x}$  and  $\mathbf{y}$  of  $\mathbf{X}$  and  $\mathbf{Y}$ ; otherwise  $\mathbf{X}$  is said to be *dependent* on  $\mathbf{Y}$

- ▶ Using the definition of conditional probability, we get an equivalent expression for the independence condition

$$P(\mathbf{X}, \mathbf{Y}) = P(\mathbf{X})P(\mathbf{Y})$$

- ▶  $\mathbf{X}$  independent of  $\mathbf{Y} \Leftrightarrow \mathbf{Y}$  independent of  $\mathbf{X}$
- ▶ Independence of a set of variables.  $X_1, \dots, X_n$  are independent iff

$$P(X_1, \dots, X_n) = \prod_{i=1}^n P(X_i)$$

## Example for Independence Testing

	<i>Toothache = true</i>	<i>Toothache = false</i>
<i>Cavity = true</i>	0.04	0.06
<i>Cavity = false</i>	0.01	0.89

- Is *Toothache* independent of *Cavity* ?

# Conditional Independence

- ▶ Let  $\mathbf{X}$ ,  $\mathbf{Y}$  and  $\mathbf{Z}$  be three disjoint sets of variables.  $\mathbf{X}$  is said to be *conditionally independent* of  $\mathbf{Y}$  given  $\mathbf{Z}$  iff

$$P(\mathbf{x}|\mathbf{y}, \mathbf{z}) = P(\mathbf{x}|\mathbf{z})$$

for all possible values of  $\mathbf{x}$ ,  $\mathbf{y}$  and  $\mathbf{z}$ .

- ▶ Equivalently  $P(\mathbf{x}, \mathbf{y}|\mathbf{z}) = P(\mathbf{x}|\mathbf{z})P(\mathbf{y}|\mathbf{z})$
- ▶ Notation,  $I(\mathbf{X}, \mathbf{Y}|\mathbf{Z})$

# Belief Networks

- ▶ A simple, graphical notation for conditional independence assertions and hence for compact specification of full joint distributions
- ▶ Syntax:
  - ▶ a set of nodes, one per variable
  - ▶ a directed acyclic graph (DAG) (link  $\approx$  “directly influences”)
  - ▶ a conditional distribution for each node given its parents:  
 $P(X_i | Parents(X_i))$
- ▶ In the simplest case, conditional distribution represented as a conditional probability table (CPT)

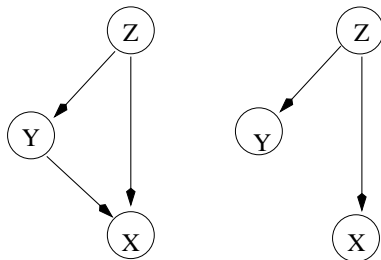
## Belief Networks 2

- ▶ DAG  $\Rightarrow$  no directed cycles  $\Rightarrow$  can number nodes so that no edges go from a node to another node with a lower number
- ▶ Joint distribution

$$P(X_1, \dots, X_n) = \prod_{i=1}^n P(X_i | \text{Parents}(X_i))$$

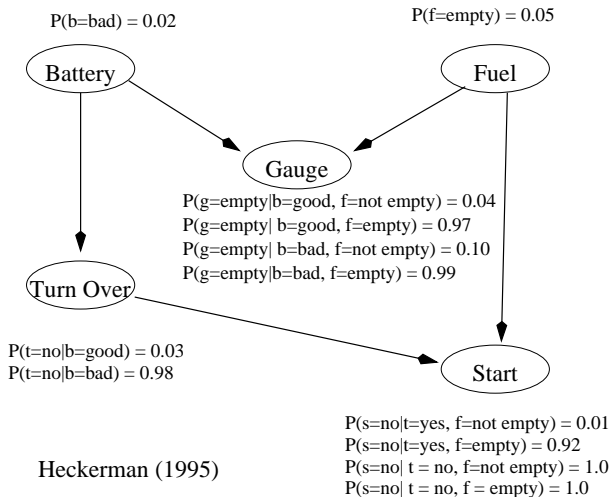
- ▶ *Missing* links imply conditional independence
- ▶ Ancestral simulation to sample from joint distribution

## Graphical example



- ▶ LHS: No independence  
 $P(X, Y, Z) = P(Z)P(Y|Z)P(X|Y, Z)$
- ▶ RHS:  $P(X, Y, Z) = P(Z)P(Y|Z)P(X|Z)$ , with  $I(\mathbf{X}, \mathbf{Y}|\mathbf{Z})$
- ▶ Note: there are other graphical structures that imply  $I(\mathbf{X}, \mathbf{Y}|\mathbf{Z})$

# Example Belief Network



- ▶ Unstructured joint distribution requires  $2^5 - 1 = 31$  numbers to specify it. Here can use 12 numbers
- ▶ Take the ordering  $b, f, g, t, s$ . Joint can be expressed as

$$P(b, f, g, t, s) = P(b)P(f|b)P(g|b, f)P(t|b, f, g)P(s|b, f, g, t)$$

- ▶ Conditional independences (missing links) give

$$P(b, f, g, t, s) = P(b)P(f)P(g|b, f)P(t|b)P(s|t, f)$$

- ▶ What is probability of  
 $P(b = \text{good}, t = \text{no}, g = \text{empty}, f = \text{not empty}, s = \text{no})?$

# Constructing belief networks

1. Choose a relevant set of variables  $X_i$  that describe the domain
2. Choose an ordering for the variables
3. While there are variables left
  - (a) Pick a variable  $X_i$  and add it to the network
  - (b) Set  $Parents(X_i)$  to some minimal set of nodes already in the net
  - (c) Define the CPT for  $X_i$

- ▶ This procedure is guaranteed to produce a DAG
- ▶ To ensure maximum sparsity, add “root causes” first, then the variables they influence and so on, until leaves are reached. Leaves have no direct causal influence over other variables
- ▶ **Example:** Construct DAG for the car example using the ordering  $s, t, g, f, b$
- ▶ “Wrong” ordering will give same joint distribution, but will require the specification of more numbers than otherwise necessary

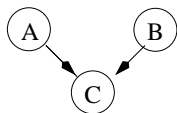
# Defining CPTs

- ▶ Where do the numbers come from? Can be elicited from experts, or learned, see later
- ▶ CPTs can still be very large (and difficult to specify) if there are many parents for a node. Can use combination rules such as Pearl's (1988) NOISY-OR model for binary nodes

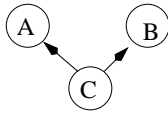
# Conditional independence relations in belief networks

- ▶ Consider three disjoint groups of nodes,  $\mathbf{X}$ ,  $\mathbf{Y}$ ,  $\mathbf{E}$
- ▶ Q: Given a graphical model, how can we tell if  $I(\mathbf{X}, \mathbf{Y}|\mathbf{E})$ ?
- ▶ A: we use a test called *direction-dependent separation* or *d-separation*
- ▶ If every undirected path from  $\mathbf{X}$  to  $\mathbf{Y}$  is **blocked** by  $\mathbf{E}$ , then  $I(\mathbf{X}, \mathbf{Y}|\mathbf{E})$

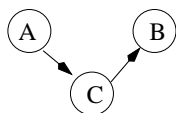
## Defining blocked



C is head-to-head



C is tail-to-tail



C is head-to-tail

A path is blocked if

1. there is a node  $\omega \in \mathbf{E}$  which is head-to-tail wrt the path
2. there is a node  $\omega \in \mathbf{E}$  which is tail-to-tail wrt the path
3. there is a node that is head-to-head and neither the node, nor any of its descendants, are in  $\mathbf{E}$

## Motivation for blocking rules

- ▶ Head-to-head  $I(a, b|\emptyset)$

$$p(a, b, c) = p(a)p(b)p(c|a, b)$$

$$p(a, b) = p(a)p(b) \sum_c p(c|a, b) = p(a)p(b)$$

- ▶ Tail-to-tail  $I(a, b|c)$

$$p(a, b, c) = p(c)p(a|c)p(b|c)$$

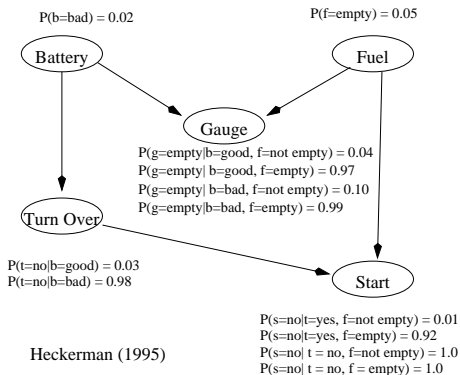
$$p(a, b|c) = p(a, b, c)/p(c) = p(a|c)p(b|c)$$

- ▶ Head-to-tail  $I(a, b|c)$

$$p(a, b, c) = p(a)p(c|a)p(b|c)$$

$$p(a, b|c) = p(a, b, c)/p(c) = p(a, c)p(b|c)/p(c) = p(a|c)p(b|c)$$

# Example



- ▶  $I(t, f|\emptyset)$  ?
- ▶  $I(b, f|s)$  ?
- ▶  $I(b, s|t)$  ?

# The Bayes Ball Algorithm

- ▶ §2.1 in Jordan handout (2003)
- ▶ Paper “Bayes-Ball: The Rational Pastime” by R. D. Shachter (UAI 98)
- ▶ Provides an algorithm with linear time complexity which given sets of nodes  $\mathbf{X}$  and  $\mathbf{E}$ , determines the set of nodes  $\mathbf{Y}$  s.t.

$$I(\mathbf{X}, \mathbf{Y} | \mathbf{E})$$

- ▶  $\mathbf{Y}$  is called the set of irrelevant nodes for  $\mathbf{X}$  given  $\mathbf{E}$

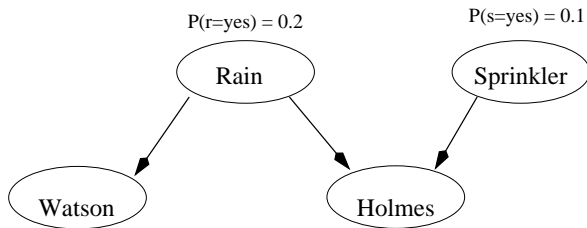
# Inference in belief networks

- ▶ Inference is the computation of results to queries given a network in the presence of evidence
- ▶ e.g. All/specific marginal posteriors e.g.  $P(b|s)$
- ▶ e.g. Specific joint conditional queries e.g.  $P(b, f|t)$ , or finding the most likely explanation given the evidence
- ▶ In general networks inference is NP-hard (loops cause problems)

## Some common methods

- ▶ For tree-structured networks inference can be done in time linear in the number of nodes (Pearl, 1986).  $\lambda$  messages are passed up the tree and  $\pi$  messages are passed down. All the necessary computations can be carried out locally. HMMs (chains) are a special case of trees. Pearl's method also applies to polytrees (DAGS with no undirected cycles)
- ▶ Variable elimination (see Jordan handout, ch 3)
- ▶ Clustering of nodes to yield a tree of cliques (junction tree) (Lauritzen and Spiegelhalter, 1988); see Jordan handout ch 17
- ▶ Symbolic probabilistic inference (D'Ambrosio, 1991)
- ▶ There are also approximate inference methods, e.g. using stochastic sampling or variational methods

# Inference Example



$$P(w=yes|r=yes) = 1$$
$$P(w=yes|r=no) = 0.2$$

$$P(h=yes|r=yes, s=yes) = 1.0$$
$$P(h=yes|r=yes, s=no) = 1.0$$
$$P(h=yes|r=no, s=yes) = 0.9$$
$$P(h=yes|r=no, s=no) = 0.0$$

- ▶ Mr. Holmes lives in Los Angeles. One morning when Holmes leaves his house, he realizes that his grass is wet. Is it due to rain, or has he forgotten to turn off his sprinkler?
- ▶ Calculate  $P(r|h)$ ,  $P(s|h)$  and compare these values to the prior probabilities
- ▶ Calculate  $P(r, s|h)$ .  $r$  and  $s$  are marginally independent, but conditionally dependent
- ▶ Holmes checks Watson's grass, and finds it is also wet. Calculate  $P(r|h, w)$ ,  $P(s|h, w)$
- ▶ This effect is called *explaining away*

# Learning in belief networks

- ▶ General problem: learning probability models
- ▶ Learning CPTs; easier. Especially easy if all variables are observed, otherwise can use EM
- ▶ Learning structure; harder. Can try out a number of different structures, but there can be a huge number of structures to search through
- ▶ Say more about this later

## Some Belief Network references

- ▶ E. Charniak “Bayesian Networks without Tears”, AI Magazine Winter 1991, pp 50-63
- ▶ D. Heckerman, “A Tutorial on Learning Bayesian Networks”, Technical Report MSR-TR-95-06, Microsoft Research, March, 1995, <http://research.microsoft.com/~heckerman/>
- ▶ J. Pearl “Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference”, Morgan Kaufmann, 1988
- ▶ E. Castillo, J. M. Gutiérrez, A. S. Hadi “Expert Systems and Probabilistic Network Models”, Springer, 1997
- ▶ S. J. Russell and P. Norvig, “Artificial Intelligence: A Modern Approach”, Prentice Hall, 1995 (chapters 14, 15)
- ▶ F. V. Jensen, “An introduction to Bayesian networks”, UCL Press, 1996
- ▶ D. Koller and N. Friedman, “Probabilistic Graphical Models: Principles and Techniques”, MIT Press, 2009