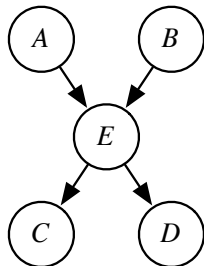


Learning a Belief Network

- If you
 - ▶ know the structure
 - ▶ have observed all of the variables
 - ▶ have no missing data
- you can learn each conditional probability separately.

Learning belief network example

Model



Data

<i>A</i>	<i>B</i>	<i>C</i>	<i>D</i>	<i>E</i>
<i>t</i>	<i>f</i>	<i>t</i>	<i>t</i>	<i>f</i>
<i>f</i>	<i>t</i>	<i>t</i>	<i>t</i>	<i>t</i>
<i>t</i>	<i>t</i>	<i>f</i>	<i>t</i>	<i>f</i>
		...		

→ Probabilities

$P(A)$

$P(B)$

$P(E|A, B)$

$P(C|E)$

$P(D|E)$

Learning conditional probabilities

- Each conditional probability distribution can be learned separately:
- For example:

$$P(E = t | A = t \wedge B = f) \\ = \frac{(\# \text{examples: } E = t \wedge A = t \wedge B = f) + c_1}{(\# \text{examples: } A = t \wedge B = f) + c}$$

where c_1 and c reflect prior (expert) knowledge ($c_1 \leq c$).

- When there are many parents to a node, there can be little or no data for each probability estimate:

Learning conditional probabilities

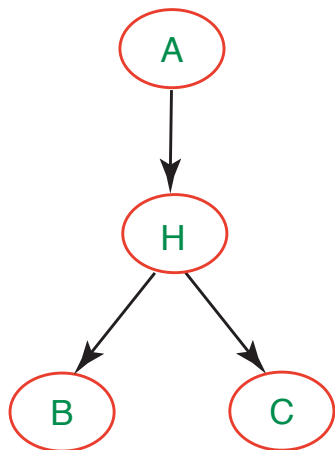
- Each conditional probability distribution can be learned separately:
- For example:

$$P(E = t | A = t \wedge B = f) \\ = \frac{(\# \text{examples: } E = t \wedge A = t \wedge B = f) + c_1}{(\# \text{examples: } A = t \wedge B = f) + c}$$

where c_1 and c reflect prior (expert) knowledge ($c_1 \leq c$).

- When there are many parents to a node, there can be little or no data for each probability estimate: use supervised learning to learn a decision tree, linear classifier, a neural network or other representation of the conditional probability.
- A conditional probability doesn't need to be represented as a table!

Unobserved Variables



- What if we had only observed values for A , B , C ?

A	B	C
t	f	t
f	t	t
t	t	f
	...	

EM Algorithm

Augmented Data

<i>A</i>	<i>B</i>	<i>C</i>	<i>H</i>	<i>Count</i>
<i>t</i>	<i>f</i>	<i>t</i>	<i>t</i>	0.7
<i>t</i>	<i>f</i>	<i>t</i>	<i>f</i>	0.3
<i>f</i>	<i>t</i>	<i>t</i>	<i>f</i>	0.9
<i>f</i>	<i>t</i>	<i>t</i>	<i>t</i>	0.1

E-step



M-step

Probabilities

$$P(A)$$
$$P(H|A)$$
$$P(B|H)$$
$$P(C|H)$$

- Repeat the following two steps:
 - ▶ **E-step** give the expected number of data points for the unobserved variables based on the given probability distribution. Requires probabilistic inference.
 - ▶ **M-step** infer the (maximum likelihood) probabilities from the data. This is the same as the full observable case.
- Start either with made-up data or made-up probabilities.
- EM will converge to a local maxima.

Belief network structure learning (I)

$$P(\text{model}|\text{data}) = \frac{P(\text{data}|\text{model}) \times P(\text{model})}{P(\text{data})}.$$

- A model here is a belief network.
- A bigger network can always fit the data better.
- $P(\text{model})$ lets us encode a preference for smaller networks (e.g., using the description length).
- You can search over network structure looking for the most likely model.

A belief network structure learning algorithm

- Search over total orderings of variables.
- For each total ordering X_1, \dots, X_n use supervised learning to learn $P(X_i | X_1 \dots X_{i-1})$.
- Return the network model found with minimum:
 - $\log P(\text{data} | \text{model}) - \log P(\text{model})$
 - ▶ $P(\text{data} | \text{model})$ can be obtained by inference.
 - ▶ How to determine $-\log P(\text{model})$?

Bayesian Information Criterion (BIC) Score

$$P(M|D) = \frac{P(D|M) \times P(M)}{P(D)}$$

$$-\log P(M|D) \propto -\log P(D|M) - \log P(M)$$

- $-\log P(D|M)$ is the negative log likelihood of the model: number of bits to describe the data in terms of the model.
- If $|D|$ is the number of data instances, there are $2^{|D|}$ different probabilities to distinguish. Each one can be described in $|D|$ bits.
- If there are $||M||$ independent parameters ($||M||$ is the dimensionality of the model):

$$-\log P(M|D) \propto$$

Bayesian Information Criterion (BIC) Score

$$P(M|D) = \frac{P(D|M) \times P(M)}{P(D)}$$

$$-\log P(M|D) \propto -\log P(D|M) - \log P(M)$$

- $-\log P(D|M)$ is the negative log likelihood of the model: number of bits to describe the data in terms of the model.
- If $|D|$ is the number of data instances, there are $|D| + 1$ different probabilities to distinguish. Each one can be described in $\log(|D| + 1)$ bits.
- If there are $||M||$ independent parameters ($||M||$ is the dimensionality of the model):

$$-\log P(M|D) \propto -\log P(D|M) + ||M|| \log(|D| + 1)$$

(This is approximately the (negated) BIC score.)

Belief network structure learning (II)

- Given a total ordering, to determine $parents(X_i)$ do independence tests to determine which features should be the parents
- XOR problem: just because features do not give information individually, does not mean they will not give information in combination
- Search over total orderings of variables

Missing Data

- You cannot just ignore missing data unless you know it is missing at random.
- Is the reason data is missing correlated with something of interest?
- For example: data in a clinical trial to test a drug may be missing because:

Missing Data

- You cannot just ignore missing data unless you know it is missing at random.
- Is the reason data is missing correlated with something of interest?
- For example: data in a clinical trial to test a drug may be missing because:
 - ▶ the patient dies
 - ▶ the patient had severe side effects
 - ▶ the patient was cured
 - ▶ the patient had to visit a sick relative.

— ignoring some of these may make the drug look better or worse than it is.
- In general you need to model why data is missing.

Causal Networks

- A causal network is a Bayesian network that predicts the effects of interventions.
- To intervene on a variable:
 - ▶ remove the arcs into the variable from its parents
 - ▶ set the value of the variable
- Intervening on a variable only affects descendants of the variable.

Causality

- We would expect a causal model to obey the independencies of a belief network.
- Not all belief networks are causal:



- Conjecture: causal belief networks are more natural and more concise than non-causal networks.
- We can't learn causal models from observational data unless we are prepared to make modeling assumptions.
- Causal models can be learned from randomized experiments.

General Learning of Belief Networks

- We have a mixture of observational data and data from randomized studies.
- We are not given the structure.
- We don't know whether there are hidden variables or not. We don't know the domain size of hidden variables.
- There is missing data.

... this is too difficult for current techniques!