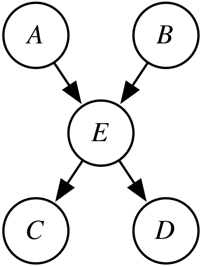
Learning a Belief Network

- 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

A	B	C	D	E
t	f	t	t	f
f	t	t	t	t
t	t	f	t	f
		$\dots$		

$\longrightarrow$ 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 little or no data for each probability estimate:

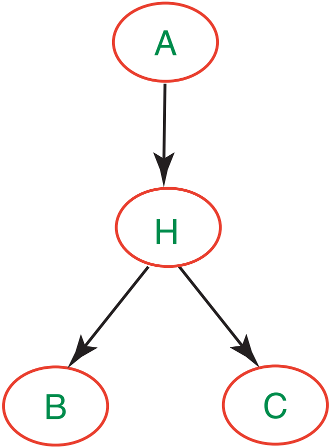
Learning conditional probabilities

- 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 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
	$\ldots$	

EM Algorithm

Augmented Data

A	B	C	H	$Count$
t	f	t	t	0.7
t	f	t	f	0.3
f	t	t	f	0.9
f	t	t	t	0.1
	$\cdots$			$\cdots$

E-step

M-step

Probabilities

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

EM Algorithm

- 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)

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

- A model here is a belief network.
- A bigger network can always fit the data better.
- $P(model)$ 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 $X_1, \ldots, X_n$ use supervised learning to learn $P(X_i | X_1 \ldots X_{i-1})$.
- Return the network model found with minimum:
  $-\log P(data|model) - \log P(model)$
  - $P(data|model)$ can be obtained by inference.
  - How to determine $-\log P(model)$?

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 different probabilities to distinguish. Each one can be described in 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.

Causality

- A causal model lets us predict the effect of an intervention.
- 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.
- We can learn causal models from randomized experimentation.

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!