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: A → B → E → C → D

Data:

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
<th>E</th>
</tr>
</thead>
<tbody>
<tr>
<td>t</td>
<td>f</td>
<td>t</td>
<td>t</td>
<td>t</td>
<td>f</td>
</tr>
<tr>
<td>f</td>
<td>t</td>
<td>t</td>
<td>t</td>
<td>t</td>
<td>t</td>
</tr>
<tr>
<td>t</td>
<td>t</td>
<td>f</td>
<td>t</td>
<td>f</td>
<td></td>
</tr>
</tbody>
</table>

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:
  \[
  P(E = t \mid A = t \land B = f) = \frac{(\#\text{examples: } E = t \land A = t \land B = f)}{(\#\text{examples: } A = t \land B = f)} + c_1
  \]
  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:

\[
P(E = t \mid A = t \land B = f)
= \frac{(\#\text{examples}: E = t \land A = t \land B = f) + c_1}{(\#\text{examples}: A = t \land B = f) + c}
\]

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

<table>
<thead>
<tr>
<th></th>
<th>$A$</th>
<th>$B$</th>
<th>$C$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>$t$</td>
<td>$f$</td>
<td>$t$</td>
</tr>
<tr>
<td>2</td>
<td>$f$</td>
<td>$t$</td>
<td>$t$</td>
</tr>
<tr>
<td>3</td>
<td>$t$</td>
<td>$t$</td>
<td>$f$</td>
</tr>
<tr>
<td>...</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
EM Algorithm

Model

A → H → B → C

Augmented Data

<table>
<thead>
<tr>
<th>A</th>
<th>B</th>
<th>C</th>
<th>H</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>t</td>
<td>f</td>
<td>t</td>
<td>t</td>
<td>0.7</td>
</tr>
<tr>
<td>t</td>
<td>f</td>
<td>t</td>
<td>f</td>
<td>0.3</td>
</tr>
<tr>
<td>f</td>
<td>t</td>
<td>t</td>
<td>f</td>
<td>0.9</td>
</tr>
<tr>
<td>f</td>
<td>t</td>
<td>t</td>
<td>t</td>
<td>0.1</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>

Probabilities

E-step

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

M-step
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 fully-observable case.

- Start either with made-up data or made-up probabilities.
- EM will converge to a local maxima.
Belief network structure learning (I)

Given examples \( e \), and model \( m \):

\[
P(m \mid e) = \frac{P(e \mid m) \times P(m)}{P(e)}.\]

- A model here is a belief network.
- A bigger network can always fit the data better.
- \( P(m) \) lets us encode a preference for simpler models (e.g., smaller networks)
- \rightarrow 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 $P(X_i \mid X_1 \ldots X_{i-1})$.
- Return the network model found with minimum:
  $- \log P(e \mid m) - \log P(m)$
  $\triangleright P(e \mid m)$ can be obtained by inference.
  $\triangleright$ How to determine $- \log P(m)$?
Bayesian Information Criterion (BIC) Score

\[ P(m \mid e) = \frac{P(e \mid m) \times P(m)}{P(e)} \]

\[ -\log P(m \mid e) \propto -\log P(e \mid m) - \log P(m) \]

- \( -\log P(e \mid m) \) is the negative log likelihood of model \( m \): number of bits to describe the data in terms of the model.
- If \( |e| \) is the number of examples, 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):
  
  \[ -\log P(m \mid e) \propto \]
Bayesian Information Criterion (BIC) Score

\[ P(m \mid e) = \frac{P(e \mid m) \times P(m)}{P(e)} \]

\[ -\log P(m \mid e) \propto -\log P(e \mid m) - \log P(m) \]

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

\[ -\log P(m \mid e) \propto -\log P(e \mid m) + ||m|| \log(|e| + 1) \]

This is (approximately) the BIC score.
Belief network structure learning (II)

- Given a total ordering, to determine \( \text{parents}(X_i) \) 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:
  - 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.
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

- An **intervention** on a variable changes its value by some mechanism outside of the model.
- A **causal model** is a model which predicts the effects of interventions.
Causality

- An **intervention** on a variable changes its value by some mechanism outside of the model.
- A **causal model** is a model which predicts the effects of interventions.
- The parents of a node are its direct causes.
- We would expect that a causal model to obey the independence assumption of a belief network.
  - All causal networks are belief networks.
  - Not all belief networks are causal networks.
Which probabilities change if we observe sprinkler on?
Sprinkler Example

Which probabilities change if we observe sprinkler on?
Which probabilities change if we turn the sprinkler on?
In a causal model:

- To intervene on a variable:
  - remove the arcs into the variable from its parents
  - set the value of the variable

- An intervention has a different effect than an observation.
Causality

In a causal model:

- To intervene on a variable:
  - remove the arcs into the variable from its parents
  - set the value of the variable

- An intervention has a different effect than an observation.
- Intervening on a variable only affects its descendants.
Causality

In a causal model:

- To intervene on a variable:
  - remove the arcs into the variable from its parents
  - set the value of the variable

- An intervention has a different effect than an observation.
- Intervening on a variable only affects its descendants.
- Can be modelled by each variable $X$ having a new parent, “Force $X$”, where $X$ is true if “Force $X$” is true and $X$ depends on its other parents if “Force $X$” is false.
One of the following is a better causal model of the world:

- \( \text{Switch}_\text{up} \rightarrow \text{Fan}_\text{on} \)
- \( \text{Switch}_\text{up} \leftarrow \text{Fan}_\text{on} \)

...same as belief networks, but different as causal networks.
One of the following is a better causal model of the world:

- Switch_up → Fan_on
- Switch_up ← Fan_on

...same as belief networks, but different as causal networks

Alspace Example: http://artint.info/tutorials/causality/marijuana.xml
Causality

One of the following is a better causal model of the world:

- Switch_up → Fan_on
- Fan_on → Switch_up

\[
\text{...same as belief networks, but different as causal networks.}
\]

- AIspace Example: [http://artint.info/tutorials/causality/marijuana.xml](http://artint.info/tutorials/causality/marijuana.xml)

- 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.
One of the following is a better causal model of the world:

\[ \text{Switch\_up} \rightarrow \text{Fan\_on} \quad \text{Switch\_up} \rightarrow \text{Fan\_on} \]

...same as belief networks, but different as causal networks

- Alspace Example: http://artint.info/tutorials/causality/marijuana.xml
- 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.
- Conjecture: causal belief networks are more natural and more concise than non-causal networks.
- Conjecture: causal model are more stable to changing circumstances (transportability)
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!