Class web page http://www.cs.ubc.ca/~murphyk
/Teaching/CS532c_Fall04/index.html

Send email to 'majordormo@cs.ubc.ca' with the contents
'subscribe cpsc535c' to join class list.
(Note: email address does not correspond to correct class number!)

Homework due in class on Monday 20th.

Monday’s class starts at 9.30am as usual.
Review: Probabilistic inference (state estimation)

- Inference is about estimating hidden (query) variables $H$ from observed (visible) measurements $v$, which we can do as follows:

$$P(h|v) = \frac{P(v, h)}{\sum_{h'} P(v, h')}$$

- Examples:
  - Medical diagnosis: $H$ diseases, $v =$ findings/ symptoms,
  - Speech recognition: $H =$ spoken words, $v =$ acoustic waveform
  - Genetic pedigree analysis: $H =$ genotype, $v =$ phenotype
Naive inference


- We observe the grass is wet and want to know how likely it was that the sprinkler caused this event.

$$P(s = 1|w = 1) = \frac{P(s = 1, w = 1)}{P(w = 1)} = \frac{\sum_{c=0}^{1} \sum_{r=0}^{1} P(s = 1, w = 1, R = r, C = c)}{\sum_{c,r,s} P(S = s, w = 1, R = r, C = c)}$$

- Query/hidden vars = \{S\}, visible vars = \{W\},
  nuisance vars = \{C, R\}.
Naive inference

- It is easy to marginalize a joint probability distribution when it is represented as a table.
- e.g., $P(X, Y) = \sum_z P(X, Y, Z)$
Graphical models

- Problems with representing joint as a big table
  - Representation: big table of numbers is hard to understand.
  - Inference: computing a marginal $P(X_i)$ takes $O(2^N)$ time.
  - Learning: there are $O(2^N)$ free parameters to estimate.
- Graphical models solve all 3 problems by providing a structured representation for joint probability distributions.
- Graphs encode conditional independence properties and represent families of probability distributions that satisfy these properties.
- Today we will study the relationship between graphs and independence properties.
**Independence properties of distributions**

- Defn: let $I(P)$ be the set of independence properties of the form $X \perp Y|Z$ that hold in distribution $P$.

<table>
<thead>
<tr>
<th>$X$</th>
<th>$Y$</th>
<th>$P(X,Y)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0</td>
<td>0.08</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>0.32</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>0.12</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>0.48</td>
</tr>
</tbody>
</table>

$$P(X = 1) = 0.48 + 0.12 = 0.6$$
$$P(Y = 1) = 0.32 + 0.48 = 0.8$$
$$P(X = 1, Y = 1) = 0.48 = 0.6 \times 0.8$$
$$P(X = x, Y = y) = P(X = x)P(Y = y) \forall x, y$$
$$\Rightarrow (X \perp Y) \in I(P)$$
$$\text{or } P \models (X \perp Y)$$
(Local) independence properties of DAGs

- Defn: let $I_l(G)$ be the set of local independence properties encoded by DAG $G$, namely:
  $$\{X_i \perp \text{NonDescendants}(X_i) \mid \text{Parents}(X_i)\}$$
  
  - i.e., a node is conditionally independent of its non-descendants given its parents.
  
  - $\text{Ancestors}(X_i) \subseteq \text{NonDescendants}(X_i)$

![Diagram of a DAG with nodes and edges]
Example of $I_l(G)$

\[
I_l(G_\emptyset) = \{(X \perp Y)\}
\]

\[
I_l(G_{X\rightarrow Y}) = \emptyset
\]

\[
I_l(G_{Y\rightarrow X}) = \emptyset
\]
I-MAPS

• Defn: A DAG $G$ is an I-map (independence-map) of $P$ if $I_l(G) \subseteq I(P)$.

• From previous example,

\[ I_l(G_{\emptyset}) = \{(X \perp Y)\} \]
\[ I_l(G_{X \rightarrow Y}) = \emptyset \]
\[ I_l(G_{Y \rightarrow X}) = \emptyset \]
\[ I(P) = \{(X \perp Y)\} \]

• Hence all three graphs are I-maps of $P$. 
Defn: \( P \) factorizes according to \( G \) if \( P \) can be written as
\[
P(X_1, \ldots, X_N) = \prod_{i} P(X_i | \text{Pa}_G(X_i))
\]

Thm 3.2.6: If \( G \) is an I-map of \( P \), then \( P \) factorizes according to \( G \).

Proof:
\[
P(X_{1:N}) = P(X_1)P(X_2|X_1)P(X_3|X_1, X_2)\ldots \text{ chain rule}
\]
\[
= \prod_{i=1}^{N} P(X_i|X_{1:i-1})
\]
\[
= \prod_{i=1}^{N} P(X_i|\text{Pa}(X_i), \text{Ancestors}(X_i) \setminus \text{Pa}(X_i))
\]
\[
= \prod_{i=1}^{N} P(X_i|\text{Pa}(X_i)) \text{ since } G \text{ is I-map of } P
\]
**Bayes nets provide compact representation of joint probability distributions**

- Thm: If $G$ is an I-map of $P$, then $P$ factorizes according to $G$.
- Corollary: If $G$ is an I-map of $P$, then we can represent $P$ using $G$ and a set of conditional probability distributions (CPDs), $P(X_i|\text{Pa}(X_i))$, one per node.
- Defn: A Bayesian network (aka belief network) representing distribution $P$ is an I-map of $P$ and a set of CPDs.
- For binary random variables, the Bayes net takes $O(N2^K)$ parameters ($K = \text{max. num. parents}$), whereas full joint takes $O(2^N)$ parameters.
- Factored representation is easier to understand, easier to learn and supports more efficient inference (see later lectures).
Water sprinkler

$P(X_{1:N}) = \prod_{i=1}^{N} P(X_i|\text{Pa}(X_i))$

Cloudy

Sprinkler

Rain

WetGrass

$P(C, S, R, W) = P(C)P(S|C)P(R|C)P(W|S, R)$
Thm 3.2.8: If $P$ factorizes according to $G$, then $G$ is an I-map of $P$.

Proof: we must show $X \perp W|U$

$$P(X, W|U) = \frac{P(X, W, U)}{P(U)} \sum_Y P(X, W, U, Y)$$

$$= \frac{P(W)P(U|W)P(X|U) \sum_Y P(Y|X, W)}{P(U)}$$

$$= \frac{P(W, U)}{P(U)} P(X|U) \sum_Y P(Y|X, W)$$

$$= P(W|U) P(X|U)$$
Minimal I-maps

- Let $G$ be a fully connected DAG. Then $I_l(G) = \emptyset \subseteq I(P)$ for any $P$.
- Hence the complete graph is an I-map for any distribution.
- Defn: A DAG $G$ is a minimal I-map for $P$ if it is an I-map for $P$, and if the removal of even a single edge from $G$ renders it not an I-map.
- Construction: pick a node ordering, then let the parents of node $X_i$ be the minimal subset of $U \subseteq \{X_1, \ldots, X_{i-1}\}$ s.t. $X_i \perp \{X_1, \ldots, X_{i-1}\} \setminus U | U$.
- Defn (revised): A Bayesian network (aka belief network) representing distribution $P$ is a minimal I-map of $P$ and a set of CPDs.
By chaining together local independencies, we can infer more global independencies.

Defn: \( X \) is \textbf{d-separated} (directed-separated) from \( Y \) given \( Z \) if along every undirected path between \( X \) and \( Y \) there is a node \( w \) s.t. either

\begin{itemize}
  \item \( W \) has converging arrows (\( \rightarrow w \leftarrow \)) and neither \( W \) nor its descendants are in \( z \); or
  \item \( W \) does not have converging arrows and \( W \in Z \).
\end{itemize}

Defn: \( I(G) = \) all independence properties that correspond to d-separation:

\[ I(G) = \{(X \perp Y | Z) : d-sep_G(X; Y | Z)\} \]
Bayes-Ball Rules

$A$ is d-separated from $B$ given $C$ if we cannot send a ball from any node in $A$ to any node in $B$ according to the rules below, where shaded nodes are in $C$.
Soundness of d-separation

- Thm 3.3.3 (Soundness): If $P$ factorizes according to $G$, then $I(G) \subseteq I(P)$.
- i.e., any independence claim made by the graph is satisfied by all distributions $P$ that factorize according to $G$ (no false claims of independence).
- Pf: see later (when we discuss undirected graphs).
Completeness of d-separation - v1

• Defn (Completeness) v1: For any distribution $P$ that factorizes over $G$, if $(X \perp Y|Z) \in I(P)$, then $dsep_G(X;Y|Z)$.

• Contrapositive rule: $(A \Rightarrow B) \iff (\neg B \Rightarrow \neg A)$.

• Defn (Completeness, contrapositive form) v1. If $X$ and $Y$ are not d-separated given $Z$, then $X$ and $Y$ are dependent in all distributions $P$ that factorize over $G$.

• This definition of completeness is too strong since $P$ may have conditional independencies that are not evident from the graph.

• eg. Let $G$ be the graph $X \rightarrow Y$, where $P(Y|X)$ is

<table>
<thead>
<tr>
<th>$A$</th>
<th>$B = 0$</th>
<th>$B = 1$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0.4</td>
<td>0.6</td>
</tr>
<tr>
<td>1</td>
<td>0.4</td>
<td>0.6</td>
</tr>
</tbody>
</table>

• $G$ is l-map of $P$ since $I(G) = \emptyset \subseteq I(P) = \{(X \perp Y)\}$.

• But the CPD encodes $X \perp Y$ which is not evident in the graph.
Completeness of d-separation - v2

- Defn (Completeness) v2: If \((X \perp Y|Z)\) in all distributions \(P\) that factorize over \(G\), then \(dsep_G(X; Y|Z)\).

- Defn (Completeness, contrapositive form) v2: If \(X\) and \(Y\) are not d-separated given \(Z\), then \(X\) and \(Y\) are dependent in some distribution \(P\) that factorizes over \(G\).

- Thm 3.3.5: d-separation is complete.

- Proof: See Koller & Friedman p90.

- Hence d-separation captures as many of the independencies as possible (without reference to the particular CPDs) for all distributions that factorize over some DAG.
P-MAPS

• Can we find a graph that captures all the independencies in an arbitrary distribution (and no more)?

• Defn: A DAG $G$ is a perfect map (P-map) for a distribution $P$ if $I(P) = I(G)$.

• Thm: not every distribution has a perfect map.

• Pf by counterexample. Suppose we have a model where $A \perp C | \{B, D\}$, and $B \perp D | \{A, C\}$. This cannot be represented by any Bayes net.

• e.g., BN1 wrongly says $B \perp D | A$, BN2 wrongly says $B \perp D$. 

\[
\begin{align*}
&\text{A} & \text{A} & \text{A} \\
&\text{B} & \text{D} & \text{D} \\
&\text{C} & \text{C} & \text{B}
\end{align*}
\]
Undirected Graphical Models

- Graphs with one node per random variable and edges that connect pairs of nodes, but now the edges are undirected.

- Defn: Let $H$ be an undirected graph. Then $sep_H(A; C|B)$ iff all paths between $A$ and $C$ go through some nodes in $B$ (simple graph separation).

- Defn: the global Markov properties of a UG $H$ are

$$I(H) = \{ (X \perp Y|Z) : sep_H(X; Y|Z) \}$$

- UGs can model symmetric (non-causal) interactions that directed models cannot.

- aka Markov Random Fields, Markov Networks.
Expressive Power

• Can we always convert directed ↔ undirected?
• No.

No directed model can represent these and only these independencies.
\[ x \perp y \mid \{w, z\} \]
\[ w \perp z \mid \{x, y\} \]

No undirected model can represent these and only these independencies.
\[ x \perp y \]
**Conditional Parameterization?**

- In directed models, we started with \( p(X) = \prod_i p(x_i | x_{\pi_i}) \) and we derived the d-separation semantics from that.
- Undirected models: have the semantics, need parametrization.
- What about this “conditional parameterization”?
  \[
p(X) = \prod_i p(x_i | x_{\text{neighbours}(i)})
  \]
- Good: product of local functions.
  Good: each one has a simple conditional interpretation.
  Bad: local functions cannot be arbitrary, but must agree properly in order to define a valid distribution.
Marginal Parameterization?

- OK, what about this “marginal parameterization”?
  \[ p(X) = \prod_{i} p(x_i, x_{\text{neighbours}(i)}) \]

- Good: product of local functions.
  - Good: each one has a simple marginal interpretation.
  - Bad: only very few pathological marginals on overlapping nodes can be multiplied to give a valid joint.
Clique Potentials

• Whatever factorization we pick, we know that only connected nodes can be arguments of a single local function.

• A clique is a fully connected subset of nodes.

• Thus, consider using a product of clique potentials:

\[
P(X) = \frac{1}{Z} \prod_{\text{cliques } c} \psi_c(x_c) \quad \quad \quad Z = \sum_X \prod_{\text{cliques } c} \psi_c(x_c)
\]

• Each clique potential \( \psi_c(x_c) > 0 \) is an arbitrary positive function of its arguments.

• The normalization term \( Z \) is called the partition function (a function of the parameters \( \psi \)) and ensures \( \sum_X P(x) = 1 \).

• Without loss of generality we can restrict ourselves to maximal cliques. (Why?)

• A distribution \( P \) that is representable by a UG \( H \) in this way is called a Gibbs distribution over \( H \).
Examples of Clique Potentials

(a)

(b)
The model implies $x \perp z \mid y$

$$p(x, y, z) = p(y)p(x \mid y)p(z \mid y)$$

We can write this as:

$$p(x, y, z) = p(x, y)p(z \mid y) = \psi_{xy}(x, y)\psi_{yz}(y, z)$$

$$p(x, y, z) = p(x \mid y)p(z, y) = \psi_{xy}(x, y)\psi_{yz}(y, z)$$

cannot have all potentials be marginals
cannot have all potentials be conditionals

The positive clique potentials can only be thought of as general “compatibility”, “goodness” or “happiness” functions over their variables, but not as probability distributions.