Probabilistic graphical models
CPSC 532c (Topics in AI)
Stat 521a (Topics in multivariate analysis)

Lecture 2

Kevin Murphy

Wednesday 15 September, 2004

Administrivia

- 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:

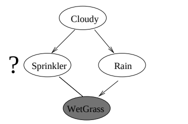
$$P(h|v) = \frac{P(v,h)}{\sum_{h'} P(v,h')}$$

- 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

- Represent joint prob. distribution $P(C,S,R,W)$ as a 4D table of $2^4 = 32$ numbers.
- We observe the grass is wet and want to know how likely it was that the sprinkler caused this event.

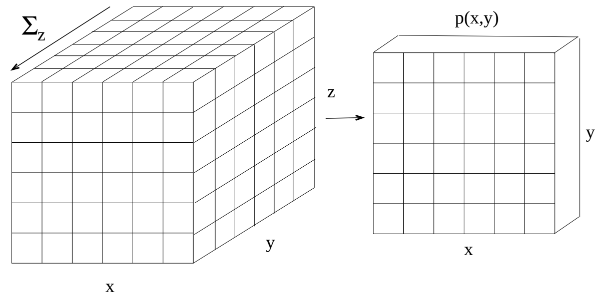
$$P(s=1|w=1) = \frac{P(s=1,w=1)}{P(w=1)}$$
$$= \frac{\sum_{c=0}^{1}\sum_{r=0}^{1} P(s=1,w=1,R=r,C=c)}{\sum_{c,r,s} P(S=s,w=1,R=r,C=c)}$$



- 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.

X	Y	P(X,Y)
0	0	0.08
0	1	0.32
1	0	0.12
1	1	0.48

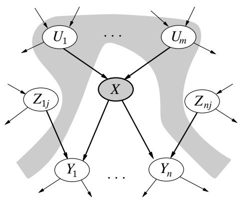
$$\begin{aligned}
P(X=1) &= 0.48 + 0.12 = 0.6 \\
P(Y=1) &= 0.32 + 0.48 = 0.8 \\
P(X=1, Y=1) &= 0.48 = 0.6 \times 0.8 \\
P(X=x, Y=y) &= P(X=x)P(Y=y) \forall x, y \\
&\Rightarrow (X \perp Y) \in I(P) \\
&\text{or } \; P \models (X \perp Y)
\end{aligned}$$

(Local) independence properties of DAGs

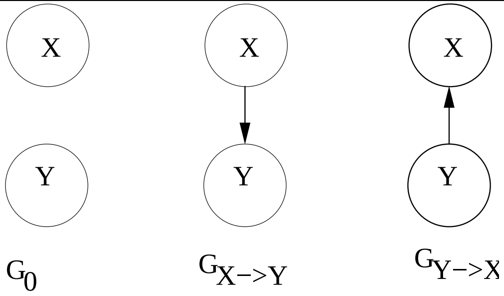
- Defn: let $I_l(G)$ be the set of local independence properties encoded by DAG G, namely:

$$\{X_i \perp \text{NonDescendants}(X_i)|\text{Parents}(X_i)\}$$

- i.e., a node is conditionally independent of its non-descendants given its parents.
- $\text{Ancestors}(X_i) \subseteq \text{NonDescendants}(X_i)$



Example of $I_l(G)$



$$
\begin{aligned}
I_l(G_\emptyset) &= \{(X \perp Y)\} \\
I_l(G_{X \to Y}) &= \emptyset \\
I_l(G_{Y \to X}) &= \emptyset
\end{aligned}
$$

I-maps

- Defn: A DAG G is an **I-map** (independence-map) of P if $I_l(G) \subseteq I(P)$.
- From previous example,

$$
\begin{aligned}
I_l(G_\emptyset) &= \{(X \perp Y)\} \\
I_l(G_{X \to Y}) &= \emptyset \\
I_l(G_{Y \to X}) &= \emptyset \\
I(P) &= \{(X \perp Y)\}
\end{aligned}
$$

- Hence all three graphs are I-maps of P.

From I-map to factorization

- Defn: P **factorizes according to** G if P can be written as
$$P(X_1, \ldots, X_N) = \prod_i P(X_i | \mathrm{Pa}_G(X_i))$$
- Thm 3.2.6: If G is an I-map of P, then P factorizes according to G.
- Proof:

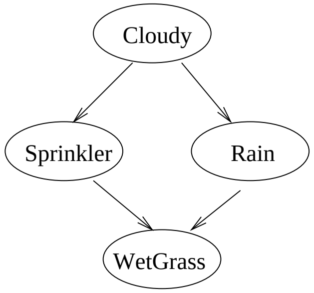
$$
\begin{aligned}
P(X_{1:N}) &= P(X_1)P(X_2|X_1)P(X_3|X_1, X_2)\ldots \text{ chain rule} \\
&= \prod_{i=1}^N P(X_i | X_{1:i-1}) \\
&= \prod_{i=1}^N P(X_i | \mathrm{Pa}(X_i), \mathrm{Ancestors}(X_i) \setminus \mathrm{Pa}(X_i)) \\
&= \prod_{i=1}^N P(X_i | \mathrm{Pa}(X_i)) \text{ since } G \text{ is I-map of } P
\end{aligned}
$$

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), $P(X_i|\mathrm{Pa}(X_i))$, 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 ($K =$ 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

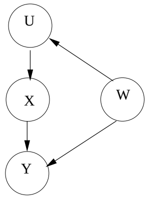
$$P(X_{1:N}) = \prod_{i=1}^{N} P(X_i|\mathsf{Pa}(X_i))$$



$$P(C,S,R,W) = P(C)P(S|C)P(R|C)P(W|S,R)$$

From factorization to I-map

- Thm 3.2.8: If P factorizes according to G, then G is an I-map of P.
- Proof: we must show $X \perp W|U$



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

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.

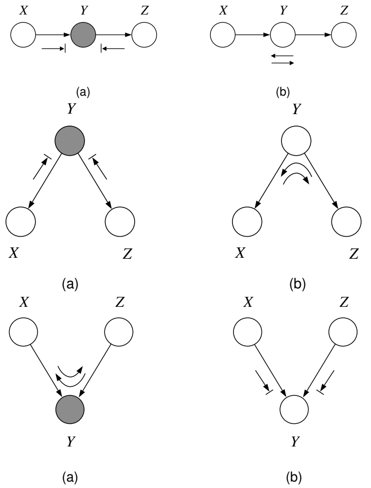
Global Markov properties of DAGs

- By chaining together local independencies, we can infer more global independencies.
- Defn: X is **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
  - W has converging arrows ($\rightarrow w \leftarrow$) and neither W nor its descendants are in z; or
  - W does not have converging arrows and $W \in Z$.
- 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 $X \to Y$, where $P(Y|X)$ is

A	$B=0$	$B=1$
0	0.4	0.6
1	0.4	0.6

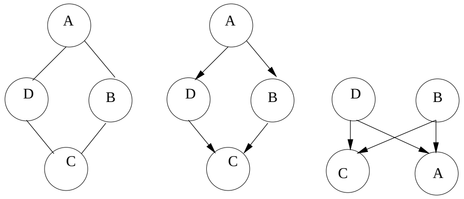
- G is I-map of P since $I(G) = \emptyset \subseteq I(P) = \{(X \perp Y)\}$.
- But the CPD encodes $X \perp Y$ which is not evident in the graph.

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.

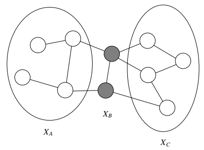
P-maps

- 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 $B \perp D$.



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 $\leftrightarrow$ undirected?
- No.

(a)

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

(b)

No undirected model can represent these and only these independencies.
$\mathbf{x} \perp \mathbf{y}$

Conditional Parameterization?

- In directed models, we started with $p(\mathbf{X}) = \prod_i p(\mathbf{x}_i|\mathbf{x}_{\pi_i})$ and we derived the d-separation semantics from that.
- Undirected models: have the semantics, need parametrization.
- What about this "conditional parameterization"?
$$p(\mathbf{X}) = \prod_i p(\mathbf{x}_i|\mathbf{x}_{\text{neighbours}(i)})$$
- 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"?

$$p(\mathbf{X}) = \prod_i p(\mathbf{x}_i, \mathbf{x}_{\text{neighbours}(i)})$$

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

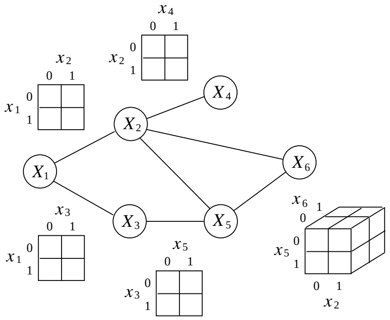
Clique Potentials

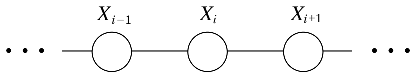
- Whatever factorization we pick, we know that only connected nodes can be arguments of a single local function.
- A *clique* is a fully connected subset of nodes.
- Thus, consider using a *product of clique potentials*:

$$\mathsf{P}(\mathbf{X}) = \frac{1}{Z} \prod_{\text{cliques } c} \psi_c(\mathbf{x}_c) \qquad Z = \sum_{\mathbf{X}} \prod_{\text{cliques } c} \psi_c(\mathbf{x}_c)$$

- Each clique potential $\psi_c(\mathbf{x}_c) > 0$ is an arbitrary positive function of its arguments.
- The normalization term Z is called the partition function (a function of the parameters ψ) and ensures $\sum_{\mathbf{x}} \mathsf{P}(\mathbf{x}) = 1$.
- 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





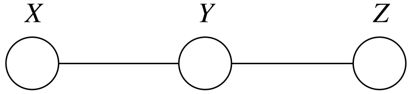
(a)

x_{i-1} \ x_i	-1	1
-1	1.5	0.2
1	0.2	1.5

x_i \ x_{i+1}	-1	1
-1	1.5	0.2
1	0.2	1.5

(b)

Interpretation of Clique Potentials



- The model implies $\mathbf{x} \perp \mathbf{z} \mid \mathbf{y}$

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

- We can write this as:

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

  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.