Conditional independence in distributions and graphs

Let $I(P)$ be the set of independence assertions of the form $(X \perp Y|Z)$ that are true for a distribution P.

Let $I(G)$ be the set of independence assertions represented by a DAG, G.

G is an I-map for P if I(G) $\subseteq I(P)$

In other words, all indepence represented in I-map G are true. (There could be some more in P that G does not reveal).

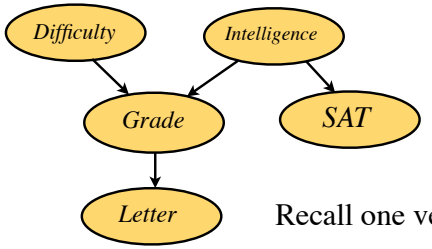
Example going from I-map to a factorization

From Kollar and Friedman

For $P(I,D,G,L,S)$, suppose $I(Graph)$ tells us

$(D \perp I,S|\varnothing)$ $(L \perp I,D,S|G)$ $(S \perp D,G,L|I)$ $(G \perp S|I,D)$

Note that $(A \perp B,C|X)$ just means $(A \perp B|X)$ and $(A \perp C|X)$



Recall one version of DAG semantics is

$X_i \perp \text{NonDescendents}(X_i) \mid \text{Parents}(X_i)$

Example going from I-map to a factorization

For $P(I,D,G,L,S)$, suppose $I(Graph)$ tells us

$(D \perp I,S|\varnothing)$ $(L \perp I,D,S|G)$ $(S \perp D,G,L|I)$ $(G \perp S|I,D)$

Note that $(A \perp B,C|X)$ just means $(A \perp B|X)$ and $(A \perp C|X)$

$$P(I,D,G,L,S) = P(I)P(D|I)P(G|I,D)P(L|I,D,G)P(S|I,D,G,L)$$

See next lecture for more comments on this example.

In particular, in this lecture I did not say that the chain rule ordering should be in lexigraphical order which makes it relatively easy to see that an arbitrary graph can be converted to the factorization which we introduced at the beginning. Further, it explains why do not care that the rules output do not include G independent of S given I (even though we can get it from the previous rule by symmetry.

Reminder about independence

$(A \perp B|\varnothing) \Rightarrow P(A|B) = P(A)$

Proof: $P(A|B) = \dfrac{P(A,B)}{P(B)} = \dfrac{P(A)P(B)}{P(B)} = P(A)$

$(A \perp B|C) \Rightarrow P(A|B,C) = P(A|C)$

Proof: $P(A|B,C) = \dfrac{P(A,B,C)}{P(B,C)} = \dfrac{P(A|C)P(B|C)P(C)}{P(C)P(B|C)} = P(A|C)$

General rule: $(A \perp B|..., C, ...) \Rightarrow P(A|... ,B, C, ...) = P(A|..., C, ...)$

Example going from I-map to a factorization

For $P(I,D,G,L,S)$, suppose $I(Graph)$ tells us

$(D \perp I,S|\varnothing)$ $(L \perp I,D,S|G)$ $(S \perp D,G,L|I)$ $(G \perp S|I,D)$

Note that $(A \perp B,C|X)$ just means $(A \perp B|X)$ and $(A \perp C|X)$

$P(I,D,G,L,S) = P(I)P(D|I)P(G|I,D)P(L|I,D,G)P(S|I,D,G,L)$

$(D \perp I|\varnothing) \Rightarrow P(D|I) = P(D)$

$(L \perp I,D,S|G) \Rightarrow P(L|I,D,G) = P(L|G)$

$(S \perp D,G,L|I) \Rightarrow P(S|I,D,G,L) = P(S|I)$

So, $P(I,D,G,L,S) = P(I)P(D)P(G|I,D)P(L|G)P(S|I)$

Interesting questions

- Does every probability distribution have a Bayesian network?

Chain rule says yes

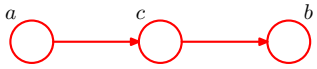
- Given the independence structure of a probability distribution, is the corresponding graph unique (ignoring isomorphisms)?

Case study of three nodes says no

- Do our graphs faithfully capture the independence structure of our distributions?

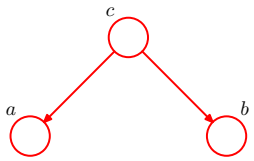
TBA

Back to case one



- Let a="smokes", c="high blood pressure", b="stroke"
- p(cla) tells you probability of having high blood pressure if you smoke (for some definition of each).

Can we distinguish case two from case one?



- Let a="smokes", b="high blood pressure", c="stroke"
- p(alc) tells you probability of being a smoker if you have high blood pressure (for some definition of each).
- Data for estimating p(cla) in first case, and p(alc) in second case cannot tell you which model you should prefer.
  - "Correlation is not causation"
- Causality implied by our generative process is about the statistics of the data, not physical causality.

Misunderstanding of probability may be the greatest of all impediments to scientific literacy.

Stephen Jay Gould

Can graphs capture all independence?

- Do our graphs faithfully capture the independence structure of our distributions?
- Recall that

  G is an I-map for P if $I(G) \subseteq I(P)$

  In other words, all independence represented in G are true.
  (There could be more independence in P that G does not reveal).

- Hence we are asking if $I(G) \equiv I(P)$
  Since $I(G) \subseteq I(P)$ this amounts to asking if $I(P) \subseteq I(G)$

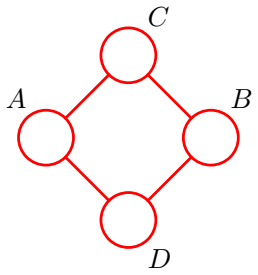
Perfection

G is an P-map for P if $I(G) \equiv I(P)$ (perfect map)

In other words, all independence represented in G are true, and there are no other independence relations.

Do all distributions have perfect maps?

Perfection may not be attainable



Suppose that we have
$(A \perp B | C, D)$
and
$(C \perp D | A, B)$

Now, draw the Bayes net
(have fun!).

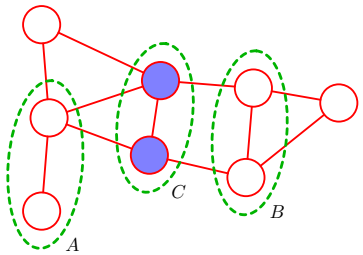
Note **no arrows**, but a link still means some probabilistic relation.

Undirected graphical models

- Also referred to as
  - Markov Networks
  - Markov Random Fields
- Nodes represent (groups of) random variables
- Edges represent probabilistic relations between connected nodes.
- We have already seen an example suggestive that arrows are not always helpful.

Undirected graphical models

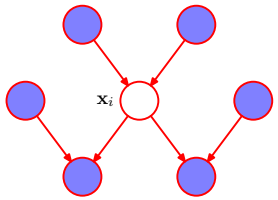
- The analog to d-separation is simper

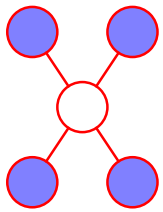


Here $(A \perp B | C)$ for all probability distributions represented by this graph.

Markov Blanket

- The Markov blanket of a node, X, is a particular set of (nearby) nodes Y where $X \perp X_i | B$ for all X_i
- For directed graphs the Markov blanket is the parents, children, and co-parents of X.
- For undirected graphs this is simply the set of nodes connected to X.



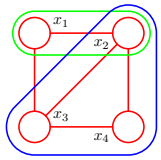


Undirected graphical models

- Bayes nets where nodes only have one parent are easily converted to undirected graphs without changing links.
- (Discussed in more detail soon)

Semantics of undirected graphical models

- Intuitively, for any two nodes, x_i and x_j, not connected by a link, $x_i \perp x_j | \mathbf{x}/\{i,j\}$.
- This leads to describing the semantics in terms of maximal cliques.
  - A clique is fully connected subset of nodes from the graph
  - A maximal clique is a clique where no node in the graph can be added to it without it ceasing to be a clique.



All parwise linked nodes are cliques. For example $\{x_1, x_2\}$ is a clique (green). However, it is not a maximal clique. $\{x_2, x_3, x_4\}$ is a maximal clique (blue). If we add another node (only x_1 is left) we no longer have a clique.

Semantics of undirected graphical models (2)

Let C index maximal cliques. Then

$$p(\mathrm{x}) = \frac{1}{Z}\prod_c \psi_C(\mathrm{x}_C)$$

where $Z = \sum_x \prod_c \psi_C(\mathrm{x}_C)$ (or $\int_x \prod_c \psi_C(\mathrm{x}_C)$) is the partition function, and $\psi_C(\mathrm{x}_C)$ are the clique potentials.

If x_i and x_j do not share an edge, then they do not share cliques.

So $p(\mathrm{x}) = \frac{1}{Z}\prod_{c(i)} \psi_C(\mathrm{x}_C) \prod_{c(j)} \psi_C(\mathrm{x}_C) \prod_{c \notin c(i) \cup c(j)} \psi_C(\mathrm{x}_C)$

Semantics of undirected graphical models (3)

We will assume that all $\psi_C(\mathrm{x}_C) > 0$.

In general, we leave the semantics of $\psi_C(\mathrm{x}_C)$ open, but for undirected graphs that come from directed graphs where each node has one parent, the semantics follows that for the directed graphs.

Since $\psi_C(\mathrm{x}_C) > 0$ we will often write $\psi_C(\mathrm{x}_C) = \exp\{-E(\mathrm{x}_C)\}$ where E() is the energy function.

Example of a Markov random field

- Consider a binary image (pixels are either black or white).
- Pixels are represented by {-1,1}.
- Suppose the image have is an underlying accurate image where some of the bits have been flipped by a noise process.