Announcements

- "midterm one" now posted soon
  - This "midterm" is no different in format from the assignments so far
- K&F chapter 3 has been posted

Independence in graphs and distributions

- A probability distribution (e.g., from your model) has certain conditional independence in its variables
- Our graphs also imply such independence assertions
- For distributions that factor as directed graphs
  - We can use d-Separation to ask if any particular one is true $\mathcal{I}(\mathcal{G})$
  - We can list *local* ones using the rule to get $\mathcal{I}_l(\mathcal{G})$

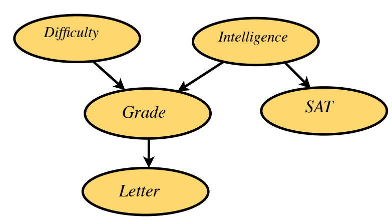
$$\text{For each } X_i : X_i \perp \text{NonDescendents}(X_i) \mid \text{Parents}(X_i)$$

Independence in graphs and distributions

- An independence assertion could be true for *some* probability distributions that factor according to a graph, but we are referring to those that are *always* true.
- The local independence properties are equivalent to the factorization (one derives the other, see K&F, chapter 3)
- We can have independence properties that are not in $\mathcal{I}_l(\mathcal{G})$ but can be found by d-separation

$$\mathcal{I}_l(\mathcal{G}) \subseteq \mathcal{I}(\mathcal{G}) \quad \text{(sometimes strict subset)}$$

Independence in graphs and distributions



Does d-separation say D and I are c.i. given L?
Does d-separation say G and S are c.i. given I?
Does d-separation say S and L are c.i. given I?

Do the local independencies agree?

Interesting questions

- Does every probability distribution have a corresponding Bayesian network?

Chain rule says yes

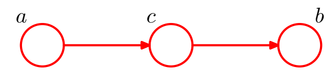
- Given the independence structure of a probability distribution, and a graph that captures them all (I(G)=I(P), is the corresponding graph unique (ignoring isomorphisms)?

Case study of three nodes says no

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

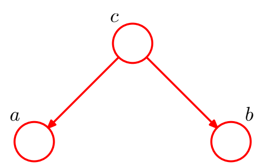
TBA

Back to case one



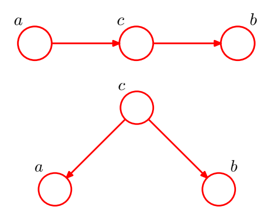
- Let a="smokes", c="high blood pressure", b="stroke"
- p(c|a) tells you the probability of having high blood pressure if you smoke (for some definition of each).

Can we distinguish case two from case one?



- Let a="smokes", c="high blood pressure", b="stroke"
- p(a|c) tells you probability of being a smoker if you have high blood pressure (for some definition of each).

Can we distinguish case two from case one?



$$p(a,b,c) = p(a)p(c \mid a)p(b \mid c) = p(a,c)p(b \mid c)$$

$$p(a,b,c) = p(c)p(a \mid c)p(b \mid c) = p(a,c)p(b \mid c)$$

(both lead to the testable $(a \perp b \mid c)$)

- Data for estimating p(c|a) in first case, and p(a|c) in second case cannot tell you which model you should prefer.
  - "Correlation is not causation"
- Causality implied by our generative (ancestral sampling) process is about the statistics of the data, not physical causality.

More on causality

- References
  - Koller and Friedman, Chapter 21 which starts on page 1009!
  - Classic book by Pearl, Causality: Models, Reasoning, and Inference, 2000
    - A version is available on-line (bayes.cs.ucla.edu/BOOK-99/book-toc.html)

More on causality

- We have been focussed on the joint distribution which is adequate (arguably optimal) for answering the queries we have studied
- In particular, we know how distributions over unknowns change due to evidence
- For many problems (e.g., computer vision and much of machine learning) this is sufficient
  - Either causes are obvious or not relevant

More on causality

- Two correlated variables can have multiple equivalent graphs hinting at **different** causal stories able to provide the **same** joint.
  - A causes B
  - B causes A
  - C causes both A and B
  - A and B cause C (and A and B are correlated by explaining away)

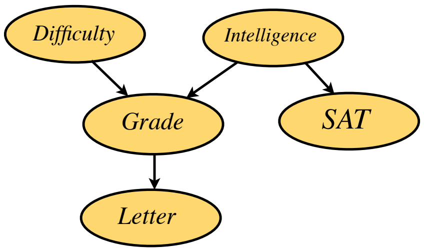
- Given a choice, we prefer the Bayes net that also represents our causal theory (if we have one)
  - More natural, easier to understand, better building block
  - Helps tell you determine whether observed statistics are consistent with your theory
    - (Covered briefly next)

Intervention

- Two Bayes nets that give the same joint distribution can differ in what they say about an intervention.
- We represent an intervention, x, as setting some subset of the variables, X, to the value, x, denoted by $do(X{=}x)$.
  - Example 1: Creating an experimental group that will not smoke
  - Example 2: Setting your grade to A by hacking into a computer
- On the surface, this might look like conditioning on X, but it is different --- the graph needs to change also
  - We need to “mutilate” the graph

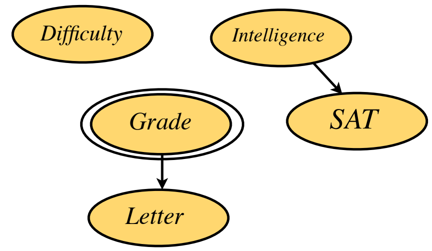
Representing Intervention

- Example one (students and grades, again)
  - Does observing grade change your belief about SAT?
- Now, suppose we intervene on the *Grade* random variable
  - E.G., we fix it by hacking into the grade computer
  - Now does observing grade change your belief about SAT?



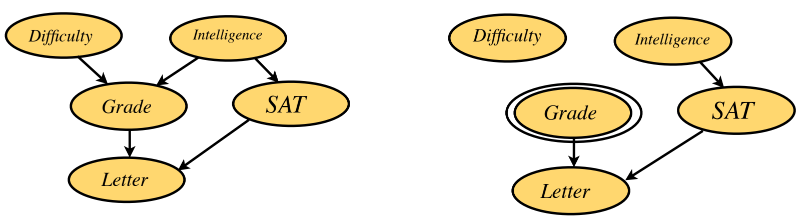
Representing Intervention

- The intervention not only conditions on the variable, it cuts the links that influence it. This is the mutilated graph.



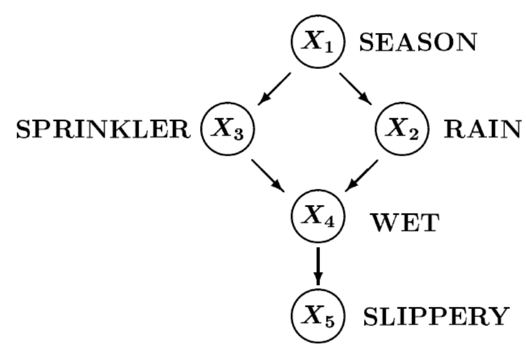
Representing Intervention

- Another example --- the student from before with a link between SAT and letter. Now we expect that the intervention does not entirely explain the letter, but that the influence of *grade* is direct (only).



Representing Intervention

- Another example --- from Pearl, 2000.
  - Consider the intervention of turning the sprinkler "on"



Representing Intervention

- Representation of the intervention of turning the sprinkler on.

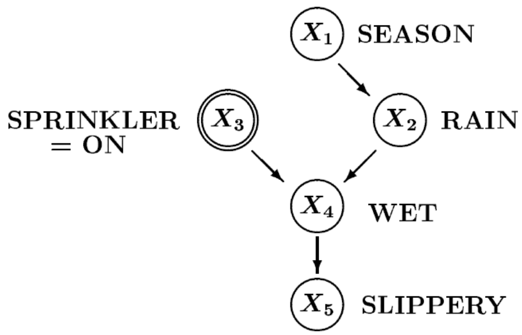
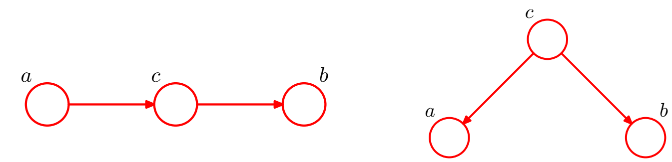


Figure 1.4: Network representation of the action “turning the sprinkler On.”

Back to smoking and high blood pressure



- a=“smokes”, c=“high blood pressure”, b=“stroke”
- Intervene on c.

- Now the two graphs are distinguishable based on data.

Back to graphs in general

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 $I(P) \subseteq I(G)$

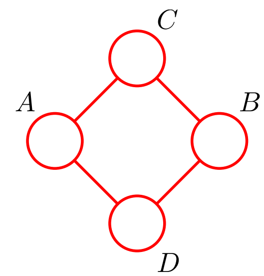
Perfection

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?

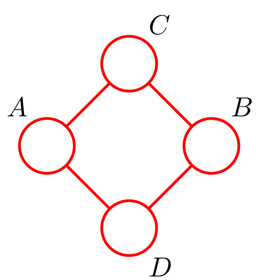
Perfection may not be attainable



The “misconception” example in K&F (pp. 82-3), where Alice, Bob, Charles, and Debbie study in pairs shown, but A and B never work together, nor do C and D.

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

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.

Interesting questions

- Does every probability distribution have a corresponding Bayesian network?

  Chain rule says yes

- Given the independence structure of a probability distribution, and a graph that captures them all (I(G)=I(P), is the corresponding graph unique (ignoring isomorphisms)?

  Case study of three nodes says no

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

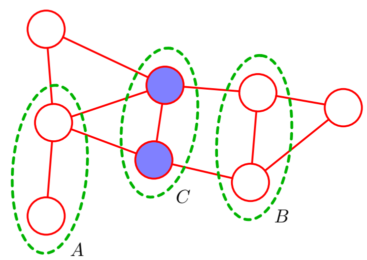
  Misconception example says no

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

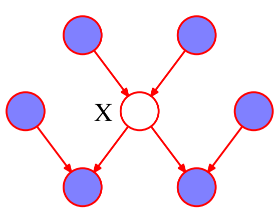
- The analog to d-separation is simper
  - Disjoint sets A and B are independent conditioned on C if all paths from nodes in A to nodes in B pass through C.

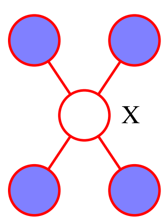


Here $\left(A \perp B \middle| C\right)$ for all probability distributions represented by this graph.

Markov Blanket

- The Markov blanket of a node, X, is a particular set of (nearby) nodes B where $X \perp X_i \middle| B$ 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, $x_i \perp x_j | \mathbf{x}/\{i,j\}$.

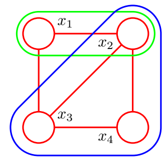
- So, $p(\ldots,x_i,\ldots,x_j,\ldots) = p(x_i|\mathbf{x}/\{i,j\})\,p(x_j|\mathbf{x}/\{i,j\})\,p(\mathbf{x}/\{i,j\})$

- This suggests that an appropriate factorization should not have factors with these (non directly linked) nodes together.

- A group of nodes that are all (fully) connected cannot be factored by the above rule.

Semantics of undirected graphical models

- So, we add nodes into factors, provided that they are all connected.

- 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.