Lecture 4: Bayesian Networks - Part II

- D-separation
- D-maps
- Perfect maps
- Markov networks
- Knowledge engineering for Bayesian networks

Recall from last time: I-maps

- A DAG G is called an I-map of a probability distribution P if P satisfies the independence assumptions implied by G:

$$I(X_i, Nondescendents(X_i)|Parents(X_i)), \forall i = 1, \dots n$$

- If G is an I-map for a distribution P, then P factorizes according to G, which means that we can represent P more compactly, in terms of local probability models
- A Bayes net representation of a distribution P is an I-map of P together with the local probability models
- Ideally, we would like a minimal I-map of P
- But some minimal I-maps are smaller than others!

Implied independencies

- Independencies between variables are important because they can help us answer queries more efficiently.
- So it would be interesting to know what conditional independencies are implied by a Bayes net structure G, based on $Markov(G)$:

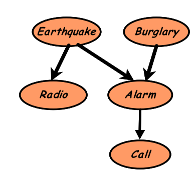
$$I(X_i, Nondescendents(X_i)|Parents(X_i)), \forall i = 1, \dots n$$

- Some independencies are trivially implied (e.g. $I(X, Y|Z) \to I(Y, X|Z)$)
- We want to know if two sets of variables X and Y are conditionally independent given evidence about a set of variables Z

Dependency flow

The intuition is that if we get evidence about a variable in Z, this evidence will get propagated along paths in the graph, where a path is a sequence of neighboring variables (not necessarily going in the direction of the arcs). This might enable or disable flow of dependency between other nodes

Example:

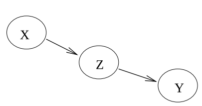


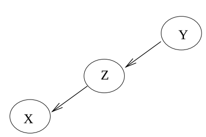
- $R \leftarrow E \rightarrow A \leftarrow B$
- $C \leftarrow A \leftarrow E \rightarrow R$

Dependency and paths in the graph

- We will consider paths from X to Y going through variables in Z.
- Knowing a value in Z can have two possible effects:
  - Enable the flow of influence from X to Y - the path becomes **active**
  - Disable the flow of influence - the path becomes **blocked**
- If paths between X and Y are always blocked, then X and Y are conditionally independent given Z; we say that X and Y are **d-separated given** Z
- We consider first neighboring nodes
- If two nodes have an arc between them, obviously they are not independent

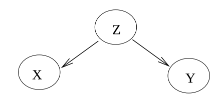
Indirect connections





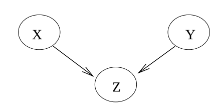
- If we do not know the value of Z, then knowing X can help us compute Y and vice versa
- But if we know Z, then knowing X does not influence what we believe about Y (because we have the more direct influence of Z)
- So X and Y are conditionally independent given Z

Common cause



- If we do not know the value of Z, then knowing X can help us compute Y and vice versa
- But if we know Z, X and Y are conditionally independent, because of the Markov assumption about the Bayes net structure

Common effect



- This is called a **v-structure**
- If we do not know anything about Z, X and Y are independent (see, e.g. $Earthquake$ and $Burglary$ in the alarm network example
- But if we know Z, then knowing something about X influences the belief about Y (through "explaining away")
- In this case, X and Y are **not conditionally independent** given Z.

D-separation in general

Let G be a Bayes net structure and let $X_1 - \ldots - X_n$ be an undirected path in G. Let Z be a subset of nodes. The path $X_1 - \ldots - X_n$ is active given evidence Z if:

- Whenever we have a v-structure $X_{i-1} - X_i - X_{i+1}$, then X_i or one of its descendents is in Z
- No other node along the path is in Z.

We say that X and Y are **d-separated given** Z, denoted $d\text{-}sep_G(X, Y|Z) = yes$, if there is no active path between X and Y given Z.

D-separation algorithm

To determine whether $d\text{-}sep_G(X, Y|Z)$, we need to enumerate all paths between X and Y and check that they are all blocked.

This can be done efficiently:

1. Traverse the graph bottom-up and mark all the nodes that are in Z or have descendents in Z. These can potentially enable v-structures
2. Do a depth-first search from X to Y, backtracking when a node is blocked. A node is blocked if either:
   (a) It is the "middle" of a v-structure and is not marked
   (b) It is in Z and it does not satisfy (a)
3. If the depth-first search succeeds, then there is an active path and $d\text{-}sep_G(X, Y|Z) = no$. Otherwise, return yes.

Soundness

Theorem: If G is an i-map of P and $d\text{-}sep_G(X, Y|Z) = yes$, then P satisfies $I(X, Y|Z)$.

Informally, any independence reported by d-separation is satisfied by the underlying distribution.

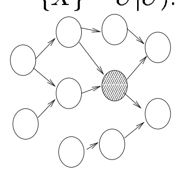
Completeness

Theorem: If $d\text{-}sep_G(X, Y|Z) = no$, then there exists a distribution P such that G is an i-map of P and P does not satisfy $I(X, Y|Z)$.

Informally, any independence not reported by the d-separation might be violated by the underlying distribution. The graph structure alone is not sufficient to determine if this is the case.

Markov blanket

Consider a node in $G = (V, E)$. Suppose we want the smallest set of nodes U such that X is independent of all other nodes in the network given U: $I(X, V - \{X\} - U|U)$. What should U be?



- Clearly, at least X's parents and children should be in U
- But this is not enough to clock v-structures; U sill also have to include X's "spouses" - i.e. the other parents of X's children

The set U consisting of X's parents, children and other parents of his children is called the **Markov blanket** of X.

Moral graphs

Given a DAG G, we define the **moral graph of** G to be an undirected graph U over the same set of vertices, such that the edge (X, Y) is in U if X is in Y's Markov blanket

- If G is an i=map of P, then U will also be an i-map of P
- But many independencies are lost when going to a moral graph

D-maps

- A graph G is a **dependency map (d-map)** of probability distribution P if $I(X, Y|Z) \rightarrow Z$ d-separates X and Y.
- Intuitively, a d-map guarantees that connected variables are indeed dependent
- This is the converse of the i-map property, which guarantees that disconnected variables are indeed independent.
- An empty graph is trivially a d-map for any probability distribution
- A complete graph is trivially an i-map for any probability distribution
- Can we get a graph that satisfies both properties?

Perfect maps

A DAG G is a **perfect map** of a distribution P if and only if it is both an i-map and a d-map. That is:

$$I(X, Y|Z \leftrightarrow d\text{-}sep(_G X, Y|Z)$$

- A perfect map captures all the independencies of a distribution
- Perfect maps are unique, up to DAG equivalence
- How can we construct a perfect map for a distribution?

Some distributions do not have perfect maps!

Example: We have two independent unbiased coins that we toss. If both coins come up the same, a bell rings with probability 2/3.

Here, there are three minimal i-maps (which?) but none is a perfect map.

Constructing Bayes nets in practice

Usually, we do not construct Bayes nets based on knowledge of the joint probability distribution P. We have some vague idea of the dependencies in the world, and we need to make that precise in a Bayes net. This involves several steps:

- Formulating the problem
- Choosing random variables
- Choosing independence relations
- Assigning probabilities in the CPDs

Example: Icy road

Note: This is taken from Nir Friedman's slides

Inspector Lestrade is awaiting his two colleagues Sherlock Holmes and Dr. Watson. He also wants to go to lunch. Having heard that Holmes has been in a car crash, he says: "Good. The road is probably coated with ice, so Watson will also crash his car." So he goes off to lunch.

How do we model this reasoning?

First step: formulate the question in probabilistic terms:

We want $P(\text{Watson crash} \mid \text{Holmes crash})$

Example: Choosing the variables

We need all random variables relevant to the problem (including those not in the evidence or the query):

- Ice - is there ice on the road?
- Holmes - has Holmes' car crashed?
- Watson - has Watson's car crashed?

In real life, we would also have to decide if the variables should have more than two values, or be continuous.

We need to make sure that we include in the Bayes net all variables that could cause "explaining away' patterns.

E.g. Could Holmes have been drunk? That would decrease the probability of the road being icy and of Watson crashing

Choosing random variables

- Variables must be *precise*. What are the values, how are they defined, and how are they measured?
  E.g. *Heart-attack* vs. *Risk-of-heart-attack*
- If the variables are continuous and we discretize them, a coarse discretization may introduce additional dependencies.
  E.g. cholesterol example
- There several kinds of variables:
  - Observable
  - Sometimes observable (e.g. medical tests)
  - Hidden - these may or may not be useful to include, depending on the other independencies that they generate

Example: Choosing the structure

- It seems that I influences both H and W.
- But should there be a more direct connection between H and W?

Choosing the structure

- Causal connections tend to make the graphs sparser. Note that causality is judged in the world, not in our inference process.
  E.g. Suppose you are drawing a Bayes net for an insurance company and you have two random variables:
  previous-accident and is-good-driver.
  In the world, the quality of the driver influences whether he/she has accidents. But the company would think about the causality in the other direction
- In general, these models are approximate. There is a trade-off between precision and the size and sparsity of the graph.

Example: Choosing the probabilities

- The probability of an icy road can be estimated based on local weather data
- The conditional probabilities should be estimated by someone who knows their driving skills (e.g. Lestrade)

Choosing numbers for the CPDs

- Conditional probabilities could come from a few sources:
  - An expert
    - People hate picking numbers!
    - Having a good network structure usually makes it easier to elicit numbers from people too.
  - An approximate analysis (e.g. in card games)
  - Guessing
  - Learning
- Bad news: In all these cases, the numbers are approximate!
- Good news: the numbers usually do not matter all that much.
- Sensitivity analysis can help in deciding whether certain numbers are critical or not for the conclusions

Important factors when choosing probabilities

- Avoid assigning zero probability to any events
- The relative values (or ordering) of conditional probabilities for $P(X|Y)$, given different values of Y is important
- Having probabilities that are orders of magnitude different can cause problems in the network

Summary

- A Bayes net represents a probability distribution using two components: a DAG G and a collection of conditional probability distributions $P(X_i|Parents(X_i))$.
- An additional requirement is that G is a minimal i-map of P
- All independencies implied by a Bayes net can be computed efficiently using the d-separation algorithm
- Perfect maps are (in some sense) the best representation of a distribution, but some distributions do not have them.