Lecture 4: Bayesian Networks

Rachael Briggs

July 31, 2014

Random Variables

- Uppercase letters $A, B, C, \ldots$ stand for random variables (e.g., the price of a pot of tea, your height in centimetres).
- Lowercase letters $a, b, c, \ldots$ stand for values of random variables (e.g., €2.50 as the price of a pot of tea; 165cm as your height in centimetres.)

  I'll also use lowercase letters to designate the event of a variable taking on a particular value (e.g., that a cup of tea costs €2.50, that you are 165cm tall).
- Bold uppercase letters $\mathbf{A}, \mathbf{B}, \mathbf{C}, \ldots$ stand for sets of random variables (e.g., {how a particular coin landed, how a particular die landed}).
- Bold lowercase letters $\mathbf{a}, \mathbf{b}, \mathbf{c} \ldots$ stand for sets of values for sets of random variables (e.g. {the coin landed heads; the die landed six}).
- I'll use $P(A_1, A_2, \ldots A_n)$ to stand for a joint probability distribution over values of $\{A_1, A_2, \ldots A_n\}$.
- I'll use $P(A_1, A_2 \ldots A_n | B_1, B_2 \ldots B_m)$ to stand for a set of conditional probability distributions: one distribution over values of $A_1, A_2 \ldots A_n$ conditional on $B_1 = b_1 \wedge B_2 = b_2 \wedge \ldots \wedge B_n = b_n$, for each possible combination of values $b_1, b_2 \ldots b_n$.
- The **product** $P(\mathbf{A}|\mathbf{C})P(\mathbf{B}|\mathbf{C})$ is a set of joint (conditional) probability distributions that assigns to each $\mathbf{a}$, $\mathbf{b}$, and $\mathbf{c}$, the probability

  $$P(\mathbf{a}, \mathbf{b}|\mathbf{c}) = P(\mathbf{a}|\mathbf{c})P(\mathbf{b}|\mathbf{c})$$
- I will say that **A** and **B** are **independent** conditional on **C**, that **C screens off B** from **A**, or that $(\mathbf{A} \perp \mathbf{B} | \mathbf{C})$ iff for every set of values **a** for **A**, **b** for **B**, and **C** for **c**,

  $$P(\mathbf{a}|\mathbf{b}, \mathbf{c}) = P(\mathbf{a}|\mathbf{b})$$

So: what is all this vocabulary going to do for us?

Encoding Information

Suppose I want to describe the relationship between two random variables: language aptitude L (let's say that this takes either a "high" value l^1 or a "low" value l^0) and score S on a standardised language test (which also takes either a "high" value s^1 or a "low" value s^0) . One way to do it is with a joint probability distribution [Koller and Friedman, 2009]:

L	S	$P(L,S)$
l^0	s^0	0.665
l^0	s^1	0.035
l^1	s^0	0.06
l^1	s^1	0.24

Another option is to use

a distribution for L:

l^0	l^1
0.7	0.3

and a pair of conditional distributions for S given L:

L	s^0	s^1
l^0	0.95	0.05
l^1	0.2	0.8

The Example: More Complicated Version

An employer is selecting job candidates partly on the basis of language aptitude L (which can be high or low). They have three kinds of data: standardised test scores S (which can be high or low), letters of recommendation R (which can be positive or lukewarm), and grades G (which can be A, B, or C). Test scores and grades are affected by language aptitude. Grades also depend on the difficulty D of the language course the candidates have taken (this can be high or low). Furthermore, it is known that the letter-writers are all lazy: they don't remember their students, but just look at the students' grades, and base their letters entirely on the grades.

If we wanted to completely specify a probability distribution over these variables, we would need a table with 96 entries! Luckily, we can encode the information from this example using a graph (see slides). Notice:

- **Nodes** represent random variables.
- **Edges** represent relations of direct dependence.
- A is a **parent** of B iff B directly depends on A. **Child** is the inverse of **parent**; **ancestor** is the ancestral of **parent** and **descendant** is the ancestral of **child**.

Markov Condition Where $\mathbf{PAR}(A)$ is the set of A's parents, then for any $\mathbf{B}$ disjoint from $\mathbf{PAR}(A)$ and not containing any descendants of A,

$P(A|\mathbf{B}, \mathbf{PAR}(A)) = P(A|\mathbf{PAR}(A))$

Chain Rule $P(A_1, A_2 \ldots A_n) = \Pi_{i=1}^{n}(P(A_i|\mathbf{PAR}(A_i)))$

D-Separation

Given a probability distribution and a graph that satisfies the Markov Condition for that distribution, when are two variables independent?

A sufficient condition: when they are **d-separated**.

- A **path** in a graph is a set of arrows such that the first has a node in common with the second, the second has a node in common the third…, and so on. (In this definition, it doesn't matter which way the arrows are pointing.)
- Suppose we have a path from X, through Z, to Y. There are only four ways this path can possibly be.
  1. $X \rightarrow Z \rightarrow Y$
     (X is an indirect cause of Y, via Z)
  2. $Y \rightarrow Z \rightarrow X$
     (Y is an indirect cause of X, via Z)
  3. $X \leftarrow Z \rightarrow Y$
     (Z is a common cause of X and Y)
  4. $X \rightarrow Z \leftarrow Y$
     (Z is a common effect of X and Y—this kind of connection is often called a **collider**.)
- A node on a path is **active** iff either it belongs to type 1-3 and its value is unknown, or it belongs to type 4 and its value or the value of one of its descendants is known. (You can think of active nodes as allowing information to pass through.)
- A path from X to Y is **active** iff all its nodes are active.
- Two (disjoint) sets of variables **X** and **Y** are **d-separated** iff there is no active path from any variable in **X** to a variable in **Y**.
- When the Markov Condition holds, any two d-separated sets of variables are guaranteed to independent. There are cases of independence without d-separation, but for every graph, there is some probability distribution that is Markovian relative to the graph, such that only the d-separated variables are independent.

From Correlation to Causation

- So far, I've given a statistical interpretation of these graphs. (I was a little sloppy with the word "cause" in the section on d-separation.) But the arrows can be given a causal interpretation. We should think that direct causal dependencies will be reflected by statistical dependencies, in the way expressed by the Markov Condition. (This rule is often called the **Causal Markov Condition**.)

- We can replace the probabilistic connections in a Bayesian network with deterministic causal connections, plus unknown "latent" variables that generate noise (though we don't have to).

- Instead of marginal probability distributions, we can introduce deterministic structural equations.

  $A = f(\mathbf{PAR}(A))$ says that A's value is determined by of the values of A's parents according to function f.

  Note: this structural equation is asymmetric. $A = B$ says that A's value is counterfactually determined by B's value; $B = A$ says that B's value is counterfactually determined by A's value.

- There's a key difference between *observing* an event, and *intervening* to make an event happen.

  - Suppose that, on my evidence, if I learn that a child goes to a private school, I should become more confident that the child will get a good grade on their university entrance exam. Is sending my child to a private school an effective way to improve the child's grade?
  - Simpson's paradox case [Koller and Friedman, 2009]: Suppose that taking drug D is correlated with recovering from nasty medical condition M, but anti-correlated with recovery among men and among women. Should a conscientious doctor prescribe drug D to her patients?

		cured	**sick**
men	**drug**	21 (70%)	9 (30%)
	no drug	8 (80%)	2 (20 %)
women	**drug**	2 (20%)	8 (80%)
	no drug	12 (40%)	18 (60%)
total	**drug**	23 (57.5%)	17 (42.5%)
	no drug	20 (50%)	20 (50%)

- Your probability for $\mathbf{x}$ conditional on *observing* $\mathbf{y}$ should be the traditional

  $$P(\mathbf{x}|\mathbf{y}) = \frac{P(\mathbf{x}|\mathbf{y})}{P(\mathbf{y})}$$

  But your probability for $\mathbf{x}$ conditional on *causing* $\mathbf{y}$ should be

  $$P(\mathbf{x}|do(\mathbf{y})) = P_{\mathbf{X}}(\mathbf{y})$$

  where $P_{\mathbf{X}}$ is the probability distribution in a *submodel* generated by replacing the structural equation for $\mathbf{X}$ with a new equation that sets $\mathbf{X} = \mathbf{x}$ (or by replacing the conditional probability distribution for $\mathbf{X}$ on its parents with a probability distribution that assigns 1 to $\mathbf{X} = \mathbf{x}$).

Actual Cause

We often want to know when one event $C = c$ caused another $E = e$ in a particular instance. Did the defendant's negligence cause the plaintiff's death? Did the fish I ate for dinner make me sick? Did Grandpa's smoking cause Grandpa's lung cancer? It would be nice to give an answer in terms of counterfactual dependencies (e.g., it seems relevant to determine whether Grandpa would've still had cancer if he hadn't smoked). Here's a popular theory that predates causal models:

Transitive Closure of Counterfactual Dependence $C_0 = c_0$ is a cause of $E = e$ iff

> $C = c$ and $E = e$, and either
>
> > $E = e$ counterfactually depends on $C_0 = c_0$,
> >
> > or there is a chain of events $C_1 = c_1$, $C_2 = c_2 \ldots C_n = c_n$ such that $E = e$ counterfactually depends on $C_n = c_n$ and for each C_i, $C_i = c_i$ counterfactually depends on $C_{i-1} = c_{i-1}$.
>
> [Lewis, 2000]

But there are counterexamples!

Pre-Emption Suzy is teaching Billy to smash windows, starting with the window of their next-door neighbour. Billy aims a rock at the window. Suzy stands by with a baseball bat to swing at the window, in the event that Billy fails to break it. But Billy hits the window and shatters it.

Intransitive Causation Hiking in a mountain pass, I see a boulder hurtling toward me, and duck to avoid it. By ducking, I manage to survive—if I hadn't ducked, I would be dead. But the boulder falling did not cause me to survive [Hitchcock, 2001]

First Pass $C = c$ causes $E = e$ iff

> $C = c$ and $E = e$
>
> and there is at least one route from C to E for which an intervention on C will change the value of E, given that other parents of C that are not on this route have been fixed at their actual values.

Unfortunately, there are still counterexamples.

Overdetermination The cafeteria at LMU is good about taking feedback. If at least one person threatens to sue them over their horrible eel pie, they will stop serving it. Catrin and I both threaten to sue. The cafeteria duly withdraws the eel pie.

Definition of Actual Cause Suppose we have variables C and E, a path ϕ from C to E, and some variables $V_1 \ldots V_n$ not on ϕ. The values $v_1 \ldots v_n$ are in the **redundancy range** for $V_1 \ldots V_n$ if,

> given the actual value of C,
>
> there is no intervention that in setting the values of $V_1 \ldots V_n$ to $v_1 \ldots v_n$, will change the (actual) value of E.

Then $C = c$ causes $E = e$ iff

> $C = c$ and $E = e$
>
> and for at least one directed path ϕ from C to E, and way of fixing by interventions all parents of E that do not lie along ϕ at some combination of values within their redundancy range, there is an intervention on C that will cause a change in the value of E [Woodward, 2003].

References

Christopher Hitchcock. The Intransitivity of Causation Revealed in Equations and Graphs. *The Journal of Philosophy*, 98(6):273, June 2001.

Daphne Koller and Nir Friedman. *Probabilistic Graphical Models*. Principles and Techniques. MIT Press, 2009.

David Lewis. Causation as Influence. *The Journal of Philosophy*, 97(4):182, April 2000.

James Woodward. *Making things happen: A theory of causal explanation*. 2003.