From conditional probabilities to causal Bayesian networks

Contents

Directed acyclic graphs are becoming an increasingly popular tool to describe and analyze causal knowledge. This short primer, which is based on the first 3 chapters of Judea Pearl's book "Causality" [1] introduced some basic concepts.

Conditional probabilities and the chain rule

While causal knowledge encompasses more than associations between variables, causal knowledge is also based on knowledge about the relationship of different variables. Such relationships can be expressed in the language of probability. The axioms of probability calculus are:

$$0 \leq P(A) \leq 1, \tag{1}$$

$$P(\text{sure proposition}) = 1, \tag{2}$$

$$P(A \text{ or } B) = P(A) + P(B) \text{ if } A \text{ and } B \text{ are mutually exclusive} \tag{3}$$

In probability theory, conditional probabilities are used to express the probability of an event A (or a variable value) given knowledge about another event or value of another variable B:

$$P(A|B) = \frac{P(A,B)}{P(B)} \tag{4}$$

That is, the probability of A given B is equal to the joint probability of A an B divided the probability of B.

By rearranging (4) one obtains the product rule, which calculates the joint probability of A and B from knowledge about the conditional probability and the unconditional probability.

$$P(A,B) = P(A|B)P(B) \tag{5}$$

The product rule (5) can be generalized to express the joint probability of more than two events E.

$$P(E_1, E_2, ..., E_n) = P(E_n|E_{n-1}, ..., E_2, E_1)...P(E_2|E_1) \tag{6}$$

A probabilistic model $P(S)$ encodes information about the relationship of events (or values of variables) S in a way that is consistent with the axioms of probability (1)-(3). That is, the sum the probabilities of all possible combinations of events must be 1.

[1] pdfs here: http://bayes.cs.ucla.edu/BOOK-2K/ch1-1.pdf, http://bayes.cs.ucla.edu/BOOK-2K/ch1-2.pdf, http://bayes.cs.ucla.edu/BOOK-2K/ch1-3.pdf

Conditional independence

Even probability models with a moderately large number of event classes or variables become quickly unwieldy when all variables are mutually dependent. Probability models are simpler if some variables X and Y are independent given knowledge of a third variable Z. Using x, y and z as realizations of X, Y and Z, X and Y are conditionally independent if

$$P(x|y,z) = P(x|z) \text{ whenever } P(y,z) > 0 \tag{7}$$

that is, if we already know Z, knowledge about Y will not provide any additional information about the value of X.

Directed acyclic graphs

A graph consists of a set of nodes which represent variables, that are connected by edges.

A directed graph has only directed edges, where the direction of an edges indicates the cause and effect. Bi-directed edges indicated existence of a common, unobserved causes for two nodes.

A path goes from node to node by following along edges. Directed paths are special cases of paths in which each step from node to node ends in an arrow. Nodes with a path between them are connected and those without a path between them are disconnected.

Graphs that contain directed paths that end in themselves (cycles) are cyclic graphs. Directed acyclic graphs are characterized by the absence of cycles.

The relationship between nodes can be described with pedigree-terms. Immediate predecessors of a node in a DAG are called *parents* and immediate successors are called *children*, parents with a common child are called *spouses*, and a family consists of a *child* node and its parents. The group of all node preceding a child is called *ancestors*. An exogenous node without parents, is called a *root* and node without children a *sink*. Trees are connected DAGs in which every child has at most one parent, and if every node also has only one child it is called a chain.

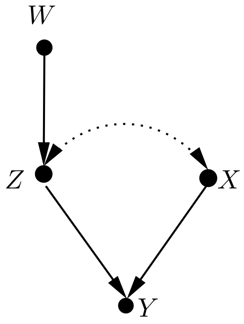

Figure 1: A simple DAG

Bayesian networks

Directed graphs are used to describe assumptions, to represent joint probability functions in an economical manner, and to facilitate the analysis of the former two. Given these functions, DAGs belog into the class of *Bayesian networks*, a term coined to emphasize

> (1) The subjective nature of input information; (2) the reliance on Bayes' conditioning as the basis for updating information; (3) the distinction between causal and evidential modes of reasoning.

Bayesian networks facilitate analysis of DAGs by using the concept of *Markovian parents*, which are defined as the group of parents PA_j that render the child X_j independent of all other variables in a DAG except the *Markovian parents*. Without the property of *Markovian parents*, the joint distribution would need to be calculated as

$$P(x_1, ..., x_n) = \prod_j P(x_j | x_1, ..., x_{j-1}). \tag{8}$$

With Markovian parents, a DAG can be constructed following the following simple algorithm:

Algorithm 1 Drawing a Bayesian network with nodes x_j using Markovian parents pa_j

```
for $x_j = 1, 2, \ldots, N$ do
    for $x_i \in pa_j$ do
        draw an edge from $x_i$ to $x_j$
    end for
end for
```

and the calculation of the probability distribution for a DAG simplifies to

$$P(x_1, ..., x_n) = \prod_j P(x_j | pa_j). \tag{9}$$

A DAG G can only be a Bayesian network of a probability distribution P, if P can be decomposed as in (9). That is, it is not possible to construct a DAG that is a Bayesian network for a probability distribution P, if parents in P are not Markovian parents[2]. If a graph P can be factorized with a function (9) that was derived from a graph G, it is said that G represents P, or that G and P are *markov relatives* or *markov compatible*.

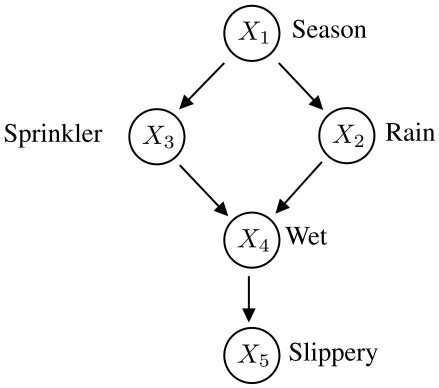

Figure 2: A Bayesian network

Using the probability distribution for the DAG / Bayesian network in Figure 2 can be calculated as

$$P(x_1, x_2, x_3, x_4, x_5) = P(x_1)P(x_2|x_1)P(x_3|x_1)P(x_4|x_2, x_3)P(x_5|x_4) \tag{10}$$

[2] Right now I cannot come up with an example P or DAG that does not have Markovian parents. One possibility would be a fully connected network, in which a node is connected to all other nodes, so that it cannot be made independent of any parents, but then one could just add all nodes in the set PA_j for this child.

d-Separation

d-Separation refers to a situation in which a path between two variables X and Y is closed after conditioning on variable(s) Z.

Paths are blocked when conditioning on a variable (here Z) in a chain ($X \rightarrow Z \rightarrow Y$) or a fork ($X \leftarrow Z \rightarrow Y$). In contrast, a path with an inverted fork ($X \rightarrow Z \leftarrow Y$) is already blocked, and conditioning would open that path. Synonyms for "blocking" a path are interrupting the flow of information in a graph, or rendering to previously dependent variables independent conditional on Z.

Conditional independence $(X \perp\!\!\!\perp Y|Z)_P$ and d-separation $(X \perp\!\!\!\perp Y|Z)_G$ are "mirror-concepts", as the former refers to independence in a probability model P and the latter to independence in a graph G.

Causal Bayesian Networks

Whereas Bayesian Networks encode independence assumptions, they do not need to be interpreted causally. However, the step by step construction of variables in a graph described in Algorithm 1 can be seen as a data-generating process, which shows why a causal interpretation of Bayesian networks is attractive and intuitive. According to Pearl,

> conditional independence judgement are byproducts of stored causal relationships

and therefore

> tapping and representing those relationships directly [is] a more natural and more reliable way of expressing what we know about the world.

One important advantage of Bayesian Networks is that they have a high degree of flexibility. this means that it is easy to adapt them (compared to adapting a probability model P) to changed causal models by adding or removing edges. This flexibility is due to the assumption that changes can be implemented locally, i.e. by considering only (*Markov*) child (x_j) parent (pa_j) pairs and by ignoring more distant ancestors and descendants (Which I think works due to the conditional independence given PA_j).

Declaring a Bayesian Causal Network a causal model (as opposed to an observational model) opens the possibility to investigate interventions, i.e. setting variables to specific values. For instance, to model a situation in which the sprinkler is always on, one changes the graph in Figure 2 by deleting the arrow from X_1 to X_3 and by setting the sprinkler to "On". The new probability distribution then will be:

$$P_{X_3=\text{On}}(x_1, x_2, x_4, x_5) = P(x_1)P(x_2|x_1)P(x_4|x_2, X_3 = \text{On})P(x_5|x_4) \tag{11}$$

Deleting the path $X_1 \rightarrow X_3$ reflects the fact that once X_3 is set to a specific value, it is independent of X_1. Setting a variable to a specific value is done with the do operator, where

$$do(X = x)$$

indicates that variable X is set to a specific value x. Differently than *doing*, *observing* the effect of $X = On$ involves calculating conditional probabilities from the full graph.

Of course, the new-gained ability of Causal Bayesian Networks to predict the results of interventions is not for free. The price is the assumption that paths in a DAG depict causal relationships.

The notion of interventions plays an important role in the formal definition of *Causal Bayesian Network*. We define $P(V)$ as a probability distribution over variables V, $P_x(V)$ as the probability distribution for an intervention in which $do(X = x)$ sets a subset $X \subseteq V$ to x, and $\boldsymbol{P_*}$ as the set of possible $P_x(V)$. Then, DAG G is a causal Bayesian network if an only if:

1. $P_x(v)$ is Markov relative to G
2. $P_x(v_i) = 1$ for all $V_i \in X$ whenenver v_i is consistent with $X = x$

3. $P_x(v_i|pa_i) = P(v_i|pa_i) textrmforallV_i \notin X$ whenenver pa_i is consistent with $X = x$

if these conditions hold, the effect of an intervention $do(X = x)$ can be calculated as

$$P_x(v) = \prod_{i|V_i \notin X} P(v_i|pa_i) \text{ for all } v \text{ consistent with } x \tag{12}$$

From this, two things can be derived:

The conditional probability $P(v_i|pa_i)$ has the same value as the effect of setting parents to a specific value $P_{pa_i}(v_i)$. Put differently, the effect of a variable having a certain value is the same when the value was observed or when it was set.

$$P(v_i|pa_i) = P_{pa_i}(v_i) \tag{13}$$

When one controls the direct causes of a child P_{pa_i} other variables s have no influence on a child:

$$P_{pa_i,S} = P_{pa_i} \tag{14}$$