J. Pearl: Causality, Chapter One

Introduction to Probabilities, Graphs and Causal Models

a summary by Severin Hacker

Throughout the whole chapter we are dealing with two important concepts: causality and probabilities. Causality means lawlike necessity ("men die"). Probabilities mean exceptionality, doubt and lack of regularity ("who wins the football world championship?"). There are two reasons why we do a probabilistic analysis of causality: (a) in natural language we often use causal expressions in situations where uncertainty is predominant ("you will fail the course because of your laziness") and (b) even the most assertive causal expressions are subject to exceptions ("Brazil will win the world championship"). Probability theory tolerates unexplicated exceptions and allows us to focus on the main issues of causality.

We will use the Bayesian interpretation of probability which states that probabilities are used to measure our degree of belief in certain propositions. For example, if A is the statement "Ted Kennedy will seek the nomination for president in year 2004" then $P(A|K)$ measures a person's subjective belief in A given a body of background information (or context) K. Everything the person knows about American politics, about Ted Kennedy and about the state of the world in general can be summarized in K. Bayesian philosophers see the conditional relationship as more basic than that of joint events and more compatible with the human mind ("How likely is it to see this *given* I see that" and not "How likely is it to see this *and* that").

A probabilistic model is an encoding of information that permits us to compute the probability of every statement S. Since every Boolean formula can be expressed as a disjunction of elementary events and since the elementary events are mutually exclusive and we know their probabilities by the joint probability function we can always compute the probability of every statement S using the additivity axiom. Thus, any joint probability function represents a complete probabilistic model. Joint probability functions allow us to check whether we have sufficient information (the probability of each elementary event can be inferred) and whether the information is consistent (the probabilities add up to 1). However, in practice they are rarely specified explicitly but through indirect representations. One such representation are Bayesian networks.

Before we deal with vertices and edges we should make clear what conditional independence means. X is *conditionally independent* of Y given Z if $P(x|y,z) = P(x|z)$ (when $P(y,z) > 0$). That is - as soon as we know Z we cannot find out more about X if someone tells us Y. Information about Y becomes irrelevant for determining X. The graphoid axioms, several more or less surprising properties of conditional independence, can be proofed by the basic axioms of probability theory.

We use Bayesian networks (a) to provide convenient (graphical) means of expressing assumptions (b) to facilitate shorter representations of joint proba-

bility functions and (c) to facilitate efficient inferences from observations. To store $P(x_1, ..., x_n)$ for n binary variables would require 2^n entries. If we know that some variable X_j is not sensitive to all predecessors of X_j but only to a small subset PA_j then we can considerably simplify the input information required to build the joint probability function. The set PA_j is then called the *Markovian parents* of X_j. If a probability function P admits this simplification trick relative to a DAG G, which dictates the sets PA_j, we say that P is *Markov relative* to G.

To test whether X is independent of Y given Z in any distribution compatible with G we need to test whether Z *blocks* all paths from X to Y (X,Y,Z are sets). This can be formalized with the *d-separation-criterion*: p is blocked by Z if and only if (a) p contains a chain $i \rightarrow m \rightarrow j$ or a fork $i \leftarrow m \rightarrow j$ such that m is in Z or (b) p contains an inverted fork $i \rightarrow m \leftarrow j$ and neither m is in Z nor any descendant of m is in Z. (a) is rather intuitive, if I tell you a direct cause m of an event j no one is really interested about indirect causes or other consequences of m which means you can forget about i . (b) is a bit tricky: if I tell you a consequence m (e.g. headache) which has two originally independent causes i and j (flu or depression) and if I tell you that you have no flu then the other cause (depression) becomes far more likely and so i and j become dependent. And if we want to avoid that they become dependent, we are not allowed to know m or any of its descendants.

Now we can connect our two worlds d-separation and conditional independence. If sets X and Y are d-separated by Z in a DAG G, then X is independent of Y conditional on Z in every distribution compatible with G. If they are not d-separated by Z in G then X and Y are dependent conditional on Z in at least one distribution compatible with G. A fundamental insight is that for determining whether a given probability P is Markov relative to a given DAG G it is necessary and sufficient that every variable be independent of all its nondescendants conditional on its parents.

As we have seen there exist many DAG's G that are Markov relative to a given P (for each ordering of variables in P there exists a DAG G). However, we are primarily interested in causal interpretation of data, therefore we do not consider variable oderings that do not respect time and causation. Pearl argues that causal relationships are more robust to change. An example is the causal relationship S_1 "Turning the sprinkler on would not affect the rain" versus its probabilistic counterpart S_2 "The state of the sprinkler is independent of the state of the rain." First, S_2 changes from false to ture when we learn what season it is (this corresponds to the (a) case in the d-separation-criterion). Second, S_2 changes from true to false when we know what season it is and we learn that the pavement is wet (which corresponds to the (b) case in the d-separation criterion). On the other hand S_1 remains true regardless of what we learn or know.

Finally, there are two conceptions of causality: (a) a stochastic one which believes that nature's law are inherently probabilistic and (b) a deterministic one which believes that randomness is due to human ignorance of underlying boundary conditions. Pearl is a strong advocate of the deterministic conception

(I'm not). He then introduces functional models/structural equations to convince us that the Laplacian model (b) is superior. One rather strange feature of functional models is that you can compute probabilities of counterfactuals. An example: Joe has taken the treatment and died, we ask for the probability Q that Joe would have died had he not been treated. Why should we ever ask such things? (One is for sure, Joe does not care at all).