Notes POC1

Artur Gaspar

May 8, 2024

Contents

1 Causality Chapter 1: Introduction **2**

1.1 Section 1.1: Introduction and review 2

1.1.1 Odds and likelyhood 2

1.1.2 Coariance, Correlation, Regression Coefficient 2

1.1.3 Axioms 3

1.1.4 Counter example for third axiom 3

1.1.5 Why conditionals are enough to specify independence 4

1.2 Section 1.2: Bayesian Networks 4

1.2.1 1.2.1) Conventions 4

1.2.2 1.2.2) Bayesian Networks 4

1.2.3 Instability of parents for non-positive distributions 5

1.2.4 1.2.3) d-separation 5

1.2.5 1.2.4) Inference with BNs 6

1.3 Section 1.3: BNs with causal directions 6

1.4 Section 1.4: Counterfactuals 6

1.4.1 Laplacian vs Stochastic model 6

1.4.2 1.4.1: Structural Equations 6

1.4.3 1.4.2: Probabilistc Predictions (and Definitions and equivalences between SCMs and BNs) 7

1.4.4 1.4.3: Interventions 7

1.4.5 1.4.4: Counterfactuals 8

1.4.6 Questions and confusions 8

1.5 Section 1.5: Some terminology 9

Chapter 1

Causality Chapter 1: Introduction

1.1 Section 1.1: Introduction and review

1.1.1 Odds and likelyhood

Odds are the fraction of probabilities. **Prior (predictive/prospective) odds** is $\frac{p(H)}{p(\neg H)}$, and the **Posterior (diagnostic/retrospective) odds** is $\frac{p(H|e)}{p(\neg H|e)}$. This is how much more likely the hypothesis is to be true than false a priori and after observing the event e.

The **Likelyhood Ratio (Risk Ratio for epidemology)** is $\frac{p(e|H)}{p(e|\neg H)}$, remembering that Likelyhood is a function of B in $p(A|B)$, while the probability is a function of A.

The formula is: Posterior Odds = Prior Odds $\times$ Likelyhood Ratio.

My interpretation of $p(H|e)$ is the probability we give to H in the world where e happens, thus if we do $\frac{p(H|e)}{\neg H|e}$ we're seeing how more probable (multiplicatively) H is to be true in this world, and if we do $\frac{p(e|H)}{p(e|\neg H)}$ we're seeing how much more likely is the event e to happen in the world in which H is true than in the world in which it's not (it's a comparison accros worlds).

I interpret the likelyhood ratio as how many more times the evidence appears in the world where H is true than in the world where it's not.

So how more likely the hypothesis is to be true than false, after we observe the event = how more likely the hypothesis was to be true before the observation was made times $\times$ how many more times the evidence appears in the world where H is true than in the world where it's not.

Odds of hypothesis after e = odds before e $\times$ how much more e happens in H than in $\neg H$.

1.1.2 Coariance, Correlation, Regression Coefficient

Covariance is the expected value of $(X - E[X])(Y - E[Y])$, distance to the averages, $cov(X, X) = var(X) = (std(X))^2$, and **Correlation** is $corr(X, Y) = \frac{cov(X,Y)}{std(X)std(Y)}$.

Regression coefficient when estimating Y using X is $corr(X, Y) \times \frac{std(Y)}{std(X)}$, which is how much Y will change by unity of X we change, if we use the line that minimizes the quadratic error of the Y estimate. I kind of interpret this as $\frac{\text{(X-unities)}}{\text{(X-unities per standard devition of X)}} \times corr(X, Y) \times std(Y) =$

(number of standard deviations of X)$\times corr(X,Y)\times std(Y)$ = (number of standard deviations of Y)$\times$ (Y-unities per standard deviations of Y) = (Y-unities). The strange thing with this interpreatation is that $corr(X,Y) = (\frac{\text{standard deviations of X}}{\text{standard devations of Y}}) = \frac{\text{standard deviations of Y}}{\text{standard deviations of X}}$, is the function that given one ammount of standard deviations returns the other one... Maybe this is a reflection of the limitations of the linearity assumption?

1.1.3 Axioms

Finally, the graphoid axioms for independence of random variables (all of them conditioned on Z, and I simplified a little bit):

1. **Symmetry**: X is independent of Y iff Y is independent of X
2. **Decomposition, Weak Union and Contraction**: X is independent of YW iff ((X is independent of Y) and (X is independent of W conditional on Y)). This is not how it's written in the book, but I think this single affirmation is equivalent to the Decomposition, Weak Union and Contraction axioms.
3. **Intersection** (only for strictly positive distributions): X is independent of W given Y and X independent of Y given W implies X independent of YW.

Summary:

1. Independence is symmetric.
2. Being independent from two things is equivalent to being independent to one alone and the other given the first one. In other words, being independent from two things means that looking at the value of one does'nt help and looking at the other after knowing the first doesn't help as well.
3. Being independent from two things (if nothing is impossible (?), maybe so we can condition on anything?) is the same as being independent from the first even if you know the second and being independent from the second even if you know the first.

If X is independent of Y, then $p(x|y,z) = p(x|z)$, we can ignore the irrelevant information.

1.1.4 Counter example for third axiom

The third axiom does not hold for instance in the following joint distribution (A, B, C are the random variables with two values each):

1. $p(a1, b1, c2) = \frac{1}{2}$.
2. $p(a2, b2, c1) = \frac{1}{2}$.
3. All other probabilities equal 0.

Then we have $p(a1|b1, c2) = 1 = p(a1|b1) = p(a1|c2) \neq p(a1) = \frac{1}{2}$ and $p(a2|b2, c1) = 1 = p(a2|b2) = p(a2|c1) \neq p(a2) = \frac{1}{2}$.

It doesn't make sense to talk about other conditional probabilities, as they are conditioned on something inexistent. We can say that A is independent of B given C, and A is independent of C given B, but A is not independent of BC.

This happens here because some values of BC are impossible, so we kind of know the value of C only by knowing B and vice-versa... So we know C iff we know B, and then after we learn one we don't need the other, but we can't ignore both.

If anything was possible, we would have $p(a1|c1) = p(a1|b1, c1) = p(a1|b1) = p(a1|b1, c2) = p(a1|c2) = k$, so $p(a1) = p(a1|c1)p(c1) + p(a1|c2)p(c2) = k(p(c1) + p(c2)) = k$, so the axiom follows.

1.1.5 Why conditionals are enough to specify independence

Just a disclaimer: $p(b1, c1) = 0 \to 0 = p(a, b1|c1) = p(a|c1) \times p(b1|c1) = p(a|c1) \times 0$, so to satisfy the independence we really just need to specify it for possible "worlds" (the values k that make $|k$ possible)...

If we write the independence with the "and" way, we get $p(a1, b1|c2) = 1 = p(a1|c1) \times p(b1|c2) = p(a1, c2|b1) = 1 = p(a1|b1) \times p(c2|b1)$, and the same for the other one, but it seems more complicated to me, the only advantage would be that we could write the zero parts, $0 = p(a1, b2|c2) = p(a1|c2) \times p(b2|c2) = 1 \times 0$. Let's try to use only the conditional version.

1.2 Section 1.2: Bayesian Networks

1.2.1 1.2.1) Conventions

skeleton of a graph is the undirected version of it.

In this book, a **path** might not follow the direction of the edges.

Family of a graph is a node and it's parents.

Root is a node without parents and **sink** a node without children.

Tree is a connected graph with at most one parent per node (one node can point to many but only one node can point to it), and **chain** is one with at most one child per node (onde node can point to only another one, but many can point to it).

1.2.2 1.2.2) Bayesian Networks

One of the main goals is to represent an joint distribution with less data, which is possible if every variable is independent of almost all others.

The **Markovian parents** of a node is a minimal set of nodes that, conditioned on them, the value of the node is independent from the value of all other nodes. It's a set of variables that we can condition on to ignore the rest when estimating the initial node, but such that we can't remove any variable from this set.

This set is unique if the joint distribution is strictly positive, and this implies an unique Bayesian Network.

I believe that if it's not strictly positive, then if we consider that the parents must be minimal, it might be impossible to draw a Bayesian Network, otherwise we accept non-minimal sets of parents *and acknoledge taht we might have more than one BN.* See the subsection "Instability of parents for non-positive distributions".

We say that G represents P, or G is compatible with P if we can decompose P with the information we extract from G (the DAG). For instance, $p(a1, b1, c1, d1, e1) = p(b1|a1)p(c1|a1)p(d1|b1, c1)p(e1|d1)$ is the decomposition for the graph with $A \to B$, $A \to C$, $B \to D$, $C \to D$ and $D \to E$.

1.2.3 Instability of parents for non-positive distributions

I think that it's possible to have more than one minimal set (of Markovian Parents) if the third graphoid axiom is not satisfied, because then we can have X independent of Y given Z and of Z given Y, but not on YZ. So we might want to require the distributions to be strictly positive...

Take for instance the following joint for A, B, C binary:

1. $p(a1, b1, c2) = \frac{1}{2}$.
2. $p(a2, b2, c1) = \frac{1}{2}$.
3. All other probabilities equal 0.

Here if we know one value we know the other two, so $\{B\}$ or $\{C\}$ are minimal markovial parents for A; $\{A\}$ or $\{B\}$ are minimal for C; and $\{A\}$ or $\{C\}$ are minimal for B. So, we kind of can't create an undirected graph that represents the dependencies well... One node will connect to other two, but it acually depends on only one (any one)...

1.2.4 1.2.3) d-separation

This is a criterion to extract the conditional independences between variables from the graph.

X, Y and Z here can be sets of more than one variable.

We say that a *path* is **d-separated** or **blocked** if either Z has a variable in the middle of the way or as a confounder, or the path has a collider which is not in Z and no descendent of the collider is in Z.

We say that Z d-separates X from Y if it does so for every path from X to Y.

It's really important to consider the descendent part! Conditioning on a variable unblocks every collider that is reachable in reverse order (following the arrows reversed) from this variable.

X and Y are d-separated by Z if and only if for all distributions compatible with the independencies of G, X and Y are conditionally independent given Z. Also, if they are not d-separated, almost all distributions make they dependent (they don't say "how much indepedent").

Selection bias, **Berskon's paradorx** or **explaining away effect** is the situation in which after conditioning on one variable we render two others dependent (knowing that one does not have a specific value lets us increase the chance of another, for instance).

Observational Equivalence is the situation in which we have two graphs such that any distribution compatible with one is also compatible with the other.

It happens iff they have the same undirected structure and the same "v-structures", which are converging arrows without a connection between their tails: $X \to Y \leftarrow Z$ but no arrow between X and Z forms a v-structure.

1.2.5 1.2.4) Inference with BNs

The book comments a bit on how we could try to estimate conditional probabilities of some variables given the observation of others. I'm not going to focus on this.

1.3 Section 1.3: BNs with causal directions

A Causal Bayesian Network is a Bayesian Network with causal directions.

We say that a distribution after an intervention is compatible with the CBN if it's Markov relative to it (we can decompose the joint with respect to the BN, or the parents make the childs independent of non-descendents), the chance of the interventions happening is one, *and the conditional probabilities remain the same for variables we didin't act on.*

The joint after the intervention can be factorized as $P(v) = \Pi_{i|V_i \notin X} P(v_i|pa_i)$, which is basically the original joint without the $P(v_i|pa_i)$ of the variables we acted on. v is a vector here (the entrances are the values of the random variables that are represented by the nodes of the CBN).

Two properties: $P_{pa_i}(v_i) = P(v_i|pa_i) =$ interventions are according to the conditionals, and $P_{pa_i,s}(v_i) = P_{pa_i}(v_i) =$ no interventions besides the one in the parents can influence a variable

Pearl argues that the advantage of causal models is to transport results to other environments and predict the results of changes that aren't purely observacional.

1.4 Section 1.4: Counterfactuals

1.4.1 Laplacian vs Stochastic model

The Laplacian one has deterministic functions and unobserverd probablistic variables, he stochastic one is more similar the Bayesian Network approach, if I understood correctly

Pearl says that this is more general than probablistic functions, but to me this just makes sense if by stochastic he doesn't mean something like a Markov Chain instead of the function, as this would certanly be more general... The BNs do not really have Markov Chains, but conditional probabilities, maybe that's what he means?

1.4.2 1.4.1: Structural Equations

Structured Equation Models are defined defining each variable as a function of the parents and unobserved variables (erros). If it's linear, then its a **Linear Structured Equation Model**.

One important point: Pearl says that it's possible to estimate counterfactuals with data and a causal model, and to test empirically whether they hold or not. I believe he will focus on how to do that in later chapters.

In the linear models, the coefficients are the variation rates per forced variation of a value, in the sense that it's how much the value would change if we changed only that value by one unity.

It's usually assumed that the error terms are independent, if they are dependent we represent a dotted double-headed arrow between the variables involved.

The hyerarchy of Causal problems defined by Pearl are:

1. **Predictions** are the "what if we found out that the value of this other variable was this?"

2. **Interventions** are the "what if we set the value of this other variable to this?"

3. **Counterfactuals** are the "what would be the value of this variable if the value of this other one was that instead of this?"

1.4.3 1.4.2: Probabilistc Predictions (and Definitions and equivalences between SCMs and BNs)

Causal Diagram is the diagram obtained by connecting the parents to the childs according to the structural equations. If this graph is a DAG, then it's **semi-Markovian**, and if the erros are independent, then it's **Markovian**. If it's semi-markovian, the joint is completely determined by the distribution on errors.

If the model is markovian, then this is a valid Causal Diagram: given the parents, a node is idependent of all other non-descendants. The proof is just to get the full graph, with the errors, then notice that we can remove the errors without losing independencies.

Pearl says that this is implied if we include every variable that might be a causa of two or more others, and that there is no correlation without causation...

The idea seems to look at the data and determine all probabilities first, even without knowing the deterministic functions (and the errors or distribution on errors) themselves... For any joint distribution compatible with a bayesian network, there is always at least one Functional Model with this same network (and Pearl mentions that usually there are infinitely many) that generates it with some values for the error/unobserved variables.

So, I think this is what he meant before, that the functional models are more general: we can encode in them anything we could encode in a BN.

1.4.4 1.4.3: Interventions

Four advantages mentioned by Pearl of using the graphical representation of Causal Models are:

1. The conditional independencies do not depend on the specific functions themselves, so if we can represent something in the causal model even with limited information, and given the model we don't need to compute anything to know whether some variables are independent given others (this is also possible with BNs, isn't it? We can also just build the graph without the probabilities and check independencies).

2. It's simpler to specify the connections, and the model has few parameters (I would argue that BNs has the same number or less parameters, the advantage to me is actually that the functions are finite, while the probability distributions are not, but then the distributions on unknowns are also infinite).

3. It's simpler to think of whether or not the parent set has all relevant variables that are a direct cause of some variable, instead of checking whether they make this variable independent of the others when we condition on them (and are a maximum set that does that). (we kind of could do this for BNs, right? But yeah, we would need to think that the independence is guaranteed, I think I agree with this one)

4. If something changes, the change might be local on some variables only, and with these models we can model this change by changing less the model, instead of recomputing everything from scratch. (This really does seem like a big advantage, if we change from one country to another the functions will change, and the conditional probabilities of the BNs change,

but the functions might be simpler. Again, the unknowns might change as well but I don't doubt at all that it's simpler to determine the unknowns than to re-estimate the conditional probabilities)

1.4.5 1.4.4: Counterfactuals

The idea is to say which variables were responsible for some result. For instance, if someone takes an experimental treatment to a disease and dies, did they die *because*, *despite* or *regardless* of the treatment?

Pearl says that we can treat counterfactuals as, instead of what would have happenned with X_1 if $X_2 = y$ instead of $X_2 = x$, what will happen if we reverse the outcome and repeat the experiment keeping everything equal except the value of X_2. This is called the *persistency* assumption.

I didn't understand why the assumption is necessary, and how exactly can we reach the conclusion for the assumption, but Pearl says that the proportion of people that died and recovered are equal with or without taking treatment, then (ignoring sampling variances) the proportion of people that died under treatment but woudn't if not treated would be the same than the proportion of people that didn't die without treatment but would under treatment. The idea seams to be that if the treatment is $x\%$ rensposible for the death of someone, then it would be $x\%$ responsible for the death of someone alive and untreated; if $x\%$ of the treated dead died because of the treatment, then $x\%$ of the alive untreated would have died if treated.

Two different situations given as examples that generate the above data but have different counterfactuals are: the treatment has no effect or half of the population has an allergy that protects them from the disease but kills them if they receive treatment. In the first case, treating someone untreated wouldn't change anything (all of the dead untreated would still be dead if treated), in the second group everyone that died under treatment was allergic and would still be alive if untreated.

The basic idea, viewing the SCM as a CBN with the unknowns explicited, is to Bayesanly-update the values of the unobserved variables given the observations, then intervene with the alternative values of whatever we want to know the alternative, then re-compute the distribution after the intervention. Viewing as Structured Equations, I think that we set the values of the observations to estimate the values of u, then set the new values of the alternative world and recompute everything. Pearl divides this into the following:

1. Abduction: basically estimate $P(u)$ from the observations.

2. Action: basically do the intervention, "bend the course of history minimally to comply with the hypothetical condition".

3. Prediction: Compute the desired probability.

Pearl says it's possible to compute estimatives without the full functinons between nodes and without knowing the distribution of unknowns, with just some assumptions of both.

1.4.6 Questions and confusions

I still am a bit confused about being able to have more than one set of parents per node if the distribution is not strictly positive... What do we do about that? What if there is a logical limitation, and an example that's better (and harder to find the problem) than just two equal variables causing another? Would everything break or is it stable to lead to an "almost zero" probability when it would be zero?

I didn't get why the assumption of $p(y|x) = \frac{1}{2}$ of (1.46) was necessary for the exercise "left for the reader". I think I will be able to do this later.

1.5 Section 1.5: Some terminology

Apparently, *probabilistic* stuff are quantities obtained from the joint, and *statistical* stuff is obtained from the joint of observables, ignoring non-observables completely.

Causal stuff are things defined in terms of a causal model.