Causal inference cheat sheet

Last compiled: September 12, 2020

1. Basic probability

- Law of total probability: $P(A) = \sum_i P(A, B_i)$ (a.k.a. marginalizing over B)
- Chain rule of probability: $P(A, B) = P(A|B)P(B)$
- Thus, $P(A) = \sum_i P(A|B_i)P(B_i)$
- Expectation: $E(g(X)) = \sum_x g(x)P(x)$
- Conditional mean: $E(X|Y) = \sum_x xP(x|y)$
- Variance: $\sigma_X^2 = E[(X - E(x))^2]$
- Covariance: $\sigma_{XY} = E[(X - E(X))(Y - E(Y))]$
- Correlation coefficient: $\rho_{XY} = \sigma_{XY}/(\sigma_X \sigma_Y)$
- Regression coefficient of X on Y: $r_{XY} = \rho_{XY}\sigma_X/\sigma_Y = \sigma_{XY}/(\sigma_Y^2)$ (for the equation $X = r_{XY}Y + c + \mathcal{N}(0, \sigma^2)$)
- Conditional independence: $(X \perp\!\!\!\perp Y|Z) \iff P(x|y, z) = P(x|z)$

The recursive decomposition of the joint distribution into parents which characterises Bayesian networks is

$$P(x_1, ..., x_n) = \prod_i P(x_i|pa_i) \tag{1.1}$$

1.1. d-separation in Bayesian networks

A path p is d-separated (or blocked) by a set of notes Z if and only if

1. p contains a chain $i \to m \to j$ or a fork $i \leftarrow m \to j$ such that the middle node m is in Z, or
2. p contains a collider $i \to m \leftarrow j$ such that the middle node m is not in Z and such that no descendant of m is in Z

where an arrow $pa_j \to x_j$ denotes part of a directed acyclic graph (DAG) in which variables are represented by nodes and arrows are drawn from each node of the parent set PA_j towards the child node X_j.

Probabilistic implications of d-separation Consequently, if X and Y are d-separated by Z in a DAG G, then $(X \perp\!\!\!\perp Y|Z)$ in every distribution compatible with G. Conversely, if X, Y, and Z are *not* d-separated by Z in a DAG G then X and Y are dependent conditional on Z in almost all distributions compatible with G (assuming no parameter fine-tuning).

2. Functional causal models

A functional causal model consists of a set of equations of the form

$$x_i = f_i(pa_i, u_i), \quad i = 1, ..., n \tag{2.1}$$

where pa_i are the set of variables (parents) that directly determine the value of X_i and U_i represents errors (or "disturbances") due to omitted factors. When some disturbances U_i are judged to be dependent, it is customary to denote such dependencies in a causal graph with double-headed arrows. If the causal diagram is acyclic, then the corresponding model is called *semi-Markovian* and the values of the variables X are uniquely determined by those of the variables U. If the error terms U are jointly independent, the model is called *Markovian*.

Linear structural equation models obey

$$x_i = \sum_{k \neq i} \alpha_{ik} x_k + u_i, \quad i = 1, ..., n \tag{2.2}$$

In linear models, pa_i corresponds to variables on the r.h.s. of the above equation where $\alpha_{ik} \neq 0$.

2.1. Counterfactuals in functional causal models: An example

Consider a randomized clinical trial, where patients are/are not treated $X \in \{0, 1\}$. We also observe whether the patients die after treatment $Y\{0, 1\}$. We wish to ask the question: did the patient die *because of* the treatment, *despite* the treatment, or *regardless* of the treatment.

Assume $P(y|x) = 0.5$, and therefore $P(y, x) = 0.25$ for all x and y. We can write two models with the same joint distribution

Model 1 (treatment no effect):

$$x = u_1 \tag{2.3}$$
$$y = u_2 \tag{2.4}$$
$$P(u_1 = 1) = P(u_2 = 1) = \frac{1}{2} \tag{2.5}$$

Model 2 (treatment has an effect):

$$x = u_1 \tag{2.6}$$
$$y = xu_2 + (1 - x)(1 - u_2) \tag{2.7}$$
$$P(u_1 = 1) = P(u_2 = 1) = \frac{1}{2} \tag{2.8}$$

Let Q=fraction of deceased subjects from the treatment group who would not have died had they not taken the treatment. In model 1, $Q = 0$ since X has no effect on Y. In model 2, subjects who died $(y = 1)$ and were treated $(x = 1)$ must correspond to $u_2 = 1$. If $u_2 = 1$ then the only way for $y = 0$ is for $x = 0$. I.e. if you are a patient for whom $u_2 = 1$ then the only way not to die is to not take the treatment, so the treatment caused your death. So $Q = 1$.

Consequence 0: joint probability distributions are insufficient for counterfactual computation

Consequence 1: stochastic causal models are insufficient for counterfactual computation

Consequence 2: functional causal models are sufficient to define and compute counterfactual statements.

2.2. General method to compute counterfactuals

Given evidence $e = \{X_{obs}, Y_{obs}\}$, to compute probability of $Y = y$ under hypothetical condition $X = x$ apply the following steps:

1. Abduction: Update the probability of disturbances $P(u)$ to obtain $P(u|e)$
2. Action: Replace the equations corresponding to variables in the set X by the equations $X = x$
3. Prediction: Use the modified model to compute the probability $Y = y$.

3. Causal Bayesian networks

Given two disjoint sets of variables X and Y, the **causal effect** of X on Y, denoted as $P(y|\hat{x})$ or $P(y|do(x))$, is the probability of $Y = y$ by deleting all equations from Eq.(2.1) where variables X are on the l.h.s., and substituting $X = x$ in the remaining equations.

This corresponds to mutilating the DAG such that all arrows pointing directly to X_i are removed. **Amputation is the difference between seeing and doing**.

For an atomic intervention, we get the *truncated factorization* formula

$$P(x_1, ..., x_n|\hat{x}'_i) = \begin{cases} \prod_{j \neq i} P(x_j|pa_j) & \text{if } x_i = x'_i \\ 0 & \text{if } x_i \neq x'_i \end{cases} \tag{3.1}$$

The $j \neq i$ denotes the removal of the term $P(x_i|pa_i)$ from Eq.(1.1) (i.e. amputation). A $do(x_i)$ is a severely limited sub-space of the full joint distribution, since the distribution only has support where the intervention variable x_i is equal to its particular intervention value x'_i, rather than a continuum of values in Eq.(1.1).

Multiplying and dividing by $P(x'_i|pa_i)$ yields

$$P(x_1, ..., x_n|\hat{x}'_i) = \begin{cases} P(x_1, ..., x_n|x'_i, pa_i)P(pa_i) & \text{if } x_i = x'_i \\ 0 & \text{if } x_i \neq x'_i \end{cases} \tag{3.2}$$

Marginalization of the above leads to the following theorem.

Adjustment for direct causes Let PA_i denote the set of direct causes of variable X_i, and let Y be any set of variables disjoint of $\{X_i \cup PA_i\}$. The causal effect of $do(X_i = x'_i)$ on Y is

$$P(y|\hat{x}'_i) = \sum_{pa_i} P(y|x'_i, pa_i)P(pa_i) \tag{3.3}$$

where $P(y|x'_i, pa_i)$ and $P(pa_i)$ are preintervention probabilities. This is called "adjusting for PA_i".

4. Inferring causal relations

- IC algorithm is for inferring causal structure given observational data when there are no latent variables
- IC* algorithm is for inferring causal structure given observational data when there are latent variables. The PC algorithm is apparently more contemporary (see Spirtes et al 2010)
- There are local criteria for potential cause and genuine cause
- Spurious association: X and Y are spuriously associated if they are dependent in some context and there exists a latent common cause, as exemplified in the structure $Z_1 \to X \to Y \leftarrow Z_2$