Math for AI Safety | Lionel Levine

Conditional Independence

September 4, 2024 | *Notes by Matthew Haulmark*

1 Conditional Independence

We'll be following Pearl's book *Causality* (Chapters 1 and 3).

Setup: Discrete random variables

$$X_1, \ldots, X_n \colon (\Omega, \mathcal{F}, P) \to \mathbb{R}$$

Joint Distribution

$$p(x_1, \ldots, x_n) := P(X_1 = x_1, \ldots, X_n = x_n)$$

The joint probability distribution is impractical when n is large. For example, computing the marginal distribution of X_1 from the joint distribution

$$p(x_1) = \sum_{x_2, \ldots, x_n} p(x_1, x_2, \ldots, x_n)$$

involves a sum with exponentially many terms (2^{n-1} terms if $X_2, \ldots, X_n$ are binary random variables).

Definition (Independence): X_1 and X_2 are *independent* if $p(x_1, x_2) = p(x_1)p(x_2)$ for all $x_1, x_2 \in \mathbb{R}$

While there's nothing wrong with this definition, it doesn't always capture how people reason intuitively about independence.

Example: Consider these two events

$$A_1 = \{\text{Tompkins county has a forest fire in 2024}\}$$
$$A_2 = \{\text{Inflation greater than 5\% in 2024}\}$$

and let $X_i = 1_{A_i}$ be the corresponding indicator random variables. (Notation: For an event A, the random variable 1_A equals 1 if A occurs and 0 otherwise.)

Accurately estimating $p(x_1)$, or $p(x_2)$, or $p(x_1, x_2)$ in this case might require specialized knowledge of weather forecasting, or macroeconomics, or both! On the other hand, it doesn't take specialized knowledge to reason that X_1 and X_2 are independent. What's going on with this kind of reasoning? Can we make it more precise than a general sense that "inflation doesn't have much to do with forest fires"? Whatever this reasoning is doing, it proceeds by *some other method* that doesn't involve computing the joint and marginal probabilities.

Definition: Random variables X and Y are *conditionally independent* given Z, if

$$p(x|y,z) = p(x|z) \qquad \text{for all } x, y, z \in \mathbb{R} \text{ such that } p(y,z) > 0.$$

Notation: $X \perp\!\!\!\perp Y|Z$.

Intuitively, this means: If we know $Z = z$, then learning that $Y = y$ does not provide any additional information about the value of X.

Example (Buses): Let T_1 and T_2 be arrival times of consecutive buses at a bus stop. Then T_1 and T_2 are dependent, but

$$T_2 \perp\!\!\!\perp T_1 \,|\, X_2$$

where X_2 is the current location of bus 2: Once we know the location of bus 2, the arrival time of bus 1 doesn't provide any additional information about the arrival time of bus 2.

To turn this example into math, let's add some assumptions: the buses move at constant speed $v = 10$ miles per hour, so $T_i = X_i/v$ where X_i is the current distance of bus i from the bus stop. And the spacing is random: X_1 and $X_2 - X_1$ are independent random variables with the exponential distribution with a mean of 5 miles. If we learn that $T_1 = 1$ minute (bus 1 is early) that's going to decrease our estimate of T_2. But if we *also* learn that $X_2 = 100$ miles (bus 2 is very far away) then the information about T_1 becomes irrelevant to our estimate of T_2.

1.1 Properties of Conditional Independence

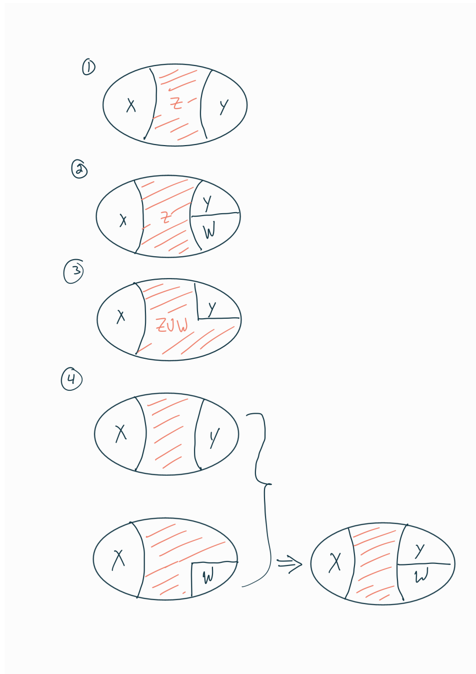
(1) **Symmetry:** $\big(X \perp\!\!\!\perp Y|Z\big) \Rightarrow \big(Y \perp\!\!\!\perp X|Z\big)$

(2) **Decomposition:** $\big(X \perp\!\!\!\perp (Y,W)|Z\big) \Rightarrow \big(X \perp\!\!\!\perp Y|Z\big)$

(3) **Weak union:** $\big(X \perp\!\!\!\perp (Y,W)|Z\big) \Rightarrow \big(Y \perp\!\!\!\perp X|(Z,W)\big)$

(4) **Contraction:** $\big(X \perp\!\!\!\perp Y|Z\big) \& \big(X \perp\!\!\!\perp W|(Y,Z)\big) \Rightarrow \big(X \perp\!\!\!\perp (Y,W)|Z\big)$

Corresponding Cartoons: There is a sense in which we can think of conditional independence in terms of blocking paths between sets.



Pearl observed: In an undirected graph $G = (V, E)$ if we let X, Y, W, X be subsets of V, and set $(X \perp\!\!\!\perp Y|Z)_G$ to mean that every path from X to Y in G passes through Z. Then properties 1-4 are satisfied. We will come back to this observation. The analogy between dependence and graph reachability turns out to be much closer when G is a *directed* graph, and $(X \perp\!\!\!\perp Y|Z)_G$ stands for something called d-separation.

We will prove Symmetry and Decomposition.

Proof of 1. Symmetry: We claim the following:

$$(X \perp\!\!\!\perp Y|Z) \iff p(x, y|z) = p(x|z)p(y|z)$$

for all $x, y, z \in \mathbb{R}$ with $p(z) > 0$.

proof of claim:

$$p(x|y, z) = \frac{p(x, y, z)}{p(y, z)} = \frac{p(x, y, z)}{p(y, z)} * \frac{p(z)}{p(z)} = p(x, y|z) * \frac{1}{p(y|z)}$$

Now, if $(X \perp\!\!\!\perp Y|Z)$ we have

$$p(x, y|z) = p(y|z)p(x|y, z) = p(y|z)p(x|z)$$

.

Proof of 2. Decomposition: Given $\big(X \perp\!\!\!\perp (Y, W)\big)|Z$. We have $p(x|y, w, z) = p(x, z)$ whenever $p(y, w, z) > 0$.

We want $p(x|y, z) = p(x|z)$ whenever $p(y, z) > 0$.

$$p(x|y,z) = \frac{p(x,y,z)}{p(y,z)} = \sum_w \frac{p(x,y,w,z)}{p(y,z)} = \sum \frac{p(x|y,w,z)p(y,w,z)}{p(y,z)} = \frac{p(x|z)}{p(y|z)} \sum p(y,w,z) = p(x|z)$$

Notice that the terms where $p(y, w, z) = 0$ do not contribute.

Some History

1985: Pearl and Paz conjectured that conditions (1)-(4) are complete. In otherwords, for any 3-place relation $\perp\!\!\!\perp$ satisfying (1)-(4) there is a probability measure P such that conditional independence with respect to P is $\perp\!\!\!\perp$.

1992: Studeny disproved the conjecture. Showed:

$$\text{If } X_0 \perp\!\!\!\perp X_i|X_{i+1} \text{ for all } i = 1, \dots, n-1,$$
$$\text{then } X_0 \perp\!\!\!\perp X_{i+1}|X_i \text{for all } i = 1, \dots, n-1.$$

Call this property S_n. It turns out that S_n is not implied by the conjunction of $S_1, \dots, S_{n-1}$. In fact, no finite set of axioms is complete for conditional independence!

2006: Simicek and 2007: Sullivant both gave a counterexamples to the conjecture in which $X_1, \dots, X_n$ are jointly Gaussian.

1.2 Unconditional Independence

Notation: $X \perp\!\!\!\perp Y|\emptyset$ means $p(x|y) = p(x)$ for all x, y such that $p(y) > 0$. Equivalently, $p(x, y) = p(x)p(y)$ for all x and y.

Question: **Which is stronger, conditional independence or unconditional independence?**

Example 1:

$$\begin{tikzcd} & C \arrow[dl] \arrow[dr] & \\ A & & B \end{tikzcd}$$

$$p(a, b, c) = p(c)p(a|c)p(b|c)$$

Is $A \perp\!\!\!\perp B|C$?

$$p(a, b|c) = \frac{p(a, b, c)}{p(c)} = p(a|c)p(b|c)$$

Is $A \perp\!\!\!\perp B|\emptyset$?... not necessarily

Example 2:

$$\begin{tikzcd} A \arrow[dr] & & B \arrow[dl] \\ & C & \end{tikzcd}$$

$$p(a, b, c) = p(a)p(b)p(c|a, b)$$

Is $A \perp\!\!\!\perp B|C$? Not in general

Is $A \perp\!\!\!\perp B|\emptyset$? Yes

<u>Answer</u>: Neither is stronger!

1.3 G-Markov distributions

Next class: we'll talk about G-Markov distributions where, G is a directed acyclic graph (generalizing the examples of three-vertex graphs above). These are also called Bayes Nets. The d-separation theorem of Pearl and Verma will allow us to reason from the graph which conditional independence conditions hold.