Causal Statistical Decision Theory|Why subjunctive probability?

David Johnston

February 20, 2020

I have too many adjectives here. I think I should either go with *consequence maps* and *stochastic consequence spaces* or *subjunctive probabilities* and *subjunctive probability spaces*. The latter is more precise, but also harder to parse before you read the definition.

1 Probability spaces and subjunctive probability spaces

In a very broad sense, a decision maker - whether they be man or machine - is someone or something that takes some kind of provocation - data, problem specification, circumstances - and chooses a decision. Were we considering beings of a suitable level of divinity, we might doubt whether all the things they might choose could be contained in a set, but it is a basic property of mortal or mechanical decision makers that there is some set D of things that they might ultimately choose. Conversely, there is a set D such that every decision that a decision maker might choose is an element of D.

That is, given some set A of provocations and D of decisions, a decision maker implements a map $A \to D$. Another distinguishing feature of a decision maker concerns the internal details of how she implements this map. We take the view that, in order to be considered a decision maker, she must implement this map by comparing the available decisions in light of the given information $a \in A$. Abstractly, she posesses some higher order function $f : A \to E^D$ which, broadly speaking, takes some piece of information $a \in A$ and returns a function f_a which determines the result of each $d \in D$ given the information a. A second higher order function $\text{ch} : E^D \to D$ considers each decision $d \in D$ along with its result $f_a(d) \in E$ and chooses one according to some criterion. Our decision maker implements $\text{ch} \circ f$.

Concretely, our decision maker might be responsible for producing a binary classifier that takes some piece of data X and returns a class in $\{0, 1\}$, leaving us with $D = \{0, 1\}^X$. The given information is a set of n labeled examples $A = X \times \{0, 1\}^n$, and f might be an *empirical risk* functional that takes data

$a \in A$ and a possible classifier $d \in D$ and returns the number of misclassified examples when d is applied to A. In such a case, ch might be the functional given by arg min with some means of breaking ties.

Another type of decision maker is one that tries to affect the world in some way. Suppose we have some set E that represents possible configurations of the world and a utility function $u : E \to \mathbb{R}$ that rates the desirability of any given state of the world. A decision maker that tries to affect the world is one whose choice function chooses the "most preferred" state of the world. This might correspond to a choice function given by utility maximisation $\text{ch} : g \mapsto \arg\max_{d \in D} u(g(d))$, for example. For example, consider a robot that only wants to switch a light on – for this robot, we might take relevant states of the world to be $E = \{\text{off}, \text{on}\}$, the utility to be $u : \text{on} \mapsto 1$ and $\text{off} \mapsto 0$, D to be sequences of signals it might send to each of its actuators, the background information $A = \{\text{switch up} \mapsto \text{light on}, \text{switch down} \mapsto \text{light on}\}$ stipulates the relationship between a switch and the light and the function $f : a \mapsto a \circ f_0$ takes the prior knowledge of whether a motor sequence $d \in D$ leaves the switch in the up or down position and then applies the specification a to this knowledge. This robots choice function could be given by utility maximisation.

As a general principle, decision makers implement some function $h : A \to D$ which factorises as $\text{ch} \circ f$ for some choice rule ch and some results model f. We take this to be a general model of something that chooses a decision by comparing available options. We are particularly interested in decision makers that try to affect the world, in particular those that use the principle of *expected utility maximisation* as their choice rule. Expected utility maximisation is a very standard rule for choosing preferred states of the world subject to some uncertainty, and we will simply accept it at the outset to facilitate a thorough exploration of the results model f.

Expected utility maximisation (henceforth "EU") necessitates that the class of results models f have a particular signature. The EU choice rule compares a set of probability distributions over states of the world E and returns a decision associated with the preferred probability distribution. Thus, denoting by $\Delta(\mathcal{E})$ the set of probability distribution over E (now equipped with σ-algebra $\mathcal{E}$), f must have the signature $f : A \to \Delta(\mathcal{E})^D$. In particular f returns stochastic functions of the type $D \to \Delta(\mathcal{E})$.

So far we have not made any structural assumptions about the set D. One such assumption that we do want to make is that stochastic decisions are possible. That is, if we can choose $d_1 \in D$ and we can choose $d_2 \in D$, then it is also possible to choose d_1 with probability α and d_2 with probability $(1 - \alpha)$ for $0 \leq \alpha \leq 1$. This is easy enough to do - we simply have to assert that there is some $d_{(\alpha,1,2)} \in D$ such that $d_{(\alpha,1,2)}$ corresponds to the choice "d_1 with probability α and d_2 with probability $(1 - \alpha)$". We would additionally like our results model f to have the property that, for all $a \in A$,

$$f_a(d_{(\alpha,1,2)}) = \alpha f_a(d_1) + (1 - \alpha) f_a(d_2) \tag{1}$$

This need not be true of all expected utility maximising decision makers - for example, suppose our decision maker is faced with a mind reader that can

perceive whether she decided on d_1 or decided on $d_{(\alpha,1,2)}$ and subsequently obtained d_1 after performing a randomisation step, and this mind reader is for whatever reason inclined to respond differently in each case. Then our decision maker would not be able to assume Equation 1 holds. However, we will proceed on the assumption that mind readers are absent and that choosing d_1 is the same as choooosing to randomise and then obtaining d_1.

Actually, I'm not completely sure why this assumption is needed, but it seems important.

To strengthen this notion, suppose $(D, \mathcal{D})$ is measurable. Recall that for all $d \in D$, $f_a(d)$ is a probability measure on E. For $G \in \mathcal{E}$, write $f_a(d; G)$ for the probability measure $f_a(d)$ evaluated at G, and $f_a(\cdot; G)$ for the map $d \mapsto f_a(d; G)$. Then we will call a result map f regular if for all $a \in A$, $G \in \mathcal{E}$ and all $\gamma \in \Delta(\mathcal{D})$ f_a, $f_a(\cdot; G)$ is measurable and there exists some d_γ such that

$$f_a(d_\gamma; G) = \int_D f_a(d'; G) d\gamma(d') \tag{2}$$

I want to introduce a vector space structure on D, then show D is closed convex and hence it is isomorphic to the set of probability measures on some D' constructed from the extreme points of D.

Note that by the assumption of measurability, we have that f_a is a Markov kernel for all $a \in A$.

For our causal analysis we work with *subjunctive probability spaces*, a generalisation of the more familiar probability spaces. The subjunctive mood is used to describe hypothetical or supposed states of the world, and subjunctive probability spaces are models that ask for some hypothetical state and give us back a probability space. Subjunctive probability spaces are also different to *conditional probability spaces* Rényi (1956) as *hypothesising* or *supposing* (that is, those things described by the subjunctive mood) are different to *conditioning*, which is more closely analogous to *focussing your attention*.

Does this definition need to be here? I didn't know what this word meant before I looked it up

We use subjunctive probability spaces because *supposition* is a core part of decision problems, one we cannot get away from. In contrast, modelling supposition with conditional probability (which would be necessary if we were to insist on using probability spaces) adds additional structure to our models which isn't clearly warranted and is potentially confusing.

Any decision problem must involve the comparison of different decisions. This comparison takes the form

- *Suppose* I choose the first decision - then the consequence would be X
- *Suppose* I choose the second decision - then the consequence would be Y
- Etc.

If we prefer X to Y, then perhaps we should choose the first decision. We may be ambivalent as to what type of thing X and Y represent, how we ought

to determine what values X and Y take or how we ought to determine what is preferable, but it is hard to do away with supposing that different decisions were taken and that these decisions come with their own consequences and still have what could reasonably be considered a decision problem. This process of supposition implicitly invokes a "subjunctive function" - we provide a hypothetical decision, and we are given a consequence of that hypothetical. Of particular interest are models that, for each hypothetical choice, return a probability distribution over space Ω. Such models are called *supposition functions* by Joyce (2000), but we will call them *consequence maps* in this work. We consider this particular type of subjunctive function - from possible decisions to probability distributions - to be attractive because we consider probability to be a sound choice for modelling uncertainty and stochasticity.

Formally, a consequence map $\mathcal{C}$ is a stochastic function or *Markov kernel* from $D \to \Delta(\Omega)$, where $\Delta(\Omega)$ represents the set of all probability distributions on Ω. One might recall that, given two random variables $\mathsf{Y} : \Omega \to Y$ and $\mathsf{X} : \Omega \to X$ on some probability space $(\mathbb{P}, \Omega, \mathcal{F})$, the probability of Y conditional on X is a Markov kernel $X \to \Delta(\mathcal{Y})$. It is possible, in general, to define a probability distribution on the expanded space $\Omega \times D$ such that $\mathcal{C}$ is a conditional probability. However, there are good reasons to keep the concepts of consequence maps and conditional probability separate. Firstly, there are technical issues such as the fact that it is not always possible to find a distribution on $\Omega \times D$ that yields $\mathcal{C}$ as a *unique* conditional probability (Hájek, 2003) (though this requires an uncountable set of possible decisions, which is not a problem we consider in this work). Secondly, it is not clear that that the way we ought to handle consequence maps is identical to the way we handle conditional probabilities. Consider an expanded set of options:

1. Suppose I choose the first decision d_1 - then the consequence would be P_1
2. Suppose I choose the second decision d_2 - then the consequence would be P_2
3. Suppose I choose either the first or second decision d_1 or d_2 - then the consequence would be ???
4. Suppose I choose d_1 with probability $0 \leq q \leq 1$ and d_2 otherwise - then the consequence would be ???

If we regard the consequence map as a conditional probability then it would be natural to consider the result of the third option to be a unique probability distribution obtained by conditioning on $\{d_1, d_2\}$, equal to $\alpha P_1 + (1-\alpha)P_2$ for some fixed $0 \leq \alpha \leq 1$. However, there is no obvious reason that these should be mixed in some fixed proportion α - it seems more appropriate to me to say that if d_1 or d_2 is chosen then the result will be P_1 or P_2.

On the other hand, the result of the fourth option seems like it should be given by $qP_1 + (1-q)P_2$, at least for most ordinary problems. If we were confronted with a mind reader who could tell the difference between us having chosen d_1

and us having randomised between d_1 and d_2 but come up with d_1 anyway then we might have reason to revise this assumption, but we will generally proceed under the assumption that such mind readers are absent. For any ambient probability measure, we can choose q not to be equal to α (as defined in the previous paragraph), and so the result will not be equal to the ambient measure conditioned on $\{d_1, d_2\}$.

In general, choosing the consequence map to be a conditional probability requires there to exist some joint distribution over the decision D and all other random variables in the problem. However, when we adopt hypotheses about what decisions we might make, we throw away key parts of this joint distribution (for example, we throw away the marginal distribution of D) and replace it with whatever we want to suppose instead. Instead of adopting a full joint distribution and then throwing away some parts to get hold of the consequence map, as we would if we were working with a probability space, a subjunctive probability space only supplies those parts which we intend to keep - i.e. in only supplies the consequence map.

References

Alan Hájek. What Conditional Probability Could Not Be. *Synthese*, 137(3): 273–323, December 2003. ISSN 1573-0964. doi: 10.1023/B:SYNT.0000004904.91112.16. URL `https://doi.org/10.1023/B:SYNT.0000004904.91112.16`.

James M. Joyce. Why We Still Need the Logic of Decision. *Philosophy of Science*, 67:S1–S13, 2000. ISSN 0031-8248. URL `www.jstor.org/stable/188653`.

A. Rényi. On Conditional Probability Spaces Generated by a Dimensionally Ordered Set of Measures. *Theory of Probability & Its Applications*, 1(1): 55–64, January 1956. ISSN 0040-585X. doi: 10.1137/1101005. URL `https://epubs.siam.org/doi/abs/10.1137/1101005`.

Appendix: