Lecture 3: Causal Decision Theory

Rachael Briggs

July 31, 2014

A Fatalist Fallacy

You have an exam next week, which you consider yourself only about 60% likely to pass but you'd like to attend this week's Blues Festival. You're trying to decide whether to study, or ditch the library for the festival. You argue as follows. There are two possible states of the world: either you'll pass your exam, or you'll fail it. No matter which state of the world you're in, attending the festival will have a better outcome than studying. So by the Principle of Strict Dominance, you should go to the festival.

	pass $P = 0.6$	**fail** $P = 0.4$
skip	fun + success $u = 5$	fun + failure $u = -5$
study	boredom + success $u = 3$	boredom + failure $u = -7$

$$EU(\textbf{study}) = 0.6 \times 12 + 0.4 \times 2 = 8$$

$$EU(\textbf{skip}) = 0.6 \times 10 + 0.4 \times 0 = 6$$

- Savage can object: "You've picked the states wrong! States have to be independent of acts!"
- No good way of representing this independence in Savage's framework, at least if "independent" denotes a probabilistic concept.
- A diagnosis by Jeffrey: your acts are sometimes evidence about which state of the world obtains. So don't use unconditional probabilities of events to calculate expected utilities; use conditional probabilities on acts.

Definitions of Expected Utility

Summing over outcomes

Neutral

$$U(a) = \sum_{o \in O} P_a(o)u(o)$$

Savage

$$U(a) = \sum_{o \in O} P(\{s \in \mathcal{S} : a(s) = o\})u(o)$$

Jeffrey

$$U(a) = \sum_{o \in O} P(o|a)u(o)$$

Furthermore, for every proposition X,

$$U(X) = \sum_{o \in O} P(o|X)U(o)$$

Summing over equivalence classes of states

Let dependency hypothesis K be a maximal set of states such that for all $s_1, s_2 \in \mathcal{S}$ and all available acts $a \in \mathcal{A}$, $a(s_1) = a(s_2)$. Let $\mathcal{K}$ be the set of all dependency hypotheses. Where a is an available act, $K(a)$ be the outcome o such that for all $s \in K$, $a(s) = o$. The Ks partition $\mathcal{S}$, so we can rewrite the definitions of expected utility as follows:

Neutral

$$U(a) = \sum_{K \in \mathcal{K}} P_a(K)u(K(a))$$

Savage

$$U(a) = \sum_{K \in \mathcal{K}} P(K)u(K(a))$$

Jeffrey

$$U(a) = \sum_{K \in \mathcal{K}} P(K|a)u(a \cap K)$$

Furthermore, for every proposition X and every partition $\mathcal{Y}$

$$U(X) = \sum_{Y \in \mathcal{Y}} P(Y|X)U(X \cap Y)$$

Sufficient conditions for Savage's definition and Jeffrey's definition to agree:

(a) acts are probabilistically independent of states, and

(b) each conjunction of the form $(a \cap s)$ (whose conjuncts a are state and an act) is compatible with exactly one outcome $o = a(s)$

Jeffrey's Framework

- Savage distinguishes acts (objects of preference), states (objects of probability) and outcomes (objects of intrinsic value); Jeffrey has:

  Ω is a set of possible worlds.

  $\mathcal{F}$ is a set of subsets of Ω, containing the unit and closed under complementation and countable intersection.

  $\gtrsim$ is a weak preference relation on $\mathcal{F}$.

  P is a probability function mapping members of $\mathcal{F}$ to real numbers in $[0, 1]$.

  U is an expected utility function mapping members of $\mathcal{F} - \{\emptyset\}$ to real numbers.

- Value is *news value*. (This helps make sense of the notion of propositions having values, even when they're propositions that have nothing to do with our agency.)

- Jeffrey proves a representation theorem: where $\gtrsim$ obeys suitable axioms, there exist a utility function U representing $\gtrsim$, and a probability function P such that for all propositions X and partitions $\mathcal{Y}$,

$$U(X) = \sum_{Y \in \mathcal{Y}} P(Y|X)U(X \cap Y)$$

- Savage's representation theorem came with a uniqueness result: P is unique, and U is unique up to positive linear transformation. If utilities are bounded (above *or* below), then Jeffrey does not get this uniqueness result.

- As with Savage, U is subject to the choice of a unit. It is conventional to set the value of the tautology at 0 (no news is not good or bad news—it is no news).

- Once we've set the values of 0 and a unit, the following equations give allowable transformations of P and U together.

  Bolker's equations:

$$P_\lambda(X) = P(X)(1 + \lambda U(X))$$

$$U_\lambda(X) = U(X)\frac{(1 + \lambda)}{1 + \lambda U(X)}$$

$$\frac{-1}{\inf\{U(X) : X \in \Omega\}} \geq \lambda \geq \frac{-1}{\sup\{U(X) : X \in \Omega\}}$$

  Note that these equations preserve ordering on

$$INT(X) =_{df} P(X)U(X)$$

- Why do we lose uniqueness? Without appeal to outcomes, we lose the ability to define “weakly likelier than”, or $(\succsim)$, in terms of preference. Instead, we get these sufficient (but not necessary) conditions for probability comparisons.

$X(\succsim)Y$ holds whenever $Z \succ (Y \cup Z) \succsim (X \cup Z) \succ X \succsim Y$, or
$Y \succsim X \succ (X \cup Z) \succsim (Y \cup Z) \succ Z$

$X(\succ)Y$ holds whenever $Z \succ (Y \cup Z) \succsim (X \cup Z) \succ X \succsim Y$, or
$Y \succsim X \succ (X \cup Z) \succsim (Y \cup Z) \succ Z$
provided at least one $\succsim$ is replaced with $\succ$

$X(\sim)Y$ holds whenever $Z \succ (Y \cup Z) \sim (X \cup Z) \succ X \sim Y$ or
$Y \sim X \succ (Y \cup Z) \sim (X \cup Z) \succ Z$

The Mysterious Case of the Noise at the Door

There is a noise outside. It’s either the mail carrier (bringing either an unwelcome bill, or a welcome letter from my friend) or a person (either the odious bill collector, or my wonderful friend in person). I hope it’s the mail. Why do I hope it’s the mail?

Option 1

bill	letter	collector	friend
$P = 0.2$ $U = -4$	$P = 0.3$ $U = 4$	$P = 0.3$ $U = -8$	$P = 0.2$ $U = 10$

Option 2 ($\lambda = 0.05$)

bill	letter	collector	friend
$P = 0.16$ $U = -5.25$	$P = .36$ $U = 3.5$	$P = 0.18$ $U = -14$	$P = 0.3$ $U = 7$

Newcomb’s Problem

An eccentric billionaire hands you a closed box containing either $1,000,000 or nothing. She offers you an additional $1,000 in an open box. There’s a catch. Before you were offered this choice, she made thorough study of your personality and predicted what you would do. You don’t know her methods, but you know that she is accurate: she picks both 90% of the accepters and 90% of the refusers correctly. The contents of the cosed box were determined by her prediction: if she predicted you’d refuse the $1,000, the box contains $1,000,000; otherwise, it’s empty. Should you take the two boxes, or only the one closed box?

	$1M in closed box	**$0 in closed box**
1-box	$P = 0.45$ $U = 1,000,000$	$P = 0.05$ $U = 0$
2-box	$P = 0.05$ $U = 1,001,000$	$P = 0.45$ $U = 1,000$

Jeffrey's Answer

$$\begin{aligned} U(\textbf{1-box}) &= P(\textbf{\$1M}|\textbf{1-box})U(\textbf{1-box} \cap \textbf{\$1M}) \\ &\quad + P(\textbf{\$0}|\textbf{1-box})U(\textbf{1-box} \cap \textbf{\$0}) \\ &= 0.9 \times 1,000,000 + 0.1 \times 0 \\ &= 900,000 \end{aligned}$$

$$\begin{aligned} U(\textbf{2-box}) &= P(\textbf{\$1M}|\textbf{2-box})U(\textbf{2-box} \cap \textbf{\$1M}) \\ &\quad + P(\textbf{\$0}|\textbf{2-box})U(\textbf{2-box} \cap \textbf{\$0}) \\ &= 0.1 \times 1,001,000 + 0.9 \times 1,000 \\ &= \text{101,000} \end{aligned}$$

- If you one-box, you will (probably) get the \$1M; if you two-box, you will (probably) find nothing in the closed box.
- Your one-boxing is *evidence* that you will end up rich: it has high *news value*. (This is why Jeffrey's theory is often called **evidential**.)
- Note the violation of the Principle of Strong Dominance.

Savage's Answer

$$\begin{aligned} U(\textbf{1-box}) &= P(\textbf{\$1M})u(\textbf{1-box}(\textbf{\$1M})) \\ &\quad + P(\textbf{\$0})u(\textbf{1-box}(\textbf{\$0})) \\ &= 0.5 \times 1,000,000 + 0.5 \times 0 \\ &= 500,000 \end{aligned}$$

$$\begin{aligned} U(\textbf{2-box}) &= P(\textbf{\$1M})u(\textbf{2-box}(\textbf{\$1M})) \\ &\quad + P(\textbf{\$0})u(\textbf{2-box}(\textbf{\$0})) \\ &= 0.5 \times 1,001,000 + 0.5 \times 1,000 \\ &= \text{501,000} \end{aligned}$$

- The closed box contains either \$1M or \$0, independent of anything you do. In either condition:

  If you were to one-box, you would end up with whatever was in the closed box.

  If you were to two-box, you would end up with whatever was in the closed box, plus \$1000.
- Two-boxing *causes* you to be richer; it has high *use value*. (This is why versions of Savage's theory are often called **causal**.)

Rewriting Savage

Counterfactual Conditionals

Probability theory lets us rewrite Savage's definition of expected utility as follows.

$$U(a) = \sum_{o \in O} \left(\sum_{K \in \mathcal{K} : K(a) = o} P(K) \right) u(o)$$

Suppose each $s \in \mathcal{S}$ is a conjunction of the form

> $\cap\{a_i \mathrel{\Box\!\!\rightarrow} o_i : a_i \in \mathcal{A}\}$
>
> "If act a_i were performed, outcome o_i would result, and . . . "

Then:

$$U(a) = \sum_{o \in O} P(a \mathrel{\Box\!\!\rightarrow} o) u(o)$$

Imaging

The standard way of updating probabilities on new information E: set your new probability for a proposition X equal to your old conditional probability $P(X|E)$, where this satisfies the

Ratio Formula

$$P(X|E) = \frac{P(X \wedge E)}{P(E)} \text{ (where defined)}$$

An alternative is

Imaging The function $\mathcal{I}$ maps each $w \in \Omega$ and $X \in \mathcal{F}$ to a probability function over $\mathcal{F}$. Say that

$$P_w^E = \mathcal{I}(w, E)$$

(We assume that for all w and E, $P_w^E(E) = 1$)

$$P^E(X) = \sum_{r \in [0,1]} \left(\sum_w P(\{w : P_w^E(X) = r\}) r \right)$$

If dependency hypotheses are sets of worlds that *image alike*, so that whenever a is an available act, and $w_1, w_2 \in K$, we have $P^a_{w1} = P^a_{w_2}$, then:

$$U(a) = \sum_{o \in O} P^a(o) u(o)$$

Chances

$$U(a) = \sum_{o \in O} \left(\sum_{K \in \mathcal{K}} P(K) ch_K(o|a) \right) u(o)$$

References

Richard C Jeffrey. *The logic of decision.* 2nd edition, 1983.

James M Joyce. *The Foundations of Causal Decision Theory.* Cambridge University Press, 2008.

David Lewis. Causal decision theory. *Australasian Journal of Philosophy*, 59(1):5–30, March 1981.

Leonard J Savage. *The Foundations of Statistics.* Courier Dover Publications, 1972.