Decision Theory notes

PHIL 735 Week 9
Matt Kotzen and Jim Pryor

November 1, 2023

1 Bayesian Decision Theory in general

- Core idea:
  - Let u be a "utility function"—i.e., a function that assigns to each possible outcome of an act a number reflecting how good or bad that outcome is.
  - Then, act A_1 is non-strictly preferred to act A_2 just in case $\sum_i p_i u_i \geq \sum_j p_j u_j$, where the p_i are the probabilities of the possible results of A_1 and the u_i are their utilities, and similarly for the p_j and u_j with respect to the results of act A_2.
  - The quantity $\sum_i p_i u_i$ is called the *expected utility* of act A_1, and similarly for act A_2.
  - Different (Bayesian-ish) decision theories disagree about how to understand and represent acts and outcomes, how to interpret the p_i, etc.

2 Savage's Decision Theory

- Savage (1954): Three components: states, acts, outcomes. For example:
  - Two possible states of the word: heads, tails
    - States are the bearers of probabilities
  - Two possible outcomes: +$10, -$10
    - Outcomes are the bearers of utilities
  - Two possible acts: bet-on-heads, bet-on-tails
    - Acts are functions from states to outcomes
      - bet-on-heads(heads)=+$10

            - bet-on-heads(tails)=-\$10
            - bet-on-tails(heads)=-\$10
            - bet-on-tails(tails)=+\$10

- To calculate the expected utility of an act: $EU(A) = \sum_i p(S_i)u(A(S_i))$
    - Supposing that utilities are linear in dollars (one utile per dollar), that $p(\text{heads}) = .6$, and that $p(\text{tails}) = .4$:
        - $EU(\text{bet-on-heads}) =$ $p(\text{heads})u(\text{bet-on-heads(heads)}) + p(\text{tails})u(\text{bet-on-heads})u(\text{bet-on-heads(tails)}) =$ $(.6)(10) + (.4)(-10) = 2$
        - $EU(\text{bet-on-tails}) =$ $p(\text{heads})u(\text{bet-on-tails(heads)}) + p(\text{tails})u(\text{bet-on-tails})u(\text{bet-on-tails(tails)}) =$ $(.6)(-10) + (.4)(10) = -2$
        - So, we should choose action bet-on-heads

3 Jeffrey's Decision Theory

- Objection to Savage's theory: Suppose that I park my car in a sketchy neighborhood and am approached by a suspicious character who offers to "protect" it for me while I am about my business; his price is \$10. Now suppose I reason as follows. There are two possible states of the world, one in which this suspicious character will smash my windshield and one in which he will not. If I pay him the \$10 then the possible outcomes here are one in which my windshield is not smashed and I am out \$10, and one in which my windshield is smashed and I am still out \$10. If I don't pay him, the possible outcomes are smashed vs. not-smashed, though I am not out \$10 in either case. Whatever the probabilities of the states are here, the expected utility of not paying him will clearly be higher on Savage's account, since the outcomes determined by that act are better in each possible state.

- Intuitively, the solution to this problem involves thinking about the probabilities of states, conditional on acts. But, for Savage, states are the (unique) bearers of probabilities, and since it doesn't make sense to think about the probability of an act, it also doesn't make sense to think about the conditional probability of a state, conditional on an act (bear in mind the Ratio Formula here).

- Instead, Jeffrey (1965) treats acts, states, and outcomes as propositions, which can all be arguments for both the probability and utility functions

- Consider all the "ways things could turn out"—i.e., a partition of the outcome space Ω into (mutually exclusive and jountly exhaustive) ω_i's

- For Jeffrey, the "evidential" expected utility of act A, $EEU(A) = \sum_i p(\omega_i|A)u(\omega_i \wedge A)$
- Jeffrey's theory can deliver the result that we ought to buy the protection in the "sketchy neighborhood" case
- Jeffrey's theory (unlike Savage's theory) is "partition-invariant": no matter how you partition Ω into ω_i (representing "ways things could turn out"), you will get the same $EEU(A)$

4 Causal Decision Theory

- The main objections to Jeffrey's "evidential" decision theory stem from a resistance to the idea that conditional probabilities like $p(\omega_i|A)$ are (always) the right probabilities to use when calculating EU
- Suppose, for example, that new research shows us that smoking does not actually cause lung cancer. As it turns out, the correlation between smoking and lung cancer is due to a common cause—there is a gene that disposes one to smoke, and also disposes one to develop lung cancer. Suppose that you know that you'll enjoy smoking somewhat, but that you *really* don't want to get cancer.
  - On the one hand: the probability of you having the gene, and thus developing lung cancer, conditional on you smoking, is quite high; thus, the EEU of smoking is quite low. And the probability of you having the gene, and thus developing lung cancer, conditional on you not smoking, is quite low; thus, the EEU of not smoking is high.
  - On the other hand: either you have the gene or you don't, and there's nothing you can do about it. So you might as well go ahead and get the "free" pleasure of smoking.
- Some responses from evidential decision theorists:
  - Accept the implication that you shouldn't smoke in the case above. Similarly, accept "one-boxing" as the rational response to the Newcomb Problem.
  - "Tickle Defense"
  - Add a necessary condition for rational action, such as the requirement that a rational action be *ratifiable*. An action is unratifiable when, on the assumption that the agent certainly performs it, another option has higher expected utility.
- Causal decision theorists insist that the right probabilities to use when calculating EU should take account of *causal* impacts of considered actions on the world, but not the merely *evidential* correlations of considered actions with states of the world

  - According to the causal decision theorist, Jeffrey's evidential decision theory goes wrong by taking account of these mere evidential correlations

- Causal decision theorists have taken different approaches in working this thought out

  - Skyrms (1980):
    - Look for a "special" partition of Ω into "causal dependency hypotheses" K_i that specify how the various outcomes that you care about depend causally on what you do.
    - Then, calculate the causal expected utility of an act as $CEU(A) = \sum_i p(K_i) \sum_j p(\omega_j | A \wedge K_i) u(\omega_j)$
    - I.e., calculate the expected evidential value of the act on each possible assumption about what the causal connections might be, and then sum up, weighting by the probability of each assumption about what the causal connection might be.
    - The thought is supposed to be that, when the causal structure is given, the tendency of an action to bring about an outcome coincides with the conditional probability of the outcome given the action
    - But, note that we lose the partition-invariance that was attractive in Jeffrey's theory
  - Gibbard and Harper (1978) propose that we calculate the CEU of an act as $\sum_i p(A \,\Box\!\!\rightarrow \omega_i) u(\omega_i)$
    - Lewis (1981) argues that the Gibbard-Harper approach is equivalent to Skyrms's, and hence also gives up on partition-invariance
  - Joyce (1999) has a more complicated proposal that makes use of "imaging" and which, he argues, restores partition-invariance

- Egan (2007) objects to causal decision theory: Johnny has devised a button which, if pressed, will kill all psychopaths. Johnny believes himself not to be a psychopath and places a high value on eliminating psychopaths from the world. And yet, Johnny believes that only a psychopath would push the button, and he values his own preservation much more than he values eliminating all psychopaths from existence. Intuitively, it seems to many, Johnny should not push the button, since that would tell him that his action is very likely to cause his own death (note the mixture of evidential and causal considerations here).

  - Causal decision theory seems to say that he should press the button

  - There are two relevant causal dependency hypotheses here: K_1: If Johnny presses the button he and all psychopaths will die; if he doesn't press it nobody will die; and K_2: If Johnny presses the button all psychopaths will die yet he will survive; if he doesn't press, nobody dies.
  - Given K_1, pressing is a terrible idea, whereas it is a good idea given K_2. As for not pressing, it's pretty much neutral, maintaining the status quo either way. Since Johnny thinks it very unlikely that he is a psychopath, however, K_2 seems much more likely, and so pressing will come out on top, since the expected good on K_2 will outweigh the very improbable badness that would result on K_1.
  - The trouble seems to stem from the fact that Johnny's action is *evidentially* relevant to which K_i obtains, but this factor is not accounted for by the causal account, since we use just $p(K_i)$ to calculate CEU.

5 Representation Theorems

- The representation theorems that most philosophers focus on show that, if an agent's preferences satisfy certain "rationality" constraints, then those preferences can be represented by a probability and utility function such that she non-strictly prefers an act A_1 to A_2 just in case A_1's expected utility is at least as high as A_2's. So, agents whose preferences satisfy the relevant constraints can be represented as expected utility maximizers.

- Representation theorems have sometimes been presented as an argument for probabilism—i.e., the view that credences are rationally bound to obey the probability calculus

- Different representation theorems appeal to different "rationality" constraints on preferences.
  - Non-strict preference is symmetric, transitive, and strongly connected (Thoma uses "complete" rather than "strongly connected")—hence, acts are totally preordered/quasi-ordered by non-strict preference
  - Some kind of "separability" or "independence" assumption, according to which the utility of a particular outcome isn't impacted by the other possible outcomes.
    - One version of a separability assumption: For all X and Y (and for all E): If E and $\neg E$ are mutually exclusive and exhaustive events, then, for all Z and W, the agent prefers $\{X \text{ if } E, Z \text{ if } \neg E\}$ to $\{Y \text{ if } E, Z \text{ if } \neg E\}$ iff they prefer $\{X \text{ if } E, W \text{ if } \neg E\}$ to $\{Y \text{ if } E, W \text{ if } \neg E\}$.

    - Another version: Let $\mathscr{L}$ be the space of lotteries over all possible outcomes. Then for all L_x, L_y, $L_z \in \mathscr{L}$ and all $p \in (0,1)$, L_x is preferred to L_y iff $p \cdot L_x + (1-p) \cdot L_z$ is preferred to $p \cdot L_y + (1-p) \cdot L_z$

- Worries about using representation theorems to establish probabilism
  - Category 1: Resistance to the "rationality" constraints on preferences
    - Thoma considers various sources of resistance
  - Category 2: Resistance to the move from the representation theorem to probabilism
    - Thoma: "The representation theorems only show that an agent who abides by the axioms of the various representation theorems can be represented as an expected utility maximiser. But this is compatible with the claim that the agent can be represented in some other way. It is not clear why the expected utility representation should be the one which furnishes the agent's beliefs and desires."
    - Hájek (2008): "Probabilism would arguably follow from the representation theorem if *all* representations of the preference-axiom-abiding agent were probabilistic representations. Alas, this is not the case, for the following 'mirror-image' theorem is equally true:

      > If all your preferences satisfy the same 'rationality' conditions, then you can be interpreted as maximizing non-expected utility, some rival to expected utility, and in particular as having credences that violate probability theory.

      How can this be? The idea is that the rival representation compensates for your credences' violation of probability theory with some non-standard rule for combining your credences with your utilities. Zynda 2000 proves this mirror-image theorem. As he shows, if you obey the usual preference axioms, you can be represented with a sub-additive belief function, and a corresponding combination rule."