Week 1 Summary: Introduction. Decision Trees.

02.229 - Decision Theory and Practice, 2019 Jan-April

Yustynn Panicker

January 28, 2019

Contents

1 Peterson CH1: Introduction **1**
1.1 Terms . . . 1
1.1.1 Self-Explanatory . . . 2
1.2 Right vs Rational . . . 2
1.3 Claim: Pragmatically, Normative Decision Theory > Descriptive Decision Theory . . . 2
1.3.1 My Problems with it . . . 2
1.4 Instrumental Rationality . . . 2
1.4.1 Is this aim always rational? . . . 3
1.5 Jean-Jacques Rousseau's `Stag Hunt` . . . 3
1.6 History of Decision Theory . . . 3
1.6.1 Period 1: Old Period (Ancient Greece) . . . 3
1.6.2 Period 2: Pioneering Period (>1650s) . . . 3
1.6.3 Period 3: Axiomatic Period (>1920s) . . . 3

2 Peterson CH2: The Decision Matrix **3**
2.1 Terms . . . 3
2.1.1 Scales . . . 4
2.2 3 Transformative Decision Rules . . . 4
2.2.1 Order-Independence (OI) . . . 4
2.2.2 The Principle of Insufficient Reason (ir) . . . 4
2.2.3 Merger of states (ms) . . . 4

3 Gilboa CH1: Feasibility and Desirability **4**
3.1 No direct causal link . . . 4
3.1.1 Zen and the Absurd (as in Camus Absurd) . . . 5
3.2 Uncertainty and Feasibility . . . 5
3.2.1 Example . . . 5
3.3 Link is mediated by information . . . 5
3.3.0.1 Example . . . 5

1 Peterson CH1: Introduction

1.1 Terms

Term	Definition/Explanation	Notes
Risk	$\mathbb{P}$(outcomes) known	

Continued on next page

Continued from previous page

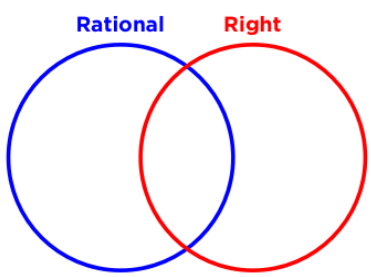
Term	Definition/Explanation	Notes
Ignorance	$\mathbb{P}$(outcomes) unknown	
Uncertainty	(Ignorance) $\vee$ (Risk $\cup$ Ignorance)	Context-dependent :(
Rational	Most reasonable outcome (ex ante)	I assume reasonability is based on available info
Right	Outcome is (at least weakly) pareto dominant. (ex post)	
Social Choice Theory	More than one decision maker	Some collective entities can be reduced to single decision makers ($\therefore$ not SCT)

1.1.1 Self-Explanatory

- Decision-maker
- Set of alternatives
- True state of the world
- Outcome
- Principle of maximizing expected value

1.2 Right vs Rational

Due to imperfect information, rationality does not necessarily correspond to rightness.



1.3 Claim: Pragmatically, Normative Decision Theory > Descriptive Decision Theory

Author claims that people behave rationally most of the time as they have good lives. Possibly flimsy argument (I don't like it).

1.3.1 My Problems with it

- `Rational` $\neq$ `right`, as seen above. E.g. maybe good lives are due to instinct rather than rationally correct behaviour
- People aren't living close to their best possible lives. Most of their lives suck. IMO people operate on habit more than reason

1.4 Instrumental Rationality

Presupposes an aim (which is external to `decision theory`)

1.4.1 Is this aim always rational?

- Widely thought that single aims cannot be evaluated in terms of rationality (though sets of aims can be irrational, e.g. inconsistent)
- John Rawls argues some aims are irrational (e.g. counting blades of grass on a courthouse lawn is too unimportant to be rational)
  - IMO, problematic argument. Importance varies according to values, values vary between people and even within the same person they are temporally inconsistent

1.5 Jean-Jacques Rousseau's `Stag Hunt`

	stag	rabbit
stag	5 , 5	0 , 3
rabbit	3 , 0	3 , 3

- Tension between risk minimization and outcome maximization
- Rational choice is solely and directly dependent on `trust`

1.6 History of Decision Theory

1.6.1 Period 1: Old Period (Ancient Greece)

- Normative decision examples instead of rules
- Followed by 1500 years of decision theory stagnation

1.6.2 Period 2: Pioneering Period (>1650s)

- Probability theory developed (Pascal and Fermat through letter correspondence)
- Some resistance by Catholic Church in normative moral theory (of course)
- 1738, `moral value` (now known as utility) was coined

1.6.3 Period 3: Axiomatic Period (>1920s)

- Attempt to make axioms from priciples of rational decision making
- 1950s was a golden age for decision theory
  - Still highly relevant to today

2 Peterson CH2: The Decision Matrix

2.1 Terms

Notation	Term	Definition/Explanation
(square in decision tree)	Choice Node	-
(circle in decision tree)	Chance Node	-
π	*Formal decision problem*, $\pi =< A, S, O >$	<A, S, O> = Acts, States, Outcomes
$t(\pi) \succeq \pi$	-	$t(\pi)$ is *at least* as reasonable as π
$t(\pi) \sim \pi$	-	$t(\pi)$ is *exactly* as reasonable as π
$a \circ b$	-	$(a \circ b)(\pi) = b(a(\pi))$
-	Transformative decision rule	Decision rule that modifies formalization of a decision problem
-	Effective decision rule	Filter that singles out some set of recommended acts
-	Rival formalizations	$\geq$ 2 formalizations of same problem that are both 1. equally reasonable and 2. strictly better than other formalizations

2.1.1 Scales

Scale	Strictly increasing	Difference information	Ratio information
Ordinal	Yes	No	No
Cardinal: Interval	Yes	Yes	No
Cardinal: Ratio	Yes	Yes	Yes

Note: Ordinal scales are invariant up to positive monotone transformations

2.2 3 Transformative Decision Rules

2.2.1 Order-Independence (OI)

If OI-condition holds for all $\pi \in \Pi$:

- $(u \circ t)(\pi) = (t \circ u)(\pi)$

2.2.2 The Principle of Insufficient Reason (ir)

If state probabilities are unknon, π may be transformed into π' in which *equal probabilities are assigned to all states*

2.2.3 Merger of states (ms)

If $\geq$ 2 states yield identical outcomes under all acts, they can be collapsed into one (with probabilities summed up, if known)

3 Gilboa CH1: Feasibility and Desirability

Can (feasibility) vs want (desirability)

3.1 No direct causal link

Usually, straightforwardly independent

3.1.1 Zen and the Absurd (as in Camus Absurd)

Under some cases (e.g. mathematicians who like challenges), feasibility alsely) appears to have direct negative causal link with desirability.

Author argues that the act of challenge is sought rather than the state of infeasibility, and thus the causal link still does not exist.

Violates Occam's razor, but makes sense

3.2 Uncertainty and Feasibility

Feasibility of states need not propogate to feasibility of states

3.2.1 Example

You can certainly perform the act try to solve a math problem for 2h, without being certain about whether the state of having it solved is indeed achievable)

3.3 Link is mediated by information

3.3.0.1 Example

- Desire: Buy strawberries
- Situation: End of day; 1 box of strawberries left over ($\therefore$ feasible)
- Thought process: Why did no one buy that last box? Does it suck?
- Decision: Don't buy, even though feasible, as the unlikeliness of the feasibility itself may be a signal about quality

Week 1 Summary: Introduction. Decision Trees.

02.229 - Decision Theory and Practice, 2019 Jan-April

Yustynn Panicker

February 4, 2019

Contents

1 Effective Decision Rules **2**
- 1.1 Maximin and Maximax . . . 2
  - 1.1.1 Maximin . . . 2
  - 1.1.2 Maximax . . . 2
- 1.2 Further Constraints: Leximin and Leximin Modifications . . . 2
  - 1.2.1 Explanation . . . 2
  - 1.2.2 Formal Definition . . . 2
- 1.3 Combination: Optimism-Pessimism Rule (a.k.a alpha-index rule) . . . 2
  - 1.3.1 Random Thought . . . 2
- 1.4 Problem: Relevance of non-extreme values [all] . . . 3
- 1.5 Problem: Unintuitive equivalence [Minimax, Maximax] . . . 3
- 1.6 Minimax Regret . . . 3
  - 1.6.1 Explanation . . . 3
  - 1.6.2 Procedure (won't formally describe) . . . 3
  - 1.6.3 Note . . . 3
  - 1.6.4 Problem: Argument from irrelevant alternatives . . . 4
    - 1.6.4.1 Example: Addition of a_5 . . . 4
    - 1.6.4.2 Counter . . . 4

2 Transformative Decision Rules **4**
- 2.1 Principle of Insufficient Reason . . . 4
  - 2.1.1 Problem: Which states should be considered? Modeling problem . . . 4
  - 2.1.2 Problem: Uniform probability assumption seems arbitrary . . . 4
    - 2.1.2.1 Counter: Symmetry . . . 4
      - 2.1.2.1.1 Problem . . . 5
  - 2.1.3 Problem: Practically, it's often not complete ignorance . . . 5
- 2.2 Randomized Acts . . . 5
  - 2.2.1 Procedure . . . 5
    - 2.2.1.1 Example . . . 5
  - 2.2.2 Potential Problem . . . 5

3 Axiomatic Analysis of the Decision Rules **5**
  - 3.0.1 Descriptions of Axioms . . . 6
  - 3.0.2 Axiomatic Analysis . . . 6

1 Effective Decision Rules

1.1 Maximin and Maximax

They're the same, just at different extreme ends.

1.1.1 Maximin

Informal Principle of choosing the act with the largest minimal outcome obtainable

Formal $a_i \succeq a_j \iff \min(a_i) \geq \min(a_j)$

1.1.2 Maximax

Informal: Principle of choosing the act with the largest maximal outcome obtainable

Formal: $a_i \succeq a_j \iff \max(a_i) \geq \max(a_j)$

1.2 Further Constraints: Leximin and Leximin Modifications

Below is for leximin. Leximax is just the opposite

1.2.1 Explanation

Description: Essentially a way to filter the dominance space from `maximin` (also an `effective decision rule`)

Procedure: Iteratively compare the next minimal outcomes under the states until you find a difference. Remove the act(s) with a lower outcome

Equivalence: The only remaining equivalent acts are acts which are equivalent in every state

1.2.2 Formal Definition

$a_i \succ a_j \iff$ there exists some positive integer n such that $\min^n(a_i) > \min^n(a_j)$ and $\min^m(a_i) = min^m(a_j)$ for all $m < n$

1.3 Combination: Optimism-Pessimism Rule (a.k.a alpha-index rule)

Informal: Weighted combination of maximin and maximax. Weight parameter α reflects optimism

Formal: $a_i \succeq a_j \iff \alpha \cdot \max(a_i) + (1-\alpha) \cdot min(a_i) \geq \alpha \cdot \max(a_j) + (1-\alpha) \cdot \min(a_j)$

1.3.1 Random Thought

You can probably combine with leximin/leximax to some degree too, to get an even better framework

1.4 Problem: Relevance of non-extreme values [all]

Particularly easy to see when the mins are close. E.g. below.

Applying the maximin principle selects option a_2, but intuitively a_1 seems far better.

Note: Example formulated for `maximin/leximin`. Just flip the logic for `maximax`

	s_1	s_2	s_3	s_4	s_5	s_6	s_7
a_1	1	0.99	99999	99999	99999	99999	99999
a_2	1	1	1	1	1	1	1

Pedantic Note: you don't know the probability distribution of the state space. Maybe a_2 is better after all...

1.5 Problem: Unintuitive equivalence [Minimax, Maximax]

Note: Example formulated for `maximin`. Just flip the logic for `maximax`

	s_1	s_2
a_1	1	99999
a_2	1	1

Under vanilla `maximin`, both acts are equally reasonable. Obviously, this is weird

1.6 Minimax Regret

1.6.1 Explanation

- Essentially an attempt to formalize the concept of `regret`

1.6.2 Procedure (won't formally describe)

Before

a_1	12	8	20	**20**
a_2	10	**15**	16	8
a_3	**30**	6	25	14
a_4	20	4	**30**	10

-30 -15 -30 -20

After

a_1	**−18**	−7	−10	0
a_2	**−20**	0	−14	−12
a_3	0	**−9**	−5	−6
a_4	−10	**−11**	0	−10

1.6.3 Note

It's not globally accepted that this concept is relevant to rational decision making. But a substantial number of theorists think it is.

1.6.4 Problem: Argument from irrelevant alternatives

- Ranking can be altered by adding a non-optimal alternative
- Breaks intuition about how a "normatively plausible decision rule must not be sensitive tothe addition of irrelevant alternatives"

1.6.4.1 Example: Addition of a_5

Table 3.14

a_1	12	8	20	20
a_2	10	**15**	16	8
a_3	**30**	6	25	14
a_4	20	4	**30**	10
a_5	−10	10	10	**39**

Table 3.15

a_1	−18	−7	−10	**−19**
a_2	−20	0	−14	**−31**
a_2	0	−9	−5	**−25**
a_3	−10	−11	0	**−29**
a_5	**−40**	−5	−20	0

1.6.4.2 Counter

- Prima facie intuition is wrong. It's "rational to compare alternatives with the entire set of alternatives".

2 Transformative Decision Rules

2.1 Principle of Insufficient Reason

Pro: Decision under ignorance → Decision under risk

2.1.1 Problem: Which states should be considered? Modeling problem

- More states means lower probability for each state (direct influence on choice strategy)
- Choosing relevant states is often not easy
- Traditional argument for `ir` is from symmetric states (e.g. dice sides). Many problems have no such symmetry

2.1.2 Problem: Uniform probability assumption seems arbitrary

Under ignorance, any probability distribution *seems to be* equally justifiable as any other. Assumption of equality seems arbitrary

2.1.2.1 Counter: Symmetry

- Assume every probability distribution is equally justifiable
- Use lens of toy 2-state case $S = \{s_1, s_2\}$
- Every probability distribution has a symmetric partner (e.g. $\{p_{s_1} = 0.6, p_{s_2} = 0.4\}$ has $\{p_{s_1} = 0.4, p_{s_2} = 0.6\}$

- Exception: Uniform distribution. Suggests uniform distribution is a collapsed state of (in this case) 2 identical probability distributions. Making it multiplicatively more reasonable as any other case (in this case, 2x more reasonable)

2.1.2.1.1 Problem

- Beautiful argument, but the assumption of the uniform distribution being an additively collapsed one is a bit dubious imo

2.1.3 Problem: Practically, it's often not complete ignorance

You generally know some things or at least have a sense of ordinal ranking for the probabilities of some of the states. Why not use it?

2.2 Randomized Acts

2.2.1 Procedure

- Create a new act with expected values as outcomes
- If your decision making strategy selects random act, then randomly choose one of those initial acts
- Note: Choosing the random act is not a choice on its own, but a procedure to arrive at an actual choice

2.2.1.1 Example Introduce random act a_3

Table 3.18

a_1	1	0
a_2	0	1

Table 3.19

a_1	1	0
a_2	0	1
a_3	1/2	1/2

2.2.2 Potential Problem

I'm assuming the random choice doesn't have to be uniformly distributed. But this opens up a whole can of worms by allowing you to tweak the probability distribution of the random function to bias it towards whatever choice you irrationally want.

3 Axiomatic Analysis of the Decision Rules

Taken directly and shamelessly from the textbook

3.0.1 Descriptions of Axioms

1. **Ordering**: $\succeq$ is transitive and complete. (See Chapter 5.)
2. **Symmetry**: The ordering imposed by $\succeq$ is independent of the labelling of acts and states, so any two rows or columns in the decision matrix could be swapped.
3. **Strict Dominance**: If the outcome of one act is strictly better than the outcome of another under every state, then the former act is ranked above the latter.
4. **Continuity**: If one act weakly dominates another in a sequence of decision problems under ignorance, then this holds true also in the limit decision problem under ignorance.
5. **Interval scale**: The ordering imposed by $\succeq$ remains unaffected by a positive linear transformation of the values assigned to outcomes.
6. **Irrelevant alternatives**: The ordering between old alternatives does not change if new alternatives are added to the decision problem.
7. **Column linearity**: The ordering imposed by $\succeq$ does not change if a constant is added to a column.
8. **Column duplication**: The ordering imposed by $\succeq$ does not change if an identical state (column) is added.
9. **Randomisation**: If two acts are equally valuable, then every randomisation between the two acts is also equally valuable.
10. **Special row adjunction**: Adding a weakly dominated act does not change the ordering of old acts.

3.0.2 Axiomatic Analysis

	Maximin	Optimism–pessimism	Minimax regret	Insufficient reason
1. Ordering	⊗	⊗	⊗	⊗
2. Symmetry	⊗	⊗	⊗	⊗
3. Strict dominance	⊗	⊗	⊗	⊗
4. Continuity	⊗	⊗	⊗	⊗
5. Interval scale	×	⊗	×	×
6. Irrelevant alternatives	⊗	⊗	–	⊗
7. Column linearity	–	–	⊗	⊗
8. Column duplication	⊗	⊗	⊗	–
9. Randomisation	⊗	–	⊗	×
10. Special row adjunction	×	×	⊗	×

Symbol	Meaning
-	Incompatible with decision rule
×	Compatible with decision rule
⊗	Necessary and sufficient for decision rule

Week 3 Summary: Probability theory and Bayes' Rule. Expected value maximization.

02.229 - Decision Theory and Practice, 2019 Jan-April

Yustynn Panicker

February 10, 2019

Contents

1 Lack of Summary of Notes **1**

1 Lack of Summary of Notes

To be frank, I already know this material (mostly from the course `40.001 Probability`). I did input a lot of it into anki, but I'd rather not expend the effort to summarize basic probability rules if that's okay.

I'm fine dropping this summary as one of my lowest scores, and this drop won't be repeated. I don't know any of the other topics well enough that I'd want to skip doing the summary.

Week 4 Summary: Preferences and utility. Expected utility maximization.

02.229 - Decision Theory and Practice, 2019 Jan-April

Yustynn Panicker

February 18, 2019

Contents

1 Terminology **1**

2 Tadelis Ch1 **2**
2.1 Completeness Axiom in Preference Relations . . . 2
2.2 Rational Choice Paradigm . . . 2
2.2.1 Assumptions . . . 2
2.2.2 Meaning . . . 2
2.2.3 `Homo Economicus` . . . 2
2.2.3.1 Fun Facts . . . 3

3 Tadelis Ch2.1-2.3 + Peterson CH4.1 - 4.3, 4.6 **3**
3.1 Simple and Compound Lotteries . . . 3
3.1.1 Meaning of Simple Lottery . . . 3
3.1.2 Compound Lotteries . . . 3
3.2 Backward Induction . . . 3
3.2.1 Relation to Dynamic Programming . . . 3
3.3 Maximizing Expectation . . . 3
3.3.1 Units . . . 3
3.3.2 Keyne's Objection . . . 4
3.4 Axiomatic Approach . . . 4
3.4.1 Axioms . . . 4
3.5 Misc . . . 4
3.5.1 `de minimis` Principle as a Solution to the St Petersburg Paradox . . . 4
3.5.2 Two-Envelope Paradox (infinite swapping reasoning) . . . 4
3.5.2.1 Set Up . . . 4
3.5.2.2 Choice . . . 4
3.5.2.3 The paradox . . . 5
3.5.2.4 Comments . . . 5

1 Terminology

Term	Definition/Explanation	Notes
Terminal node	Decision tree, the leafs where payoffs are attached	
Payoff function	Function that maintains transitivity of pre-existing preference relations	

Continued on next page

Continued from previous page

Term	Definition/Explanation	Notes
Degenerate lottery	A deterministic situation (probably pointlessly) modelled as a lottery	Probabilities of states arising based off chosen actions would be 1 or 0
Intensity	Essentially, magnitude	E.g. intensity of preference is not invariant to positive monotonic transforms (it's cardinal, not ordinal)
Risk neutral	Indifferent between a lottery and sure situation given that the expected payout is the same	
Risk averse	Strictly prefers a sure situation to a lottery, given that the expected payout is the same	
Risk loving	Strictly prefers a lottery to a sure situation, given that the expected payout is the same	Not as ridiculous at it seems *prima facie*. In situations where you're wealthy/have nothing substantial to lose, risk is fun.
Risk neutral	Indifferent between a lottery and sure situation given that the expected payout is the same	
Degenerate lottery	A deterministic situation (probably pointlessly) modelled as a lottery	Probabilities of states arising based off chosen actions would be 1 or 0

2 Tadelis Ch1

2.1 Completeness Axiom in Preference Relations

Complete iff any two outcomes can be ranked by preference relation

2.2 Rational Choice Paradigm

2.2.1 Assumptions

Full understanding of the following:

1. All possible acts
2. All possible outcomes
3. Mappings for every state/outcome pair
4. His rational preferences (payoffs) over outcomes

2.2.2 Meaning

Rational if payoff-maximizing.

Manner of maximization is unclear in the text (is it expectation maximization?) but various methods were covered in `Week 2`.

2.2.3 `Homo Economicus`

Means "economic man". Essentially, the perfect rational agent.

2.2.3.1 Fun Facts

- Wikipedia describes this as a "caricature" and "mythical species", although in a way that seems to be a tool, rather than pejoratively
- Oppositionally, see homo reciprocans ("reciprocating man").
  - "cooperative actors who are motivated by improving their environment."
- Sometimes used by non-economists to critique economic approaches (distilled down to the "people aren't rational" argument)

3 Tadelis Ch2.1-2.3 + Peterson CH4.1 - 4.3, 4.6

3.1 Simple and Compound Lotteries

3.1.1 Meaning of Simple Lottery

- Stochastic state distribution
- Probability of states are contingent upon performed acts, which suggest that acts are not necessarily seen as independent of states. This is different from what we learnt in class (assumption of causal independence)

3.1.2 Compound Lotteries

- A simple lottery after a simple lottery (multi-level decision tree)
- Can be transformed into a simple lottery (take each path as a single decision, then consider terminal nodes only)

3.2 Backward Induction

It's easiest to work backwards then dealing with multi-stage decisions trees.

3.2.1 Relation to Dynamic Programming

It's one strategy for it, but certainly not the only one (e.g. you could just as easily work forwards in shortest path problems)

3.3 Maximizing Expectation

3.3.1 Units

Unit	Explanation/Notes
Monetary value	$$$
Value	Monetary value, adjusted for diminishing marginal 'value' of money to the agent
Utility	Yes. Not well defined, but POV is important. Allows for non-monetary evaluation

Utility, value, monetary value, payoff

3.3.2 Keyne's Objection

"In the long run we are all dead".

Essentially, a criticism of relying on LLN.

3.4 Axiomatic Approach

Arrives at the principle of maximizing utility without using LLN (thereby sidestepping Keynes' problem).

3.4.1 Axioms

1. If all outcomes of an act have utility u, then the utility of the act is u
2. For acts, utility relations follow dominance relations (e.g. strict dominance implies higher utility)
3. Every decision probable can be transformed into a rival formalization with equiprobable (possibly repeated) states, which maintain the utilities of all acts
4. Setup: 1) two equiprobable outcomes 2) The better outcome is made slightly worse Claim) Overall utility of act can be preserved by adding some amount of utility to the other outcome

These 4 axioms lead to the following conclusion: the utility of an act = its expected utility

3.5 Misc

3.5.1 `de minimis` Principle as a Solution to the St Petersburg Paradox

Roughly put, sufficiently improbable outcomes should be ignored

3.5.2 Two-Envelope Paradox (infinite swapping reasoning)

- No established solution exists.
- This is **not** the same as the Monty Hall problem

3.5.2.1 Set Up

- Two seemingly identical envelopes (A and B)
- 1 envelope contains twice as much as the other
- You don't know which is which, can only choose 1

3.5.2.2 Choice

- Make a decision
- Given chance to swap decision

3.5.2.3 The paradox

- $\mathbb{E}[swap] = \frac{1}{2} \cdot 2x + \frac{1}{2} \cdot \frac{x}{2} = \frac{5}{4}x$, so logically, swap
- Given choice again just before picking
- $\mathbb{E}[swap] = \frac{1}{2} \cdot 2y + \frac{1}{2} \cdot \frac{y}{2} = \frac{5}{4}y$, so logically, swap
- Given choice again, just before picking
- Repeat ad infinitum

3.5.2.4 Comments The problem is not to find an alternate formulation of the given situation. The problem is to find the issue with this current formulation, which gives nonsensical results.

Week 5 Summary: Discounting the future. The value of information. Biased choice.

02.229 - Decision Theory and Practice, 2019 Jan-April

Yustynn Panicker

February 25, 2019

Contents

1 Discounting the Future **2**
- 1.1 Misc . . . 2
  - 1.1.1 Example case . . . 2
- 1.2 Strategy . . . 2
  - 1.2.1 Quickly calculating optimal act . . . 2
    - 1.2.1.1 **TODO** ??? Why can we assume concavity? . . . 2

2 The Value of Information **2**

3 Biased Choice **2**
- 3.1 Possible Fallacies . . . 2
  - 3.1.1 Framing effects . . . 3
  - 3.1.2 Endowment effect . . . 3
  - 3.1.3 Sunk Costs . . . 3
- 3.2 Heuristics . . . 3
  - 3.2.1 Representative Heuristic . . . 3
  - 3.2.2 Availability Heuristic . . . 3

4 Multiple Selves as a Framework to Understand Irrational Behavior (Ainslie) **3**
- 4.1 General Argumentation Strategy . . . 3
- 4.2 Subtle vs Visceral Rewards . . . 4
- 4.3 Status Quo: The "Shopping List" Definition of Value . . . 4
- 4.4 Problems . . . 4
  - 4.4.1 Many rewards function as punishments as well . . . 4
  - 4.4.2 Abstemoniousness in situations where we could very well reward ourselves . . . 4
  - 4.4.3 Events serving as rewards are changeable . . . 4
  - 4.4.4 Many apparent rewards cannot be produced by direct effort . . . 4
- 4.5 Failed reductionist resolution attempts using the field of psychology . . . 4
  - 4.5.1 `Classical Conditioning` as a bridging mechanism (between 'subtle' and 'visceral' rewards) 4
  - 4.5.2 `Elementary Drives for Game-like Rewards` as a bridging mechanism (between 'subtle' and 'visceral' rewards) . . . 5
  - 4.5.3 Empirically, people often fail to max expected income/min costs . . . 5
  - 4.5.4 People consent to undergo pain . . . 5
  - 4.5.5 Empirically, discount factor seems to be very, very high . . . 5
- 4.6 Resolution using `multiple selves` (demonstrated by `intrapersonal bargaining`) . . . 5
  - 4.6.1 General rule of thumb . . . 5
    - 4.6.1.1 Devices that aid compromise . . . 5

1 Discounting the Future

1.1 Misc

Problem is known as `choosing consumption over time`. It's a formalization of why greedy approaches don't work.

1.1.1 Example case

Budgeting. You've got \$X to last you n time. Obviously, you can't just spend it straight away because you'll be screwed in the long run (almost immediate run, in this case)

1.2 Strategy

Essentially, do the following where:

- discount factor $\delta \in [0, 1]$
- x_1 represents the act of spending \$x$_1$
- \$X is the total amount of money you have

$$\max_{x_1} u(x_1) + \delta u(X - x_1)$$

1.2.1 Quickly calculating optimal act

If we assume u is concave, we just find the max point and solve for x_1, depending on the specifics of what your u function was in the first place. This is:

$$u'(x_1) = \delta u'(K - x_1)$$

1.2.1.1 TODO ??? Why can we assume concavity?

Idk yet.

2 The Value of Information

1. Take the act that gives the max expected value without information
2. Take the act that gives the max expected value with information
3. `2 - 1` gives you the value of information.

Note: Value can indeed be negative (maybe having the info would bias you to act badly)

3 Biased Choice

Microecons stuff, basically.

3.1 Possible Fallacies

Note that these are not definitely fallacies and often can be argued either way.

3.1.1 Framing effects

The effect the frame/representation has on decision

3.1.2 Endowment effect

Tendency to value what we ave more than what we do not yet have

Possible Reasons

- Information (more info once we own it)
- Transaction costs
- Habit formation (linked to transaction costs)

3.1.3 Sunk Costs

Yes.

3.2 Heuristics

Central idea is that in order to deal with overwhelming info, we resort to heuristics (which may give rise to the 'fallacies') (idea from Kahneman and Tversky)

3.2.1 Representative Heuristic

- You have a representation of a thing, and use this to govern your estimation of a probability distribution
- Leads to problems where $\mathbb{P}(AB)$ is seen as more likely than $\mathbb{P}(A)$ on its own if you think B is more in line with the 'representation'

3.2.2 Availability Heuristic

- If you can think of an example of it happening, you assume it's more likely
- Leads to problems of anecdotal evidence/whatever example you have the brainpower to conjure up being used

4 Multiple Selves as a Framework to Understand Irrational Behavior (Ainslie)

4.1 General Argumentation Strategy

Ainsle is essentially proposing a new psychological, reductionist framework of interpreting irrational behaviour: modelling each individual as multiple, internally rational selves. He contrasts this with other reductionist attempts from psychology which he claims have failed.

4.2 Subtle vs Visceral Rewards

This is his terminology. I don't think it's standard. Also, was a little tough to decipher so I'm not sure I'm correct in this.

Visceral: basal, biological/survival rewards (visceral)

Subtle: other rewards (emotional etc)

4.3 Status Quo: The "Shopping List" Definition of Value

Essentially, "that which a person will pay money for is good"

- Also being turned to by psychologists as a method of defining reward
- Glaring problem (in depth below) is that people often behave irrationallly

4.4 Problems

4.4.1 Many rewards function as punishments as well

4.4.2 Abstemoniousness in situations where we could very well reward ourselves

4.4.3 Events serving as rewards are changeable

E.g. diminishing returns, fashion, etc

4.4.4 Many apparent rewards cannot be produced by direct effort

Described as "states which usually occur as by-products", and "lost in any systematic attempt to attain them".

Examples given are:

- sleep
- laughter
- happiness
- dignity

Personally, I see no problem here and disagree that systematic attempts are not possible. I don't see the issue with a sequence of steps being needed in order to achieve these states. Particularly since we've seen that multi-level decision trees can be linearized into a decision matrix.

4.5 Failed reductionist resolution attempts using the field of psychology

4.5.1 `Classical Conditioning` as a bridging mechanism (between 'subtle' and 'visceral' rewards)

Atomic units here are the visceral rewards, which are composed with each other in complex manners.

- Attempts to model subtle rewards as a composition of visceral ones.
- Problematic as the modelling complexity and verification of the model are pretty untenable
  - e.g. do you like computer science because someone told you you're good at computers when you were a kid? Not easy to prove it, and not necessarily correct either

4.5.2 `Elementary Drives for Game-like Rewards` as a bridging mechanism (between 'subtle' and 'visceral' rewards)

Atomic units here are the elementary drives for game-like rewards, which are composed with each other in complex manners.

- Attempt in the 60s by Fowler (1967) and Hunt (1963)
- Problem: Tough to extricate properties of situation needed to satisfy these drives
- Declared a failure by Coombs and Avrunin (1977)

4.5.3 Empirically, people often fail to max expected income/min costs

- Value of money varies a lot
  - Different people value money differently
  - People treat money differently in different situations (e.g. you're more okay losing $100 when spending $1,000,000$ on a house than you would be normally)

4.5.4 People consent to undergo pain

4.5.5 Empirically, discount factor seems to be very, very high

4.6 Resolution using `multiple selves` (demonstrated by `intrapersonal bargaining`)

Atomic unit here is multilpe, internally-rational selves for each person, which interact through `intrapersonal bargaining` (basically the cartoon trope of angel/devil on your shoulders arguing over what you should do)

This is the crux of what the chapter is talking about. The author thinks it's very promising. IMO it's a great and promising approach, but he hasn't developed it well.

4.6.1 General rule of thumb

Agent will try to obtain "situations which permit a stable compromise between his long- and short-term interest"

4.6.1.1 Devices that aid compromise Note that this is supposedly descriptive. These devices supposedly act in concert to produce a prima facie irrationality.

- Extrapsychic devices (physical/social constraints which limit future choice space)
- Control of attention (philosophy of "out of sight -> out of mind")
- Control of emotions (e.g. "positive thinking")
- Private rules (grouping temptations into sets, to up the ante of breaking rules about giving in to any single temptation)

Week 6 Summary: Paradoxes of choice. Prospect theory. Biases in probabilistic reasoning.

02.229 - Decision Theory and Practice, 2019 Jan-April

Yustynn Panicker

March 4, 2019

Contents

1 Statistics-Related **1**
1.1 Biased Samples . . . 1
1.1.1 Winner's Curse . . . 1
1.1.2 Regression to Mean . . . 1
1.1.2.1 Slot machine example . . . 1
1.1.2.2 Why it's a biased sample example . . . 1
1.2 Confounding Conditional Probabilities . . . 1
1.2.1 Joke (Illustration of lack of noticing independence) . . . 1
1.3 Classical vs Bayesian Statistics . . . 1
1.3.1 Classical Statistics . . . 1
1.3.2 Bayesian Statistics . . . 2

2 Prospect Theory **2**
2.1 Misc . . . 2
2.2 Prospect Value . . . 2
2.3 On the Violations of Transitivity and Completeness . . . 2
2.3.1 Transitivity . . . 2
2.3.2 Completeness . . . 2
2.4 Usefulness of Descriptive Decision Theory . . . 3
2.5 Bounded Rationality . . . 3

1 Statistics-Related

1.1 Biased Samples

Often, bias arises in sampling methodology

- Psychological biases (e.g. framing effects, question order can influence answers)
- Sample may not be representative

1.1.1 Winner's Curse

This is a good example of sampling from a biased population.

Essentially, in auctions where lots have uncertain value, the winning bids tend to be higher than the unbiased estimates (because of course, the winner is the highest bid, not the most accurate one). Often times, these winners lose money (therefore they are 'cursed' with winning)

1.1.2 Regression to Mean

Anomolous behaviour should not be expected to continue, but rather to regress back to the mean.

1.1.2.1 Slot machine example If a slot machines keeps giving out wins, it'll regress back to the mean of not giving out wins. The 'lucky streak' thing is nonsense, statistically speaking)

1.1.2.2 Why it's a biased sample example We tend to go to restaurants that we had a good experience at, but that experience has a luck factor (it's not just the skill of the chef!). Our habit is thus formed based on data that was likely noised up by positive luck.

1.2 Confounding Conditional Probabilities

I really like the joke example given by Gilboa, so I'll list it here. The essence is that people are generally bad with conditioning probabilities.

1.2.1 Joke (Illustration of lack of noticing independence)

`Suppose you are about to board a flight and you fear the plane might be blown up by a bomb. An old joke suggests that you take a bomb on the plane with you, because the probability of two bombs is really low.`

Obviously, your choice to bring a bomb or not has no influence over other people bringing a bomb. These things are independent.

1.3 Classical vs Bayesian Statistics

1.3.1 Classical Statistics

- Point estimation
- Interval estimation

- Useful when dealing in "objectivity" (dubious objectivity, but ykwim)

1.3.2 Bayesian Statistics

- Allows for probabilistic treatment of μ (classical does not)
- Characterized by Bayesian updating

2 Prospect Theory

2.1 Misc

- Descriptive
- Kahneman and Tversky's attempt to reconcile readily-observed violations of expected utilty

2.2 Prospect Value

$$\sum_{i=1}^{n} w(p_i) \cdot w(u_i)$$

where u_i is the utility of outcome i, and p is the probability of it arising.

2.3 On the Violations of Transitivity and Completeness

2.3.1 Transitivity

Essentially, May observed cyclical prefernces in roughly 20% of study participants (21.5% for the original 306-participant study, 18.5% for another similar one).

For reference, here are the choices.

Table 14.2

	Ticket 1–30	Ticket 31–60	Ticket 61–100
A	$54	$0	$0
B	$24	$24	$0

Table 14.3

	Ticket 1–30	Ticket 31–60	Ticket 61–100
B	$24	$24	$0
C	$4	$4	$4

2.3.2 Completeness

It's suggested that our choices may be inherently probabilistic, which means that they inevitably violate completeness.

The implication is that completeness is descriptively incorrect as a choice of axiom, which is pretty interesting.

2.4 Usefulness of Descriptive Decision Theory

The primary issue is that there's a divorce between the correct thing to do, and what people actually do. This is observable from Hume's law (an `ought` cannot be derived from an `is`).

So the question is: why study descriptive decision theory at all?

1. Instrumental value for people whose outcomes depend on how others act (e.g. politicians. Could act in accordance with normaitivity, but the descriptive actions of other people is generally out of their control).
2. Intrinsic value (knowing how humans behave is intrinsically important, blah blah. True.)
3. Can help us be aware of what fallacies to avoid, etc

2.5 Bounded Rationality

Some decision theorists argue that rationality is a limited resource. The extreme case of this is to argue that normative decision theory is useless.

While the `bounded rationality` thing makes sense, the discarding of normativity is ridiculous, IMO. Fully agree with Peterson here: normativity gives us a goal post to move towards, and we have the potential to become more rational.

Week 9 Summary: Social preferences and choice.

02.229 - Decision Theory and Practice, 2019 Jan-April

Yustynn Panicker

April 8, 2019

Contents

1 Goal of Social Choice Theory **1**

2 Descriptive Issues **1**

3 Why not just use utility for SWF? **1**

4 Terminology **1**

5 Arrow's Impossibility Theorem **2**
5.1 Implication 2

6 Sen's Construction of Liberalism as Incompatible With Pareto **2**
6.1 Sen's Minimal Condition of Liberalism 2
6.2 Nozick's Criticism 2

7 Harsanyi's Utilitarian Theorems **2**
7.1 General 2
7.2 Issues/Limitations 2

1 Goal of Social Choice Theory

- Analyze collective decision problems

The real goal here is to find a proper reduce function (known in the field as a `Social Welface Function`, or SWF). that takes as input a set of individual preference orderings `G` and outputs a `social preference ordering` S

Term	Meaning
SWF	Social Welfare Function
G	set of `individual preference orderings`
S	`social preference ordering`

$$SWF : G \mapsto S$$

2 Descriptive Issues

- Cyclical preferences (known as the `voting paradox`)

3 Why not just use utility for SWF?

- Interpersonal utility comparisons are an intractable problem

4 Terminology

Term	Meaning	Remarks
Decisiveness (Gilboa calls this unanimity)	Decisive if everyone in the group has the same preference ordering	Can be decisive w.r.t some social state pair (e.g. (a,b)) or decisive overall
Non-dictatorship	No single individual of a group of people can make the group	Libertarianism seems inescapably to have dictators then
Ordering (property)	For any combination of individual preference orderings, the social preference ordering must be `complete`, `asymmetric` and `transitive`	
Asymmetric preference relation	Effectively, it means a strict preference relation (no indifference)	
Independence of irrelevent alternatives (property)	Pairwise preference relations are robust to changes in the set of alternatives	This can be seen as a more general statement of the problem of irrelevant alternatives in minimax regret
Pareto	The group of all individuals in society is decisive	Unclear on what society means, but I'm guessing any arbitrary group of >1 person is fine

5 Arrow's Impossibility Theorem

5.1 Implication

No SWF exists such that it meets the following conditions:

- Non-dictatorship
- Ordering
- Independence of Irrelevant Alternatives
- Pareto

6 Sen's Construction of Liberalism as Incompatible With Pareto

6.1 Sen's Minimal Condition of Liberalism

It's a very, very low bar.

Essentially, it's liberal iff $\exists \geq 1$ preference relation pair where a single individual's preference determines the entire group's preference for that relation pair

Obviously, this cannot then be pareto.

6.2 Nozick's Criticism

- Liberalism is not a SWF, but a constraint on the feasible set of alternatives

7 Harsanyi's Utilitarian Theorems

7.1 General

His overall idea is that rational individuals can represent represent the utility of a social state on an interval scale.

(recall that an interval scale is one in which $a - b$ holds meaning, but a/b does not)

7.2 Issues/Limitations

Mostly lie on his ethical premise that *equal treatment of all individuals* is desirable.

Problems are:

1. Accurate interpersonal comparisons of utility is intractable
2. It's normative, does not stem from descriptive premises

Personally, I don't think these problems are big deals. You can't run away from the need for some sort of ethical axiom, and you don't need a perfect comparison function for utility. I'd expect you to be more right than you are now if you implement this system, even if your comparison function isn't perfect. You don't need to chase perfection - just chase improvement.

Decision Theory Week 10:

Philosophy of Probability

- Subjective Viewpoints
  - General idea
    - Probability is exists only in the mind, and is a measure of the strength of belief
  - Main challenge
    - Finding a systematic procedure to derive PDFs
      - You're essentially tryna convert qualitative information into quantitative information
    - Attempt: Savage's Representation Theorem
      - Interesting axioms
        - Sure-thing principle
          - Obviously, contentious due to Allais paradox.
      - Result
        - $\exists$ a set of utility functions (closed under positive linear transforms, meaning that interval information is preserved), such that an agent *can be described* as acting with accordance to the principle of maximizing utility
        - Important points
          - Utility functions can be seen as interval scales
          - Only claims the agent appears to follow principle of maximizing utility, not whether or not the agent actually does follow it
    - General issues
      - State-dependent utilities
        - E.g. choice between getting $100 if a bomb that is next to you gets disarmed VS getting $100 if a bomb that is next to you explodes
        - Regardless of strength of belief of bomb going off (state), the obvious choice is to bet on the situation where you don't die as $100 doesn't help when you're dead
        - Why simply excluding state-dependent utilities is not feasible
          - You can't tell what is state-dependent and what is not
      - Framing choices is difficult (which information is relevant?)
        - E.g. framing effects of monetary payoffs. For instance, convert $ -> € and have the exchange rate be different for each state
      - Emotionally inert agents can hold beliefs of varying strength
        - Gave rise to minimal subjectivism
    - Attempt: Dutch Book Theorem
      - Dutch Book
        - Combination of bets that is certain to lead to a loss
      - Theorem
        - Dutch book exists $\iff$ DM's beliefs violate any of Kolmogrov's axioms
- Misc
  - Fair Price
    - Price you'd pay for a gamble such that you'd be indifferent to choosing either role (bookie or better)
  - Bayesian probability can be seen as subjective
- Objective Viewpoints
  - Classical interpretation
    - General idea
      - Probability space can be divided into equally possible events. Use that to derive PDF
        - E.g. biased die where P(6) = 2*P(1) = 2*P(2)
          - Equally possible events of P(1 OR 2) = P(6)
    - Problem
      - Difficult to define "possible" without circular reference to "probable"
    - Proponents
      - Laplace
      - Pascal
      - Bernoulli
      - Leibniz
  - Frequentist interpretation
    - Original
      - General idea
        - Observe real-life instances of a random variable. Probability of event E is
          - $\frac{\text{num of observed occurrences of E}}{\text{num of all occurrences of random variable}}$
      - Problem
        - How do you determine which "reference class" (set of observations) an event belongs to?
    - Venn's Addition
      - Modification
        - Split between true PDF and perceived PDF
          - Ontological claim that there exists a true probability distribution for any universe, which is the limiting frequency of the events
            - Note that this is abstract af. It's an infinite event generator, so obviously, doesn't exist in our finite universe
      - More Problems
        - Does this limiting frequency even exist?
          - It's infinite, our universe is finite
        - If it does exist, how do you gain access to it?
          - It's infinite, so any non-infinite size reference class has no reason to be representative. E.g. 100 sequential tails flips does not HAVE to mean that the coin is unfair
        - How do you repeat an experiment exactly?
          - You can never have the exact same conditions
  - Propensity interpretation
    - General idea
      - Probability measures the P"propensity", "disposition" or "tendency" of an object to give rise to a certain effect
    - Development
      - Initially made by Popper, but he mixed it with frequenist interpretation
      - Contemporary philosophers have constructed pure versions
    - Problems (very similar to freq problems)
      - What does propensity mean?
      - Why would propensity exist in the first place?
      - How do you justify deriving propensity from a finite series of observations?
      - Humphrey's Paradox
        - Essentially, arrow of time seems to matter.
          - E.g. P(late | train is delayed) makes sense. But saying P(train is delayed | late) does not, because the future shouldn't affect the past
  - Epistemic/Logical interpretation
    - General idea
      - Probability is a scalar measure of evidence
    - Problem
      - Sometimes, we want probability to be purely guess-based. This makes that impossible
      - Evidence for A, measured with P(A), is not necessarily evidence for P(!A)
        - Problem is Kolmogrov's axiom of the entire universe having P(1). Necessarily, then, P(!A) = 1 - P(A), but there may not be evidence supporting this.
        - The logic is somewhat similar to Humphrey's paradox
- Misc
  - Ecumenical approach claims that "probability" is an overloaded word, and can be decomposed into different subfields
    - E.g. ecumenical approach may claim that some subset of probability assertions should be classified under propensity interpretation, some should be under subjective interpretation, etc

Week 11 Summary: Causal, evidential and functional decision theory.

02.229 - Decision Theory and Practice, 2019 Jan-April

Yustynn Panicker

May 18, 2019

Contents

1 Meta **1**

2 Causal Decision Theory (CDT) **1**

3 Evidential Decision Theory (EDT) **1**
3.1 Problems . . . 1

4 Functional Decision Theory (FDT) **2**
4.1 (My) Analysis . . . 2
4.1.1 Rewarding Irrationality . . . 2
4.1.2 Miscellaneous Thoughts . . . 2

5 Some Paradoxes **2**
5.1 Newcomb's Box (predictor) . . . 2
5.2 Psychopath Killer Button . . . 3
5.3 Death in Damascus (`no stable acts`) . . . 3
5.4 Psychological Twin Prisoner's Dilemma . . . 3

1 Meta

This week I'm trying for a more reflective and personal rather than descriptive summary. As such, I'm writing what I think and how I interpret things rather than summarizing. I'm referring to the text less as an exercise. I quite possibly will get things wrong.

2 Causal Decision Theory (CDT)

See acts and outcomes as causally independent always. Do **not** acknowledge that the DM's estimated probability distribution does not reflect the objective, true probability distribution.

The problem here is obvious - this assumption that your probability should not be influenced by what you *would* do in some cases is nonsense.

3 Evidential Decision Theory (EDT)

See acts and outcomes as causally dependent (at least sometimes). This is due essentially to incomplete information. See this in a Bayesian way - the very choice of an act may reveal more information about the probability of an outcome (e.g. the psychopath example).

Distinguishing itself from CDT, this does indeed acknowledge that the probability distribution in the initial decision model is the subjective one, not the underlying objective one.

Peterson mentioned that this comes in many flavors, but only went into detail about the Bayesian one. Just based off the name, I don't know what others there could be.

3.1 Problems

In my opinion, none of these problems are very strong. This seems to be the most rational approach.

The crux of the issues lies with the denominator here:

$$P(X|Y) = \frac{P(X \wedge Y)}{P(Y)}$$

Where X is the outcome and Y is the act, this implies that you know your own probability of performing an act $P(Y)$ (in the denominator).

I don't know about this. It's not strictly necessary to have the denominator (as we've seen in Bayesian inferencing). We always know the denominator is between 0 and 1, so we just add multiply the learning rate by some constant. But this is not a strong counter-argument - there's the need to choose the constant, and how on earth do you do that?

I'm dissatisfied with Peterson's counterarguments to this problem. They seem to imply that it is indeed not possible for an agent to estimate their own probability distribution.

Maybe a stronger counter-argument is that of course we can model our own choices probabilistically, as we very clearly have models of ourselves in our own heads. We do. This is the basis of self-reflection and consciousness in the first place. It may or may not be an accurate model of ourselves, but it is odd to deny that we have one. Thus, whatever argument gets flung against having models of ourselves, it is simply descriptively true that we *do* have these self-models.

Another possible line of argument is that probability is an overloaded term (use the ecumenical approach from last week's philosophy of probability).

Also, one of the largest issues is simply that we will infinitely recurse upon and update our subjective probabilities in order to maximize our expected utility. This seems strange to me. This converges upon a final model of ourselves that decides with probability 1 to perform the action with the maximal expected utility. This means you've got a model of yourself as the rational self. It's rational. What's the issue? I feel like I'm misunderstanding something here, but I don't know what it is.

4 Functional Decision Theory (FDT)

I'm having a lot of trouble understanding this. As far as I can tell, it's useful for scenarios where the outcome depends on a skilled predictor's prediction of the decision maker's decision.

Yes, that's quite a head-warpey timey-wimey approach, so let me try to deconstruct the character of it.

1. Assume the predictor will always predict correctly (for simplicity)
2. Assume that the prediction is set in stone

At this stage, you have the ability to choose. While the predictor has no ability to change the past, the predictor had (in my simplification) perfect knowledge of your future choice at the time of its prediction.

Thus you choose to behave in the way the predictor said would have led to maximum utility. E.g. you one-box. Even in the transparent Newcomb version, you *always* one-box.

4.1 (My) Analysis

4.1.1 Rewarding Irrationality

In my opinion, this is irrational. As Yudkowsky mentions, there's a criticism that these types of decision problems (the paradoxes FDT resolves better) reward irrationality. I agree with that criticism. Which in turn might mean that this better performance suggests that FDT is an irrational decision theory.

4.1.2 Miscellaneous Thoughts

This sort of metacognition-esque method makes me think about those attempts by an AI to model a human's model of the AI itself. This sort of thing is useful for making te AI give the human more predictable output (easier for the AI to do if it's trying to understand what you would expect). The way the human thinks the AI will perform becomes a sort of implicit control mechanism. The difference here is that the control mechanism is explicit - you one-box because Newcomb's predictor explicitly lets you know that if you one-box, you will be more likely to get the huge reward.

So the use for this might simply be in controlling agents. It doesn't have to be about rational decision making.

Another way of putting this is that it ensures trust, which is probably useful for protracted games.

5 Some Paradoxes

5.1 Newcomb's Box (predictor)

- 2 boxes, choice to none, both, box 1 or box 2
- Box 1 always has $1000 in it
- Box 2 has either $1M or $0

- Reliable predictor has already predicted a 99% chance that, should you pick ONLY box 2, box 2 contains $1M

Should you take box 2, or both boxes? Obviously, the predictor cannot affect the boxes anymore.

5.2 Psychopath Killer Button

- Belief that anyone who would press the button is a psychopath
- Press the button, killing all psychopaths. You get +ve reward
- If you die, you get more -ve reward than the magnitude of +ve reward from killing all psychopaths

What do you do? Obviously, based on your own belief about pressing the button implying psychopathy, you should not do it

5.3 Death in Damascus (`no stable acts`)

- You meet Death accidentally
- He is surprised to see you and mentions he will visit you tomorrow
- You want to escape death and have the option of moving cities. Do you stay or move?

It is said there is no stable act here, simply because whatever you choose, death will have predicted it and thus you should have done the opposite.

There's clearly no rational act here. Seriously. Damned if you do, damned if you don't. I agree with Peterson's stated criticism - this lies outside the bounds of rationality.

5.4 Psychological Twin Prisoner's Dilemma

- Prisoner's Dilemma, but, of course, the other person is your psychological twin
- Clearly, it is prudent to assume the other person does what you will do

This breaks CDT as it assumes act-state independence. The challenge is this: how do you get the agents to cooperate (and therefore maximize reward) in a strategy that is rational for all games, not just the prisoner's dilemma?

Week 12 Summary: Applications: I. Pascal's Wager. II. Discounting. III. The value of life.

02.229 - Decision Theory and Practice, 2019 Jan-April

Yustynn Panicker

April 15, 2019

Contents

1 Pascal's Wager **1**
1.1 Main Arguments . . . 1
1.1.1 Argument from Superdominance . . . 1
1.1.2 Argument from Expectation . . . 1
1.2 Some Premises and Attacks on Them . . . 1
1.2.1 Positive probability for the existence of God . . . 1
1.2.2 For the other three outcomes other than the (believe, God exists) outcomes, the utility values are finite . . . 1
1.2.3 Being rational, you must maximize utility . . . 1
1.2.4 The decision matrix is the same for everyone . . . 1
1.3 Attacks Against Validity . . . 1
1.3.1 Mixed Strategies . . . 1
1.3.2 Definition: Regularity of Probability Function . . . 2
1.3.2.1 My Confusion . . . 2
1.3.3 All Actions Have Infinite Utility . . . 2
1.4 Meaning of Belief . . . 2

2 Qalys and Discounting **2**
2.1 Qalys . . . 2
2.2 Discounting . . . 2
2.2.1 Misc . . . 2

1 Pascal's Wager

1.1 Main Arguments

These are just the arguments meant to favor Pascal's Wager, without the counters. Personally, I think the arguments aren't great, but I'm probably also biased against them since I'm an atheist.

1.1.1 Argument from Superdominance

You lose nothing by believing, but you stand to gain a lot. Therefore, you should believe.

1.1.2 Argument from Expectation

As believing in God when God is real gives infinite utility, this act will always have a higher expectation

1.2 Some Premises and Attacks on Them

1.2.1 Positive probability for the existence of God

- Full atheists may maintain the claim that $P(God) = 0$

1.2.2 For the other three outcomes other than the (believe, God exists) outcomes, the utility values are finite

- Many Gods Objection
- Notion of infinite utility is suspect
- Would God really be unforgiving?

1.2.3 Being rational, you must maximize utility

- (point to Allais paradox, Ellsberg paradox) Maximizing expected utility may not be sufficient

1.2.4 The decision matrix is the same for everyone

- No it's not
  - E.g. people lose different things by restricting themselves to believing in God (e.g. if I were already living a relatively Christian life, it would cost me less to believe in God than someone who embodies Satanistic beliefs)

1.3 Attacks Against Validity

1.3.1 Mixed Strategies

- You could perform a random choice (between believing and not believing). The expected utility would still be infinity

1.3.2 Definition: Regularity of Probability Function

Probability function is `regular` iff it assigns probability 1 to logical truths, and 0 to contradictions

1.3.2.1 My Confusion I don't really get what this has to do with anything. I think he's arguing that one should stay open-minded, or else that one cannot be dogmatic about beliefs (e.g. dogmatically believe in God, as Pascal suggests) and still be rational.

The point here is that (assuming rationality necessitates regularity) you must regard any choice as a gamble. Not sure why. Something to do with not being dogmatic?

1.3.3 All Actions Have Infinite Utility

This is a bit of a weird one.

The idea here is that if you believe any act will *eventually* result in a non-zero chance that you believe in God, then *any* act (including choosing not to believe in God when presented with the wager) has the same expected utility of ∞ (because probability of eventually believing in god $\times\infty$ reward)

1.4 Meaning of Belief

It seems strange to think that a perfunctory belief in God satisfies the condition allowing you to go to heaven.

2 Qalys and Discounting

2.1 Qalys

Quality-adjusted life years. These are used by doctors to determine actions for people's lives The primary question is *should* future goods be discounted. Generally speaking, we do discount them. But *should* we?

2.2 Discounting

The thrust of the issue here is whether discounting *should* be done. Obviously, it often is done (in RL, in economics, etc).

2.2.1 Misc

- Economists are generally concerned with commodities (and tend to discount)
- Philosophers are generally concerned with well-being (and tend to be against discounting)

Author assumes that future well-being should not be discounted, and that this view is the norm.

Decision Theory Week 13: Superintelligence and the AI alignment problem.

- Cognitive Superpowers and the Superintelligent Will
  - Warning: Do not anthropomorphize AI
    - Intelligence
    - Motivation
  - Considerations
    - Competing agents with contrary goals to the AI in question
      - If there are no competing agents, Bostrom claims there is nothing standing in the way of a sufficiently intelligent malevolent agent taking over the world
  - 3 approaches to predicting AI motivation
    - 1. Predictability through design
    - 2. Predictability through inheritance
      - E.g. intelligent agent is made by digitally cloning human brain, then it *may* be that it has human motivation
    - 3. Predictability through convergent instrumental reasons
      - See convergent instrumental convergence thesis
  - Orthogonality Thesis
    - Intelligence and final goals are more or less independent
  - Instrumental Convergence Thesis
    - Some instrumental values are useful for a very wide range of goals, so they are effectly always present in intelligent enough agents
    - Why this is so worrying
      - Regardless of the explicit goal programmed in, intelligent enough agents will want to do things that *may* infringe on human interests
      - E.g. An agent tasked with finding the largest prime number may start to ensure its own survival at the expense of humanity, so that it may continue searching for this prime number in perpetuity
- Misc
  - What intelligence means here
    - Not rationality or reason
    - Means-end reasoning, prediction, planning, etc
  - Singleton
    - From set theory: a set with only one element
    - Refers to an AI agent that is the sole, completely autonomous decision maker
    - The general argument Bostrom uses is that AI converges into a singleton, either by us giving a well-connected agent more and more power or by it taking that power itself through social manipulation etc (in order to fulfil whatever goal it has)
- Motivating examples of specific AI gone wrong
  - Riemann hypothesis catastrophe
  - Paperclip AI
- General thoughts
  - Why does everybody hate on this book?
    - One common criticism is that Bostrom has insufficient knowledge about AI, so he's wrong. It's ad hominem so that should be ignored - his reasoning is strong.
    - A better criticism is that general AI may not even become a thing, and if it does it'll take more significant breakthroughs in tech. This doesn't resolve all the issues he brings up, though. He talks about specific AI a lot.
    - My personal feeling is that Terminator made it very, very hard for anyone who's into AI to take this seriously. It gave this weird malevolent strawman of what AI being bad looks like, and we all know it won't be like that. So whenever we hear arguments favoring AI danger, we switch off and assume that people are arguing from ignorance. It's unfair, but that's my hypothesis for why people who supposedly know what they're talking about dismiss stuff like the AI Alignment Problem.
- Personal Thoughts
  - The ideas of motivation and intelligence are pretty dubiously applied to algorithms. These are human constructs to make sense of the chaos - it does not seem obvious that it is useful to frame the AI in a way that phrases it as a being with values and intelligence
- Concerns
  - It is impossible to predict good behaviour based on the track record of the system in its 'juvenile' stage
    - Algo will be evaluating a wider strategy space when it is 'mature'
  - Malignant Failure Models
    - Perverse instantiation
      - E.g.
        - "Make us smile more" -> "Paralyze facial muscles to make constant smiles"
          - "Make us smile more without paralyzing our facial muscles" -> "Stimulate motor cortex to force us to smile more"
- Control Methods (all suck)
  - Capability control
    - Boxing methods
      - Reduce the solution space, or else limit its interaction with the real world
    - Incentive methods
    - Stunting methods
      - Limit intelligence
    - Tripwires
  - Motivation selection
    - Direct specification
    - Domesticity
      - Design the agent with severely limited ambition and activities
    - Augmentation
      - Start from a non-superintelligent agent which is benevolent/human-like, and make that base superintelligent
- Types of AI (4 castes)
  - 4 castes
    - Oracle
      - Q&A system
    - Genie
      - Command-executing system
    - Sovereign
      - Open ended, autonomous operation
    - Tool
      - Not meant to exhibit goal-directed behavior
  - Thoughts
    - Interesting divergence from the traditional supervised/unsupervised/reinforcement learning trichotomy in ML. I like it. Purpose-based rather than method-based.