Оператор актуальности и проблема вероятности, или *Что узнала Спящая Красавица?*

Виктор Горбатов, 2012
vic-gorbatov@ya.ru

Байесовская вероятность

- Байес определял вероятность как степень уверенности в истинности суждения
- Для определения степени уверенности в истинности суждения *при получении новой информации* используется теорема Байеса:

$$P(A|B) = \frac{P(B|A)\,P(A)}{P(B)}$$

Центрированные миры, пропозиции, вероятности

- Принципиальное различие между знанием *de dicto* и знанием *de se* (о себе, от первого лица)
  - Льюис: пример с двумя богами
  - Хофштадтер: пример с видеокамерой

- Центрированный мир (Льюис, 1979) – мир, привязанный к определенному субъекту и моменту времени: *<мир, субъект, момент времени>*
- Центрированная пропозиция – множество центрированных миров
- Знать вероятность центрированной пропозиции – значит знать вероятность того, что твой центрированный мир входит в соответствующее множество центрированных миров

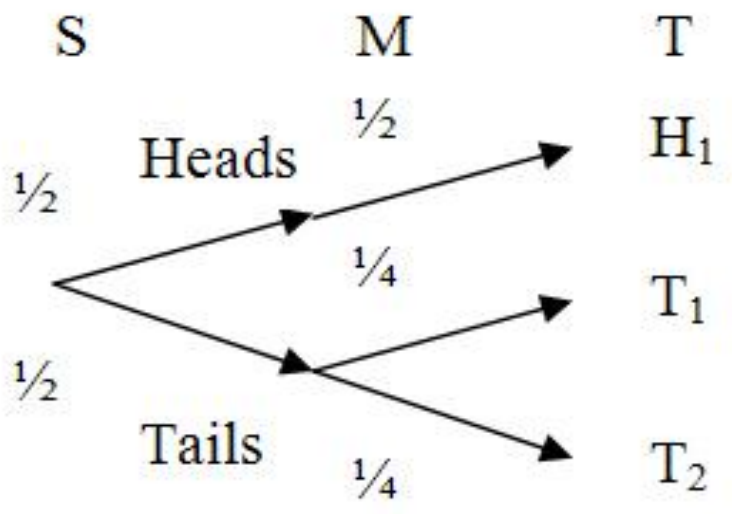
Центрированная кондиционализация

- Что изменится в байесовской теории принятия решений, если мы будем рассматривать не de dicto, а de se убеждения?
  - Льюис: «Очень мало. Мы заменим пространство миров на пространство центрированных миров, или на пространство обитателей миров. Все остальное останется прежним.»
- Вы берете свои старые вероятности
- Во всех мирах, несовместимых с полученными свидетельствами, устанавливаете вероятность на 0
- Нормализуете вероятности в оставшихся мирах (чтобы в сумме они составляли 1)

Представьте, что вы – Спящая Красавица (СК)

- Исследователи погружают вас в сон
- В течение двух дней, пока длится эксперимент со сном, вас будят один или два раза
- Количество пробуждений зависит от подбрасывания монеты (орел – один раз, решка – два)
- После каждого пробуждения вам вкалывают медикамент, заставляющий вас забыть, что вы просыпались
- Когда вы проснетесь, как вы оцените вероятность выпадения орла?

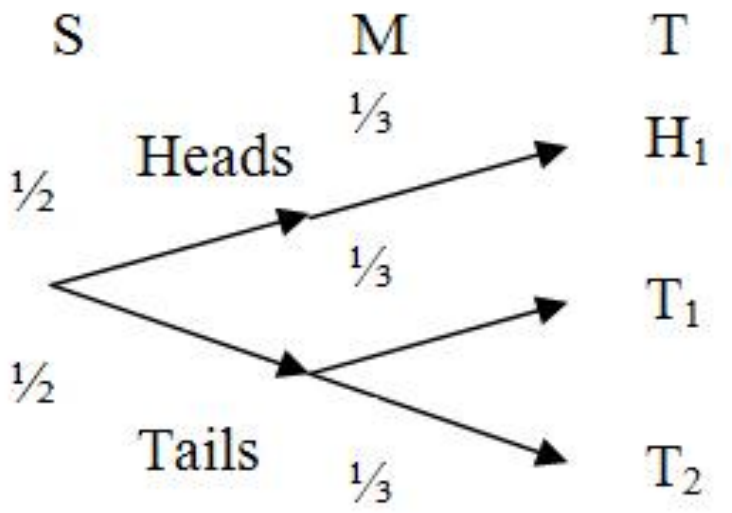
Позиция «двоечников»



$$P(H_1) \neq P(T_1) = P(T_2)$$

- Конечно, ½ ! Вы знаете, что изначально монета правильная, следовательно, вероятность выпадения орла была ½
- После пробуждения вы не получили *никакой новой информации*
- Следовательно, вероятность осталась ½

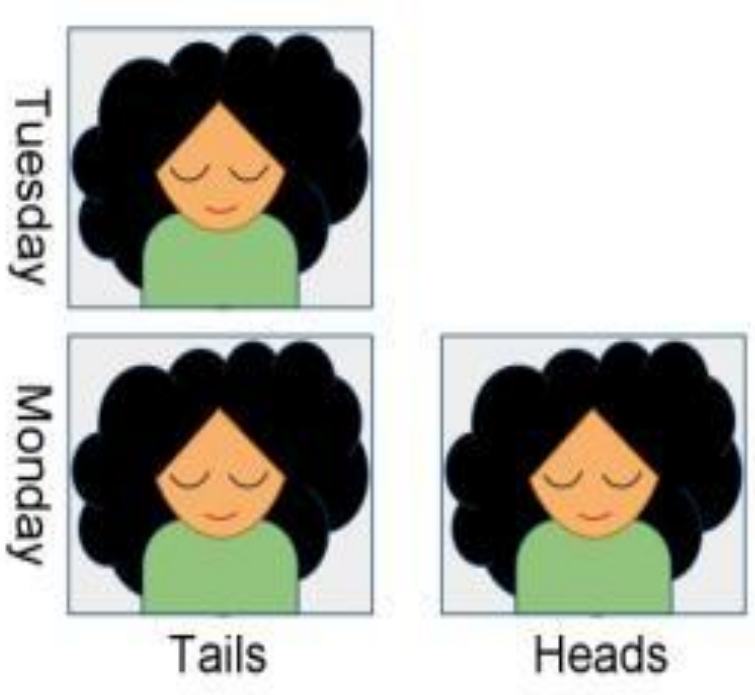
Позиция «троечников»



$$P(H_1) = P(T_1) = P(T_2)$$

- Конечно, ⅓!
- Представим, что эксперимент провели множество раз
- Тогда *в долгосрочной перспективе* пробуждений с выпавшим орлом будет ⅓
- Следовательно, даже в однократном эксперименте эта вероятность равна ⅓

Адам Элга



- Если после пробуждения вы узнали, что монета выпала решкой, то для вас это означает, T_1 или T_2 , а поскольку они субъективно неразличимы, вы припишете им равную вероятность: $P(T_1 | T_1$ или $T_2) = P(T_2 | T_1$ или $T_2)$
- Значит, даже до пробуждения $P(T_1) = P(T_2)$
- Если после пробуждения вы узнали, что это ПН, то для вас это означает, что H_1 или T_1 , и тогда $P(H_1 | H_1$ или $T_1) = P(T_1 | H_1$ или $T_1)$
- Значит, даже до пробуждения $P(H_1) = P(T_1)$
- Таким образом, $P(H_1) = P(T_1) = P(T_2)$

Откуда взялась 1/6?

- Элга: пробудившись , СК не узнала ничего нового (она изначально знала, что в любом мире проснется в понедельник)
- Различие в том, что в ВС она не рассматривала свое нахождение во времени как фактор, релевантный вопросу о вероятности P(H), а в ПН рассматривает
- Как получилось, что этот фактор теперь стал релевантным?
- Является ли информация о том, что он стал релевантным, новым знанием?

Принцип рефлексии

- Бас ван Фрассен:
  - Любой агент, который уверен, что *завтра* будет приписывать вероятность x пропозиции R (хотя он не получит *никакой новой информации* и не испытает *никаких когнитивных катаклизмов* в течение указанного времени), должен *уже сегодня* приписывать вероятность x пропозиции R

Ответ Льюиса

- Элга: приписываемая вероятность изменилась, следовательно, принцип рефлексии нарушен
- Льюис: принцип рефлексии не может быть нарушен, следовательно, вероятность не изменилась
  - На самом деле, СК *узнала кое-что новое*, проснувшись
  - Элга смешивает две разные вероятности: до и после сообщения о том, что *сейчас ПН*, а это можно сделать, лишь *зная наперед будущее*
  - $P(T_1) \neq P_+(T_1)$

Что узнала Спящая Красавица?

- Рут Вайнтрауб, «Простое решение»
- СК изначально знала, что ее разбудят в ПН
- Проснувшись, она узнала, что ее разбудили *сейчас* (т.е. что ПН *сейчас*)

- Как оператор актуальности делает знание de se релевантным относительно знания de dicto?

Логика актуальности и двумерная семантика

- Оператор актуальности

  $M \models_{w,v} Ap$, е.т.е. $M \models_{w,w} p$

  – Семантический смысл оператора A в том, что он делает мир референции миром оценки
- Необычные свойства оператора A:
  – Пусть p является контингентным апостериори тогда
  – Ap – *необходимое апостериори*
  – $p \leftrightarrow Ap$ – *контингентное априори**

Второе измерение

- Пусть обычные миры – это онтологические альтернативы, отвечающие за необходимость/ контингентность
- Тогда центрированные миры – это познавательные альтернативы, отвечающие за априорность/ апостериорность

- Предметом априорного знания не всегда бывают необходимые пропозиции, а апостериорного – контингентные

Чалмерс: байесианство несоместимо с референциализмом

- Оливия исследует наследственное заболевание D
- Известно, что $P(D|\alpha(x))=0,1$, $P(D|\beta(x))=0,2$ $P(D|\alpha(x)\ \&\ \beta(x))=0,9$
- Утром она обнаружила ген α у доктора Джекила, а вечером – ген β у мистера Хайда. Как она оценит вероятность того, что Хайд болен D?
- Чему равна вероятность J=H? 1 или, скажем, 0,001?
- Иногда новые свидетельства не повышают субъективную вероятность, когда мы это ожидаем, и наоборот, повышают, когда мы этого не ждем
- Объектом субъективной вероятности являются не вторичные, а первичные интенсионалы

2D-матрица для «Джекил=Хайд»

Миры оценки / *Миры референции*	w_1	w_2	w_3
$<w_1, a, t_1>$	1	1	1
$<w_2, b, t_2>$	0	0	0
$<w_3, c, t_3>$	0	0	0

2D-матрица для «Я здесь сейчас»

Миры оценки / *Миры референции*	w_1	w_2	w_3
$<w_1, a, t_1>$	1	0	0
$<w_2, b, t_2>$	0	1	0
$<w_3, c, t_3>$	0	0	1

2D-матрица для p

Миры оценки / *Миры референции*	w_1	w_2	w_3
$<w_1, a, t_1>$	1	0	1
$<w_2, b, t_2>$	0	0	1
$<w_3, c, t_3>$	1	0	0

2D-матрица для Ар

Миры оценки / *Миры референции*	w_1	w_2	w_3
$<w_1, a, t_1>$	1	1	1
$<w_2, b, t_2>$	0	0	0
$<w_3, c, t_3>$	0	0	0

2D-матрица для p ↔ Ap

Миры оценки / *Миры референции*	w_1	w_2	w_3
$<w_1, a, t_1>$	1	0	0
$<w_2, b, t_2>$	0	1	0
$<w_3, c, t_3>$	0	0	1

Проблема «твистеров»

- Мы видим, что *априорное* предложение может быть *контингентным*, но как оно может стать при этом *апостериорным*?
- Использование обычных твистеров ничего не дает – они оставляют диагональ 2D-матрицы без изменения (меняется горизонталь на вертикаль)
- Нм нужен твистер, который менял бы горизонталь на диагональ
- Столнейкер: таким «твистером» может служить прагматический понятый оператор утверждения
- В разных контекстах утверждаемое содержание может быть повернуто к нам либо *фактуальной*, либо *семантической* стороной, хотя разделить их в общем случае невозможно

Анафема первичным интенсионалам

- Р. Столнейкер: «Было бы прекрасно, если бы мы имели нейтральный язык с внутренним образом детерминированной семантикой – язык, который не требовал бы никаких фактуальных допущений для его интерпретации, но при этом мог бы дать полное описание действительного мира, и всех возможных миров. (…) Но я не думаю, что все это возможно. (…) Семантические и фактуальные вопросы оказываются переплетены, и в этом заключается проблема»

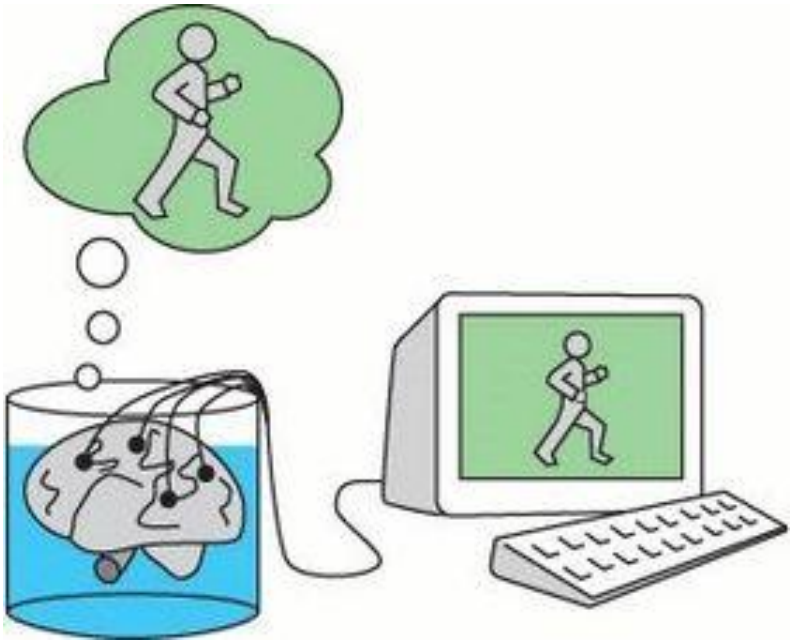
Крис Мичем: раздельная кондиционализация

- Вы берете гипотетическое начальное распределение вероятностей по мирам (hp)
- Во всех центрированных мирах, несовместимых с полученными свидетельствами, устанавливаете вероятность на 0
- Затем нормализуете вероятности в оставшихся мирах, чтобы в сумме они давали 1, но при этом были пропорциональны hp
- Наконец, вы нормализуете вероятности центрированных альтернатив внутри каждого мира, чтобы они в сумме давали вероятность этого мира, но при этом были пропорциональны hp

Окрашенная комната

- Если выпала решка, то первый раз вы проснетесь в черной комнате, второй – в белой
- Если выпал орел, то монету подбрасывают снова, чтобы определить, в какого цвета комнате вы проснетесь в понедельник
- Если вы открыли глаза и увидели, что комната черная, то какова вероятность орла (Н)?
- Для Льюиса и Элги она останется той же, что и в воскресенье: ½ (ведь элиминирована половина Н-миров и половина Т-альтернатив)
- Половина Н-миров была элиминирована, но ни одного Т-мира (исчезла половина Т-альтернатив, но не миров)
- Крис Мичем: вероятность орла будет ⅓

Аргумент «множественные мозги»



- Рассмотрим гипотезу о том, что вы – «мозг в банке». Естественно приписать ей очень невысокую, но при это ненулевую вероятность. Допустим, далее, что в вашем мире постоянно производятся мозги в банках, находящиеся в состояниях, субъективно неотличимых от ваших. Если Элга прав, то вероятность первоначальной гипотезы будет стремиться к 1.

Аргумент «ученый-садист»

- Рассмотрим гипотезу о том, что вы находитесь в мире, где некий ученый каждую секунду создает n мозгов в состояниях, субъективно неотличимых от ваших. А еще полсекунды спустя он их все уничтожает. Естественно приписать ей очень невысокую, но при это ненулевую вероятность. Но если Льюис прав, вероятность первоначальной гипотезы будет стремиться к 0.

Аргумент "Up-and-Down"

- Ученый подбрасывает монетку – если выпадет решка, в полдень следующего дня он создаст n мозгов в состояниях, субъективно неотличимых от вашего, а вечером того же дня уничтожит половину из них. Если выпадет орел, никакие мозги не создаются и не уничтожаются. Какова вероятность орла?
  - Если прав Элга, она стремится к 1
  - Если Льюис – к 0
  - Если Мичем, она останется равна ½

Аргумент «варьируемые мозги»

- Допустим, что миры делятся на два вида: в нормальных у вас нет «мозговых дублеров», зато в странных на каждое ваше состояние в нормальном мире есть мозг, чье состояние субъективно неотличимо. Так что если вы в следующую секунду станете есть шоколадное мороженое, то исключите множество нормальных миров, где вы этого не делаете. Но при этом не будет исключен ни один «странный» мир, так как в каждом из них есть мозг, состояние которого субъективно неотличимо от того, как вы едите шоколадное мороженое. Если прав Мичем, вероятность того, что вы в «странном мире», будет возрастать относительно вероятности оказаться в нормальном.

Заключение

- Проблема СК имеет много идейных параллелей (проблема рассеянного водителя, квантовое бессмертие, сильный антропный принцип)
- Прежде чем использовать некоторые мысленные эксперименты, надо научиться строить другие мысленные эксперименты, позволяющие грамотно оценивать вероятности выводов, получаемых из первых
- У нас пока нет единой и окончательной теории принятия решений для СК, но уже есть некоторые логические критерии, позволяющие среди различных концепций разумным образом выбрать лучшие кандидатуры на эту роль