Normative Model I

April 22, 2020

Language

Let O be the set of all stones (objects), and use A, R, R' to represent a stone's role in a causal interaction - A stands for the *agent*, R for the pre-interaction state of the *recipient*, and R' the post-interaction state of the recipient, also named the *result*.

Feature space $F = \{L, S, \ldots\}$ consists of *lightness* L - the color shadings, and *sidedness* S - number of edges of the polygons. Note that this feature set may grow as the experiment setting gets richer.

For this model, lightness L takes value l_1, l_2, l_3, l_4 along the lightness scale from light to dark. Specifically, l_1 stands for *light*, l_2 for *medium*, l_3 for *dark*, and l_4 for *very dark*, as used in the experiment. Sidedness S takes value $p_3, \ldots, p_7$, where each subscript represents the number of edges for a polygon (hence p).

We define a value-reading function $v(o, f) = u, o \in O, f \in F$ and u is the value of feature F for object o.

For our current model, each stone has two features - lightness L and sidedness S. Hence for each object $o \in O$ there are two kinds of value-reading functions $v(o, L)$ and $v(o, S)$. For simplicity, we will use $L(o)$ instead of $v(o, L)$ to read stone o's lightness L, and $S(o)$ instead of $v(o, S)$ to read stone o's sidedness S.

With these definitions, the language of this task consists of

- Atomic sentences:
  - $L(A), L(R), L(R'), S(A), S(R), S(R')$: read stone feature values
  - $1, \ldots, 7$: index numbers
  - $l_1, \ldots, l_4, s_3, \ldots, s_7$: feature values
- Relations:
  - $=, \neq$: compare if values match. Eg. $L(A) = l_2, S(A) \neq S(R')$.
  - $>, <$: compare values within the same feature by comparing their subscripts. Eg. $l_1 < l_2, S(A) > S(R)$.
  - $+, -$: plus and minus operations on feature value subscripts. Eg. $s_1 + 1 = s_2, L(R') = L(R) - 1$.
  - $\models$: *if* the cause conditions (the part before $\models$) satisfy, result effects (the part after $\models$) follow. Eg. $(c(R) = c(A)) \models (c(R') = c(R))$. If the cause conditions are not satisfied, we assume that the recipient stone does not change. See below for legit candidates of a cause condition or result effect.

- Grammar:
  - If ϕ, ψ, χ are an atomic sentences, the following compositions are *basic sentences* of the language: $(\phi = \psi), (\phi \neq \psi), (\phi > \psi), (\phi < \psi)$, $(\phi + \psi = \chi), (\phi - \psi = \chi)$
  - If α, β are basic sentences, $[\alpha, \beta]$ is the conjunction of α and β. Number of conjuncts in such conjunctions ≥ 2.
  - A causal hypothesis is of the form $[\alpha_1, \ldots, \alpha_k] \models [\alpha_{k+1}, \ldots, \alpha_n]$, where
    * Either the cause condition is $\top$ - *any*, or each basic sentence of $\alpha_1, \ldots, \alpha_k$ contains A or R, but not R'.
    * Each basic sentence of $\alpha_{k+1}, \ldots, \alpha_n$ contains R'.

Examples

- $[(s(A) \neq s(R)), (c(A) \neq c(R))] \models [(c(R') = c(A)), (s(R') = s(A))]$
  If the agent stone and the recipient stone have different shapes and colors, then the recipient stone will turn into the same as the agent stone; otherwise the agent will have no effect on the recipient stone.
- $(s(A) = s_3) \models (s(R') = s(R) + 1)$
  If the agent stone is a triangle (regardless of its color), then the recipient stone shape's number of edges increases by 1; otherwise the agent will have no effect on the recipient stone.

Hypotheses

Let's restrict the complete hypothesis space by only allowing meaningful sentences up to step 1.

For $\alpha_{k+1}, \ldots, \alpha_n$: let α_{k+1} be $(s(R') = x)$, where x takes value from $s_3, \ldots, s_7, s(A), s(R), s(A)+1, s(A)-1, s(R)+1, s(R)-1$; then, replace $=$ with $\neq, >, <$. Apply the same procedure for $c(R')$ to compose α_{k+2}. With 4 color shadings and 5 shapes this amounts to $(4+6) \times 4 + (5+6) \times 4 = 84$ effects.

As for the cause conditions part, i.e., $\alpha_1, \ldots, \alpha_k$, there are several options.

- Restrain from using *if*, the cause conditions part can be trivialized to just $\top$, meaning that properties of the agent and recipient stones do not matter; just being the agent or being the recipient suffices to produce the specified effect.
- Applying the procedure used for producing the effects to produce cause conditions. For example, let α_1 be $(s(A) = x)$ where x takes value from $s_3, \ldots, s_7, s(R), s(R)+1, s(R)-1$, then replace $=$ with $\neq, >, <$. Let α_2 be $(c(A) = x)$ and apply the same procedure. Similar for $(s(R) = x)$ and $(c(R) = x)$. Note that $(s(A) = s(R))$ is equivalent to $(s(R) = s(A))$, therefore when generating cause conditions using the $(s(R) = x)$ form, duplicates must be removed.

- A complex reasoner can combine multiple causal hypotheses instead maintaining only one - in a *if ... else ...* manner.

Learning

Let $\langle A, R, R' \rangle$ be a complete data point d. For a causal hypothesis h, if $s(A), c(A), s(R), c(R)$ make the cause conditions true and with $s(R'), c(R')$ they make the result effects true, then $P(d|h) = 1$. If $s(A), c(A), s(R), c(R)$ fails to satisfy the cause conditions, recipients should remain as they are (because no causes are posed onto them), hence $P(d|h) = 1$ if $s(R') = s(R)$ and $c(R') = c(R)$, and $P(d|h) = 0$ otherwise.

Thus, upon observing a complete data point d,

$$P(h|d) = \frac{P(d|h)P(h)}{\sum_{h_i \in H} P(d|h_i)P(h_i)}$$

Assuming a flat prior for the first data point, our Bayesian learner updates hypothesis space 6 times sequentially upon observing the six learning shots.

Generalization

Upon observing a partial data point $d' = \langle A, R \rangle$, the complete data point d^* normalizes over 20 possible R's.