[Magic Stones] Normative model v2

Bonan Zhao b.zhao@ed.ac.uk

January 20, 2020

Consider a magic stone task consisting of a magic stone *Agent* (A), a normal stone *Target* (T), and the effect that Target T turns into *Result* (R), a causal generalization task by nature is to infer causal rule(s) from one data point $A \to (T \to R)$, and apply the inferred causal rule(s) to new pairs of $\langle A', T' \rangle$.

By classical logic, $A \to (T \to R) \Leftrightarrow (A \wedge T) \to R$, we therefore consider the antecedent and the consequent separately, taking the agent stones and target stone as a fixed-order pairs.

According to the task description, each stone $\mathbf{S}$ has two properties: *color* ($\mathbf{S}_c$) and *shape* ($\mathbf{S}_s$), each of which takes value from a set of three elements. Without loss of generality, we let our normative model construct causal rules for each property following the same kind of logic, and use rules from both properties to construct what comes out as the result stone. In other words, observing $\langle A = A_c \wedge A_s, T = T_c \wedge T_s \rangle$, the normative model will produce some color-rule $f_c(\langle A_c, T_c \rangle) = R_c$, and some shape rule $f_s(\langle A_s, T_s \rangle) = R_s$, as a result the prediction one makes will be $R = R_c \wedge R_s$.

Below I will lay out how to construct such causal rules normatively, allowing equality $=$, negation $\neg, \neq$, and conjunction $\wedge$. It is possible to include disjunction $\vee$ into the model, triggering a potentially infinite number of causal rules. However, I will take disjunction out of the current normative model, reasons explained elsewhere. I will give examples of what rules this model can construct if we do allow disjunctions as this note proceeds.

Hypotheses

We will use a standard Bayesian framework to formalize both inference and generalization. In this framework, a color rule and a shape rule together form a hypothesis $h \in \mathbf{H}$, which will then be used with observed data-points for updating or predicting.

As mentioned above, we assume all the properties follow the same kind of logic to produce causal rules. Therefore, consider property $p \in \{color, shape\}$, a causal rule for p takes $\langle A_p, T_p \rangle$ as input, and outputs a value for R_p.

To facilitate this formalization, let me introduce one more helper notion $v_p()$. v stands for *value*. For a given property p, $v_p()$ samples one exact value from all the elements that belong to this property. $v_p(\mathbf{S}) = \mathbf{S}_p$, with stone $\mathbf{S}$ as the argument, returns the value of property p for stone $\mathbf{S}$. Since we will be talking about causal rules for the same property throughout this note, we can ignore subscript p when there is no ambiguity from context.

For the antecedent part of the rules, there are 3 groups of atomic conditions that we can use to formulate causes. They are: agent-wise $v(A) = v()$ and $v(A) \neq v()$, target-wise $v(T) = v()$ and $v(T) \neq v()$, and relative-wise $v(A) = v(T)$ and $v(A) \neq v(T)$. Combining rules from these three groups gives a complete list of 12 antecedent conditions (with group 3 being not compulsory):

For the antecedent part of the rules

1. $(v(A) = v()) \wedge (v(T) = v())$
2. $(v(A) = v()) \wedge (v(T) \neq v())$
3. $(v(A) \neq v()) \wedge (v(T) = v())$
4. $(v(A) \neq v()) \wedge (v(T) \neq v())$
5. $(v(A) = v()) \wedge (v(T) = v()) \wedge (v(A) = v(T))$
6. $(v(A) = v()) \wedge (v(T) \neq v()) \wedge (v(A) = v(T))$
7. $(v(A) \neq v()) \wedge (v(T) = v()) \wedge (v(A) = v(T))$
8. $(v(A) \neq v()) \wedge (v(T) \neq v()) \wedge (v(A) = v(T))$
9. $(v(A) = v()) \wedge (v(T) = v()) \wedge (v(A) \neq v(T))$
10. $(v(A) = v()) \wedge (v(T) \neq v()) \wedge (v(A) \neq v(T))$
11. $(v(A) \neq v()) \wedge (v(T) = v()) \wedge (v(A) \neq v(T))$
12. $(v(A) \neq v()) \wedge (v(T) \neq v()) \wedge (v(A) \neq v(T))$

The consequent part has 10 cases

1. $v()$
2. $\neg v()$
3. $v(A)$
4. $\neg v(A)$
5. $v(T)$
6. $\neg v(T)$
7. $\neg v() \wedge \neg v(A)$
8. $\neg v() \wedge \neg v(T)$
9. $\neg v(A) \wedge \neg v(T)$
10. $\neg v() \wedge \neg v(A) \wedge \neg v(T)$

There are $12 \times 10 = 120$ causal rules for one property, hence $120^2 = 14,400$ hypotheses in the hypothesis space for this normative model, if we allow only equality, negation, and conjunction. Allowing disjunction could produce complex rules $(\neg v() \vee (v(A) \wedge \neg v(T))) \models v() \vee v(A) \vee \neg(v(A) \wedge v(B))$, which easily blow-up the space. One reason why we skip it for now.

Likelihood of a hypothesis generating a data point is therefore $P(d|h) = \frac{1}{n}$ where n is the number of all the data-points hypothesis h can produce. Note that two theories can produce the same Result stone given the same input stone pairs, but with (very likely) different likelihoods.

Bayesian update

Following standard Bayesian update, assuming we start with a equal distribution of hypotheses, the first encounter of the training task updates the space by

$$P(h|d) = \begin{cases} \frac{P(d|h)P(h)}{\sum_i P(d|h_i)P(h_i)} & \text{if } h \text{ produces } d \\ 0 & \text{otherwise} \end{cases} \tag{1}$$

And when asked to make the predictions, posterior predicted is calculated by

$$P(d^*) = \frac{\sum_i P(d^*|h_i)P(h_i)}{\sum_j \sum_i P(d_j^*|h_i)P(h_i)}$$