Bayes I

Or: How to be precise with uncertainty

- *The Language of Thought: computational cognitive science approaches to category learning*
- Who: Fausto Carcassi
- When: Sommer semester 2022

Where are we now?

- We have now learned about how to:
  - Write a formal grammar for a specific cognitive domain, e.g. music
  - Write an interpretation function for it that gives each sentence in the grammar a meaning, compositionally.
- This is cool as it allows us to *generate* objects from the domain randomly.
- However, we can't really do anything useful with this.
- What we want to do is go the other way:
  - Start from some object(s) in the domain
  - Infer what sentence(s) in the LoT generated it / what grammar
- For this, we are going to need how to go from a generative process and some observations to the probability of hidden causes: Bayesian inference!

Interpretations of probability

- It is tempting to say: probability is anything satisfying the probability axioms.
- Kolmogorov axioms:

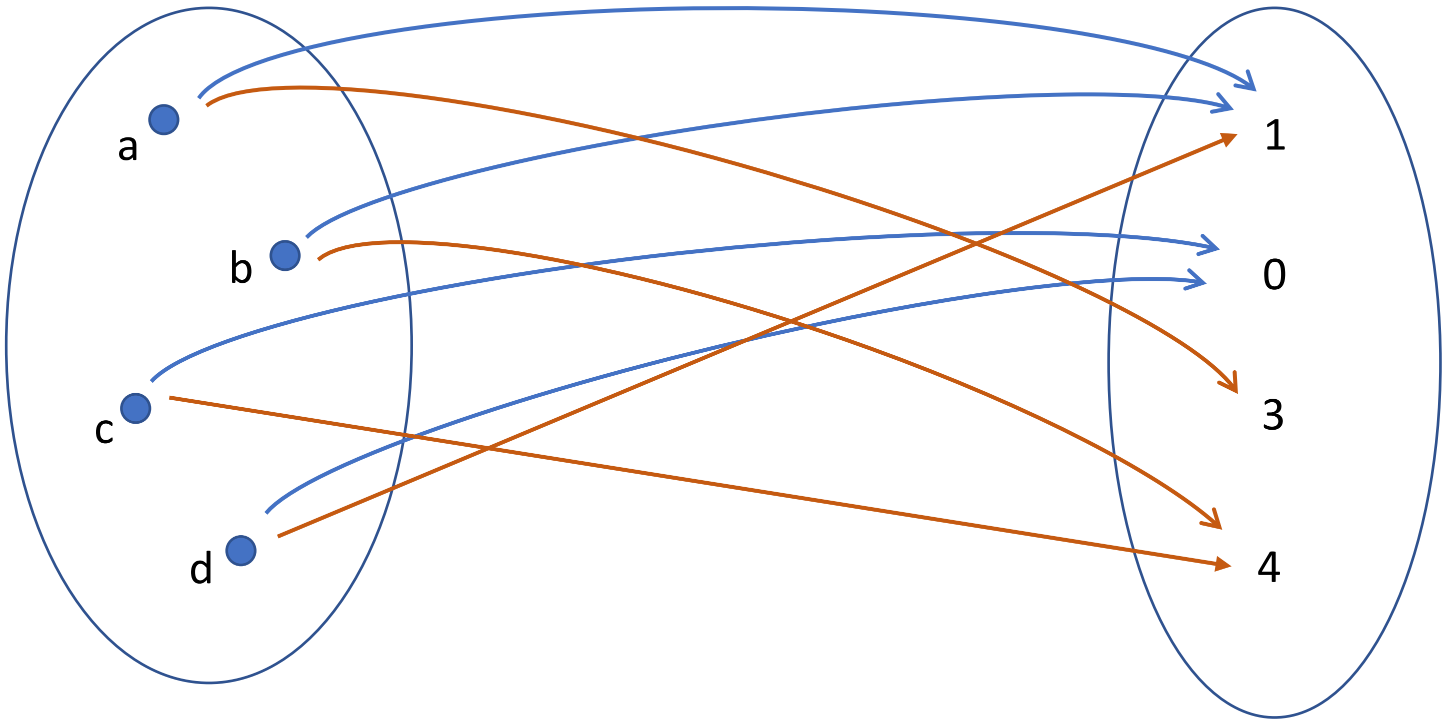
1. (Non-negativity) $P(A) \geq 0$, for all $A \in F$.
2. (Normalization) $P(\Omega)=1$.
3. (Additivity) $P(A \cup B)=P(A)+P(B)$ for all $A,B \in F$ such that $A \cap B=\emptyset$

- However, this is not very satisfying: we can give a semantics to the Kolmogorov axioms with things that are clearly not probabilities, e.g. normalized weight.
- And there are other axiomatizations of probability.
- It seems like we need to first decide on some notion of probability to then formalize it.

Three main interpretations (SEP)

- *Classical / logical / evidential*: An epistemological concept, which is meant to measure objective evidential support relations. For example, “in light of the relevant seismological and geological data, California will probably experience a major earthquake this decade”.
- *Frequentist*: A physical concept that applies to various systems in the world, independently of what anyone thinks. For example, “a particular radium atom will probably decay within 10,000 years”.
- *Subjective:* The concept of an agent’s degree of confidence, a graded belief. For example, “I am not sure that it will rain in Canberra this week, but it probably will.”
- Typically, Bayesian probability is associated with the subjective view!

From joint to conditional

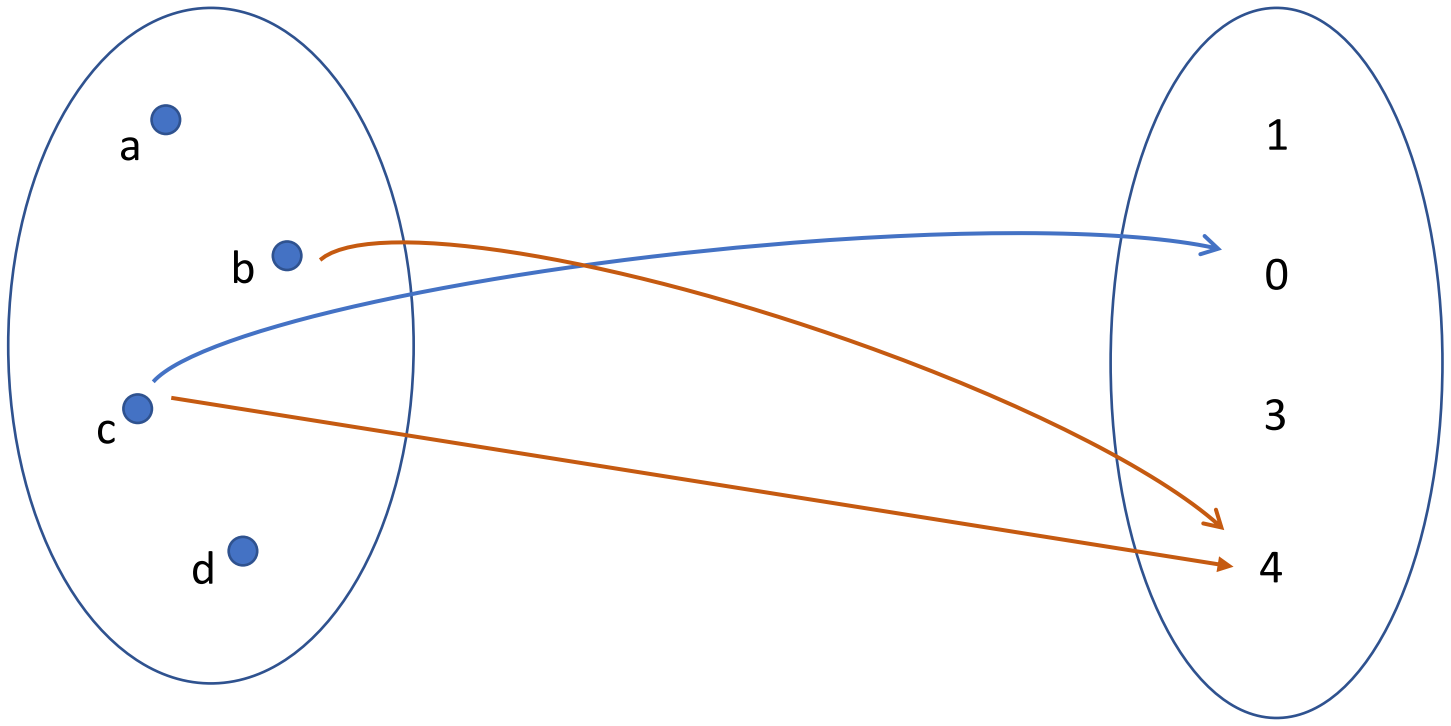


$$P(X = x, Y = y) = P(\{\omega \mid X(\omega) = x \wedge Y(\omega) = y\})$$

$$P(X = 0, Y = 4) = P(\{c\})$$

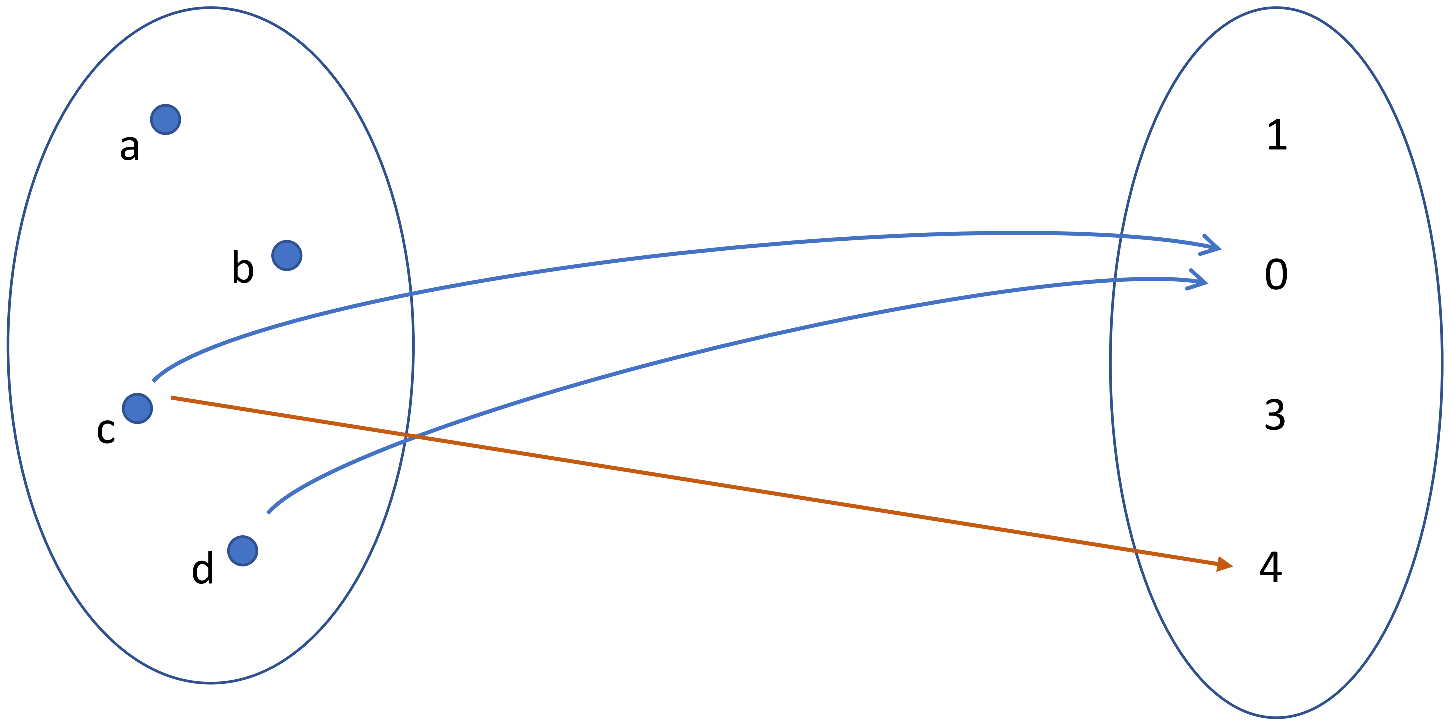
$$P(X = x \mid Y = y) = \frac{P(\{\omega \mid X(\omega) = x \wedge Y(\omega) = y\})}{P(\{\omega \mid Y(\omega) = y\})}$$

From joint to conditional



$$P(X = 0 \mid Y = 4) = \frac{P(\{c\})}{P(\{b, c\})}$$

From joint to conditional



$$P(Y = 4 \mid X = 0) = \frac{P(\{c\})}{P(\{c, d\})}$$

Conditional probability to Bayes theorem

- We can see that the numerator in the calculations for P(X=x|Y=y) and P(Y=y|X=x) is the same, namely: the set of events that satisfy both!
- But the denumerator changes in the two expressions:

$$P(X=0 \mid Y=4) = \frac{P(\{c\})}{P(\{b,c\})} \qquad P(Y=4 \mid X=0) = \frac{P(\{c\})}{P(\{c,d\})}$$

  - P(X=x|Y=y) it's the probability of the set of events that are y
  - P(Y=y|X=x) it's the probability of the set of events that are x
- Going from one to the other gives us Bayes theorem:

$$\underbrace{P(Y=y \mid X=x)}_{\text{Start}}\underbrace{P(X=x)}_{\text{Expand}}\underbrace{\frac{1}{P(Y=y)}}_{\text{Shrink}} = \underbrace{P(X=x \mid Y=y)}_{\text{End!}}$$

A motivating example

- Suppose that we have a bag with an infinite number of marbles.
- n% of the marbles are blue, 1-n% are red.
- Suppose we take 20 marbles out of the bag.
- We know from a couple weeks ago how to calculate the probability of getting exactly m blue marbles as a function of n.
- But suppose we don't know n. Rather, we get a number m of blue marbles and we want a posterior over possible proportions n.
- Can we write this with conditional probability notation?
- Conceptually, what we need to do is go from one conditional probability, namely P(m blue marbles | n) to another, namely P(n | m blue marbles)

Bayes' theorem, a simple derivation

EBERHARD KARLS
UNIVERSITÄT
TÜBINGEN

- To do this, we can use Bayes theorem
- There is also a simple derivation of Bayes theorem you can keep in mind.
- First, note that from the definition of conditional probability we can write the joint in two different ways:

$$\begin{aligned} P(H\&D) &= P(H \mid D)P(D) \\ &= P(D \mid H)P(H) \\ P(H \mid D)P(D) &= P(D \mid H)P(H) \\ P(H \mid D) &= \frac{P(D \mid H)P(H)}{P(D)} \end{aligned}$$

The components of Bayes theorem

- Three ingredients in Bayes theorem:

$$P(H \mid D) = \frac{\overbrace{P(D \mid H)}^{Likelihood}\,\overbrace{P(H)}^{Prior}}{\underbrace{P(D)}_{Evidence}}$$

- The likelihood is the probability of the data *given* the hypothesis (as a function of the hypothesis though!)
    - How to interpret it?
- The prior is the probability of the hypothesis NOT conditioned on the data
    - How to interpret it?
- The evidence is the probability of the data NOT condition on an H.
    - How to interpret it?

The components of Bayes theorem

- Three ingredients in Bayes theorem:

$$P(H \mid D) = \frac{\overbrace{P(D \mid H)}^{Likelihood}\overbrace{P(H)}^{Prior}}{\underbrace{P(D)}_{Evidence}}$$

- Let's think what happens when we change the components individually.
- Note that you can rewrite the evidence as a sum! Which one?
    - This means that if we calculate the numerators for all hypotheses and put them in a vector, and then we normalize the vector (divide it by its sum), we don't need to explicitly calculate the evidence.
    - If the space of hypotheses is infinite, it's often easy to calculate the numerator and hard or impossible to calculate the denominator!
- Is this all clear?

Bayes' theorem vs Bayesian update

- Usually, we apply Bayes theorem to calculate P(H | D), where:
  - The hypothesis is something about the world we can't observe directly
  - The data is something we can observe directly
- You can think of an application of Bayes theorem as a way of updating one's model of the world when new data comes in.
- A prior and a posterior then are relative to *one update*
- So we can think of one application of Bayes' theorem as an update in the state of knowledge given some data
- This gives a very natural way of thinking about the way humans could update their picture of unknown quantities given a stream of new evidence.

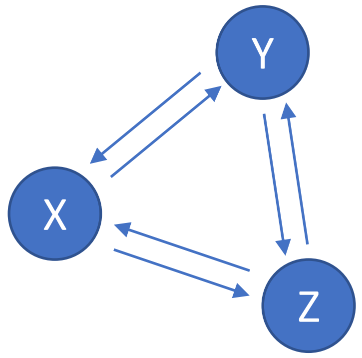
Applying Bayes' theorem to example

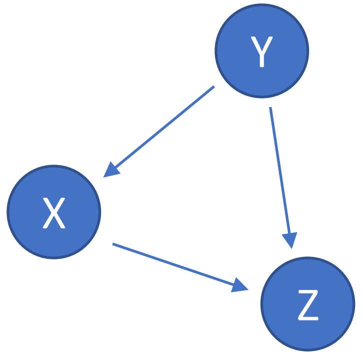
- Example 1:
  - Suppose we got the following sample from the bag above:
  - 4 blue marbles, 6 red marbles
  - Let's calculate the posterior of n
- Example 2:
  - Now I observe one more red marble.
  - What happens?
- A person tells good jokes 30% of the time, alright jokes 30% of the time and bad jokes 40% of the time. Their friend laughs 10% of the time when it's a good joke; 3% of the time when it's okay; and 7% of the time when it's bad.
  - What is the probability it was a bad joke if their friend laughs?
  - What is the probability it was an okay joke if their friend doesn't laugh?

Causal graphs

EBERHARD KARLS
UNIVERSITÄT
TÜBINGEN

- Imagine we have a bunch of random variables X, Y, Z
- This induces a joint distribution P(X, Y, Z)
- We can factor this in various equivalent ways, e.g.
  - P(X, Y | Z) P(Z)
  - P(Y | X, Z) P(X | Z) P(Z)
  - Etc.
- We know that least some of these conditional probs will depend on the way the variables *causally* influence each other.
- In a causal graph, we have a node for each variable, and we draw an arrow from A to B iff A causally influences B.
  - E.g. P(X, Y, Z) = P(Z|Y)P(X|Y)P(Z|X)
- We can distinguish between seen and unseen variables!





Thinking in generative terms

- The Bayesian approach is *generative*. This means that we imagine the data as being generated by some (unseen) mechanism.
- In practice, we start with a *joint* over data and hypotheses:
  - P(data, hypothesis)
  - The hypothesis is a combination of values for all the unseen variables
- Which then factorizes into prior and likelihood
  - P(data|hypothesis)P(hypothesis)
- The prior is the distribution we give to the unconditional random variables in the generative mechanism, the likelihood is defined by all the conditional probabilities.
- We can also give a value to all the unconditional variables in the generative model and do a 'forward pass' through the model, i.e. calculate P(D|H)

Bayesian inference in formal grammars

- How do you think we could use Bayesian inference to learn a sentence in a grammar (the LoT) given some observations?
- What's the prior, likelihood, evidence, and posterior?
- What is going to be the practical problem with this?

- Next week we're going to see how we can partially solve this problem!

Summary

- This week we have seen a little bit about Bayesian inference.
- In particular, starting from the concept of a conditional distribution, we have seen how to go from one conditional $P(D|H)$ and a prior $P(H)$ to a posterior $P(H|D)$
- This is basically the fundamental idea of Bayesian inference. Everything else is an elaboration on this.
- A big problem with Bayesian inference is computational: we need clever algorithms to actually find the posterior.
- Next week we are going to see one such algorithm, the Metropolis-Hastings algorithm.
- We'll also see how to apply it to some basic cognitive examples.