Computational Cognitive Science

Lecture 9: A Bayesian model of concept learning

Chris Lucas

School of Informatics

University of Edinburgh

October 18, 2022

Reading

- "Rules and Similarity in Concept Learning" by Tenenbaum (2000; link)

Cognition as Inference

The story of probabilistic cognitive modeling so far:

- models assign probabilities to human behaviors/judgments
  - e.g., prob. of assigning category A to item i in the GCM
- We can use $P(y|\theta, \mathcal{M})$ to
  - estimate psychologically interpretable parameters
  - compare and evaluate models

Cognition as Inference

Today we'll discuss a model where probabilities aren't just a useful tool, but rather have a *cognitive status*.

Cognition as probabilistic inference

Many recent models assume that probabilities and estimation are cognitively real – we estimate and represent something like probabilities. Why?

1. people act as if they have degrees of belief or certainty
2. humans must deal constantly with ambiguous and noisy information
3. experimental evidence: People exploit and combine noisy information in an adaptive, graded way

Cognition as probabilistic inference

1. People act as if they have degrees of belief or certainty

Example:

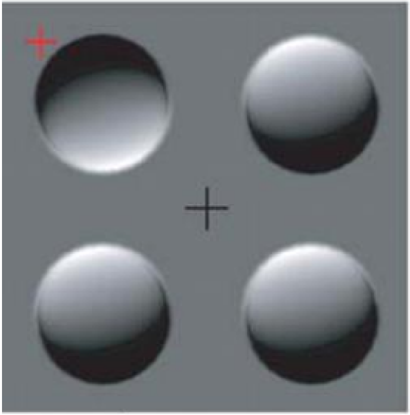
- Alice has a coin that might be two-headed.
- Alice flips the coin four times, it comes up HHHH.

Consider the following bets:

- Would you take an even bet the coin will come up heads on the next flip?
- Would you bet 8 pounds against a profit of 1 pound?
- Would you bet your life against a profit of 1 pound?

Cognition as probabilistic inference

2. Humans must deal constantly with ambiguous and noisy information, e.g.,



"A common light-prior for visual search, shape, and reflectance judgments" (Adams, 2007)

Cognition as probabilistic inference

2. Humans must deal constantly with ambiguous and noisy information, e.g.,

“Infant Pulled from Wrecked Car Involved in Short Police Pursuit”

http://languagelog.ldc.upenn.edu/nll/?p=4441

Cognition as probabilistic inference

3. People exploit and combine noisy information in an adaptive, graded way, e.g.,

- Estimating motor forces and visual patterns from noisy data
- Combining visual and motor feedback
- Learning about cause and effect in unreliable systems
- Learning about the traits, beliefs and desires of others from their actions
- Language learning

Cognition as probabilistic inference

How do people represent and exploit information about probabilities?

Intuitively:

- our inferences depend on observations, but also on *prior beliefs*;
- as more observations accrue, estimates become more reliable;
- when observations are unreliable, prior beliefs are used instead.

Today we will discuss a model built on these intuitions.

Concepts vs. Categories

Tenenbaum (2000) addresses the question of how people quickly learn new "number concepts":

- concepts can be concrete categories (dog, chair), or more vague ("healthy level" for a specific hormone, "ripe" for a pear);
- here, we will focus on number concepts ("odd number", "between 30 and 45").

Generalization is a key feature of concept learning: given a small number of positive examples, determine which other examples are also members of the concept.

Generalization

Given some examples of a concept, determine which other things belong to that concept. Two basic strategies:

- rule-based generalization: find a rule that describes the examples and apply it: *deterministic predictions;*
- similarity-based generalization: identify features of the examples and the new item, and decide based on how many features are shared: *probabilistic predictions.*

People's judgments are consistent with both strategies, but in different circumstances.

Generalization

Tenenbaum presents a Bayesian model of concept learning:

- the model can exhibit both rule-based and similarity-based behavior;
- but it is not a hybrid model: it uses only one mechanism, rules and similarity are special cases;
- explains how people can generalize from very few examples;
- *Bayesian hypothesis averaging* is a key feature of the model.

The model is trained on data from *number concept* learning.

The Number Game

I think of a "number concept" (a subset of numbers 1–100), e.g.,

- odd numbers;
- powers of two;
- numbers between 23 and 34.

I choose some examples of this concept at random and show them to you:

- $\{3, 57\}$;
- $\{16, 2, 8\}$;
- $\{25, 31, 24\}$.

You guess what other numbers are also included in the concept.

Experimental Design

Subjects are told how the game works.

Then, a few examples of the concept are presented:

- class I trials: only one example;
- class II trials: four examples, consistent with a simple mathematical rule;
- class III trials: four examples, similar in magnitude.

Subjects then rate the probability that other numbers (randomly chosen from 1–100) are also part of the concept.

Class I Trials

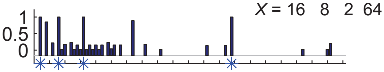
Only one example is given (16 or 60). Results:

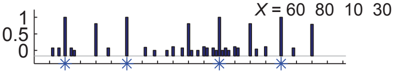
- responses fairly uniform, but slightly higher ratings for similar magnitude, similar mathematically;
- even numbers (both), powers of two (16), multiples of ten (60).

Notes: stars show examples given; missing bars are not zero, just were not queried.

Class II Trials

Four examples were given, consistent with a simple mathematical rule ({16, 8, 2, 64} or {60, 80, 10, 30}). Results:

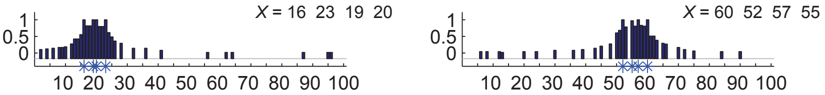




- responses reflect most specific rule consistent with examples, other numbers have a probability near zero;
- these rules are not the only logical possibility: {16, 8, 2, 64} could be "even numbers", for example.

Class III Trials

Four examples were given that didn't follow a simple rule, but were similar in magnitude ({16, 23, 19, 20} or {60, 52, 57, 55}). Results:



- responses reflect similarity gradient by magnitude;
- low probability for number more than a fixed distance away from the largest or smallest example.

Bayesian Model

Given data $X = \{x^{(1)}, \ldots, x^{(n)}\}$ sampled from concept C, we want to determine $P(y \in C|X)$ for new data point y.

As in many inference problems, a hidden variable (C) determines the inference, but we don't know C, so we will average over it:

$$P(y \in C|X) = \sum_{h \in H} P(y \in C|C = h)P(C = h|X)$$

To compute the *posterior* $P(C = h|X)$, we need to decide:

- What is the hypothesis space $\mathcal{H}$?
- What is the prior distribution over hypotheses?
- What is the likelihood function?

Hypothesis Space

In theory, all possible subsets of numbers 1–100.

The full space too large; we consider only salient subsets:

- subsets defined by mathematical properties: odds, evens, primes, squares, cubes, multiples and powers of small numbers, numbers with same final digit;
- subsets defined by similar magnitude: intervals of consecutive numbers.

Total: 5083 hypotheses.

Prior $P(C = h)$

First, assign a probability to each type of hypothesis:

- $P(C \text{ is defined mathematically}) = \lambda$;
- $P(C \text{ is defined as an interval}) = 1 - \lambda$.

Use $\lambda = \frac{1}{2}$.

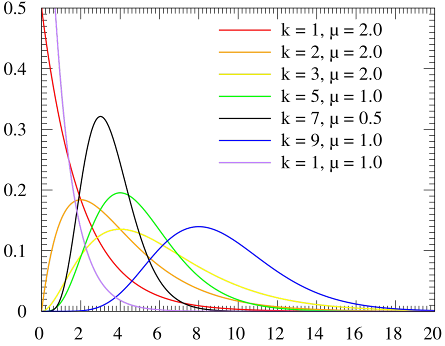
Prior $P(C = h)$

Within mathematical hypotheses:

- all are equally probable (why?)

Within interval-based hypotheses:

- medium-sized intervals are more probable than small or large intervals. Via Erlang distribution: $P(h) \propto (|h|/\sigma^2)e^{-|h|/\sigma}$.



[1]

[1] wikipedia; ikamusumeFan. $\mu = \sigma = \frac{1}{\lambda}$, $k = 2$

Likelihood $P(X|C = h)$

Assume examples are sampled uniformly at random from C.

For hypothesis h containing $|h|$ numbers, each number in h is drawn as an example with probability $1/|h|$, so:

$$P(X = x^{(1)} \dots x^{(n)}|h) = \begin{cases} \frac{1}{|h|^n} & \text{if } \forall j, x^{(j)} \in h \\ 0 & \text{otherwise} \end{cases}$$

Ex. For h = "multiples of five", $|h| = 20$, $P(10, 35|h) = 1/20^2$.

Size principle: for fixed data, smaller hypotheses have higher likelihood than larger hypotheses. As data increases, smaller hyps have exponentially higher likelihood than larger hypotheses.

Inference over Posterior

Draw inferences by averaging over hypotheses:

$$P(y \in C|X) = \sum_{h \in H} P(y \in C|C = h)P(C = h|X)$$

$P(y \in C|C = h)$ is either 0 or 1.

The posterior $P(C = h|X)$ is computed using Bayes' rule, with likelihood and prior as defined above:

$$P(C = h|X) = \frac{P(X|C = h)P(C = h)}{P(X)}$$

Alternative Models

Similarity model (SIM):

- Ignore size principle, just look at *consistency* with hypotheses.
- Posterior doesn't sharpen as sample grows

Effectively a 0/1 likelihood:

$$P(X = x^{(1)} \dots x^{(n)} | h) = \begin{cases} 1 & \text{if } \forall j, x^{(j)} \in h \\ 0 & \text{otherwise} \end{cases}$$

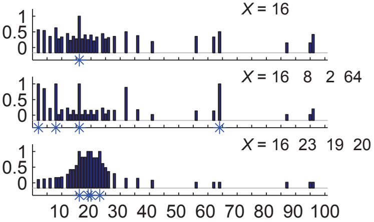
Alternative Models

Rule-based model (MIN):

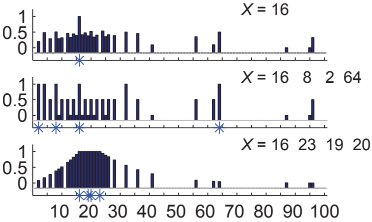
- replaces hypothesis averaging with MAP estimate: always choose the highest probability hypothesis;
- since priors are weak, guided by likelihood: always selects the smallest (most specific) consistent rule (size principle);
- reasonable when this rule (hypothesis) is much more probable than all others (Class II);
- not reasonable when many hypotheses have similar probabilities (Class I and III).

Results

Humans:

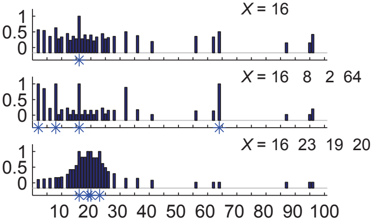


Similarity-based model:

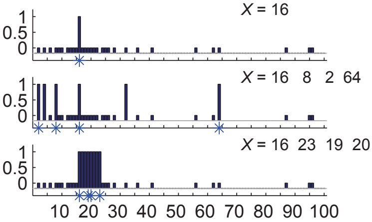


Results

Humans:

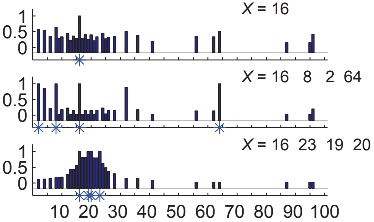


Rule-based model:

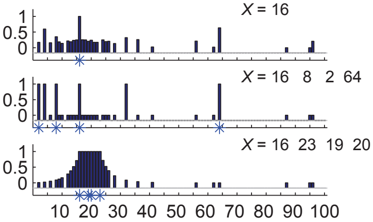


Results

Humans:



Bayesian model:



Conclusions

- Previous work suggested two different mechanisms for concept learning: rules or similarity
- no explanation for why one of them is used in any given case
- Bayesian model suggests these are two special cases of a single system implementing Bayesian inference
- this results stems from an interaction between:
  - *hypothesis averaging:* yields similarity-like behavior when many hypotheses have similar probability
  - *size principle:* yields rule-like behavior when one hypothesis is much more probable than others