Generalization, similarity, and Bayesian inference

Tenenbaum & Griffiths, 2001

Presented by Lisa Pearl

The Basic Question

- Them: Given one example x with consequential property c, how do you determine if example y also has c?
- Equivalent for Us: Given one utterance x with consequential property c = "in the language", how do you determine if utterance y is also in the language?
  (*useful for production, for instance, or determining if the current hypothesized language is the correct one given x and y as input*)

Basic Scenarios (Them)

- Doctor's Dilemma: If a hormone level of 60 yields a healthy patient, what other values in the range of 0 to 100 also yield a healthy patient?
- Hungry Birdie: If a worm with skin pigmentation of 60 is good to eat, what other values of skin pigmentation in the range of 0 to 100 also mean a worm is good to eat?

Shephard's (1987, 1994) Ideal Generalization Problem

Given an encounter with a single stimulus that can be represented as a point in some psychological space and that has been found to have some particular consequence, what other stimuli in that space should be expected to have the same consequence?

Assumption: Answer is interval in continuous psychological space, i.e. "between 50 and 70"

Basic Question for any stimulus y: Does y fall in that interval?

T & G's take on Shephard

- Shephard deals with generalization from a single encountered stimulus, and assumes the stimulus can be represented as points in a continuous metric psychological space
- …But more interesting problems often involve inferences from multiple examples, or from stimuli that are not easily represented in spatial terms (*i.e. acquiring appropriate grammar from E-language input*)

T & G observation

- When you have multiple input stimuli, the likelihood of a particular generalization depends on what the realm of hypotheses is

Ex: Input = {60, 30, 50}

- Hypothesis realm = hormone levels
  - 47 likely to be better than 80 for "healthy patient"
- Hypothesis realm = mathematical concepts
  - 80 likely to be better than 47 for "shared concept"

But back to Shephard…

- Shephard's (1987) formulation of the problem of generalization:
- Given: one example x with consequential property c
- Assumptions:
  - x can be represented as a point in continuous psychological space
  - C corresponds to some region of that space = consequential region (all points in region have property c)
- Task: find **p(y ∈ C | x)**
  - Probability that y has property c (is an element of C), given that we've just seen x

Three Questions

1) Them: What consitutes the learner's knowledge about the consequential region?
   Us: What does a learner know about a grammar/parameter option/set of utterances that can be parsed?

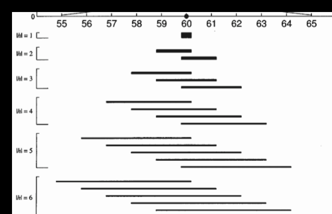
2) Them: How does the learner use that knowledge to decide how to generalize?
   Us: How does the learner use the grammar/parameter option to classify input?

3) Them: How can the learner acquire that knowledge from the example encountered?
   Us: How does the input get parsed & assigned to a particular grammar/parameter option?

Learner's Knowledge About C

- Them: Represented as a probability distribution p(h|x) over an a priori-specified hypothesis space H of possible consequential regions, where h ∈ H.
  - Us: H = a priori-specified space of possible grammars/parameter options [UG]; h = single grammar/parameter option
- One and only one element of H ($h_{correct}$) is assumed to be true
- Them: Using Shephard's (1964) suggestion that H is made up of connected subsets of psychological space (i.e. intervals for hormone levels/pigementation)

Example H, for |h| <= 6



H contains all these h

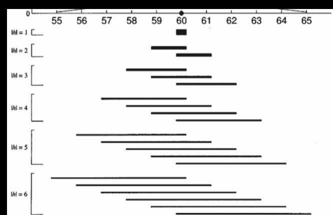
But before we see x…

- A Reasonable Question: What is the state of H before observing x (prior probability of any h ∈ H = p(h))?
- Us: What is specified by UG before a learner gets any input? Is p(h) higher for some grammars/parameter options than others?

After seeing x…

- p(h|x) = posterior probability, the probability that h = $h_{correct}$ after seeing x
- Us: The probability that one grammar/parameter option is the correct one after seeing x

p(h|x) = Degree of "belief" that h is true

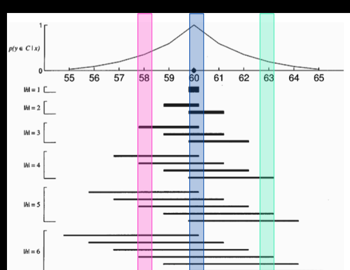


- Height of bar indicates level of "belief" (also uniform distribution over interval)

How does the learner use x to generalize?

- Generalization function
  $p(y \in C \mid x)$
- Computed by summing probabilities p(h|x) of all hypothesized consequential regions that contain y (*hypothesis averaging*)

$P(y \in C \mid x)$, x = 60



Using the input

- How do we update p(h) to p(h|x)?
- Use Bayes' rule:

$$p(h|x) = \frac{(p(x|h) * p(h))}{p(x)}$$

Using the input

- How do we update p(h) to p(h|x)?
- Use Bayes rule:

likelihood of seeing x, given this h

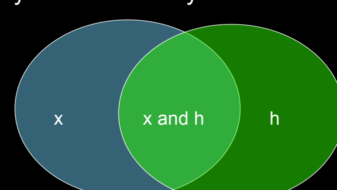
prior probability

$$p(h|x) = \frac{(p(x|h) * p(h))}{p(x)}$$

likelihood of seeing x, given **H**

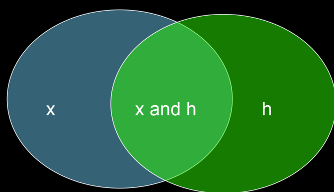
So how does Bayes' rule work exactly?

- Let's do an example. Suppose, in our world of possible events, we have two events x = "Lisa watches the movie Labyrinth" and h = "It's too hot to read outside". They may occur separately or concurrently.



So how does Bayes' rule work exactly?

- p(x) and p(h) occuring in general is given.
- What is the probability that Lisa watches the movie Labyrinth, given that it's too hot to read outside? [p(x|h)]

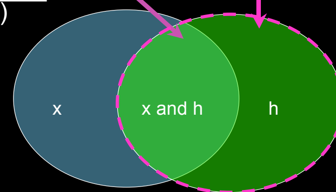


So how does Bayes' rule work exactly?

- Want p(x and h) given that h is true

So p(x|h) =

$$\frac{p(x \text{ and } h)}{p(h)}$$



So how does Bayes' rule work exactly?

- $p(x|h) = \dfrac{p(x \text{ and } h)}{p(h)}$

Therefore

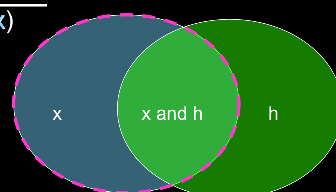
p(x|h) * p(h) = p(x and h)

So how does Bayes' rule work exactly?

- We can do the very same thing for p(h|x)

So p(h|x) =

$$\frac{p(x \text{ and } h)}{p(x)}$$



So how does Bayes' rule work exactly?

- $p(h|x) = \dfrac{p(x \text{ and } h)}{p(x)}$

Therefore
p(h|x) * p(x) = p(x and h)

From before:
p(x|h) * p(h) = p(x and h)

Therefore
p(h|x) * p(x) = p(x|h) * p(h)

$$p(h|x) = \frac{p(x|h) * p(h)}{p(x)}$$

How do we get these probabilities?

$$p(h|x) = \frac{(p(x|h) * p(h))}{p(x)}$$

p(h) = given by knowing about H (given by UG)

p(x) = given all the $h \in H$, sum the weighted probabilities for seeing x in them

so $p(x) = \Sigma_{h' \in H} p(x|h')*p(h')$

The likelihood, p(x|h)

- Shephard (1987): weak sampling
  - p(x|h) = 1 if $x \in h$, 0 otherwise
- Tenenbaum (1997, 1999): strong sampling
  - p(x|h) = 1/|h| if $x \in h$, 0 otherwise

  (works out to same as weak sampling if |h| = 1, such as if H = set of competing grammars or parameter values)

  - h could be thought of as a set of utterances generated by a grammar. (E-language of a grammar) Then, |h| > 1 makes sense.

The impact of strong sampling

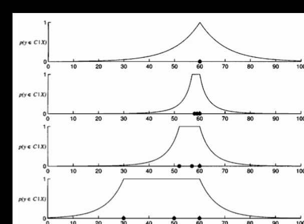
- p(x|h) depends on |h|, so more specific hypotheses (smaller intervals, sets of relevant numbers, or utterances) receive higher probabilities.

= “size principle”

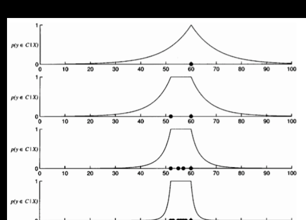
What about multiple inputs?

- What if we have more than one example to generalize from? (language: some memory for previous input)
- Example variability comes into play
  - Given: {60, 57, 52}
  - Task: generalize to interval
    - p(70 is in C) is less here than if given {60, 50, 30}
    - p(70 is in C) is more here than if given {60, 58, 59}
- Number of examples comes into play
  - Given: {60, 52, 57, 55}
  - Task: generalize to interval
    - p(70 is in C) is less here than if given {60, 52}
    - p(70 is in C) is more here than if given {60, 52, 57, 55, 58, 55, 53, 56}

Effects of Variability



Effects of Number of Examples



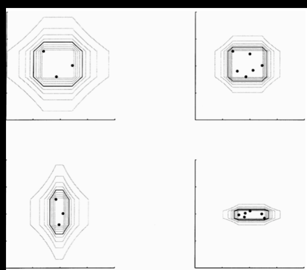
Extension of Theory

- Given: $X = \{x_1, \ldots, x_n\}$
- Task: Is $y \in C$?
- $P(y \in C \mid X)$ = sum of p(h|X) for all h that contain y

Use Bayes’ again:

- $p(h|X) = \dfrac{p(X|h) * p(h)}{p(X)}$

Also works for 2 dimensions



But what if the points *aren't* in a continuous metric psychological space?

- For instance, where "objects are represented in terms of presence or absence of primitive binary features" = conjunctive feature structures
- Consequential subsets = all objects sharing different conjunctions of features
- Parameterized grammars, anyone?

Discontinuous Points

- Example for Them: Numbers sharing certain mathematical concepts (i.e. "even number", "divisible by 3", etc.)
- Example for Us: Utterances sharing certain parameter values (i.e. "OV order") or able to be parsed by certain grammars (i.e. "g1")
- Input: 60, "We must Labyrinth watch"
- Task: p(y also shares this concept/is in the language)

Discontinuous Points

- Solving the Math Task: seems based on similarity (how many math properties are shared)
- Solving the Language Task: seems based on utterance similarity (how many parameter values are shared for grammar, if parameter value is shared or not for individual parameter)

Solving the Math Task

- Identify each mathematical property that the learner knows about with a possible consequential subset in H, then calculate similarity for each y to the input

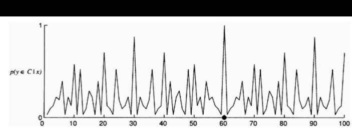


Figure 5. Bayesian generalization in the number game, given one example $x = 60$. The hypothesis space includes 33 mathematically consequential subsets (with equal prior probabilities): even numbers, odd numbers, primes, perfect squares, perfect cubes, multiples of a small number (3–10), powers of a small number (2–10), numbers ending in the same digit (1–9), numbers with both digits equal, and all numbers less than 100.

Solving the Language Task

- Identify each grammar that the learner knows about with a possible consequential subset in H (which contains the various grammars used for parsing…or the collection of utterances each grammar could parse), then calculate similarity for each y to the input

Calculating Similarity: Tversky's (1977) Contrast Model

- Similarity of x to y is a function of the features shared by both x and y, as well as the features exhibited by x but not y, and the features exhibited by y and not x (parts in common and parts not in common)
- Size principle applicability: certain kinds of features should receive higher weights in similarity comparisons, if they belong to fewer objects
- Example of size bias in action for numerical cognition:
- {60} = "even", "multiple of 10"
- {60, 80, 10, 30} = "multiple of 10"

Size Principle With Language?

- Certain parameters/parameter values should receive higher weights in similarity, if they can parse fewer utterances (perhaps given preference in parsing ambiguous utterances)
- Possible Implementation: Subset Learnability (assume the subset value, until you're forced into the superset)
- Numerical equivalent: {60, 80, 10, 30, …,42}

A Caveat…

- Size principle is tempered by prior probability
- Example (Them):
  - Concept = "all multiples of 10, except 20 and 70"
  - More specific than "all multiples of 10", so this might be predicted to be more probable…but doesn't seem to be true psychologically ({60, 30, 10, 80})
  - Why not? Because this concept has a lower prior than "all multiples of 10"
- A way to mathematically express dislike for rules + exceptions? ("Its prior is low!")

About Unsupervised Learning

- "A set of objects tends to cluster together (behave similarly)" = increases learner's prior probability that the subset is likely to share some important but as-yet-unencountered consequence
- Us:
  - Structure alternations based on specific lexical items (transitive/intransitive/ditransitive verbs, raising verbs, control verbs) or registers (subject-dropping in casual speech - "Want some?")
  - Syntactic-bootstrapping: same syntactic frame = similar semantics