Statistical Learning Mechanisms

by Sydney Bishop

2.1

designing informative models

In order for our modeling results to tell us something about how children develop syntactic knowledge (i.e., for a model to be an informative model and not just an interesting programming exercise), we need to believe that the model **reasonably approximates** aspects of a child's acquisition process. The way to do this is to make sure the **model components** are **cognitively plausible.** That is, we make reasonable assumptions for what's actually going on during the acquisition process in children when implementing each model component.

cognitive plausibility

- Theoretical research
  - defines parts of **initial state** (developing grammar, UG contents) -> impacts data in modeled child's acq intake
  - **inference** + **target knowledge state**
- Corpus analysis
  - defines **data intake,** analyzes child directed speech
  - quantifies aspects of data intake

- Experimental research
  - Defines **initial state** - parsing and extralinguistic abilities -> influences **perceptual intake** and production systems
  - helps determine inference abilities and learning periods for modeled child
- cognitive plausibility is still difficult to achieve.. If no empirical data-> **principled decision**

2.2

levels of explanation

levels of explanation

- computational-level
  - is it possible to reach a specific **target state** given a specific **initial state** + **data intake**?
  - concerned only with completing the computation accurately
    - human inference vs model may differ

- algorithmic
  - represents the steps children go through during **inference** -> sensitive to constraints (limited memory or time)
- implementational
  - how cognitive computation of acq is implemented in the brain
  - what do the steps of acquisition look neurally, in our brain?

2.3

inference

01. counting

02. reinforcement learning

03. tolerance principle

04. bayesian updating

counting

- a way of determining what actually makes it to the inference process
- translated into probabilities
- smoothing required - not observed ≠ never can occur
  - IRL, children get finite data and make generalizations
  - smoothing implemented via pseudocounts

0.5 pseudocount for each option

$$p(+wh) = \frac{count_{+wh}+0.5}{count_{+wh}+0.5+count_{-wh}+0.5} = \frac{count_{+wh}+0.5}{count_{+wh}+count_{-wh}+2*0.5} = \frac{0+0.5}{0+0+2*0.5} = \frac{0.5}{2*0.5} = 0.5$$

$$p(-wh) = \frac{count_{-wh}+0.5}{count_{+wh}+0.5+count_{-wh}+0.5} = \frac{count_{-wh}+0.5}{count_{+wh}+count_{-wh}+2*0.5} = \frac{0+0.5}{0+0+2*0.5} = \frac{0.5}{2*0.5} = 0.5$$

1 +wh-movement

$$p(+wh) = \frac{count_{+wh}+0.5}{count_{+wh}+0.5+count_{-wh}+0.5} = \frac{count_{+wh}+0.5}{count_{+wh}+count_{-wh}+2*0.5} = \frac{1+0.5}{1+0+2*0.5} = \frac{1.5}{2} = 0.75$$

$$p(-wh) = \frac{count_{-wh}+0.5}{count_{+wh}+0.5+count_{-wh}+0.5} = \frac{count_{-wh}+0.5}{count_{+wh}+count_{-wh}+2*0.5} = \frac{0+0.5}{1+0+2*0.5} = \frac{0.5}{2} = 0.25$$

having lots of data yields similar to unsmoothed prob

$$p(+wh) = \frac{count_{+wh}+0.5}{count_{+wh}+0.5+count_{-wh}+0.5} = \frac{count_{+wh}+0.5}{count_{+wh}+count_{-wh}+2*0.5} = \frac{700+0.5}{700+300+2*0.5} = \frac{700.5}{1001} = 0.6998$$

$$p(-wh) = \frac{count_{-wh}+0.5}{count_{+wh}+0.5+count_{-wh}+0.5} = \frac{count_{-wh}+0.5}{count_{+wh}+count_{-wh}+2*0.5} = \frac{300+0.5}{700+300+2*0.5} = \frac{300.5}{1001} = 0.3002$$

reinforcement learning

- linear reward scheme:
  - reward function
    - probability increased
  - penalized
    - probability decreased
- involves a learning rate parameter (γ) that determines how quickly beliefs updated
- can predict convergence behavior with unambiguous data

the tolerance principle

- how many exceptions to a potential generalization can be ‘tolerated’ and be useful for child
- determines when a child adopts a rule based on data counts
- optimizing retrieval time
- threshold: N - N/ln(N),
    - N = num items rule can apply to

“a formal approach for determining when child would choose to adopt a generalization or “rule” to account for a set of items Yang (2005, 2016). This principle is based on algorithmic-level considerations of knowledge storage and retrieval in real time, incorporating how frequently individual items occur, the absolute ranking of items by frequency, and serial memory access. It’s a sophisticated way of counting things that’s been applied to the acquisition of both morphosyntactic and syntactic knowledge”

bayesian updating

$$P(h|D) = \frac{P(D|h)*P(h)}{P(D)} = \frac{P(D|h)*P(h)}{\sum_{h' \in H} P(D|h')*P(h')} \propto P(D|h) * P(h)$$

"Bayesian updating is another kind of probabilistic inference mechanism, and it involves both prior assumptions about the probability of different options (typically referred to as hypotheses) and an estimation of how well a given hypothesis fits the data."

Key components:

- Hypotheses Space
- Posterior Probability P(h/D)
- Likelihood P(D/h)
- Prior Probability P(h)
- normalizes all hypotheses

Significance:

- allows modeling of how children generalize syntactic structures based on existing data
- links abstract theories w practical lang acq scenarios