Preference learning

Daniel Hsu

COMS 6998-7 Spring 2025

Story so far

Q: How well can we capture structure of natural language?

- Shannon's N-gram model and basic learning objective
  - Very easy to understand, but quite limited in power
- Neural language models
  - Exploit word embeddings + neural computation models
  - Some coarse- and fine-grained understanding of what they can do
  - Fairly robust understanding of learnability/generalization theory
  - Very preliminary understanding of efficient training algorithms (commensurate with how well we understand training for other neural nets)

LLM training

- Pre-training
  - Shannon's objective (a special case of "self-supervised learning")
- ~~Training~~
- Post-training
  - Supervised fine-tuning
    - Labeled examples (prompt x, response y)
    - Train model to predict desired responses to prompts
  - Instruction-tuning
    - Labeled examples (instructions i, prompt x, format f)
    - Train model so predicted responses match the desired format as instructed
  - Both of these "post-training" methods are (just) standard supervised learning

E.g., questions are not always followed-up with correct answers in natural language text

Handle custom instructions and response formats

Preference-tuning

How to make a language model polite?

- Solution 1: supervised fine-tuning on prompts with polite answers
  - Requires a polite person to write these polite answers
- Solution 2: supervised fine-tuning that rewards polite answers
  - Requires a polite person to judge whether answers are polite or not
    - How polite is polite enough? What is politeness level 7?
    - People tend to be better at comparing answers than giving absolute grades
  - Use pairwise preference comparisons to learn a reward function, which in turn is used with supervised learning

Reward model

- Classical models (like BTL): parameterized by quality score $w_i \geq 0$ for each item i

$$\Pr(i \succ j) = \frac{w_i}{w_i + w_j}$$

- Models with features: each item i has a feature vector v_i, and (log) quality score is $\log w_i = \langle \theta, v_i \rangle$ for some model parameter vector θ

$$\Pr(i \succ j) = \frac{1}{1 + \exp\big(-\langle \theta, v_i - v_j \rangle\big)}$$

- Models with context-dependent features:

$$\Pr(i \succ j | x) = \frac{1}{1 + \exp\big(-\langle \theta, \phi(v_i, x) - \phi(v_j, x) \rangle\big)}$$

- …