CS/INFO 6742: NLP and Social Interaction, Fall 2021

Nov. 9, 2021: Lecture 19 reference sheet, extended example of language-model development

---

1 Considerations for the structure of lecture

1. I want to show you how you might develop a language model that suits a language-analysis problem you face.
2. The fewer ~~hidden~~ parameters in a model, the "easier" the problem of inferring those values from data.

2 Motivating example: modeling small-talk vs. non-small talk

Let's consider a generative story like the following:

1. Pick a sentence length ℓ.
2. Pick a sequence of ℓ *states*: where the two possible state types are st for small talk, nst for not small-talk
3. For each state, pick a word according to that state's distribution over single words.

Example; we might decide we're going to say a five-word sentence, where the first word and the 4th and 5th words are going to be small-talk words.

2.1 Ideas for further refinement

- st might have a higher probability of ...
- st might have a higher probability of ...
- st might have a higher probability of

2.2 Sample data

Written "vertically" instead of "horizontally" to leave room to write on the sides.

Two ~~sentences:~~

hi
i
agree
thanks
bye

hi
sell
hi [some stock ticker symbol]
now
thanks

length modeling... (§2, #1)

"middling"

- pick a length ℓ vs. pick "long" vs. "short"
  (need to pick) [preset a threshold]

unrealistic — but direct estimation is easy.

* "picking" your length categories w.r.t. phenomena of interest.
→ manual. "prior"
→ induced by data...

how come no other class worries about length modeling? (b/c no one has had one). q: moot.

example if you are not careful about modeling:

$w_1 w_2 \cdots w_n$ ← a word.

$p(w_1 w_2 \cdots w_n) = \prod_i p(w_i)$

~~p(cats) = 1~~

$\phi_{cat} = 1$

$p(cat) = 1$.

$p(catcat) = 1 \times 1 = 1$

these are not the same p's!

~~2 diff's~~ ~~p~~

→← contradiction.

> what about priors? (length).

~~smoothing~~:

< to be more accurate - especially when sample seems limited or unrepresentative >

- there are mathematically convenient priors:

  ex: multinomial → Dirichlet prior (justified?)

- interpolating:) take as ~~prior~~ additional knowledge source an LM built on generic English. P_{LNG}

mike: P_{st} free parameter $\alpha \in [0,1]$ $p(w) = \alpha P_{st}(w) + (1-\alpha) P_{LNG}(w)$

other ways to combine 2 LMs?

backoff: if you had an indicator that your special LM was good or not:

rely on $P_{st}()$ when it's good

rely on $P_{LMG}()$ when it's not good.

~~how +~~ the details are in ~~defining~~ indicator, and making normalization work.

~~P(w)~~. ~~"p(w)"~~ ~~const~~

ex: "switch" might be frequency of the word.

---

how do you evaluate, say, $\alpha = .6$ vs. $\alpha = .9$

- see which assigns more accurate probabilities
- you can often check: 500 words from [the data you want to model] } held-out (not in inference ~~data~~)
  500 words from [from a very diff. sample. reasonable.]

[aside: don't compare probs of two samples of diff. lengths.

why? $P(w_i) = \prod_i p_i(w)$

longer sentences usually less probable.]

- $P_{\alpha=.6}$ (you want to model) | $P_{\alpha=.6}$ (you don't want to model)

$P_{\alpha=.9}$ (you want to model) | $P_{\alpha=.9}$ (out-of-domain data)