CS/IS 6742: NLP and Social Interaction, Fall 2017

Oct. 19, 2017: Reference sheet for lecture 17, "How different are two language models?"

---

1 Main concepts (but not the outline!) for today.

In many settings, we want to have at hand a way to compute a single number that tells us how different one language "source" (e.g., Republicans in 2016) is from another (e.g., Democrats in 2016, or Republicans in 2006).

We can characterize a language source as coming from an underlying/hidden class of *language models (LM)s*. We'll focus on basic classes today — *there are much more sophisticated options*.

We estimate any latent values for a particular LM from appropriate language samples.

There are several ways to measure the difference between two (estimated) distributions.

2 Estimating a multinomial's categorical distribution.

Assume a fixed non-empty finite vocabulary $V = \{v_i\}$.

The multinomial has the following parameters:

- L, the number of draws (the sample length)
- $\overrightarrow{\phi} \in \Re^{|V|}$, where $\sum_i \phi_i = 1$ and for all i, $\phi_i \geq 0$. This *categorical* distribution on just V (not V^*) specifies probabilities on the sides of the "die" whose sides are labeled with the vocabulary items v_i.

We are given a sample $S = w_1 \ldots w_L$, $w_k \in V$, and collect the counts S_i for each word v_i.

Maximum-likelihood estimate: find $\overrightarrow{\phi}$ that maximizes

$$\text{(some constant with factorials?)} \times \prod_i \phi_i^{S_i} \qquad \text{...?} \tag{1}$$

Maximum a posteriori estimate: assuming Dirichlet prior's parameter vector $\overrightarrow{\alpha}$ is fixed, find the $\overrightarrow{\phi}$ that maximizes

$$\text{Prob}(\overrightarrow{\phi}\,\text{drawn according to}\,\overrightarrow{\alpha}) \cdot \text{(some constant with factorials?)} \cdot \prod_i \phi_i^{S_i} \qquad \text{...?} \tag{2}$$

3 Language models and "single-word" distributions

In general, a (proper) language model is a (proper) probability distribution over V^*, all possible sequences (of any possible length) of words drawn from V, repeats allowed.

We will restrict attention to language models where it is sensible to consider an induced distribution, a "single-word" distribution, on just V.

The implied generative story induces a probability distribution over V^L.[1]

(How can we get a full language model on V^* from this?)

4 Measuring the difference between two "single-word" distributions

Restrict attention to $q(\cdot)$ and $r(\cdot)$ over finite sample space V.

---

[1] Strictly speaking, what the multinomial distribution is canonically considered as probability over word-count vectors of length n, not over word sequences. Hence the multiplicative term with all the factorials you're used to seeing in the formal definition of the multinomial distribution.