lec21

Tuesday, November 16, 2021 12:22 PM

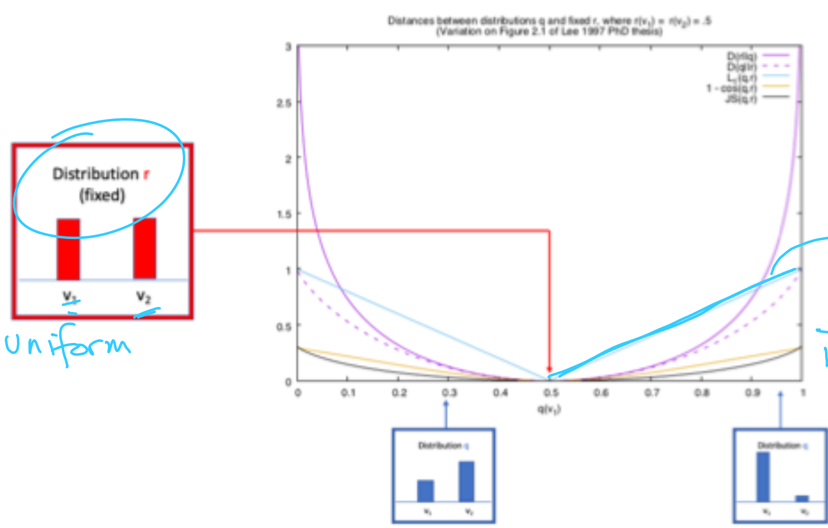
lec21

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

Nov. 16, 2021: Lecture 21: distances between language models (cont.)

1 Entropy/surprisal-based distance functions

We restrict attention to proper distributions $q(\cdot)$ and $r(\cdot)$ over finite "vocabulary" $V = \{v_i\}$. We write q_i and r_i for $q(v_i)$ and $r(v_i)$.



$L_1(q,r)$

$\sum_i |q_i - r_i|$

$L_2(q,r) = \sqrt{\sum_i (q_i - r_i)^2}$

The *surprisal*[1]:

$$-\log(r_i) = \log \frac{1}{r_i} \tag{1}$$

can be thought of as how *surprised* we should be from the perspective of using r as a model to see v_i, or r's *surprisedness* or *surprisingness* for v_i. The base of the log is customarily taken to be 2, which makes this surprisingness number interpretable as the best choice of number of bits of information to encode v_i under distribution r over V.

1.1 Cross-entropy (asymmetric)

what if q & r are the same?

If we considered the "reference" distribution to be q, then the *cross-entropy*

$$H(q||r) = \sum_i q_i \boxed{\log \frac{1}{r_i}} \text{ taking } 0 \log 0 \text{ to be } 0. \tag{2}$$

$$\sum_i q_i \log \frac{1}{q_i}$$

$$= -\sum_i q_i \log q_i$$

is the expected surprisedness for r with respect to reference distribution q.[2]

1.2 KL-Divergence – asymmetric, w/ good reason

$$\boxed{D(q||r) = \sum_i q_i \log \frac{q_i}{r_i}} \tag{4}$$

if $q_1 = 1$ so $q_2 = 0$

$1 \cdot \log(1) + 0 \cdot \log 0$

$= 0$

q: flat & r spiky

vs.

q spiky & r flat.

But if $q_1 = 1/2$, $q_2 = 1/2$

$= \frac{1}{2}\log(2) + \frac{1}{2}\log(2)$

$= 1/2 + 1/2 = 1 \neq 0$

[1] According to Wikipedia, the term was coined in Tribus, 1961, *Thermostatics and Thermodynamics*.

[2] *How you often see this in papers:* If the "reference" distribution is taken to be the one induced from the empirical counts from a sample $S = w_1 w_2 \ldots$, where each $w_k \in V$ and the length of the sample is L, then this can be refactored as:

$$\hat{H}_S(r) = \frac{1}{L} \sum_{k=1}^{L} \log \frac{1}{r(w_k)} \tag{3}$$

1.3 Jensen-Shannon divergence

See Lin, Jianhua. 1991. Divergence measures based on the Shannon entropy. *IEEE Transactions on Information Theory* 37(1): 145-151. Let $\text{avg}_{q,r}$ be the average distribution between q and r.

$$JS(q,r) = \frac{1}{2}\left[D(q||\text{avg}_{q,r}) + D(r||\text{avg}_{q,r})\right] \tag{5}$$

1.4 Skew divergence

See Lee, Lillian. 1999. Measures of distributional similarity. In *Proceedings of the ACL*, 25-32.

$$\text{skew}_\beta(q||r) = D(q||\beta \cdot r + (1-\beta)q) \tag{6}$$

Values used include $\beta = .99$.

2 Distance functions where there's a geometry on the words

The 1-Wasserstein distance, earth-mover's distance, word-mover's distance.

Assume you have a distance function over "words" — in particular, over word *embeddings*.

From Wikipedia entry:

$$\text{Wass}(q,r) = \inf_s E(d(V, V')) \tag{7}$$

where the expectation is taken over *all joint distributions s over V and V' that has marginals q and r respectively.* "inf" is the infimum.

The Wikipedia page describes the "dirt-moving" metaphor.

So, $H(q||q)$ is not necessarily 0.

$\left[H(q||r) = \sum_i q_i \log \frac{1}{r_i} \right.$ $r_i = 1$, $r_i = 0$

When is $H(q||r)$ really, really big? $\rightarrow$ $0 \cdot \log 0 \triangleq 0$

if $q_i \neq 0$ & $r_i = 0$, really, really big

~~what~~ when is $H(q||r)$ minimized:

Fix q. what r minimizes?

$$\frac{\partial}{\partial r_j}\left[\sum_i q_i \log \frac{1}{r_i} + \lambda \underbrace{\left(\sum_i r_i - 1\right)}\right]$$

Lagrange multiplier for constraint.

$$= q_j \frac{\partial}{\partial r_j} \log \frac{1}{r_j} + \lambda \cdot 1 = -q_j \log r_j + \lambda = -q_j \frac{1}{r_j} \frac{\partial r_j}{\partial r_j} + \lambda$$

set to 0: $\lambda = \frac{q_j}{r_j} \Rightarrow r_j = \frac{q_j}{\lambda}$.

$\lambda = \sum_i r_i$, normalisation.

r minimizes $H(q||r)$ when $r = q$.

The value ~~of~~ at that point is $H(q||q) = \sum_i q_i \log \frac{1}{q_i}$

"~~norm~~ "rescaled" $H(q||r)$:

$$\underline{H(q||r)} - \underbrace{\sum_i q_i \log \frac{1}{q_i}}$$