On the distribution of word pairs

Julaiti Arapat and Yoav Freund

January 19, 2015

1 Model definition

Each word is associated with an empirical distribution over words. For each pair of words (or word clusters) we define a distance, the distance is the probability that the two empirical distributions have been sampled from the same distribution.

2 Translating corpus into distribution vectors

We use the letters w, u, v and w_i, u_j etc. to denote words. The universe of all words is denoted W.

Let $n(w)$ denote the number of times the word w occurs in the corpus.

We will associate a distribution $\vec{p}(w)$ with every word w that appears in the corpus (at least some minimal number of times). we will denote by $p(w,u)$ the probability that $\vec{p}(w)$ assigns to the word u.

Denote by w_i, $i = 1, 2, \ldots, n(w)$ the ith occurance of the word w. we associate with w_i a distribution $\vec{p}(w_i)$ over the words that appeared in the context of w. If there is only one word in the context (for example, when relating main verb to main noun) then the distribution is concentrated on the single word. If there are several words, such as when using all other words that appear in the same sentence, then the distribution is created by counting the number each word $v \in W, v \neq w$ appears in the the sentence. The importance of the normalization is to give all sentences, regardless of length, equal importance.

The distribution that is associated with the word w is the average over the distribution over all of the occurances:

$$\vec{p}(w) \doteq \frac{1}{n(w)} \sum_{i=1}^{n(w)} \vec{p}(w_i)$$

3 Measuring the deviation of a word from the average distribution

Let the total number of words in the corpus be $n(W) \doteq \sum_{w \in W} n(w)$ Define the overall average context distribution as

$$\vec{p}(W) \doteq \frac{1}{N} \sum_{w \in W} n(w)\vec{p}(w) = \frac{1}{N} \sum_{w \in W} \sum_{i=1}^{n(w)} \vec{p}(w_i)$$

We define the KL divergence as

$$KL(\vec{p}(w)||\vec{p}(W)) \doteq \sum_{u \in W} p(w,u) \log \frac{p(w,u)}{p(W,u)}$$

We define the *score* of the word w as $n(w)KL(\vec{p}(w)||\vec{p}(W))$ this expression comes from the fact that the Sanov bound states that the probability that the distribution $\vec{p}(w)$ is a result of sampling $n(w)$ times from the distribution $\vec{p}(W)$ is

$$\exp(-n(w)KL(\vec{p}(w)||\vec{p}(W)))$$

The number of occurances is important! The more samples we have, the larger the impact of the KL divergence.

4 Measuring the deviation between pairs of distributions

The idea behind the JS divergence is to bound the probability that the $\vec{p}(w)$ and $\vec{p}(u)$ are a result of sampling from the same (unspecified) distribution.

The line of reasoning is as follows. First, find the single distribution $\vec{q}$ that is most likely to have generated both samples. Then, compute the score for each of the two samples using the KL divergences of each sample from $\vec{q}$ and the counts for each sample.

Given $\vec{q}$, the scores associated with the two distributions $\vec{w}$ and $\vec{u}$ is

$$JS - score(w,u) = n(w)KL(\vec{p}(w)|\vec{q}) + n(u)KL(\vec{p}(u)|\vec{q})$$

[JA: 1] With the definition of $\vec{q}$ as bellow, I think the $JS - score$ should be

$$JS - score(w,u) = n(u)KL(\vec{p}(w)|\vec{q}) + n(w)KL(\vec{p}(u)|\vec{q}),$$

because in case $n(w) >> n(u)$, we want the $JS - score$ to be close to $KL(\vec{p}(u)|\vec{q})$. If we go other way around, it will be close to 0.

It is not hard to show (Julaiti, try doing that!) that the $\vec{q}$ that minimizes the JS score is the weighted average of $\vec{p}(w)$ and $\vec{p}(u)$:

$$\vec{q} = \frac{n(w)\vec{p}(w) + n(u)\vec{p}(u)}{n(w) + n(u)}$$

Note that in this definition of the JS score the weighting of the two terms is not necessarily $1/2, 1/2$ but depends on the counts for each word.

I believe that if one of the counts is much larger than the other than the JS-score converges to the KL divergence $KL(w||v)$ where w is the rare word and v is the common word.

Note that when both words appear a large number of times then small KL-divergences are enough to distinguish them while if they have only a small number of counts then they will have a small score even if tghe KL divergences are large.

5 Computing the fractal dimension

We are interested in measuring the average pointwise dimension. For a particular point x we want to calculate the amount of mass as a function of the radius, where mass is measured in terms of the probability of words, i.e. the sum of the counts of the words that are at distance up to epsilon from the particular point.

I think that we want to use this weighting also when we take the average over different centers.

Suppose $d(w,u)$ is the distance between the distributions associated with the words w and u as defined above (sqrt of the JS distance). Then what we want to estimate is suppose the word u and the word v are chosen independently at random according to the distribution defined by the frequencies of the word. What is the CDF of the distance between u and v.