Flipped Classroom Questions on Brown Clustering and Word2Vec

Michael Collins

Question 1: Assume the Brown clustering set-up. We have a corpus, and define

$$f(u, v)$$

for any word pair (u, v) to be the number of times the bigram (u, v) is seen in the data.

In addition we define

$$f_1(u) = \sum_v f(u, v) \quad f_2(v) = \sum_u f(u, v)$$

Next, assume we have some clustering function C that maps any word in vocabulary u to a cluster $C(u) \in \{1 \dots K\}$. Here K is the number of clusters.

Define the following counts:

$$g(c, c') = \sum_{u:C(u)=c} \sum_{v:C(v)=c'} f(u, v)$$

I.e., $g(c, c')$ is the number of times we see the cluster bigram (c, c') in the data, under the function C. In addition define

$$g_1(c) = \sum_{c'} g(c, c') \quad g_2(c') = \sum_c g(c, c')$$

Under these definitions, given emission parameters $e(\cdot|\cdot)$ and transition parameters $q(\cdot|\cdot)$, the log-likelihood of the training data is

$$Q(C, e, q) = \sum_{u,v} f(u, v)[\log e(v|C(v)) + \log q(C(v)|C(u))]$$

The emission and transition parameters that maximize this function are

$$e(v|C(v)) = \frac{f_2(v)}{g_2(v)} \quad q(C(v)|C(u)) = \frac{g(C(u), C(v))}{g_1(C(u))}$$

Question: If we define the objective function for the clustering function C as

$$Q(C) = \max_{e,q} Q(C, e, q)$$

then show that

$$Q(C) = \sum_{c,c'} g(c, c') \log \frac{g(c, c')}{g_1(c) g_2(c')} + G$$

where G is a constant.

Question 2 (Follows Goldberg and Levy, 2014)

Assume we have some distribution $p(u, v)$ over word bigrams, and that $p_1(u)$ and $p_2(v)$ are the two marginal distributions:

$$p_1(u) = \sum_v p(u, v) \quad p_2(v) = \sum_u p(u, v)$$

Assume in addition that for each word w in the vocabulary, we have vectors θ'_w, θ_w in $\mathbb{R}^d$. We use Θ', Θ to denote the full matrices of embedding parameters. The objective function used to train Θ', Θ is then

$$L(\Theta', \Theta) = \sum_{u,v} \left[p(u, v) \log \frac{\exp\{\theta'_u \cdot \theta_v\}}{1 + \exp\{\theta'_u \cdot \theta_v\}} + K p_1(u) p_2(v) \log \frac{1}{1 + \exp\{\theta'_u \cdot \theta_v\}} \right]$$

Now assume that there is some setting for Θ such that for all u, v,

$$\theta'_u \cdot \theta_v = \log \frac{p(u, v)}{p_1(u) p_2(v)} - \log K$$

Assume in addition that for all u, v,

$$p(u, v) + K p_1(u) p_2(v) > 0$$

Question: Show that under the two assumptions above, if we define

$$\Theta'^*, \Theta^* = \arg\max L(\Theta', \Theta)$$

then for all u, v,

$$\theta'^*_u \cdot \theta^*_v = \log \frac{p(u, v)}{p_1(u) p_2(v)} - \log K$$

Hint: For any value of $q \in [0, 1]$, if we define

$$p^* = \arg \max_{p \in [0,1]} \left(q \log p + (1 - q) \log(1 - p) \right)$$

then $p^* = q$