Report

Zeljko Kraljevic

June 28, 2016

1 General

$$C = \sum_t \sum_a \sum_{b \in C_a^{(t)}} \left(\underbrace{log\sigma\left(x_b^{(t)} \cdot y_a^{(t)}\right)}_{u_p} + \sum_c \underbrace{log\sigma\left(-x_b^{(t)} \cdot y_c^{(t)}\right)}_{u_n} \right) \tag{1}$$

Setup when training:

- I normalize the time of the whole dataset to be between 0-10 for all datasets
- We don't have alterations currently, everything is trained all the time

2 Subsampling

I still subsample frequent words using $P(w_i) = 1 - \sqrt{\frac{1}{f(w_i)}}$. I also subsample documents in the following way:

- From a training set I always take a fixed number N of (a, b) pairs
- From every document I take a fixed number M of (a, b) pairs limiting the number of paris having the same target with K.
- When choosing a pair from document the closer the words in it are the higher the chance of it being choosen.
- Depending on N, M I calculate the probability of taking a document so that the whole dataset is always equaly present in the subsampled training set.

3 Clustering

NOTE: All of this aren't really probabilities because they are not in the range 0-1 and the sum is not one, maybe we should denote them differently.

The basic formula used for clustering is:

$$p(c \mid d) \approx \prod_{w_i \in d} p(c \mid w_i) f_c(t_i) \tag{2}$$

Which is changed into:

$$p(c \mid d) \approx \sum_{w_i \in d} log(p(c \mid w_i) f_c(t_i)) \tag{3}$$

for reasons of having a lot of words in the documents. Once this is calculated the document is clustered with:

$$doc\_cluster = \arg\max_{c \in C} p(c \mid d) \tag{4}$$

4 Time Prediction

Time prediction is similar to clustering except I don't use the time limiting function:

$$p(c \mid d) \approx \sum_{w_i \in d} log(p(c \mid w_i)) \tag{5}$$

Now because this values are always negative (and I don't know an other way) I do:

$$p'(c \mid d) \propto \frac{1}{\mid p(c \mid d) \mid} \tag{6}$$

Once I have this probability I do weight average to predict time:

$$predicted\_time = \frac{\sum_{c \in C} p'(c \mid d) t_c}{\sum_c p'(c \mid d)} \tag{7}$$

5 Finished Tests

Notes:

- Cap
- tau=0
- regularization

Results are in /develop/results/

Notes	Dataset	Iterations	clusters	Tau	Name
Without reg or cap	NIPS	500	300	1	normal
Without reg or cap	NIPS	500	300	0	normal_tau
Without reg, normalization	NIPS	500	300	1	normalization

6 Running Tests

Tests that are currently running, approximately it takes one day for a test to finish.

Notes	Dataset	Iterations	clusters	Tau	Folder
tau=0.01	NIPS	500	300	0	normal_tau_small
tau=0.01	Tweets	500	500	0	tweets_tau_small

7 TODO

- Try using aleterations, not so easy to implement