HW 2: CBOW and Word2Vec

Andrei A Simion

September 20, 2025

Negative Sampling for CBOW

In class we looked at the Skip-Gram and CBOW models and we looked at Negative Sampling. In the Skip-Gram model, we want to predict the the outside words from the center word. Negative Sampling removed the softmax dependency, which is expensive. The upshot is for a (w_c, w_o) pair we have

$$p(w_o|w_c) = \frac{\exp b_{w_o}^\intercal a_{w_c}}{\sum_{j=1}^{|\mathcal{V}|} \exp b_{w_j}^\intercal a_{w_c}}$$

and replace this by

$$p(w_o|w_c) = (\frac{1}{1 + \exp -b_{w_o}^\intercal a_{w_c}}) E_{w_k \sim p_{sample}(w)}[\prod_{k=1}^{K} \frac{1}{1 + \exp b_{w_k}^\intercal a_{w_c}}]$$

You can consider the expectation by: "Draw K random samples from the set V, with probability $p_{sample}(w)$". For CBOW, we want to predict the inner word from the words around it. Thus, if $m = 1$, for example, we have

$$p(w_c|w_{c-1}, w_{c+1}) = \frac{\exp b_{w_c}^\intercal a_{avg}}{\sum_{j=1}^{|\mathcal{V}|} \exp b_{w_j}^\intercal a_{avg}}$$

In this case, a_{avg} is the average a vector of the words w_{c-1}, w_{c+1}. The first goal is to submit what the objective for Negative Sampling would look like for CBOW. I.e., for the above example, what would it look like? Please submit a formula with justification. Your next goal is to take the notebook I give you and, using the hints and the notebook for Skip-Gram in class, implement the Negative Sampling Approach for CBOW. Can you print out the associated vectors for the validation words? Are they related, in turn, to each validation word.

Mathematical Problems

Problem 1 Let w be some word in the vocabulary $\mathcal{V}$ and let e_w be it's one-hot encoding (pretend the word is actually integer w, we might have $itos[w]$ = "cat" for example depending on how we set up the hash map between words and integers). Explain why $B^\intercal e_w = b_w \in \mathbb{R}^d$ and why this multiplication selects the w^{th} column of $B^\intercal$. Remember, if $B \in \mathbb{R}^{|\mathcal{V}| \times d}$ then $B^\intercal \in \mathbb{R}^{d \times |\mathcal{V}|}$.

Problem 2 Assume you do CBOW and Skip-Gram with negative sampling. Assume $m = 1$. Which method, on average, will get more training samples? Suppose there are 3 sentences with 7, 8, and 11 tokens. How many training sampling (positive training samples), will each method get. Draw a picture of a sentence with token counts and think about the number of samples each method gives. This is why Skip-Gram is used more often. It is more "sample efficient": you get more training data per Corpus.

Problem 3 In class we looked at the formula for the Skip-Gram for 1 sample (w_c, w_o) and got

$$\mathcal{L}(A, B) = -\log p(b_{w_o}|a_{w_c})) = -b_{w_o}^\intercal a_{w_c} + \log \sum_{w \in \mathcal{V}} \exp b_w^\intercal a_{w_c}$$

Then, we said that the gradients were as below. Prove this. Also, explain why $\frac{\partial \mathcal{L}}{\partial a_{w_c}}$ can be be interpreted as a difference between a hard guess and an expected value.

$$\frac{\partial \mathcal{L}}{\partial b_{w_o}} = -a_{w_c} + \frac{a_{w_c} \exp b_{w_o}^\intercal a_{w_c}}{\sum_{u \in \mathcal{V}} \exp b_u^\intercal a_{w_c}}$$

$$\frac{\partial \mathcal{L}}{\partial a_{w_c}} = -b_{w_o} + \sum_{w \in \mathcal{V}} b_w \frac{\exp b_w^\intercal a_{w_c}}{\sum_{u \in \mathcal{V}} \exp b_u^\intercal a_{w_c}}$$

Problem 4 Suppose we have a universe of words w and for each one we have a vector u. For a fixed word w, we'd like to find the word r such that $||u_w - u_r||^2$ is minimal. On the other hand, we can also find the word s such that $u_w^\intercal u_s$ is maximal. Are these r and s necessarily the same? What conditions on the vectors $\{u\}$ guarantee that these two problems are the same? The condition should be very clean and easy to explain.