Report homework 2

Simon Lietar, Pierre Cornilleau, Aymeric Behaegel

October 22, 2024

1 Preliminary questions

1. (a) For $c \in C_+$ we should maximize $\sigma(\mathbf{c} \cdot \mathbf{w})$ in order to minimize $-\log(\sigma(\mathbf{c} \cdot \mathbf{w}))$ and so minimize the loss.

   (b) On the contrary we should minimize $\sigma(\mathbf{c} \cdot \mathbf{w})$ in order to minimize the loss if $c \in C_-$.

   From a geometrical point of view, this induces a linear separation between the points of C^+ and those if C^-: indeed, maximize (resp. minimize) $\sigma(\mathbf{c} \cdot \mathbf{w})$ amounts to maximize (resp. minimize) $\mathbf{c} \cdot \mathbf{w}$, and if $\mathbf{c} \cdot \mathbf{w} > s^+$ (resp. $\mathbf{c} \cdot \mathbf{w} < s^-$) for $c \in C^+$ (resp. for $c \in C^-$), there is a linear separation between C_+ and C_- (provided that $s^+ \geqslant s^-$).

2. (a) The idea of contrastive learning, for classification, is to map the input into a space where the Euclidian distance (or any other simple distance) would be coherent with the semantic distance for classification: if two inputs are mapped nearby in this new space then they must belong to the same class. No (restrictive) assumption is made on the function that maps the input, so we also choose one that is robust to geometric distortions.

   (b) Here Y plays the role of the characteristic function $\mathbf{1}_{c \in C^+}$ in our equivalent formula.

   (c) The analog of E_W would be $\sigma(\mathbf{c} \cdot \mathbf{w})$. It is different because $\sigma(\mathbf{c} \cdot \mathbf{w})$ is a correlation score in the form of a probability whereas E_W is a distance.

   (d) The analogs of the functions L_G and L_I are, respectively, the functions $\sigma \mapsto -\log(\sigma)$ and $\sigma \mapsto -\log(1-\sigma)$.

2 Implementation and results

We used the Adam optimizer with a learning rate of 0.001 and a batch size of 1200. There is no good explanation for these choices other than they worked well in practice and with a satisfactory convergence speed. We used a loss with a sum reduction in order to give a larger weight to the negative samples as those are K times more numerous than positive samples. We compute validation accuracy by measuring the percentage of context words that are classified correctly, i.e. $c \cdot w > 0$ for positive and $c \cdot w < 0$ for negative context words, with the same procedure for generating c and w described as in the exercise questions.

Our results on `word2vec` training show that the loss converges well after a few epochs and that the accuracy increases to a value higher than 90%, showing that our model has correctly captured information in the embeddings.

For classification, we observe that word2vec embeddings, whether frozen or not, perform worse than the embeddings learned specifically for the classification task.

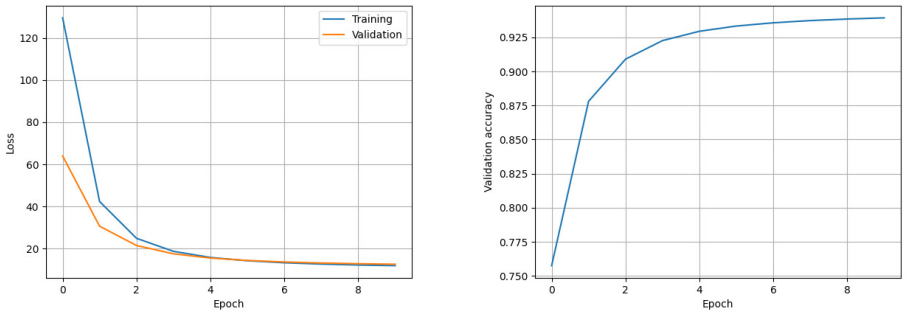

Figure 1: Word2Vec (sum) loss and accuracy.

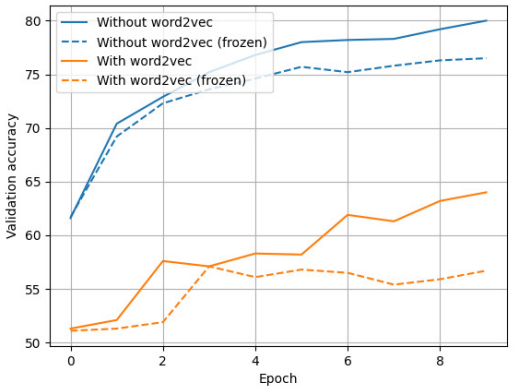

Figure 2: Classification accuracy

3 Ablation study

parameters	R = 3, K = 5	R = 3, K = 10	R = 6, K = 5	R = 6, K = 10
validation accuracy	56.6600	62.4600	58.3600	57.8400

For some reason, those parameters seem to perform poorly in regard to what we tested.

4 Conclusion

During this homework, we learned:

- how to preprocess text data, in complement to the previous homework;
- how to build a word2vec model and assess its performance;
- how to build a classification model based on embeddings, with and without pre-trained embeddings.