Ecole Nationale Supérieur Paris-Saclay
Master Mathématiques Vision Apprentissage

MP 2 - Deep Learning for Natural Language Processing

OREISTEIN Pierre

Cours: Deep Learning

January 2018

2 - Multilingual word embedding

Question: Using the orthogonality and the properties of the trace, prove that, for X and Y two matrices:

$$W = \underset{W \in O_d(\mathbb{R})}{\operatorname{argmin}} \|WX - Y\|_F = UV^T; \qquad \text{with } U\Sigma V^T = SVD(YX^T).$$

Proof.

$$\begin{aligned}
W^* &= \underset{W \in O_d(\mathbb{R})}{\operatorname{argmin}} \|WX - Y\|_F \\
&= \underset{W \in O_d(\mathbb{R})}{\operatorname{argmin}} \|WX - Y\|_F^2 \\
&= \underset{W \in O_d(\mathbb{R})}{\operatorname{argmin}} \langle WX - Y,\ WX - Y \rangle_F \\
&\quad \text{with } \langle \cdot, \cdot \rangle_F \text{ corresponding to the scalar product associated with the norm of Frobenius.} \\
&= \underset{W \in O_d(\mathbb{R})}{\operatorname{argmin}} \|WX\|_F^2 + \|Y\|_F^2 - 2\ \langle WX,\ Y \rangle_F \\
&\quad \text{we know that: } \|WX\|_F^2 = tr((WX)^T WX) = tr(X^T W^T WX) = tr(X^T X), \\
&= \underset{W \in O_d(\mathbb{R})}{\operatorname{argmin}} \|X\|_F^2 + \|Y\|_F^2 - 2\ \langle WX,\ Y \rangle_F \\
&= \underset{W \in O_d(\mathbb{R})}{\operatorname{argmax}} \langle WX,\ Y \rangle_F \\
&= \underset{W \in O_d(\mathbb{R})}{\operatorname{argmax}} tr((WX)^T Y) \\
&= \underset{W \in O_d(\mathbb{R})}{\operatorname{argmax}} tr((W)^T Y X^T) \\
&= \underset{W \in O_d(\mathbb{R})}{\operatorname{argmax}} \langle W,\ YX^T \rangle_F \\
&\quad \text{and we know that } SVD(YX^T) = U\Sigma V^T \\
&= \underset{W \in O_d(\mathbb{R})}{\operatorname{argmax}} \langle W,\ U\Sigma V^T \rangle_F \\
&= \underset{W \in O_d(\mathbb{R})}{\operatorname{argmax}} \langle UWV^T, \Sigma \rangle_F, \quad \text{by properties of the trace.} \\
&\quad \text{But we have that } UWV^T \text{ is also orthonormal because it is a product of orthonormal matrices, then.} \\
&= \underset{W \in O_d(\mathbb{R})}{\operatorname{argmax}} \sum Z_i \Sigma_i, \quad \text{with } Z = UWV^T. \\
&\leq \underbrace{\|Z\|_F}_{=1,\ \text{as Z is orthornormal}} \|Y\|_F, \quad \text{by inequality of Cauchy-Schwarz} \\
&\quad \text{Hence, the previous trace is maximal when }, Z = I, ie\ \ W* = U^T V
\end{aligned}$$

□

3 - Sentence classification with BoV

Question: What is your training and dev errors using either the average of word vectors or the weighted-average?

Below is summarised the errors (100% - accuracy in %) on the training set and the dev set for a model with the average vectors or the weighted-average vectors.

	Dev Set	Training Set
Average	66%	59%
Weighted-average	62%	57%

4 - Deep Learning models for classification

Question: Which loss did you use? Write the mathematical expression of the loss you used for the 5-class classification?

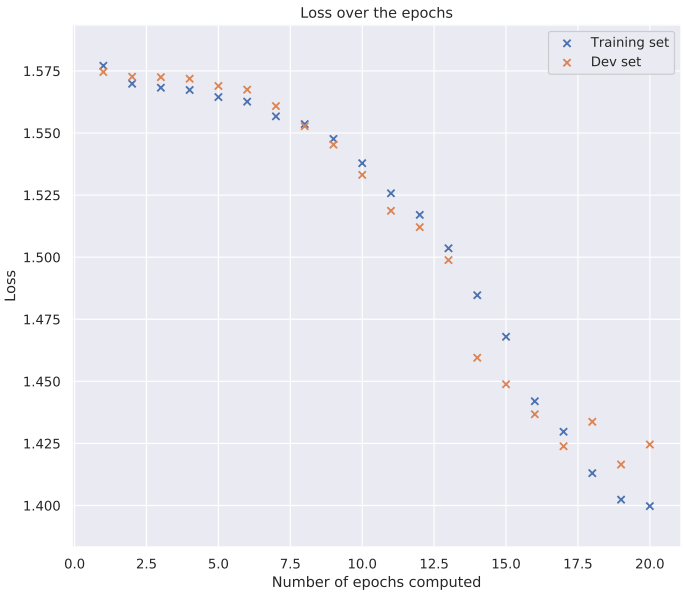
Here we face a problem of multiclass classification. So, we used the cross-entropy loss named "categorical_cross_entropy" in Keras.
This loss aims to minimise the empirical cross entropy:

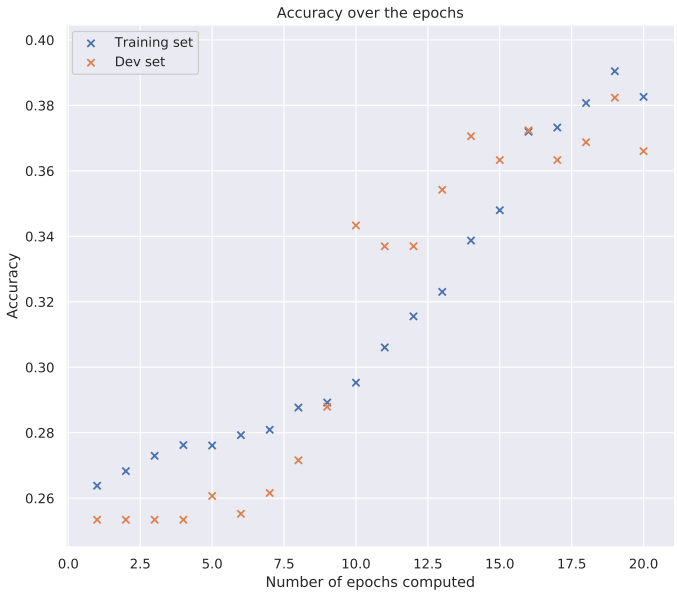
$$\sum_{i=1}^{N_{\text{train}}} \sum_{c=1}^{5} -\mathbb{1}_{\{y_i=c\}} * log(\hat{p}_{i,c})$$

With :

- N_{train}: The size of the training set.
- c: Represent a class.
- y_i: The true label of the sample i.
- $\hat{p}_{i,c}$: The probability predicted for the sample i for the class c.

Question: Plot the evolution of train/dev results w.r.t the number of epochs.





Question: Be creative: use another encoder. Make it work! What are your motivations for using this other model?

Here, we get our inspiration from the course slides and so we used an already existing implementation of an Attention-based Bi-LSTM neural network. The code can be found there:

`https://github.com/TobiasLee/Text-Classification`.

However, we did not succeed to improve the results.