école
normale
supérieure
paris–saclay

Master Mathématiques, Vision, Apprentissage
2018-2019

MP 2 - Deep Learning

Grégoire Boussac

1 Multilingual word embeddings

Question :

Minimizing $||WX - Y||_F$ is equivalent to minimizing $||WX - Y||_F^2$.

But :

$$||WX - Y||_F^2 = Tr((WX)^T WX) - 2Tr((WX)^T Y) + Tr(Y^T Y)$$
$$||WX - Y||_F^2 = Tr(X^T X) + Tr(Y^T Y) - 2Tr((WX)^T Y)$$

because $W^T W = Id$, because $W \in O_d(\mathbb{R})$.

The two first terms are constant. Thus, minimizing $||WX - Y||_F$ is equivalent to maximizing $Tr(X^T W^T Y)$.

Let's use the SVD of $Y^T X$:

$$Tr(X^T W^T Y) = Tr(W^T Y X^T) = Tr(W^T U \Sigma V^T) = Tr(\Sigma V^T W^T U)$$

But $U, V, W \in O_d(\mathbb{R})$, thus $P = V^T W^T U \in O_d(\mathbb{R})$,
and $\arg\min_{W \in O_d(\mathbb{R})} ||WX - Y||_F = \arg\max_{W \in O_d(\mathbb{R})}(Tr(\Sigma P))$.

$P^* = Id$ maximizes this last quantity, because Σ is a symetric matrix with positive coefficients, and P is orthogonal.

Thus, $V^T W^T U = P^* \implies W^* = UV^T$.
Hence the result.

2 Sentence classification with BoV

Question :

What we can see from our experiments is that using idf weighted-average degrades the performance of the logistic regression.

Accuracy	**Train**	**Test**
Average	0.489	0.436
IDF weighted-average	0.291	0.301

The performances are summarized in the table above.

3 4 - Deep Learning models for classification

Question :

I used the categorical cross-entropy which is well suited for multiclass classification. The formula of this loss is :

$$-\frac{1}{N}\sum_{i=1}^{N}\sum_{c=1}^{5}\nVdash_{y_i\in C_c} logp(y_i \in C_c)$$

Question :

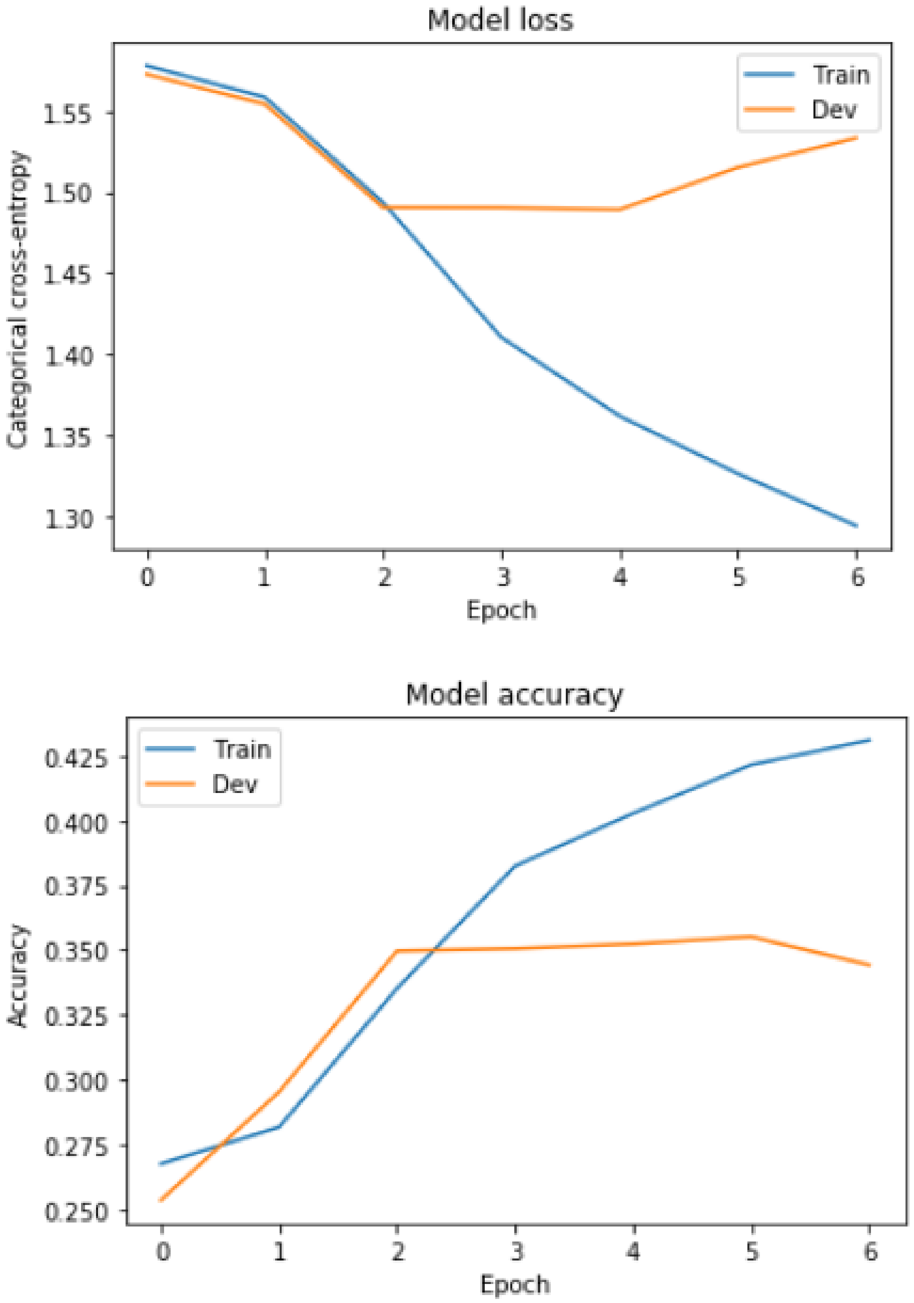

Figure 1 – Performances