Pierre Osselin
pierre.osselin@gmail.com

Project 1 : NLP

Deep Learning - Spring 2020
February 7, 2020

Question 1

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

$$W^* = \underset{W \in O_d(\mathbb{R})}{\arg\min} ||WX - Y||_F = UV^T, \text{ with } U\Sigma V^T = \text{SVD}(YX^T)$$

Answer :
We know that

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

It comes that minimizing $||WX - Y||^2$ according to $W \in O_d(\mathbb{R})$ is equivalent to maximizing $Tr(X^TW^TY)$ hence:

$$W^* = \underset{W \in O_d(\mathbb{R})}{\arg\max} Tr(X^TW^TY)$$

Now:

$$Tr(X^TW^TY) = Tr(W^TYX^T) = Tr(W^TU\Sigma V^T) = Tr(\underbrace{V^TW^TU}_{\in O_d(\mathbb{R})}\Sigma) \leq Tr(\Sigma)$$

The last inequality comes from the Von Neumann's trace inequality knowing that the singular values of an othogonal matrix are all equal to one, and that Σ is a diagonal atrix with positive elements. The upper bound is verified for $V^TW^TU = I \Leftrightarrow W = UV^T$.

Question 2

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

Answer :

	Average Word Vectors	Weighted Average Word Vectors
Train Accuracy	0.46746	0.46875
Dev Accuracy	0.41689	0.41871
Cmax	11.721	5.736

Table 1: Logistic Regression Results

The results are very similar, but a little bit better for the weighted average.

Question 3

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

Answer :
We used the categorical cross-entropy loss defined as follows:

$$\mathcal{L}(\dagger, \hat{\dagger}) = -\frac{1}{N}\sum_{i=1}^{N}\sum_{c=1}^{M} y_{i,c} \times log(p_{i,c}) \tag{1}$$

Where N is the size of the batch, M the number of classes, y a binary indicator of the class and p the probability prediction for the class by the model.

Question 4

Plot the evolution of the train/dev results with respect to the number of epochs.

Answer :

We trained our model on different values of the hyperparameter giving the dimension of the embedding. We obtained the following results:

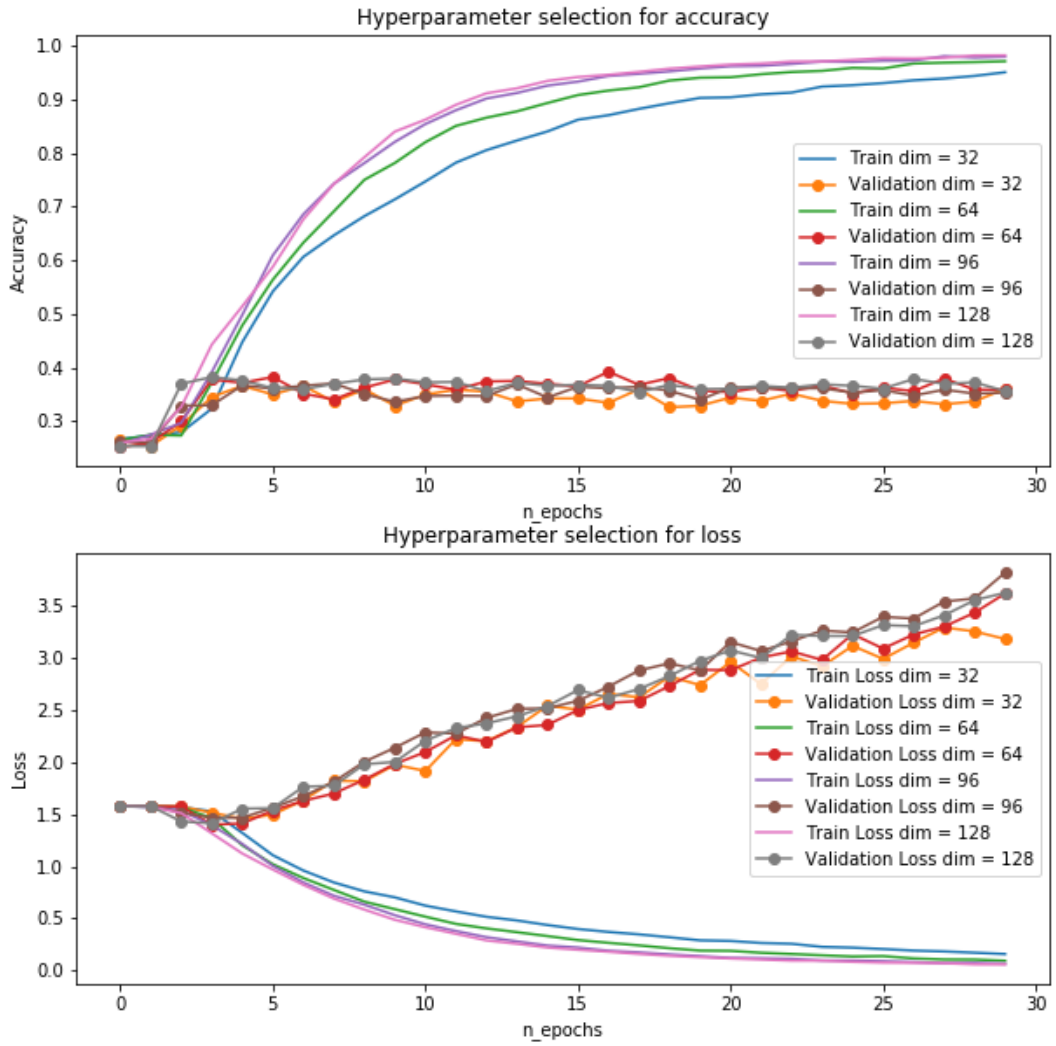

Figure 1: Results Hyperparameter Tuning

The model quickly overfit after 5 epochs in the range of parameters considered, and the validation accuracy does not significantly improve. Taking the dimension of the embedding to be 64 gives the best historical validation error of 39.2%.

Question 5

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

Answer :
In our last model we will use a convolutional network for sentence classification, as described in [1], taking similar hyperparameters. The main rationale behind the use of such models is that they use the concept of a "convolution", a sliding window or "filter" that passes over the texts, identifying important features and analyzing them one at a time, then reducing them down to their essential characteristics, and repeating the process. They also showv strong empirical results accross many different datasets. In our application we managed to achieve slightly above 40% validation accuracy, which is slightly better that our previous model.

References

[1] Ye Zhang and Byron Wallace. "A sensitivity analysis of (and practitioners' guide to) convolutional neural networks for sentence classification". In: *arXiv preprint arXiv:1510.03820* (2015).