Deep Learning for NLP - Project

Adil Rhoulam

arhoulam@ens-paris-saclay.fr

1) Using the orthogonality and the properties of the trace, prove that, for X and Y two matrices: $W^* = argmin_{W \in O_d(R)} ||WX - Y||_F = UV^T$, with $U\Sigma V^T =$ SVD(YX^T)

Answer: We have : $||WX - Y||_F^2 = ||WX||_F^2 + ||Y||_F^2$ - 2 $< WX, Y >_F$

Since $W \in O_d(R)$, we get : $||WX||_F^2 = Tra(X^T W^T WX) = Tra(X^T X) = ||X||_F^2$

then : $argmin_{W \in O_d(R)} ||WX - Y||_F = argmax_{W \in O_d(R)} < WX, Y >_F$

and : $< WX, Y >_F = Tra(X^T W^T Y) = Tra(YX^T W^T)$

Using the SVD of YX^T : $YX^T = U\Sigma V^T$, with $U, V \in O_d(R)$ and Σ is a diagonal matrix with non-negative real numbers on the diagonal

We get : $< WX, Y >_F = Tra(U\Sigma V^T W^T) = Tra(\Sigma V^T W^T U)$

Since : $U^T WV \in O_d(R)$ we get : $argmax_{W \in O_d(R)} Tra(\Sigma V^T W^T U) = argmax_{W \in O_d(R)} Tra(\Sigma W^T)$

Since Σ is diagonal : $[\Sigma W]_{i,i} = \sum_{k=1}^{p} [\Sigma]_{i,k} [W]_{k,i} = [\Sigma]_{i,i} [W]_{i,i}$

which gives : $Tra(\Sigma W^T) = \sum_{i=1}^{p} [\Sigma]_{i,i} [W]_{i,i}$

Since $W \in O_d(R)$, we have : $\forall j, \sum_{i=1}^{p} [W]_{i,j}^2 = 1$ which gives $-1 \leq [W]_{i,j} \leq 1 \quad \forall i, j$

We conclude that : $argmax_{W \in O_d(R)} < WX, Y >_F \leq \sum_{i=1}^{p} [\Sigma]_{i,i} = Tra(\Sigma)$.

This maximum is attainable when $W = UV^T$, since $< UV^T X, Y >_F = Tra(Y^T UV^T X) = Tra(XY^T UV^T) = Tra(V\Sigma U^T UV^T) = Tra(\Sigma)$

Finally, $UV^T = argmax_{W \in O_d(R)} < WX, Y >_F = argmin_{W \in O_d(R)} ||WX - Y||_F$
with $U\Sigma V^T =$ SVD(YX^T)

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

Answer: Using the average of word vectors and the Logistic Regression classifier, we get an accuracy of:

- On train set : 47.94 %
- On dev set : 43.6 %

Using the weighted-average of word vectors and the Logistic Regression classifier, we get an accuracy of:

- On train set : 48.68 %
- On dev set : 42.51 %

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

The loss that I used is the 'categorical_crossentropy'. Its mathematical expression for the 5-class classification is:
$L(\hat{y}, y) = -\frac{1}{5}\sum_{i=1}^{5}(y_i \log \hat{y}_i + (1-y_i)\log(1-\hat{y}_i))$