Deep Learning for Natural Language Processing : Project
Adrien Benamira

Instructor:
TA: Abdulkadir Çelikkanat

January 10, 2019

1 Monolingual embeddings

cf code

2 Multilingual word embeddings

Let's demonstrate the orthogonal procrustes Theorem :

The Frobenius norm can be defined as:

$$\|A\|_{\mathrm{F}} = \sqrt{\sum_{i=1}^{m}\sum_{j=1}^{n}|a_{ij}|^2} = \sqrt{\operatorname{trace}(A^*A)} = \sqrt{\sum_{i=1}^{\min\{m,n\}}\sigma_i^2(A)} \tag{1}$$

and because we have that minimisising a norm an the same norm square is the same problem, we get thank to :

$$\|A+B\|_{\mathrm{F}}^2 = \|A\|_{\mathrm{F}}^2 + \|B\|_{\mathrm{F}}^2 + 2\langle A, B\rangle_{\mathrm{F}} \tag{2}$$

That :

$$argmin_{WW^T=I}(\|WX-Y\|_{\mathrm{F}}^2) = argmin_{WW^T=I}(\|WX\|_{\mathrm{F}}^2 + \|Y\|_{\mathrm{F}}^2 - 2\langle WX, Y\rangle_{\mathrm{F}})$$

We use the orthogonal nature of W (that is, $W^{\mathsf{T}}W = WW^{\mathsf{T}} = \mathbf{I}$) and the cyclic nature of the trace ($\operatorname{trace}(XYZ) = \operatorname{trace}(ZXY)$) to show :

$$\|WX\|_{\mathrm{F}}^2 = \operatorname{trace}\left(X^{\mathsf{T}}W^{\mathsf{T}}WX\right) = \operatorname{trace}\left(X^{\mathsf{T}}X\right) = \|X\|_{\mathrm{F}}^2, \tag{3}$$

and because $\|X\|_{\mathrm{F}}^2$ and $\|Y\|_{\mathrm{F}}^2$ don't depend on W, we have :

$$argmin_{WW^T=I}(\|WX-Y\|_{\mathrm{F}}^2) = argmax_{WW^T=I}(\langle WX, Y\rangle_{\mathrm{F}}) \tag{4}$$

Using the definition of the scalair product and the cyclic propriety of the trace, we have :

$$\langle WX, Y\rangle_{\mathrm{F}} = \operatorname{trace}\left(Y^{\mathsf{T}}WX\right) = \operatorname{trace}\left(WXY^{\mathsf{T}}\right) \tag{5}$$

We have : $U\Sigma V^T = \mathsf{SVD}(YX^T)$
and so :

$$\langle WX, Y\rangle_{\mathrm{F}} = \mathrm{trace}\left(WV\Sigma U^T\right) = \mathrm{trace}\left(U^T WV\Sigma\right) = \mathrm{trace}\left(Z\Sigma\right) \tag{6}$$

with $Z = U^T WV$. We have Z orthogonal and because Σ is a diagonal matrix,

$$\mathrm{trace}\left(Z\Sigma\right) = \sum \sigma_{i,i} z_{i,i} \tag{7}$$

So minimizing$W$ for $(\|WX - Y\|_{\mathrm{F}}^2)$ is equivalent to maximizing Z for $\sum \sigma_{i,i} z_{i,i}$.

And because Z is orthogonal, the norm of every row is 1 and so the maximum for $z_{i,i}$ is 1.

$$Z = U^T WV = I_n \tag{8}$$

and Finally :

$$W^* = UV^T \tag{9}$$

3 Sentence classification with BoV

3.1 Question

Accuracy on training and validation set for C = 0.1

	Word vector	Weigth average	Word vector normalized	Weigth average normalized
Trainning	0.418	0.445	0.469	0.458
Validation	0.3906	0.391	0.405	0.395

4 Deep Learning models for classification

4.1 Question

$$\frac{-1}{N} \sum_{i}^{N} \sum_{c}^{C} \mathcal{I}_{y_i \in C_c} \log\left(y_{predict}^{(i)}\right) \tag{10}$$

The double sum is over the observations i, whose number is N, and the categories (classes) c, whose number is C

4.2 Question

cf code

4.3 Question

Reférence : Convolutional Neural Networks for Sentence Classification, Yoon Kim