1 Multilingual word embeddings

Let X, Y be two matrices in $\mathbb{R}^{d \times m}$ where m is the size of the vocabulary. We seek to minimize

$$\begin{aligned} &\min_W \|WX - Y\|_F^2 \\ &\text{s.t. } W^T W = I_d \end{aligned} \tag{1}$$

Since $\|WX\|_F = \|X\|_F$ for any $W \in O_d(\mathbb{R})$, this is equivalent to

$$\begin{aligned} &\max_W \langle WX, Y \rangle_F = \langle W, YX^T \rangle_F \\ &\text{s.t. } W^T W = I_d \end{aligned}$$

Writing the SVD decomposition $YX^T = U\Sigma V^T$ with U, V orthogonal and $\Sigma \geq 0$ diagonal, we have $\langle W, YX^T \rangle_F = \langle U^T W V, \Sigma \rangle_F$. The matrix $W' = U^T W V$ is also orthogonal, so by the Cauchy-Schwarz inequality

$$\langle W, YX^T \rangle = \langle W', \Sigma \rangle \leq \operatorname{Tr} \Sigma$$

with equality iff $W' = I_d$ i.e. $W = UV^T$.

2 Sentence classification with BoW

I trained the logistic regression model, with the regularization factor C varying in $\{0.01, 0.02, 0.05, 0.1, 0.2, 0.3, 0.5\}$.

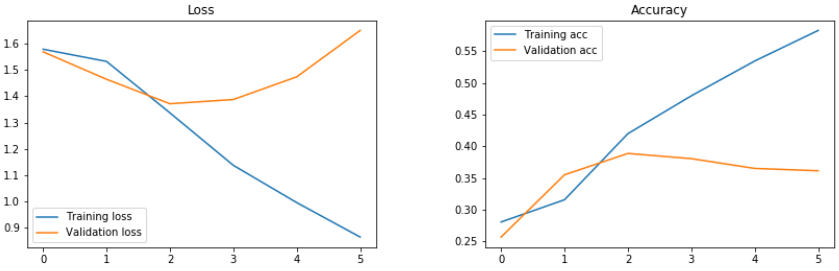
Without IDF: best parameter is $C = 0.02$, train accuracy is 0.471 and dev accuracy is 0.414.

With IDF: best parameter is $C = 0.2$, train accuracy is 0.457 and dev accuracy is 0.392.

3 Deep learning models for classification

I used the categorical cross-entropy loss

$$L(y, \hat{y}) = -\sum_{i=1}^{n} y_i \log(\hat{y}_i) \tag{2}$$


Loss
Training loss
Validation loss
1.6
1.5
1.4
1.3
1.2
1.1
1.0
0.9
0
1
2
3
4
5
Accuracy
Training acc
Validation acc
0.55
0.50
0.45
0.40
0.35
0.30
0.25
0
1
2
3
4
5