Document distances using optimal transportation

Pierre Stock

ENS Cachan

pierre.stock@ens-paris-saclay.fr

January 9, 2017

Presentation of the problem

- Compute a fast and reliable similarity measure between documents
- **Applications**: Sentiment analysis, song identification, multilingual document matching
- **Common approach**: BoW representations
- **Issue**: BoW representations are often almost **orthogonal** between documents (words have synonyms)
- **Solution**: Optimal transportation distance parametrized by a **ground metric** (e.g. word2vec).
- **Issue**: Computationally intensive, scales in $\mathcal{O}(d^3 \log(d))$.

Proposed method

- Regularize the transportation problem using an entropic term to get policy $\pi^\star$

$$T_\lambda(p,q) = \min_{\pi \in \Pi(p,q)} \langle \pi, C \rangle - \lambda E(\pi) \tag{P}$$

- Recast this problem using the Kullback-Leibler divergence

$$T_\lambda(p,q) = \lambda \min_{\pi \in \Pi(p,q)} \mathrm{KL}(\pi|\xi) = \lambda P^{\mathrm{KL}}_{\Pi(p,q)}(\xi) \tag{P'}$$

  where $\xi = e^{-C/\lambda}$

- Solve the transportation problem using iterative Bergman projections on

$$S_1 = \{\pi \in \mathbf{R}_+^{n\times n} \mid \pi \mathbf{1}_n = p\} \quad \text{and} \quad S_2 = \{\pi \in \mathbf{R}_+^{n\times n} \mid \pi^T \mathbf{1}_n = q\}$$

  and by noticing that $\Pi(p,q) = S_1 \cap S_2$.

Proposed method

- Closed forms of the projections on those sets

$$P_{S_1}^{\mathrm{KL}}(\pi) = \mathrm{diag}\left(p \oslash (\pi \mathbf{1}_n)\right)\pi \quad \text{and} \quad P_{S_2}^{\mathrm{KL}}(\pi) = \pi \mathrm{diag}\left(q \oslash \left(\pi^T \mathbf{1}_n\right)\right)$$

- Distance then computed as

$$d(p, q) = \langle \pi^\star, C \rangle$$

Algorithm 1 Sinkhorn's algorithm

```
input p, q, lambda, C, niter
xi = exp(−C ∗ C / lambda)
b = ones(1, size(p))
for i = 1..niter do
    a = p / (xi ∗ b)
    b = q / (xi.T ∗ a)
end for
pi = diag(a) ∗ xi ∗ diag(b)
d = sum(C ∗ pi)
return d
```

Main contributions

- GPU-ready implementation of Sinkhorn's algorithm to comptute document distances in Python + Tensorflow, available online including preprocessing the documents (tokenization, stop words)
- iPython notebook containing word2vec demonstrations (similarities, algebric relationships between words, PCA), influence of λ and the regularization parameter on a toy example
- Experiments on Reuters dataset ran on AWS GPUK80 (after some setup)

Numerical findings

Toy example on 4 NYT articles (2 sports, 2 politics)

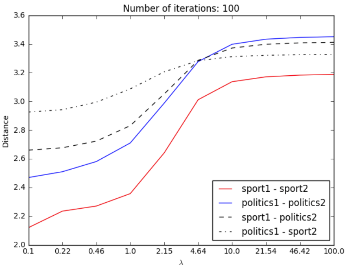

Figure: Influence of λ

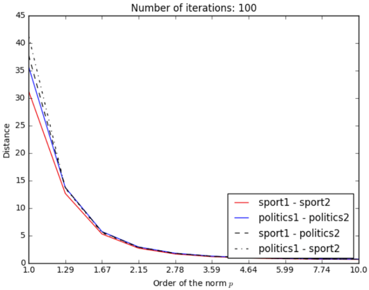

Figure: Influence of p

Reuters dataset, 6,000 (2000) train (test) samples labelled to 51 categories, k-NN error for different values of k

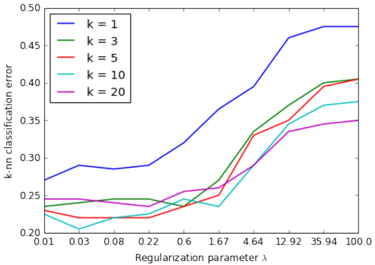

Figure: Influence of λ

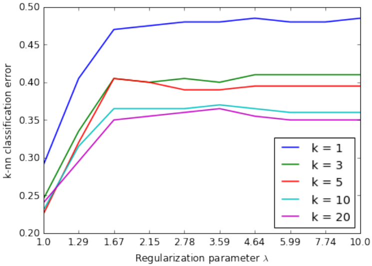

Figure: Influence of p

Conclusion & Perspectives

Entropic regularization of optimal transportation problems presents appealing properties

- **Theoretical properties**: convexity, unicity, smooth optimal policy
- **Practical properties**: Bergman iteration scheme, stability

Main conclusions

- Influence of λ: has to be tuned carefully
- Influence of p: small values of p for which the norm is still convex are preferable

Limitations

- Sinkhorn's algorithm becomes **unstable** when $\lambda \to 0$ or $p \to 0$ ($e^{-C/\lambda}$)
- BoW features lose the **ordering** of the words

Possible extensions

- Improve the **fixed dictionary**
- Try **cosine** similarity for the cost matrix