Flowing gradient through SVD

Bc. Vladimír Macko

RNDr. Radim Řehůřek, Ph.D

February 14, 2018

Overview

- Introduction to the problematics
- Problem outline
- Our work
- Results
- Plans

Document classification

This was a terrible movie

- create representation for words
- create representation for document
- predict

Word vectors

Local representation

One hot encoding

Distributed representation

Count based

Factorization of co-occurrence matrix
LSA (SVD)

Prediction based

Trained neural network
Skip gram

Count vs. prediction

Prediction

- extremely popular
- huge performance gains
- less memory demanding

Count

- less hyperparameters
- easier to “train”
- teoreticaly based

Count vs prediction

Glove vectors as explicit factorization

- Neural word embedding as implicit matrix factorization [Levy and Goldberg, 2014]

Hyperparameters matter

- Improving distributional similarity with lessons learned from word embeddings [Levy et al., 2015]

Does not work well on small datasets

- Comparative study of LSA vs Word2vec embeddings in small corpora [Altszyler et al., 2016]

LSA problems

- Sensitive to preprocessing
- Sensitive to weights
- Unsupervised and can forget things

Current solutions

- Preprocessing
- Weight - Mutual information [Wu et al., 2017], [Deng et al., 2014]
- Supervised weights: TF-KLD [Ji and Eisenstein, 2013], [Lan et al., 2009]

Our system

Baseline

Co-occurrence matrix, rescale weight, factorization, prediction
Training the predictor

Our

Co-occurrence matrix, rescale weight, factorization, prediction
Compute gradient with respect to the weights

LSA used in similar manner in [Ionescu et al., 2015]

Gradient descent

- Co-occurrence matrix M
- Weight vector t
- SVD: $U\Sigma V^T$
- Simple classifier: $\sigma(x\theta + b)$

- Reweighted matrix $M \circ t$
- SVD decomposition $U\Sigma V^T$
- Compute embedding $x = d \circ tU$
- Train classifier $\hat{y} = \sigma(x\theta + b)$
- Compute error $E = \frac{1}{2}(\hat{y} - y)^2$
- Compute derivation $\frac{\partial E}{\partial t} = (\hat{y} - y)\sigma(\hat{y})(1 - \sigma(\hat{y}))\Theta U$
- Update weights: $t = t - \alpha\frac{\partial E}{\partial t}$

Evaluation

Datasets from SentEval [Conneau et al., 2017]

- Custumer review dataset
- Movie review
- Subjective vs objective
- Opinion polarity

SVD + logistic regression

Figure 1: Precision of baseline on CR dataset for multiple tries

TFIDF + SVD + logistic regression

Figure 2: Precision of baseline on CR dataset for multiple tries

SVD + LR + gradient

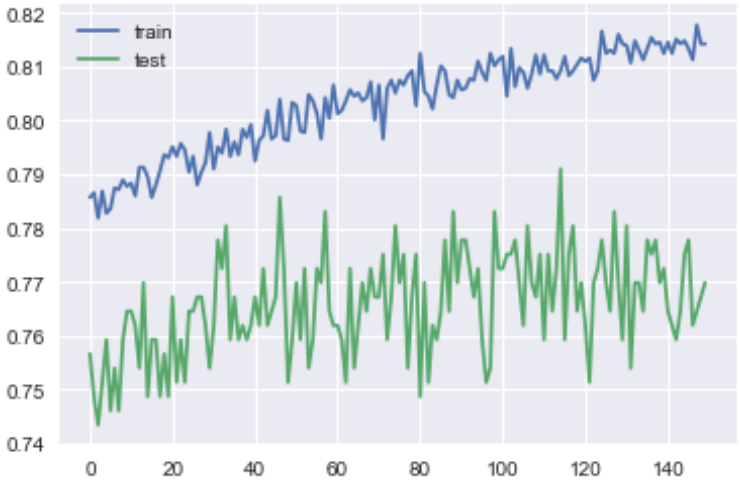

Figure 3: Precision of weight improving on CR dataset for multiple epochs

TFIDF + SVD + LR + gradient

Figure 4: Precision of tfidf weight improving on CR dataset for multiple epochs

SVD + LR + gradient

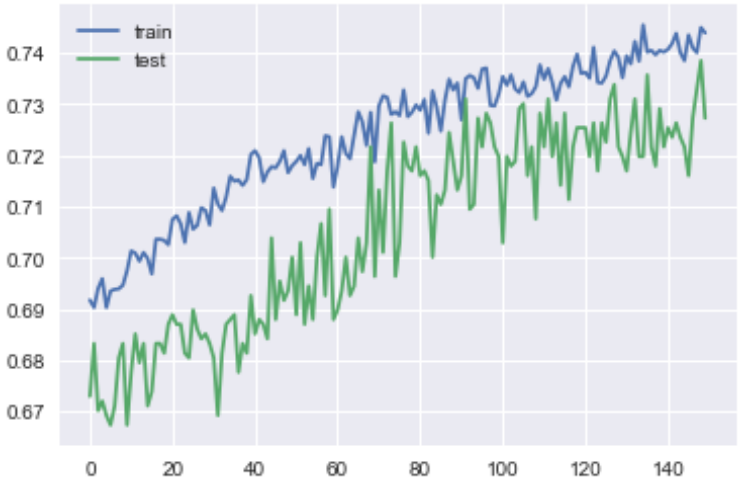

Figure 5: Precision of weight improving on MR dataset for multiple epochs

TFIDF + SVD + LR + gradient

Figure 6: Precision of tfidf weight improving on MR dataset for multiple epochs

SVD + LR + gradient

Figure 7: Precision of weight improving on MPQA dataset for multiple epochs

TFIDF + SVD + LR + gradient

Figure 8: Precision of tfidf weight improving on MPQA dataset for multiple epochs

SVD + LR + gradient

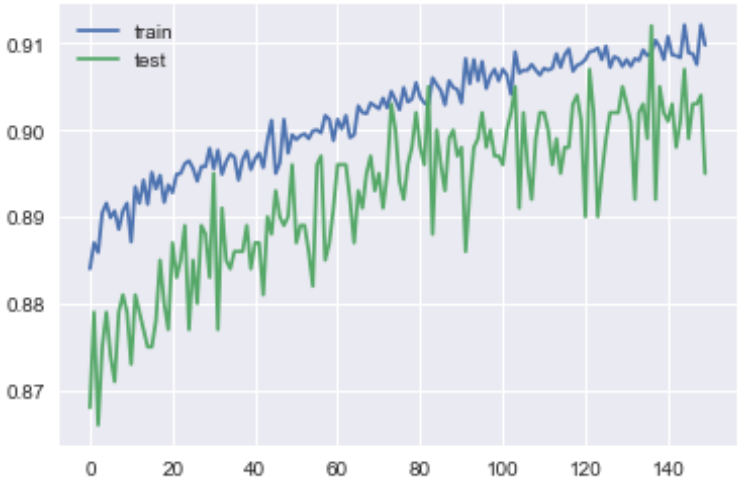

Figure 9: Precision of weight improving on SUBJ dataset for multiple epochs

TFIDF + SVD + LR + gradient

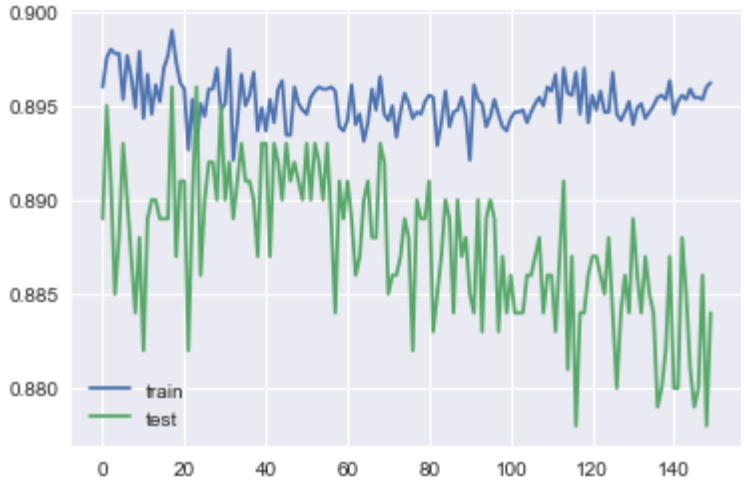

Figure 10: Precision of tfidf weight improving on SUBJ dataset for multiple epochs

Plans

- Proper exploration of results
- Extend to bigrams
- Try transfer learning
- Try to extract the formula
- Try more complicated classifiers
- Try stochastic gradient [Brand, 2006]

Thank you for your attention

Literature I

[Altszyler et al., 2016] Altszyler, E., Sigman, M., Ribeiro, S., and Slezak, D. F. (2016).

Comparative study of lsa vs word2vec embeddings in small corpora: a case study in dreams database.

arXiv preprint arXiv:1610.01520.

[Brand, 2006] Brand, M. (2006).

Fast low-rank modifications of the thin singular value decomposition.

Linear algebra and its applications, 415(1):20–30.

[Conneau et al., 2017] Conneau, A., Kiela, D., Schwenk, H., Barrault, L., and Bordes, A. (2017).

Supervised learning of universal sentence representations from natural language inference data.

arXiv preprint arXiv:1705.02364.

Literature II

[Deng et al., 2014] Deng, Z.-H., Luo, K.-H., and Yu, H.-L. (2014).
A study of supervised term weighting scheme for sentiment analysis.
Expert Systems with Applications, 41(7):3506–3513.

[Ionescu et al., 2015] Ionescu, C., Vantzos, O., and Sminchisescu, C. (2015).
Training deep networks with structured layers by matrix backpropagation.
arXiv preprint arXiv:1509.07838.

[Ji and Eisenstein, 2013] Ji, Y. and Eisenstein, J. (2013).
Discriminative improvements to distributional sentence similarity.
In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 891–896.

Literature III

[Lan et al., 2009] Lan, M., Tan, C. L., Su, J., and Lu, Y. (2009).

Supervised and traditional term weighting methods for automatic text categorization.

IEEE transactions on pattern analysis and machine intelligence, 31(4):721–735.

[Levy and Goldberg, 2014] Levy, O. and Goldberg, Y. (2014).

Neural word embedding as implicit matrix factorization.

In *Advances in neural information processing systems*, pages 2177–2185.

[Levy et al., 2015] Levy, O., Goldberg, Y., and Dagan, I. (2015).

Improving distributional similarity with lessons learned from word embeddings.

Transactions of the Association for Computational Linguistics, 3:211–225.

Literature IV

[Maas et al., 2011] Maas, A. L., Daly, R. E., Pham, P. T., Huang, D., Ng, A. Y., and Potts, C. (2011).

Learning word vectors for sentiment analysis.

In *Proceedings of the 49th annual meeting of the association for computational linguistics: Human language technologies-volume 1*, pages 142–150. Association for Computational Linguistics.

[Pennington et al., 2014] Pennington, J., Socher, R., and Manning, C. (2014).

Glove: Global vectors for word representation.

In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543.

[Wu et al., 2017] Wu, H., Gu, X., and Gu, Y. (2017).

Balancing between over-weighting and under-weighting in supervised term weighting.

Information Processing & Management, 53(2):547–557.