CSIC5011 Mini Project 1

A Dive Into NIPS Words

GU Hanlin, HUANG Yifei, SUN Jiaze
Department of Mathematics HKUST

Introduction & Objectives

One of the main tasks in Natural Language Processing is to embed words as vectors, so that they can be manipulated mathematically. Using the NIPS data set, we aim to accomplish the following objectives:

A. Visualize word embeddings by various reduction and manifold learning techniques

B. Improve the results by first classifying words and documents into different *topics*

C. Explore the relations between these *topics*, and predict future topic trends

NIPS Words Data Set

- Term-document matrix X of size 11463×5804
- Word Frequencies from 1987-2015 NIPS Conference papers
- Rows represent words, columns represent documents
- Sparse (around 90% entries are 0)

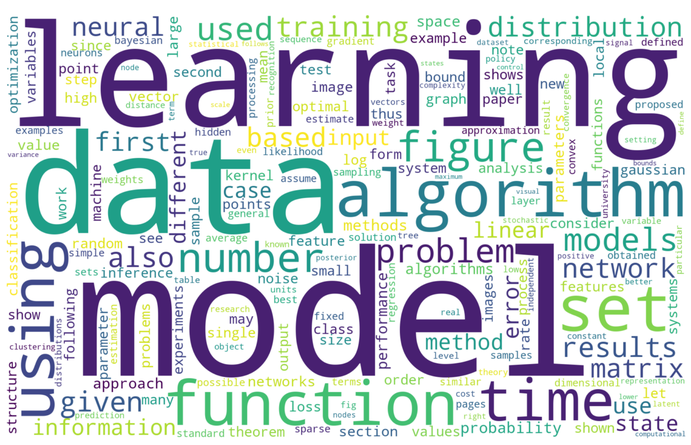

Figure 1. A word cloud generated based on the frequency of appearances

Methodology

I. Visualize the original data X, using 8 reduction methods: PCA, MDS, ISOMAP, LLE, Modified LLE, Hessian LLE, Laplacian Eigen Map, and LTSA

II. Improve the results in I. by first applying TF-IDF on X

III. Classify the data into *topics* using Latent Dirichlet Analysis (LDA)

IV. Fit a multivariable regression on the *topics* obtained from III., as well as ARIMA for time series analysis

Reduction & Visualization

- In Figure 2, we can see various manifold learning methods do not seem to show much variation in the data, while PCA and MDS perform relatively better.

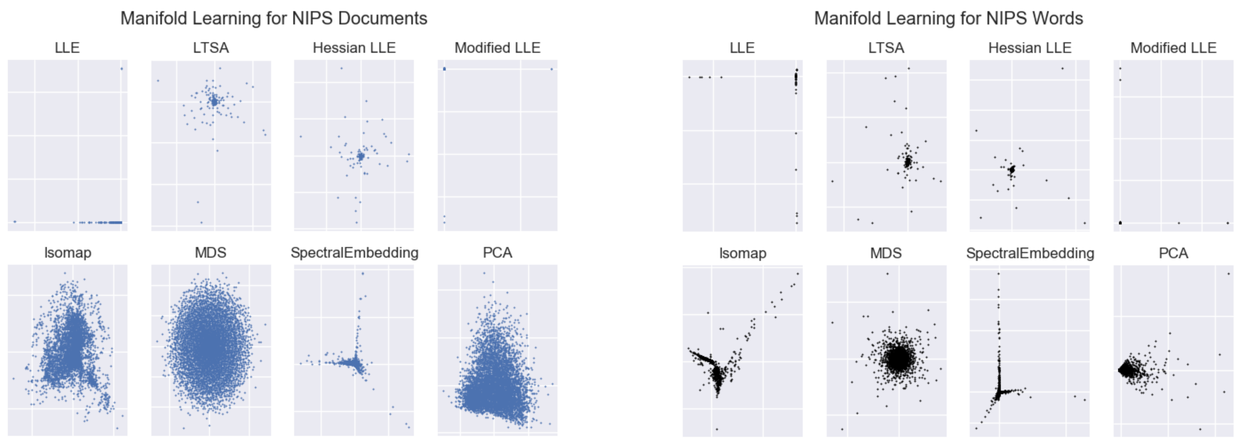

Figure 2. Visualization of document (left) and word (right) embeddings in 2D, using the original data set.

- The TF-IDF transform assigns each entry of X a weight inversely proportional to the number of documents in which the corresponding word appears.

Figure 3. Visualization of document (left) and word (right) embeddings in 2D, after the TF-IDF transform.

- By classifying the data into 5 *topics*, LDA outputs two matrices of size 11463×5 and 5×5804, corresponding to words and documents respectively.

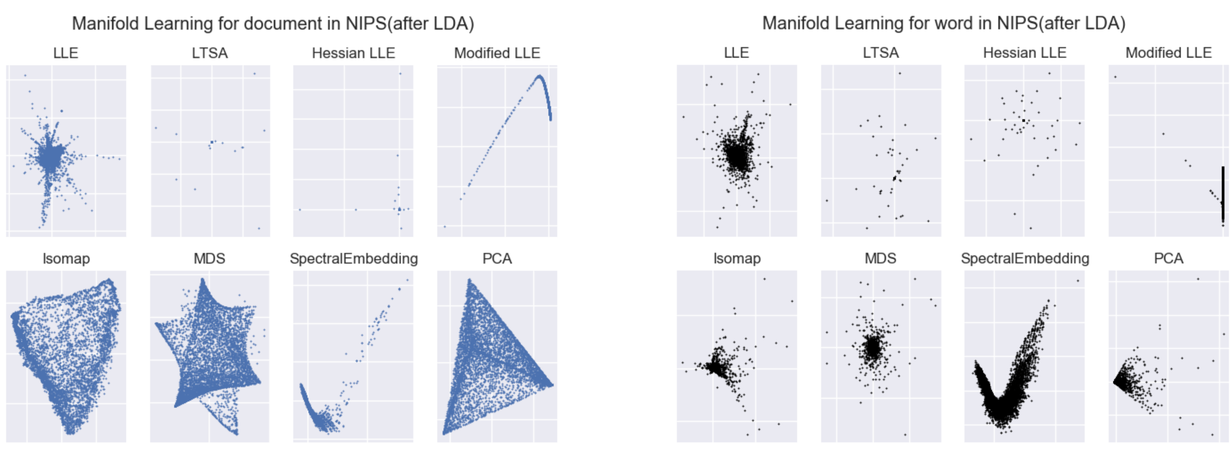

Figure 4. Visualization of document (left) and word (right) embeddings in 2D, after LDA.

Linear Regression On *Topics*

As a result of LDA, words follow multinomial distributions under each *topic*, thus we can name the 5 *topics* by their top words: ***Algorithm***, ***Learning***, ***Model***, ***Network***, and ***Image***. We take the document matrix produced by LDA, and sum up the columns by year, obtaining a 5×29 matrix, whose transpose is used as the design matrix. The fitted parameters (slopes only) are summarized below.

		Covariates (*Topics*)				
		Algorithm	*Learning*	*Model*	*Network*	*Image*
Dependent Variables (*Topics*)	*Algorithm*		1.3308	0.9638		
	Learning	0.3439				
	Model	0.5053				0.3710
	Network	-0.1702				-0.9015
	Image		-0.32446	0.2801	-0.1898	

Forecasting Future *Topic* Trends

Using the same matrix for regression, we also performed time series analysis for *topics*. Using the unit root test (ADF test), we observed the series was not stationary. After estimating two parameters: p, the level truncation of ACF, and q, the level trailing of PACF, we were able to fit an ARIMA model as shown in Figure 5. The Ljung–Box test shows that the residue of the forecast is white noise, suggesting that the model is valid.

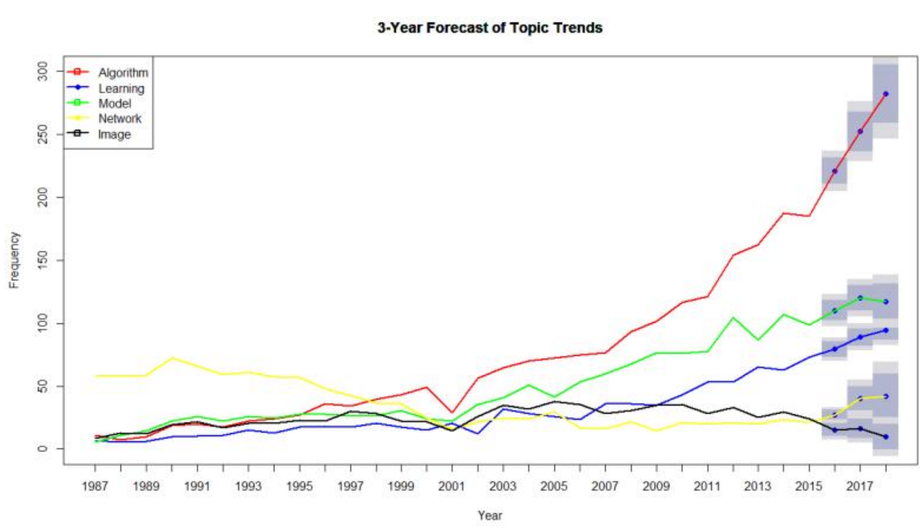

Figure 5. *Topic* trends from 1987 to 2018. The trend for 2016-2018 are predicted, with 80% and 95% confidence limits.

Analysis & Conclusion

- Results from manifold learning on the original data matrix show that these methods do not effectively capture the data pattern (even after applying the TF-IDF transform), possibly due to that the original data do not lie in a low dimensional manifold.
- Reduction and manifold learning after LDA show drastic improvements over the previous results, especially for documents. MDS and PCA produced strikingly regular patterns with faces and edges, suggesting that documents might indeed be combinations of *topics*.
- Regression on *topics* shows that correlations do exist, statistically speaking. For example, ***Algorithm*** is positively influenced by ***Learning*** and ***Model***, whereas ***Image*** is negatively affected by ***Learning*** and ***Network***.
- ARIMA predicts that ***Algorithm*** and ***Learning*** will continue their current increasing trend, whereas ***Image*** will gradually decline. An interesting observation comes from ***Network***, which has been in decline since 1990, but is projected to increase in the future. This surprisingly coincides with the recent popularity of artificial neural network and deep learning.

Individual Contribution

- GU Hanlin: Regression and Time Series Forecast
- HUANG Yifei: LDA, its analysis and visualization
- SUN Jiaze: Visualization of original data and TF-IDF

References

- D. JURAFSKY AND J. H. MARTIN, Speech and language processing: An introduction to natural language processing, computational linguistics, and speech recognition (third edition draft). https://web.stanford.edu/~jurafsky/slp3/ed3book.pdf, 2017.
- V. PERRONE, P. A. JENKINS, D. SPANO, AND Y. W. TEH, Poisson random fields for dy- namic feature models. https://archive.ics.uci.edu/ml/datasets/NIPS+Conference+ Papers+1987-2015, 2016.
- Y. YAO, *A mathematical introduction to data science*. http://math.stanford.edu/~yuany/ course/reference/book_datasci.pdf, 2017.