MATH 6380J Mini-Project 1: Realization of Recent Trends in Machine Learning Community in Recent Years by Pattern Mining of NIPS Words

Chan Lok Chun lcchanac@ust.hk
Department of Life Science, HKUST

1. Introduction

In this study, analysis has been done on a collection of words appear in journal articles published on NIPS from 1987 to 2015. Sparse PCA and linear regression on the frequency distribution of the words being used, are employed to identify the recent trends in machine learning-related research fields.

2. Hypothesis

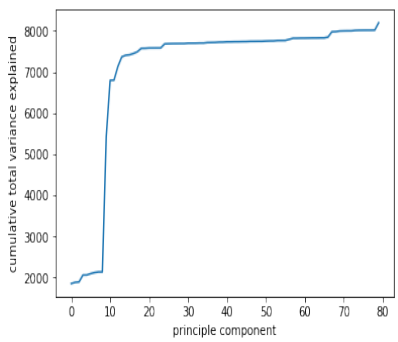
Judging on the fact that deep learning has been thriving in recent decade, it is expected that the frequency of appearance of words related to this field should have a notable increasing trend.

3. NIPS word dataset

The NIPS word dataset contains the word count of 11463 words in 5812 journal articles collected from 1987-2015.

Preprocessing:
In the analysis, the trend of words' appearance frequencies through time is interested. Therefore, average is taken over the dataset by year to obtain the relative frequency count of each word in each year. The resultant dataset becomes 11463 (words) X 29 (years).

Clustering:
Instead of analyzing 11463 words at the same time, it is believed that many words are correlated (e.g. jargons from the same research area), doing clustering could reveal some hidden structure of the dataset.
In this case, sparse PCA was chosen for its robustness in p>>n scenario.



By calculating the adjusted total variance [1] explained by the PCs, it is identified that the top 30 sparse PCs explains majority of the variations in the dataset. The top 30 PCs were then selected for further analysis

Fig.1 Top 80 PCs for explained total variance

4. Analysis

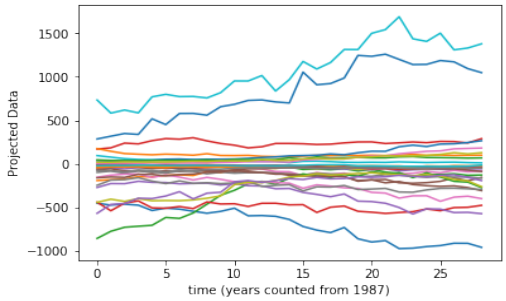


Fig. 2 Projected data of the top 30 PCs against time

To identify the recent rising trends, the PCs which could explain largely the variations in recent years are wanted. In order to do this, the data are first projected to each PC respectively, plotted against time (Fig.2) then perform linear regression to extract the slopes. The aforementioned desired PCs were identified by looking for those which give non-zero slope (tested at 99.9% significance level). After the screening, 1035 words out of 11463 are recovered from the extracted sparse PCs.

Rounds of linear regression were then performed on the relative frequencies of those 1035 words with respect to the 29 time points (1987-2015). Top 50 words with greatest slopes were selected out and seen as the top 50 increasingly frequent words used in NIPS articles in the recent years.

The list of words were then analyzed manually to check which subfield of machine learning community do they belong to.

6. Conclusion

The "top 50 rising words" were analyzed, to our surprise, there is no key words strongly related to "deep learning" appears in the list. While there are up to 36% of them belongs to "Bayesian statistics" and 12% belongs to "optimization". However, when plotting the time series data of certain deep learning-related words, they do show obvious increment in frequency with time. While it is further observed that all top-30 sparse PCs do not include these words. It is conjectured that the dataset may have certain complicated underlying structure which cannot be clustered effectively by linear projection method like sparse PCA, such that the signals of certain important features are masked in some low PCs.

7. References

[1] Zou, H., Hastie, T., & Tibshirani, R. (2006). Sparse Principal Component Analysis. Journal of Computational and Graphical Statistics, 15(2), 265-286. doi:10.1198/106186006x113430