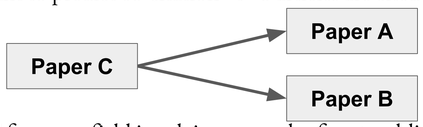
Making Research Easier: PubMed Co-Citations

Andrew Guan and Trevor Tsue
CS 229 - Machine Learning, Stanford University

Motivation

- Given that researchers uses an article in a paper, what are other useful articles?
- Current PubMed "similar articles" features depend only on NLP techniques of paper abstracts.
- We want to predict co-citation - two articles are cited by a third.



- Useful for every field involving research of past publications (i.e. nearly all disciplines)

Data

- **PubMed**
  - Abstracts, date, authors, and titles
  - Implemented string parsers to read data
- **PubMed Developer API**
  - Retrospective Citation Information
  - HTTP request, urlib, and XML ElementTree
- **Google Scholar API**
  - Cited by information for publications

Features

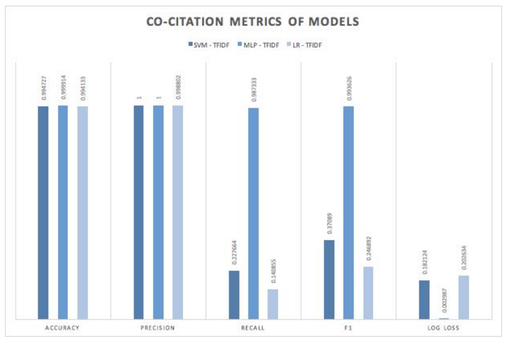
- **Vectorized Abstracts**
  - Word-stemming via Porter2
  - Count, TF-IDF Vectors
- **Feature Vectors** - 2 Papers
  - Cosine, L_2, Jaccard Distances of Abstracts
  - Number shared authors and citations
  - Publication date difference Feature Template

Results

	Training	Test
Logistic Regression	0.9940	0.9941
Support Vector Machine	0.9945	0.9947
Neural Net	0.9999	0.9999
Number of Samples	2,026,935	868,686

- **Metrics**
  - tp = true positive , fp = false positive, fn = false negative

$$Precision = \frac{tp}{tp+fp}$$
$$Recall = \frac{tp}{tp+fn}$$
$$F1 = 2 * \frac{precision * recall}{precision + recall}$$
$$logloss = -logP(yt|yp) = -(ytlog(yp) + (1-yt)log(1-yp))$$



Models

- **Input** - Feature vector comparing 2 papers
- **Output** - Co-cited (1) or not (0)
- **Logistic Regression**
  - Linear model logistic classifier with L_2 penalization
- **Support Vector Machine**
- **Neural Net**
  - Multilayer Perceptron (MLP) with 100 layers

Discussion

- All three models performed well in terms of accuracy and precision. This is likely because the data is very sparse, i.e. there are very few positives within the dataset (~0.7%). Especially by simply examining metadata about shared authors, LR and SVM could basically almost always output 0, produce few false positives with high accuracy.
- However, we are ultimately interested in minimizing false negatives rather than false positives, so Recall is what we are more interested in. The neural net model performed very well on this metric in contrast with the linear regression and SVM models. However, neural nets are prone to overfitting; we plan to investigate how neural nets perform across different fields.

Future Directions

- How well do the models we've trained work across different fields?
- What are some other models that we can try fitting to our data?
- How can we improve the preprocessing of the data by using a better stemming algorithm or using more features?

References

- Turney, Peter D. and Patrick Pantel. "From Frequency to Meaning: Vector Space Models of Semantics." Journal of Artificial Intelligence Research, vol. 37, 2010, 141-188.
- Ramos, Juan. "Using TF-IDF to Determine Word Relevance in Document Queries." 2003.
- Lin, Jimmy and W John Wilbur. "PubMed related articles: a probabilistic topic-based model for content similarity." BMC Bioinformatics, 8:423, 2007. doi:10.1186/1471-2105-8-423

Contact Andrew Guan - guana@stanford.edu, Trevor Tsue - ttsue@stanford.edu