Extractive Summarization of EHRs with Naive Bayes

Jeremy Goldwasser, Irene Li and Dragomir Radev
Department of Computer Science, Yale University, New Haven, CT

LILY Lab

Introduction

Lengthy, repetitive Electronic Health Records (EHRs) are cited as a leading factor in physician burnout. Moreover, time spent reading unnecessarily long EHRs is time that could be spent with patients, an inefficiency that wastes both time and money. We propose a method of generating summarizations of EHRs that extract the sentences with only the most salient information. We fit Naive Bayes models on medical research papers, using article bodies as the main text and abstracts as their summarizations. Applied to documents like EHRs, these models classify a predetermined fraction of the sentences as part of the summary. While their efficacy is somewhat underwhelming, the steps taken in building them may prove useful in other contexts.

Materials and Methods

We used EHRs from the MIMIC-III (Medical Information Mart for Intensive Care III) corpus. To access them, we completed a bioethics training course. In addition, we trained the model with medical research papers from the Pubmed Central Corpus. This corpus, associated with the U.S. National Library of Medicine and the National Institute of Health, is a public dataset of millions of articles.

To preprocesses the Pubmed papers, we first parsed their abstracts and bodies into plaintext files. We then created "merged" files that contained both the abstract and body.

To fit the Naive Bayes model, we created a feature vector for every sentence in each conjoined article file. (Here, we used sentences from 16,000 articles.) Four features were taken into account: average word frequency, average inverse sentence frequency, number of nouns, and length. Each sentence's class label was determined by whether or not it was also present in the conjoined abstract file. Once these vectors had been built, a Gaussian Naive Bayes model could be quickly fit.

The bodies of each paper are much longer than their abstracts (almost 20 times longer, on average). However, because most EHRs are relatively short, their summaries will contain a larger proportion of the text. To account for this discrepancy, we fit a second Naive Bayes model using only the "Introduction" section of each article body, utilizing 24,000 articles. This was also an imperfect solution: the introductions often focus on background material and are not accurately summarized by the abstract.

	Training Accuracy	Test Accuracy	Expected Accuracy from Random Guessing	Accuracy on other test dataset
Trained with Just Body Intro	70.0%	71.8%	64.9%	90.0%
Trained with Entire Body	91.0%	90.9%	89.9%	70.0%

Table 1. binary classification results on PubMed articles. The model predicts whether or not each sentence is part of the abstract (the summary).

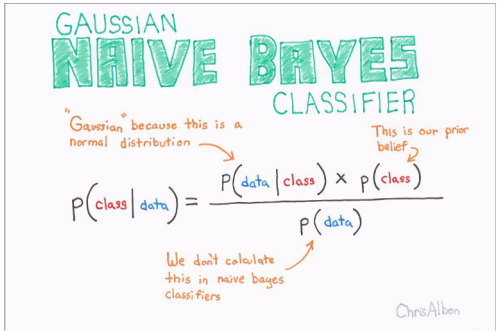

Figure 1. Naive Bayes Overview. Predictors are assumed to be independent, so P(data | class) can be expressed as the product of all the P (feature | class) terms

Results

Table 1 shows the training and test accuracy for each of Naive Bayes model. As we can see, each model produces results that are better than random guessing, but not much better. We calculate the percentage of sentences that are correctly classified as part of the abstract or the body.

The model using entire article bodies has higher accuracy not because it is better, but because with longer bodies, a lower proportion of its sentences are part of the summary. Testing the models on each other's data, we see that the "just intro" model gets only 0.1% above random results, whereas the "entire body" model performs almost as well as the "just intro" model. This suggests that it is a more versatile model.

As we can see, neither model produces great results. Unsurprisingly, when we generated summaries on the EHRs, they generally picked out random sentences.

Conclusio

In this work, we fit a Naive Bayes model to perform extractive summarization on PubMed research papers, then applied it to EHRs. The results are better than random, but not phenomenal. However, we will continue to use PubMed for summarization, as it is a large and clean corpus in the medical domain.

Moving forward, we could improve this Naive Bayes model by taking more features into account. In addition to number of nouns, for example, we could use number of verbs or adjectives. We could also try a BERT-based summarization model, training on one article at a time rather than conjoining all of them.

Acknowledgements

I am tremendously grateful to Irene Li, Dragomir Radev, and the entire LILY Lab for their assistance and direction.