Zipf's Law and Relational Learning: An Investigation of a Surprising Correlation

Leo Behe and Zachary Wheeler
Mentors: Dr. William Pottenger and Dr. Christie Nelson

Introduction

- What we are covering:
- Background (Leo)
  - Relational learning
  - Character n-grams
  - Prior research, hypotheses, methodology
- Research details and goals (Zach)
  - Bayesian classification, microtext
  - Zipf’s law
  - Evaluation of previous results

Supervised machine learning

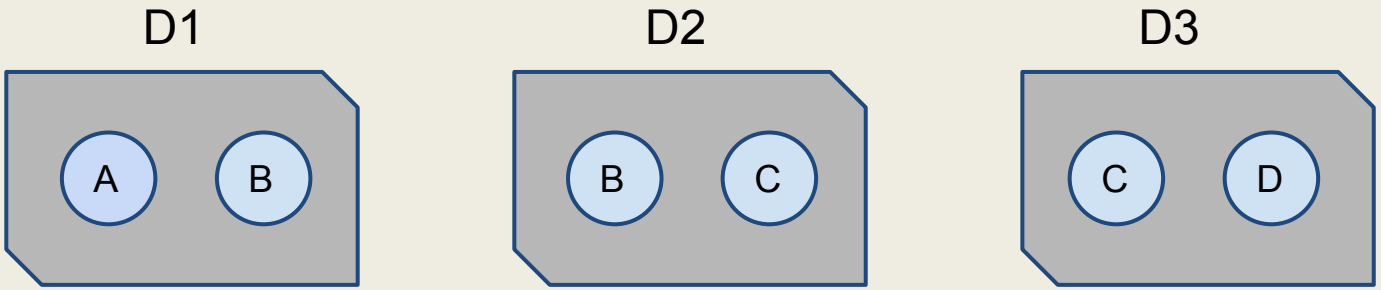
- Train, then use
- “Labeled” training set

Background: Relational Learning

- Traditional machine learning makes the assumption that data is independent and identically distributed (IID)
  - This is simpler and requires less spacetime complexity
- However, real-world data is *not* IID- instances within a feature space are heterogeneous and depend on other features
- Relational learning works by leveraging higher-order relations between instances in a feature space
  - Does not make the IID assumption

Relational Learning- Textual Example

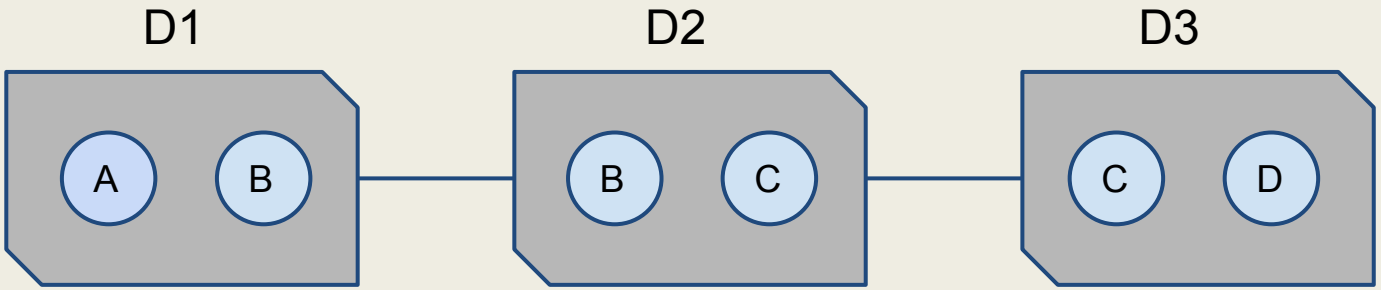
Traditional machine learning:



Note: We will be using documents and words as examples of instances and features in this presentation.

Relational Learning- Example

Relational machine learning:



- Relational learning algorithms have been shown to outperform equivalent IID learning algorithms in experiments, especially with small training sets.

Bayesian Classification

- Popular family of algorithms for problems such as document classification
  - eg spam filters
- Based on Bayes' Rule in statistics
- "Naive" Bayes ignores word order, assumes IID
  - Exploits predictive power of word frequencies
- Higher-Order Naive Bayes (HONB)
  - Exploits word pair co-occurrence chains

Microtext

- Text messages, twitter posts, etc
- Classification is a difficult problem
  - Misspellings, short documents
- HONB performance often beats Naive Bayes

Character N-Grams: An interesting feature space

methinks it is like a weasel → me|th|in|ks|_i|t_|is|_l|ik|e_|a_|we|as|el

- Data is noisy (typos, grammatical errors, scan errors)
- Character n-grams reduce the impact of noise
  - N can be any number; n = 2, shown above, produces “bigrams”
  - Different values of n can increase or decrease effectiveness of learning algorithms
- N-grams can also overlap, increasing the number of features for a given document or dataset

Classification using character n-grams

- A high rate of accuracy (roughly 80%) has been demonstrated in classification of text using n-grams
- Character n-grams are more tolerant of errors in data
  - When whole words are used, a character error renders an entire word useless- the damage is reduced when multiple n-grams replace the word
- Heavily degraded physical documents digitized by OCR (Optical Character Recognition) were improved by character n-grams
  - 19% improvement in Character Error Recognition

Field of interest: using higher-order algorithms for prediction

- Creating ground truth- i.e., labeling datasets for prediction is laborious to do manually
- Learning algorithms using relational data reduce the amount of training data required
  - However, HONB is more computationally intensive during training, but has the same complexity during prediction
- So… When should algorithms leveraging relational data be used?

Prior research

- In 2013, research was conducted which compared a well-known learning algorithm, Naive Bayes, with a relational learning model, Higher-Order Naive Bayes (HONB).
  - This research used a microtext dataset with character n-grams of various values of n.
- The results suggested a possible correlation between the Zipfian distribution of the feature set and the performance of HONB.

Zipf's Law

- Frequency of a word is inversely proportional to its rank
- Linearity on a log-log plot of frequency to rank
- Property of most natural languages
- Corresponds to the Zipfian distribution in statistics

Prior Results

- Bigrams and trigrams not Zipfian

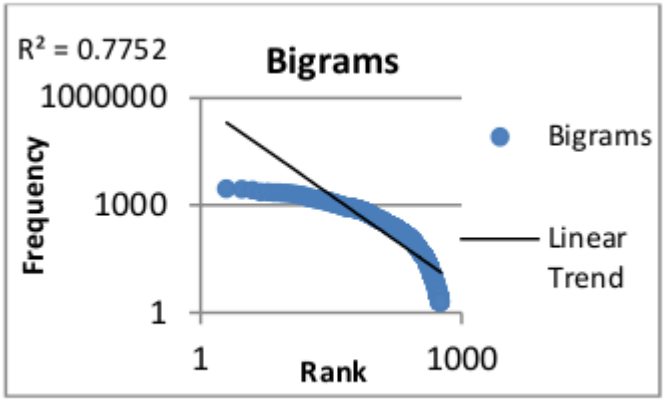


Figure 1: Non-Zipfian Character Bigrams

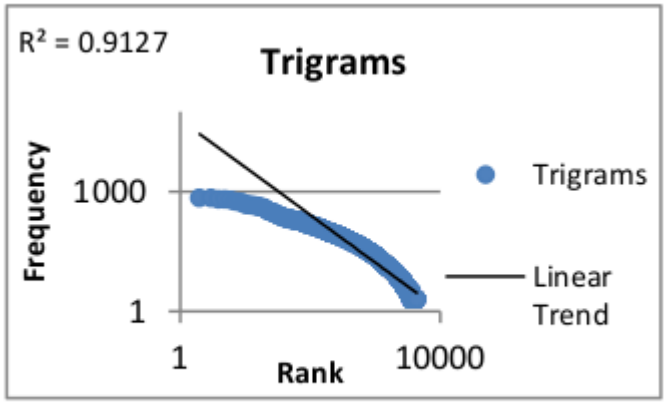


Figure 2: Non-Zipfian Character Trigrams

Prior Results

- 4-grams, 5-grams were Zipfian

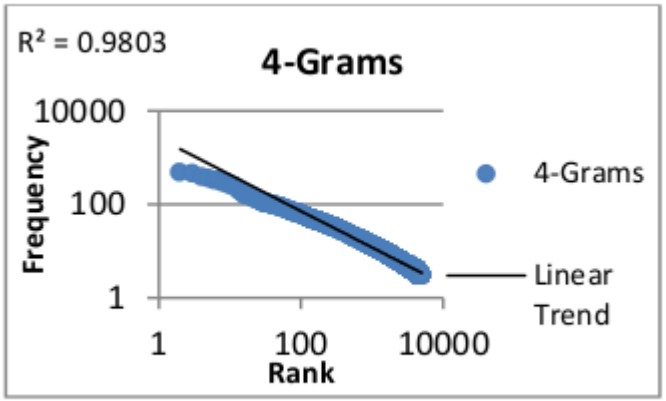


Figure 3: Zipfian Character 4-Grams

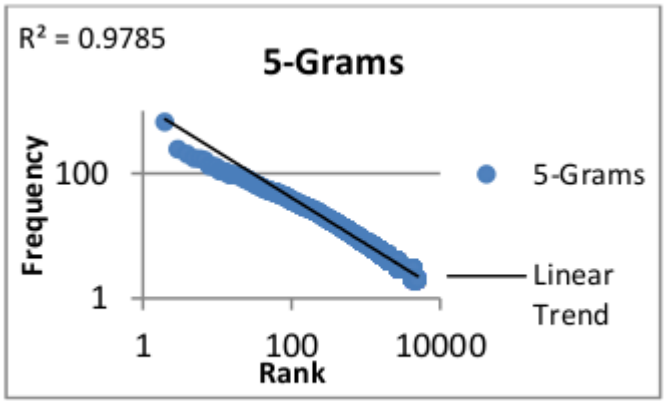


Figure 4: Zipfian Character 5-Grams

Prior Results

- 6-grams were Zipfian, but 7-grams were not

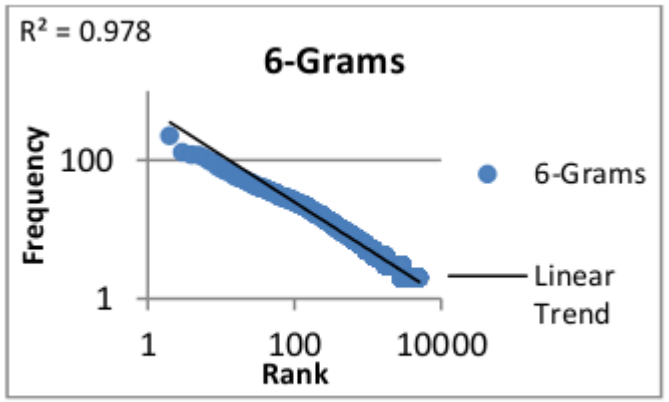


Figure 5: Zipfian Character 6-Grams

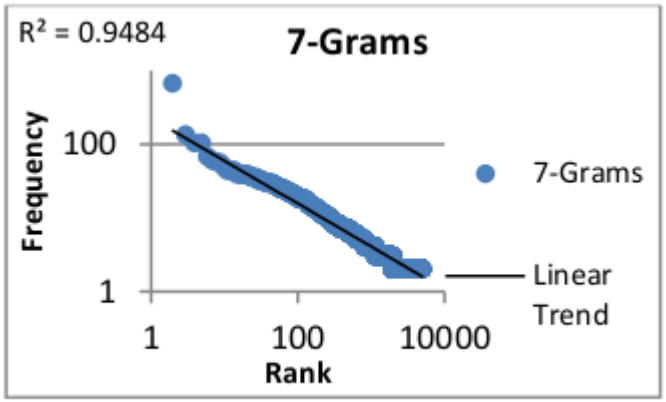


Figure 6: Non-Zipfian Character 7-Grams

Prior Results

- HONB performance seems correlated with Zipf's law
  - Performed best on 4-grams and 5-grams
- Insight into when relational features are valuable?

Literature Search

- *Ganiz, M.C.; George, C.; Pottenger, W.M.* - **Higher Order Naive Bayes: A Novel Non-IID Approach to Text Classification.** (2011, IEEE Transactions on Knowledge and Data Engineering)
- *Dutta, S.; Sankaran, N.; Sankar K., P.; Jawahar, C.V.* - **Robust Recognition Of Degraded Documents Using Character N-Grams.** (2012, 10th IAPR International Workshop on Document Analysis Systems)
- *Unpublished research by Knopp, B.D.; Nelson, C.; Pottenger, W.M.* - **Modeling Microtext with Character n-grams and Higher Order Learning.**
- *Cavnar, W.B.; Trenkle, J.M.* - **N-Gram-Based Text Categorization.** (1994, 3rd Annual Symposium on Document Analysis and Information Retrieval)

Hypotheses

- There is a correlation between the mean square error measurement of how Zipfian a dataset's feature distribution is and the classification performance of HONB.
- The effectiveness of HONB operating on a dataset composed of character n-gram features shows a statistically significant improvement as the character n-gram distribution approaches a Zipfian distribution.

Experimental Design

- Confirm prior results pointing to a correlation between Zipfian distribution and higher-order learning performance.
- Run HONB against additional datasets of varying Zipfian distribution.
- Compare the performance of HONB to the Zipfian distributions of the datasets.
- Formulate a theoretical explanation for the correlation between Zipfian distribution and HONB performance.

Datasets

- We will use the Haitian Earthquake Social Media Dataset to confirm previous results
- Other textual datasets such as the Twenty Newsgroups dataset will be used to move beyond microtext
- We will move on to datasets without n-grams, and possibly numeric datasets as well
  - Aim to determine how broad the Zipfian correlation is

In Summary- Moving Forward

- Confirm, disprove, or refine correlation
- Find a theoretical explanation
- Create simple test to predict value of relational features

Any questions?

Thanks for listening!