HIGHER ORDER LEARNING FOR CLASSIFICATION IN EMERGENCY RESPONSE

Hannah Keiler
Mentors: William M. Pottenger and Christie Nelson
DIMACS REU 2012
Final Presentation

Problem/Purpose

- **Problem** – Classifying data by hand is time consuming and potentially subjective. It is especially problematic in a state of emergency.

- **Purpose** – A potential discovery that topic modeling with Higher Order Latent Dirichlet Allocation (HO-LDA) outperforms LDA in providing a model used to classify data that informs emergency responders. Additionally, we wish to determine the optimal number of topics used in HO-LDA/LDA to provide a model – a novel area in topic modeling applications to emergency response.

Latent Dirichlet Allocation: Overview

- LDA is a **topic model**, which is a statistical model for discovering the unobserved topics that explain why documents are similar
- The observed data are the words of each document
  - Given a collection of documents, the posterior distribution of the hidden variables given the words determines the decomposition of the hidden topics of the collection.
- Example:

Topic1
Yarn
Milk
Whiskers

Topic2
Fetch
Hound
Puppy

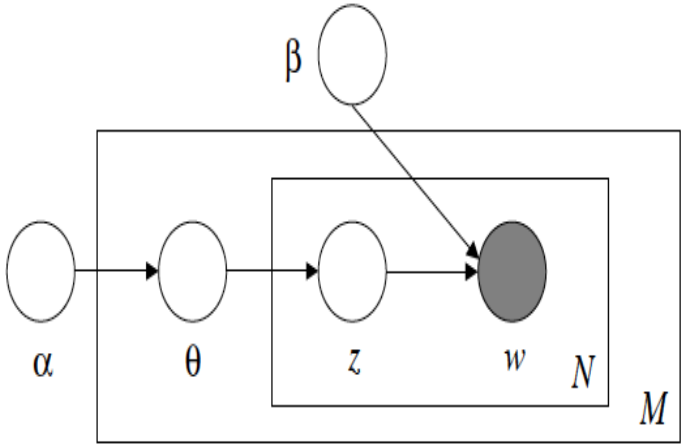
Topic3
Bowl
Water

Document
Today I played fetch using a ball of yarn with my new puppy.

Blei, D. M., Ng, A. Y., Jordan, M. I. (2003)

Latent Dirichlet Allocation

- To generate a corpus of documents using the Dirichlet Model:
  - Choose θ ~ Dirichlet(α)
  - For each of the N words w_n:
    - Choose a topic z_n ~ Multinomial(θ)
    - Choose a word w_n from $p(w_n \mid z_n; \beta)$, a multinomial probability conditioned on the topic z_n and parameterized by the topic distributions β



Latent Dirichlet Allocation: Inference

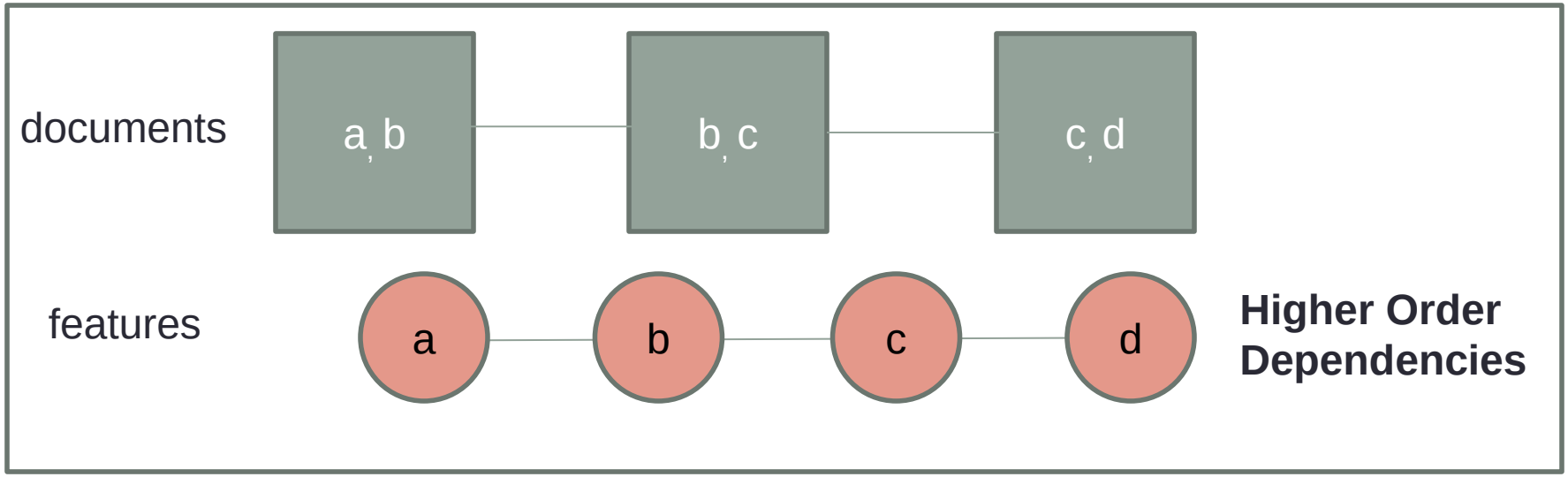
- In order to perform inference using a Gibbs Sampling algorithm, a standard algorithm, the conditional probability of occurrence of a topic for a word in the corpus (given the topic labels of the words) is used:

$$P(z_i = j | \mathbf{z}_{-i}, \mathbf{w}) \propto \frac{n_{-i,j}^{(w_i)} + \beta}{n_{-i,j}^{(\cdot)} + W\beta} \frac{n^{(d_i)} + \alpha}{n_{-i,\cdot}^{(d_i)} + T\alpha}.$$

  - Used for sampling the topic for the term *w* at position *i*. The term $n_{-i,j}^{(w)}$ corresponds to the number of occurrences of the term *w* that are assigned to the topic *j*, not including the current (i^{th}) occurrence.
- Basically, a word is assigned to a topic with probability proportional to its frequency of occurrence in that topic.

Higher Order LDA – A Novel Approach

- LDA only considers relationships between observations within individual data instances (documents) while disregarding the dependencies that link observations across documents



- As part of our novel approach, we modified the Gibbs-sampling formula of LDA by replacing feature frequencies in topics with their higher order frequencies– we replaced these counts with higher path counts, $c_{i,j}$, for the feature *w* in topic *j*.

Nelson, Christie, Pottenger, William M., Keiler, Hannah, and Grinberg, Nir. (2012)

Hypothesis

- Hypothesis: The use of Higher Order Topic Modeling of text from social media and Nuclear Detection data will efficiently inform disaster responders.
  - Hypothesis: The use of HO-LDA for topic modeling will outperform LDA
  - Hypothesis: Determining optimal number of topics to be used in LDA/HO-LDA will improve the classification metrics: accuracy, precision, recall, F-measure.

Summer Progress

- Extensive literature search on the optimal number of topics for topic modeling and clustering
- Looked at various preprocessing methods
  - Nuclear Detection data
  - Text data from the 2010 Haitian Earthquake
- A novel approach to topic modeling, HO-LDA, was applied to two different types of data
  - Multinomial numeric Nuclear Detection Data to classify radioisotopes
  - Social media text data from 2010 Haitian Earthquake to classify type of emergency response needed

Optimal Number of Topics

- One parameter for LDA and HO-LDA is the number of topics
- Determining the optimal number of topics is an open problem in topic modeling, and clustering, a related field.
- A literature search was conducted on determining the optimal number of topics or clusters for background and inspiration for future work:
  - **Clustering Gain** - *Jung, Y., Park, H., Du, D., Drake, B. L. (2003).*
  - **Gap Statistic** - *Tibshirani, R., Walther, G., Hastie, T. (2001).*
  - **Jump Algorithm** - *Sugar, C. A., James, G. M. (2003)*
  - **Hierarchical Topic Models** - *David M. Blei, Thomas L. Griffiths, Michael I. Jordan (2010)*
  - **Chinese Restaurant Process** - *Blei, D. M. ,Griffiths, T. L., Jordan, M. I (2004)*

Optimal Number of Topics: The Chinese Restaurant Process

- The CRP can be used as a prior for the number of mixture components in a mixture model
- The idea:
  - Have a Chinese Restaurant with an infinite number of tables
  - A person enters a restaurant
    - Sits at an occupied table with probability proportional to the number of people already there
    - Sits at a new table with probability proportional to a parameter α_o
  - In our case, the tables represent the topics. The number of tables is the number of topics.

Blei, D. M. ,Griffiths, T. L., Jordan, M. I (2004)

Optimal Number of Topics: The Chinese Restaurant Franchise

- What if there are multiple mixtures and you wish to share mixture components across mixtures?
  - For example, you have a set of documents, each document contains topics, and you want to be able to have the same topics across documents.
- Idea: Chinese Restaurant Franchise
  - Analog of the CRP
  - Have a CRF that has a global menu for all the restaurants
  - Now, the first customer who sits at a table orders a dish
  - The dish is shared among the subsequent members arriving at the table
  - The dish can be at multiple tables in multiple restaurants
  - Here, the dish corresponds to the topic, and the restaurant to the document.

Teh, Y. W., Jordan, M. I, Beal, M. J., Blei. (2006)

2010 Haitian Earthquake

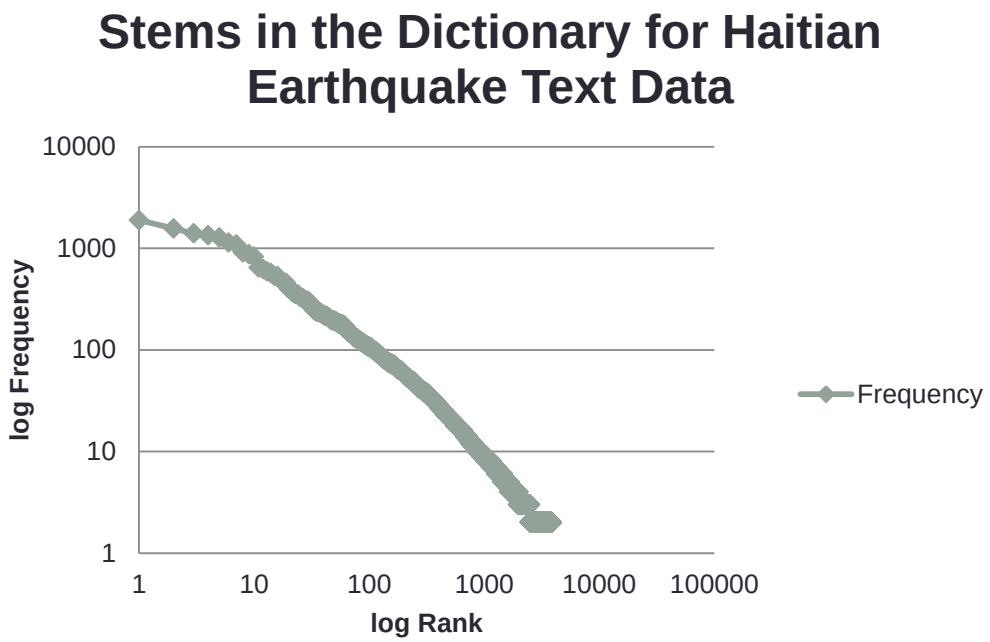
- Data is originally 3,598 texts from social media sources sent in the wake of the 2010 Haitian Earthquake
  - http://haiti.ushahidi.com/
  - 3,561 of these are in English or were translated into English. The rest are in French or Haitian Creole.
  - Example: "*THERES A BRIDGE IN DELMAS 31 THAT IS DAMANGED, THERES NO FOOD AND NO GAS YET*"
- Volunteers were responsible for manually translating and categorizing the incoming information
  - Examples of categories: *Water Shortage, Food Shortage, Medical Supplies, Collapsed Structure, Fuel Shortage*
- We want to apply LDA and HO-LDA to identify these categories automatically

Haitian Earthquake: Approach

- Preprocess
  - Remove common words (and, the, or,…), punctuation, and words occurring no more than once
  - Stem these words to make a dictionary of stems
  - Returns a vector for each document containing the dictionary words in the sentence represented as their numerical labels. This is the input for LDA and HO-LDA
    - **Method 1**: Use all stems in a word. For example, represent the word “children” by the numbers corresponding to “child” and “children”
    - **Method 2**: Use only the longest stem in each word.

Zipf's Law

- States that in a natural language corpus, the frequencies of occurrences of words are inversely proportional to their rank
  - Namely, the log-log plot is linear
- The dictionary for the Haiti text data follows Zipf's law



Haitian Earthquake: Approach

- Preprocess
- LDA/ HO-LDA
  - Choose number of topics
  - Choose Sample Size
    - 100%, 50%, 33%, 25%, 20%
    - Randomized and Stratified
- A decision tree algorithm for classification (J48) in Weka, a machine learning workspace
  - Accuracy, Precision, Recall, F-Measure
  - Both overall metrics and per class metrics

Haitian Earthquake: Results

- 100% with multiple stems: Accuracy
- These are preliminary results, which are inconclusive.
- I switched gears to work on a paper in classifying Nuclear Detection data, a similar project

Nuclear Detection

- An important problem in national security
- Classifying Nuclear Isotopes at seaports
- Classification is done on data from a handheld CZT(Cadmium Zinc Telluride)-based radiation detector, Interceptor™
  - A Thermo Scientific handheld Spectroscopic Personal Radiation Detector

Nuclear Detection: Approach

- Apart from the preprocessing, the same as for the Haitian Earthquake project
- LDA/ HO-LDA
  - Choose number of topics
  - Choose Sample Size
    - 100%, 50%, 33%, 25%, 20%
    - Randomized and Stratified
- J48 and Naïve Bayes in Weka
  - Accuracy, Precision, Recall, F-Score
  - Both overall and per class

Nuclear Detection: Approach

- Preprocess:
  - The data is originally 302 gamma ray spectrum files (instances) each with 1024 numeric channels for spectrum treated as attributes
  - **Method 1, *Multinomial Feature*:** Represent the data as draws from a multinomial distribution – this treats the ND data as if it were document data, which is what LDA normally expects
    - Transform each instance by turning it into a vector where the coordinates of the original vector are instead represented by their index the number of times of the coordinate value
      - Example: (3,4,0,1) → (0,0,0,1,1,1,1,3)
    - The transformed data is the input into LDA and HO-LDA This is done both for all 1024 attributes and for 39 attributes selected using attribute selection in Weka
  - **Method 2, *Maximum Channel Value*:** treated the highest number of the individual channel readings as the number of attributes

Nuclear Detection: Multinomial Feature Results

- 1024 Attributes, 100% of the data

ACCURACY T-TEST RESULTS – HO-LDA VS LDA WITH J48 CLASSIFIER FOR MULTINOMIAL FEATURE CREATION WITH 1024 ATTRIBUTES

#Topics	Accuracy HO-LDA Avg.	Accuracy HO-LDA Standard Deviation	P VALUE	Accuracy LDA Avg.	Accuracy LDA Standard Deviation
5	0.714	0.020	0	0.695	0.018
10	0.772	0.013	0	0.791	0.018
20	0.754	0.014	0	0.785	0.015
50	0.761	0.016	0	0.778	0.016
100	0.820	0.010	0	0.755	0.017

ACCURACY T-TEST RESULTS - HO-LDA VS LDA WITH NAÏVE BAYES CLASSIFIER FOR MULTINOMIAL FEATURE CREATION WITH 1024 ATTRIBUTES

#Topics	Accuracy HO-LDA Avg.	Accuracy HO-LDA Standard Deviation	P VALUE	Accuracy LDA Avg.	Accuracy LDA Standard Deviation
5	0.686	0.010	0	0.630	0.009
10	0.793	0.010	0	0.802	0.008
20	0.840	0.008	0	0.866	0.006
50	0.818	0.010	0	0.840	0.008
100	0.821	0.006	0.016	0.815	0.011

Nelson, Christie, Pottenger, William M., Keiler, Hannah, and Grinberg, Nir. (2012)

Nuclear Detection: Multinomial Feature Results

- 39 Attributes, 100% of the data

ACCURACY T-TEST RESULTS - HO-LDA VS LDA WITH J48 CLASSIFIER FOR MULTINOMIAL FEATURE CREATION WITH 39 ATTRIBUTES

#Topics	Accuracy HO-LDA Avg.	Accuracy HO-LDA Standard Deviation	P VALUE	Accuracy LDA Avg.	Accuracy LDA Standard Deviation
5	0.629	0.016	0	0.706	0.020
10	0.683	0.014	0	0.670	0.010
20	0.690	0.007	0	0.700	0.020
50	0.679	0.015	0	0.638	0.011

ACCURACY T-TEST RESULTS - HO-LDA VS LDA WITH NAÏVE BAYES CLASSIFIER FOR MULTINOMIAL FEATURE CREATION WITH 39 ATTRIBUTES

#Topics	Accuracy HO-LDA Avg.	Accuracy HO-LDA Standard Deviation	P VALUE	Accuracy LDA Avg.	Accuracy LDA Standard Deviation
5	0.607	0.012	0	0.748	0.007
10	0.765	0.007	0.013	0.760	0.006
20	0.786	0.010	0	0.763	0.012
50	0.756	0.009	0	0.769	0.011

Nelson, Christie, Pottenger, William M., Keiler, Hannah, and Grinberg, Nir. (2012)

Nuclear Detection: Multinomial Feature Results

- Experiments were also performed using 39 attributes for training samples of size 20%, 25%, 33%, and 50%.
- These results were similar to the previously shown, and provide evidence that for Multinomial Feature creation LDA and HO-LDA perform similarly.

Nuclear Detection: Maximum Channel Value Results

- J48 classifier on 100% of the data:

#Topics	Accuracy HO-LDA Avg.	Accuracy HO-LDA SD	P VALUE	Accuracy LDA Avg.	Accuracy LDA SD
5	0.675	0.014	0	0.610	0.012
10	0.688	0.015	0	0.641	0.014
20	0.654	0.015	0.021	0.644	0.017
50	0.665	0.014	0	0.605	0.015
100	0.565	0.015	0	0.529	0.015

- For sample sizes 20%, 25%, 33% and 50% HO-LDA statistically significantly outperformed LDA with all four metrics for 5, 10, and 20 topics, with the exception of 50% 5 topics with Accuracy and Recall metrics.
- For Maximum Channel Value Feature Creation, it is evident that HO-LDA produces better topic models than LDA

Nelson, Christie, Pottenger, William M., Keiler, Hannah, and Grinberg, Nir. (2012)

Conclusions and Future Work

- Developed HO-LDA
- Looked into the optimal number of topics question
- Looked at HO-LDA in two data sets:
  - Haitian Earthquake text data, in progress
  - Nuclear Detection Data

- Wish to incorporate a version of the Chinese Restaurant Franchise schema for topic number selection into LDA and HO-LDA for the Haitian Earthquake text data
- Continue running trials for the Haitian Earthquake text data with various sample sizes
  - Look at classification metrics for each class

Conclusion

- **Our research from the projects presented will be published/presented at the following venues**:
  - Nelson, Christie, Pottenger, William M., Keiler, Hannah, and Grinberg, Nir. "Nuclear Detection Using Higher-Order Topic Modeling." 2012 IEEE International Conference on Technologies for Homeland Security. Waltham, MA. 13-15 Nov 2012.
  - Nelson, Christie, Pottenger, William M., and Keiler, Hannah. "Nuclear Detection using Higher-Order Topic Modeling." Domestic Nuclear Detection Office and National Science Foundation (DNDO-NSF) 5th Annual Academic Research Initiative (ARI) Grantees Conference . Lansdown, VA. 23-25 Jul 2012.
  - Nelson, Christie, Pottenger, William M., and Keiler, Hannah. "Higher-Order Latent Dirichlet Allocation with Nuclear Detection Data." INFORMS 2012 Annual Meeting. Phoenix, AZ. 14-17 Oct 2012.
  - Nelson, Christie, Pottenger, William M., and Keiler, Hannah. "Real Time Optimization of Emergency Response Using Social Media Data." INFORMS 2012 Annual Meeting. Phoenix, AZ. 14-17 Oct 2012.

- **In addition, we hope to submit to:**
  - AAAI 2013 Spring Symposium on Analyzing Microtext, Stanford, CA. 25-27 Mar 2013.

References

- Blei, D. M., Griffiths, T. L., & Jordan, M. I. (2010). The nested Chinese restaurant process and Bayesian nonparametric inference of topic hierarchies. *Journal of the ACM*
- Blei, D.M., Griffiths, T. L., Jordan, M. I., & Tenenbaum, J. B. (2004). Hierarchical topic models and the nested Chinese restaurant process. *Advances in Neural Information Processing Systems*
- Blei, D., and Lafferty, J. (2009). Topic Models. *Text Mining: Classification, Clustering, and Applications*
- Blei, D., Ng, A. Y., Jordan, M. I. (2003) Latent Dirichlet Allocation. *Journal of Machine Learning Research*
- Celikyilmaz, A., Hakkani-Tur, D. (2010) A hybrid Hierarchical Model for Multi-Document Summarization. *Proceedings of the 48th Annual Meeting for the Association of Computational Linguistics.*
- Chaturvedi, S. K., Swami, D. K., Singh, G. (2011) Dirichlet Distribution with Centroid Model (DDCM) based Summarization Technique for Web Document Classification. *COMPUTE '11 Proceedings of the Fourth Annual ACM Bangalore Conference.*
- Caragea, C., et al.. (2011). Classifying Text Messages for the Haiti Earthquake. *Proceedings of the 8th International ISCRAM Conference.*
- Ganiz, M. C., Lytkin, N. I., Pottenger, W. M. (2009). Leveraging Higher Order Dependencies Between Features for Text Classification. *Proceedings of the European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases .*
- Hoffman, M., Blei, D., Cook, P. (2008). Content-based musical similarity computation using the Hierarchical dirichlet process. *International Conference on Music Information Retrieval*
- Hong, L., Davison, B. D. (2010) Empirical Study of Topic Modeling in Twitter. *1st Workshop on Social Media Analytics (SOMA '10)*
- Jordan, M. I. (2005). Dirichlet Processes, Chinese Restaurant Processes, and all that. *22nd Annual International Conference on Machine Learning (ICML)*
- Jung, Y., Park, H., Du, D., Drake, B. L. (2003). A Decision Criterion for the Optimal Number of Clusters in Hierarchical Clustering. *Journal of Global Optimization*
- Nelson, Christie and Pottenger, William M., Grinberg, Nir. (2011). Nuclear Detection Using Higher Order Learning. In the proceedings of the *IEEE International Conference on Technologies for Homeland Security*.
- Sugar, C. A., James, G. M. (2003). Finding the Number of Clusters in a Dataset: An Information-Theoretic Approach. *Journal of the American Statistical Association*
- Teh, Y. W., Jordan, M. I, Beal, M. J., Blei, D. M. (2006). Hierarchical Dirichlet Processes. *Journal of the American Statistical Association*
- Thermo Scientific. *User Manual Interceptor™ Spectroscopic Personal Radiation Detector*
- Tibshirani, R., Walther, G., Hastie, T. (2001). Estimating the Numbers of Clusters in a Data Set Via the Gap Statistic. *Journal of the Royal Statistical Society: Series B (Statistical Methodology).*
- Zipf, G. (1932). Selective Studies and the Principle of Relative Frequency in Language. *Harvard University Press*.
- haiti.ushahidi.com

Metrics

- Accuracy $\frac{\text{True positives}}{\text{Total}}$
- Precision $\frac{\text{True positives}}{\text{True positives + False positives}}$
- Recall $\frac{\text{True positives}}{\text{True positives + False negatives}}$
- F-Score $\frac{2*\text{precision}*\text{recall}}{\text{precision + recall}}$