Extended Topic Models with Numerical Features

Gökhan Çapan, Ali Caner Türkmen

March 23, 2016

Introduction: Topic Models

- *Unsupervised* learning, recover *latent* topics in documents
- Can be thought of as *clustering*.
- **Key Assumption:** Topics lead to distinct word distributions. Intuitive.

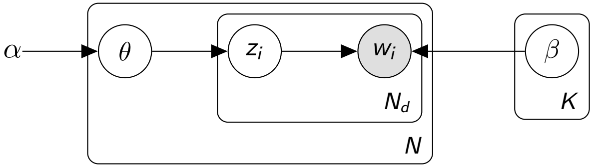
Latent Dirichlet Allocation

- Rich, probabilistic mixed membership model due to [Blei et al., 2003]
- Widely adopted and extended
- Assumes a fixed number of topics in a corpus
- **Key Idea:** A document includes words from multiple topics (in contrast with clustering)

LDA - Generative Model

- For each document;
  - Topic proportions vector is drawn ($\theta \sim \mathit{Dirichlet}(\alpha)$)
  - For each word in the document
    - A topic is drawn from topic proportions ($z_i \sim \mathit{Multinomial}(\theta)$)
    - The word is drawn from topic ($w_i \sim \mathit{Multinomial}(\beta_{z_i})$)

LDA - Bayesian Network Representation

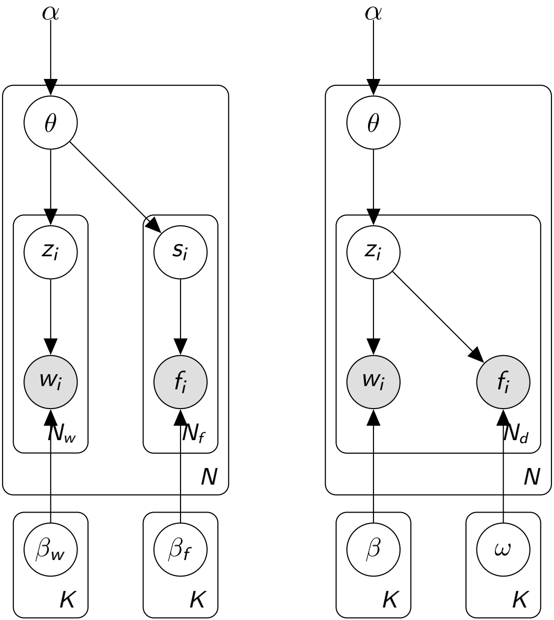


- Note the difference with document clustering (z is outside of the plate in that case, each word of a document comes from a single cluster), which is referred to as *mixture of unigrams*

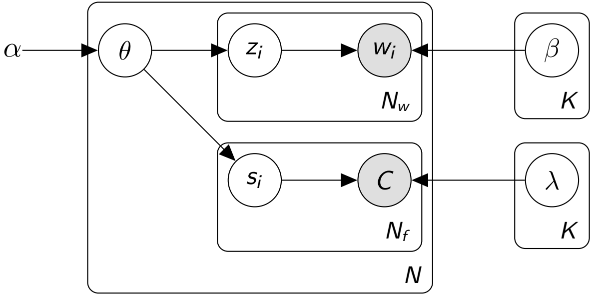
LDA - Multi Modal (Aspect) Variants

- A topic generates not only words, but also other modalities
- The features can be paired with words themselves (e.g. sentiment, polarity, word sense disambiguation)
- The topics can generate other aspects
- Some examples:
  - [Putthividhy et al., 2010] in CVPR.
  - [Roller and Im Walde, 2013] in ACL/EMNLP.
  - [Troelsgaard et al., 2014] in NIPS workshop.

Multi-Modal LDA Variants



Proposed Model (Tentative)



- Here, C is an $|L|$ dimensional vector of Poisson random variables, where L is the set of numerical features

Data Set and Features

- **Data Set:** News articles sampled from Anadolu Agency website. 1337 documents (can be expanded), 3000 tokens after adjusting for document frequency.
- **Features:** Complexity features such as word count, sentence count, average sentence length, comma count. (TBD)
- **Novel Features:** Etymological counts. Count the number of words from their etymological origins. Number of Arabic, Farsi, French words, etc. Source: TR Wiktionary Database Dump.

Learning

- **Variational inference:** Derive approximate inference algorithms based on a decoupling of the original model OR Variational EM-like procedures to find parameter estimates.
- Gibbs sampling assuming appropriate priors (tentative, out of scope for this project)

Conclusion

We propose two key contributions

- Put the topic modeling problem in an extended LDA framework, with numerical features
- Use etymological counts for the Turkish language

Thank You!

gokhan.capan [at] boun.edu.tr
caner.turkmen [at] boun.edu.tr

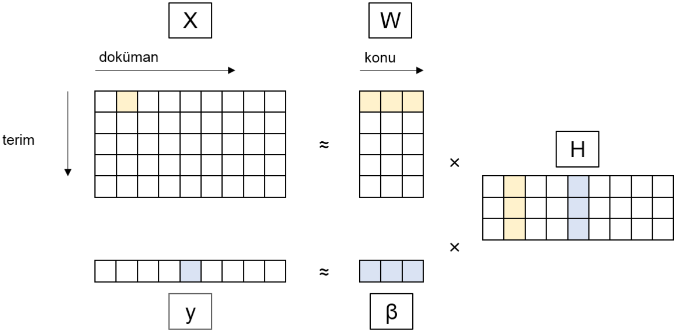
Blei, D. M., Ng, A. Y., Jordan, M. I., and Lafferty, J. (2003).
Latent dirichlet allocation.
Journal of Machine Learning Research, 3:2003.

Putthividhy, D., Attias, H. T., and Nagarajan, S. S. (2010).
Topic regression multi-modal latent dirichlet allocation for image annotation.
In *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*, pages 3408–3415. IEEE.

Roller, S. and Im Walde, S. S. (2013).
A multimodal LDA model integrating textual, cognitive and visual modalities.
In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1146–1157.

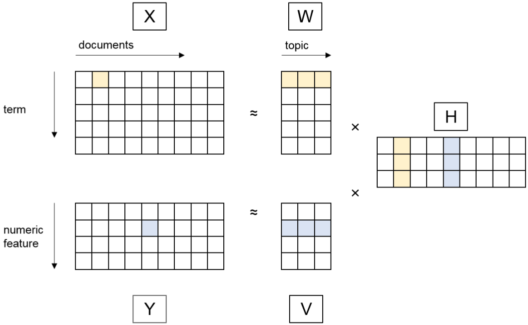
Troelsgaard, R., Jensen, B. S., and Hansen, L. K. (2014).
A Topic Model Approach to Multi-Modal Similarity.
arXiv preprint arXiv:1405.6886.

Coupled Matrix Factorization for Recovering Topics



- Represent data with well-known algebraic structures
- Jointly guide topic assignments from multimodal datasets, in a probabilistically sound framework
- Easily extensible to semi-supervised learning, kernel methods
- (T. and Cemgil, 2016)

Extended Coupled NMF for Topic Learning with Count Features



- $y_{ij}|W,H \sim GPO(\sum_t w_{it}h_{tj}, \phi)$
- $x_{kj}|V,H \sim GPO(\sum_t v_{kt}h_{tj}, \gamma)$
- Guide topic modeling with numeric count data
- Can assume priors $p(W), p(H), p(V)$ for Bayesian learning