## **Probabilistic Topic Models**

Piyush Rai

#### Topics in Probabilistic Modeling and Inference (CS698X)

March 13, 2019



Prob. Modeling & Inference - CS698X (Piyush Rai, IITK)

## **Topic Models**

- Given a collection of objects (each object a set of discrete tokens), find topics in the collection
- Can also be used to represent each object as how much each topic is present in that object

Real world example:

• Some topics inferred from a collection where each object is a document (and tokens are words)

music band songs rock album jazz pop song singer night	book life novel story books man stories kove children family	art museum show exhibition artist artists painting contury works	game knicks points team season play games night coach	show film television movie series says life man character know
theater	clinton	stock	restaurant	budget
play	bush	market	sauce	tax
production	campaign	percent	menu	governor
show	gore	fund	food	county
stage	political	investors	dishes	mayor
street	republican	funds	street	billion
broadway	dole	companies	dining	taxes
director	presidential	stocks	dinner	plan
musical	senator	investment	chicken	legislature
directed	house	trading	served	fiscal

LDA analysis of 1.8M New York Times articles:

The New Hork Times

• Also applicable to other types of data, beyond text documents, e.g.,

- Image collection: Each image is a "document" which is a bag of "visual words"
- Customer-purchase data: Each customer is a "document" and items purchased are the "words"

Pic courtesy: David Blei

Prob. Modeling & Inference - CS698X (Piyush Rai, IITK)

## **Topic Models**

- For concreteness, let's focus our discussion on topic models for text
- Topic models are based on assigning words in each document to clusters/topics (each cluster of words represents a "topic"): Essentially, this is word clustering with many documents!



• Formally, a "topic" is a distribution over tokens (a prob. vector with length = # of unique tokens)

## **Topic Models**

• The input for the topic modeling problem will be the raw contents of the documents and the goal is to infer all the unknowns of the model (topics, topic proportions for each document, etc.)





## **Applications of Topic Models**

- Can be used to learn topic-based representation for each document
- These representation are compact (think dimensionality reduction) and semantically meaningful



#### Seeking Life's Bare (Genetic) Necessities

- Also makes it easy to cluster data naturally by (learned) topics
- Can also be used to understand evolution of a corpora (e.g., how topics drift over time "Dynamic Topic Models" e.g., word prominent in one topic now may not be so after 10 years later)

## A Toy Mixture Model for Words in a Document

- Recall that topic modeling is like clustering the words of the documents
- Consider a toy K component multinoulli mixture model for the words of a single document

- "Multinoulli" because the observations (words here) are discrete tokens
- Assume the overall vocabulary consists of V unique words/tokens
- Assume each word  $w_{d,n} \in \{1, \ldots, V\}$  belongs to a cluster  $z_{d,n} \in \{1, \ldots, K\}$
- $\theta_d$  is the mixture proportion vector (summing to 1) for the document
- $\phi_k$  denotes the parameters of mixture component k
- In a Bayesian version,  $\theta_d$  and  $\phi_k$  will have priors with hyperparams  $\alpha$  and  $\eta$ , respectively

## A Toy Mixture Model for Words in a Document



- The generative story will be
  - Draw the K topic vectors  $\{\phi_k\}_{k=1}^K$  from a V-dim Dirichlet

$$\phi_k \sim \mathsf{Dirichlet}(\eta, \ldots, \eta) \qquad k = 1, \ldots, K$$

• Draw the mixing proportion vector  $\theta_d$  from a K-dim Dirichlet

$$heta_{ extsf{d}} \sim \mathsf{Dirichlet}(lpha, \dots, lpha)$$

• For each word 
$$w_{d,n}$$
,  $n = 1, \ldots, N_d$ 

$$z_{d,n} \sim \text{multinoulli}(\theta_d)$$
  
 $w_{d,n} \sim \text{multinoulli}(\phi_{z_{d,n}})$ 



# Latent Dirichlet Allocation (LDA)

- $\, \circ \,$  Let's consider a more "real" case. We have a collection of D>1 documents
- The previous toy model can be extended easily to handle this



- All docs share the same set of topics  $\phi_1, \ldots, \phi_K$  but have different topic proportions  $\theta_1, \ldots, \theta_D$
- Equivalent to jointly learning D multinoulli mixture models (one per document)
  - Information shared acorss documents via global variables  $\{\phi_k\}_{k=1}^K$ , and global hyperparams  $\alpha$  and  $\eta$
- Each topic is a distribution over V tokens  $(\phi_k)$ , each document is a distrib. over K topics  $(\theta_d)$
- This is basically the Latent Dirichlet Allocation (LDA) model (Blei et al, 2003)
- Generative story is identical to the single doc. toy example but now  $\theta_d$  for each doc.

## **Some LDA Precursors**

• Mixture of unigrams models: Each document having a single topic, all words from that topic



- Unlike the above mixture of unigrams model, LDA is an example of "admixture model" (all words of the document not drawn from the same topic but from a mixture of topics)
- Probabilistic Latent Semantic Analysis: Like LDA, each word in a doc has its own topic



- However, unlike LDA, PLSA (Hofmann (2001) is not a Bayesian model
  - As a result, LDA is less prone to overfitting





### Inference for LDA



• The goal is to infer the posterior distribution over all the latent variables

$$p(\mathbf{Z}, \Theta, \Phi | \mathbf{W}, \alpha, \eta) = \frac{p(\mathbf{W} | \Phi, \mathbf{Z}) p(\mathbf{Z} | \Theta) p(\Phi | \eta) p(\Theta | \alpha)}{p(\mathbf{W} | \alpha, \eta)}$$
(assuming hyperparams  $\alpha, \eta$  are fixed)

• Z: Topic assignments of all words across all documents

•  $\Theta = \{\theta_1, \dots, \theta_D\}$ : Topic proportion vectors; each  $\theta_d$  is a *K*-dim vector

•  $\Phi = \{\phi_1, \ldots, \phi_K\}$ : The K topics; each  $\phi_k$  is a V-dim vector

•  $\mathbf{W} = {\mathbf{w}_{d,n}}, d = 1, \dots, D, n = 1, \dots, N_d$ : Collection of all words in all the documents

## Inference for LDA



• A wide variety of inference methods exist for LDA (VI, Variational EM, MCMC, batch, online, ...)

• LDA is a locally conjugate model (recall that the model uses Dirichlet/multinoulli distr.)

• The likelihood and priors are all exponential family distributions

- MCMC, in particular Gibbs sampling, is very easy to derive due to local conjugacy
- Local conjugacy also allow deriving a simple VB algorithm (the original LDA paper<sup>†</sup> used VB)
- Note: Can even collapse some variables and do collapsed Gibbs or collapsed VB

<sup>&</sup>lt;sup>†</sup>Latent Dirichlet Allocation (Blei et al, 2003)

Prob. Modeling & Inference - CS698X (Piyush Rai, IITK)

## LDA Evaluation

- Many ways to test how well LDA performs on some data
- Extrinsic measures: Perform LDA and measure performance on some external task (e.g., accuracy of a document classification model using the topic based document representation)
- Perplexity is another intrinsic measure
- Given a test collection of M document, the perplexity is defined as

$$perplexity(D_{\text{test}}) = \exp\left\{-\frac{\sum_{d=1}^{M}\log p(\mathbf{w}_{d})}{\sum_{d=1}^{M}N_{d}}\right\}$$

.. where  $\boldsymbol{w}_d$  collectively denotes all words in a test document d

- We set aside some documents as a test set and compute the average marginal likelihood  $p(w_d)$  the model assigns to the words in these held-out documents
- A high average probability (low perplexity) implies a good fit to the data

## LDA as (Bayesian) Non-negative/Poisson Matrix Factorization



Assume X is V × D word-document count matrix (X<sub>v,n</sub> = no of times word v appears in doc d)
Assume Φ = [φ<sub>1</sub>,..., φ<sub>K</sub>] is V × K, and Θ = [θ<sub>1</sub>,..., θ<sub>D</sub>] is K × D, both are non-neg matrices

 $\,\circ\,$  Then LDA can be seen as doing a Poisson matrix factorization^{\dagger} of the matrix  ${\bm X}$ 

 $\mathbf{X} \sim \mathsf{Poisson}(\Phi \Theta)$ 

- ${\scriptstyle \bullet}$  Notation here: Poisson applied element-wise on  $\Phi\Theta$  rate matrix
- Note that word-to-topic assignments  $z_{d,n}$  don't appear in this formulation
- Note: Even if priors on  $\phi_k$  and  $\theta_d$  are not Dirichlet (e.g., each entry of vectors  $\phi_k$  and/or  $\theta_d$  has some other non-neg prior<sup>†</sup>, such as gamma, the Bayesian Poisson NMF equivalence still holds)

<sup>&</sup>lt;sup>†</sup>Sec 4 and 5 of "Beta-Negative Binomial Process and Poisson Factor Analysis" (Zhou et al, 2012)

#### LDA is easy to extend



Several extensions of LDA (too many to list here :)). Some include

- LDA with document labels (supervised LDA)
- LDA for multimodal data (images and text) and multilingual data
- LDA with topic hierarchies/correlations
- LDA with time-evolving topics
- LDA with HMM (takes into account word order)
- LDA for graphs

## LDA: Some Other Improvements

• Learning the number of topics using nonparametric Bayesian models, e.g.,

- Hierarchical Dirichlet Process (HDP) topic models
- Faster inference methods, e.g., online (SVI), amortized VI, SGLD, etc
- Deep Topic Models (e.g., using variational autoencoders for document-word count vectors)



