10-708: Probabilistic Graphical Models, Spring 2015

15 : Case Study: Topic Models

Lecturer: Eric P. Xing *Scribes: Xinyu Miao,Yun Ni*

1 Task

Humans cannot afford to deal with a huge number of text documents (e.g., search, browse, or measure similarity). We need new computational tools to help organize, search and understand these vast amounts of information. To this end, machine learning researchers have developed **Probabilistic Topic Modeling**, a suite of algorithms that aim to discover and annotate large archives of documents with thematic information, and thus help us on varieties of tasks with documents. (Blei, 2012)

One task we can do with topic models is **Document Embedding**. In a problem of Document Embedding, we want to have a mapping: $\mathbf{D} \to \mathbf{R}^d$, where $\mathbf{D}$ is the spaces of documents and $\mathbf{R}^d$ is Euclidean Space. Document Embedding enable us to compare the similarity of two documents, classify contents, group documents into clusters, distill semantics and perspectives, etc.

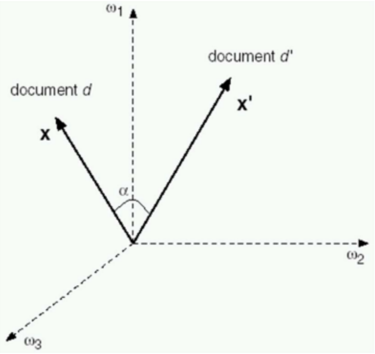


Figure 1: Visualization of Document Embedding

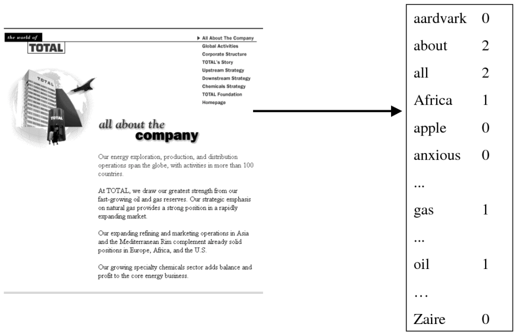
The other tasks for topic modeling include summarizing the data using topics, and visualizing how topics changes over time, and modeling user interest using topics.

2 Data Representation

Data representation defines the input and output of topic models. Generally speaking, we have two ways of representing a documents:

- **Linear Sequence of Words** In Linear Sequence of Words representation, each document are linearized into a long word vector.

- **Bag of Words** Bag of words is an orderless high-dimensional and sparse representation. Each document is represented by frequencies of words over a fixed vocabulary.



Our energy exploration, production, and distribution operations span the globe, with activities in more than 100 countries.

At TOTAL, we draw our greatest strength from our fast-growing oil and gas reserves. Our strategic emphasis on natural gas provides a strong position in a rapidly expanding market.

Our expanding refining and marketing operations in Asia and the Mediterranean Rim complement already solid positions in Europe, Africa, and the U.S.

Our growing specialty chemicals sector adds balance and profit to the core energy business.

aardvark	0
about	2
all	2
Africa	1
apple	0
anxious	0
...	
gas	1
...	
oil	1
…	
Zaire	0

Figure 2: Bag of Words Representation

These two methods of document representing have different advantages and disadvantages:

- Linear Sequence of Words is hard to perform mechanical computational comparison.
- Linear Sequence of Words lacks dimensional correspondence.
- Bag of Words maps all the documents the same dimensional space, which make the problem comparable.
- Bag of Words ignores the order of the words.
- In Bag of Words, sometimes vocabulary is too large so that it is not effective for browsing and not efficient for text processing tasks such as searching, document classification, and similarity measuring.

In topic modeling, usually we prefer Bag of Words to Linear Sequence of Words because of its advantages in dimensional correspondence.

Another important topic of data representation is semantic modeling. Rather than associating each group of documents with one topic, each group exhibits multiple components in different proportions. This is a more structured way of browsing the collection, where we can easily find similar documents.

3 Model

We now introduce topic models. Topic models organize unstructured document collection into topic simplex which involves both Topic Discovery and Dimensionality Reduction. The process of generating a document is as follows:(Blei & Lafferty, 2009)

```
Draw $\theta$ from the prior;
for each word $n$ do
    Draw $z_n$ from $multinomial(\theta)$;
    Draw $w_n | z_n, \{\beta_{1:k}\}$ from $multinomial(\beta_{z_n})$;
end
```

Algorithm 1: Generating a document using topic model

We can choose two different priors for θ in topic models. If we choose Dirichlet distribution as the prior, the model is called **Latent Dirichlet Allocation**(LDA). Inference for LDA is usually efficient because Dirichlet distribution is the conjugate prior for categorical distributed θ. However, LDA can only capture variations in each topic's intensity independently. (Blei et al., 2003) If we choose Logistic Normal distribution as prior for θ, the Model is called **CTM** or **LoNTAM**. CTM is able to capture the intuition that some topics are highly correlated and can rise up in intensity together. However, inference is hard for CTM because Logistic Normal distribution is not a conjugate prior for categorical distribution.

We often differentiate Topic Modeling with other subspace analysis methods such as **Latent Semantic Indexing**, because they use the same form of matrix decomposition. They differs in the types of matrix that is decomposed:

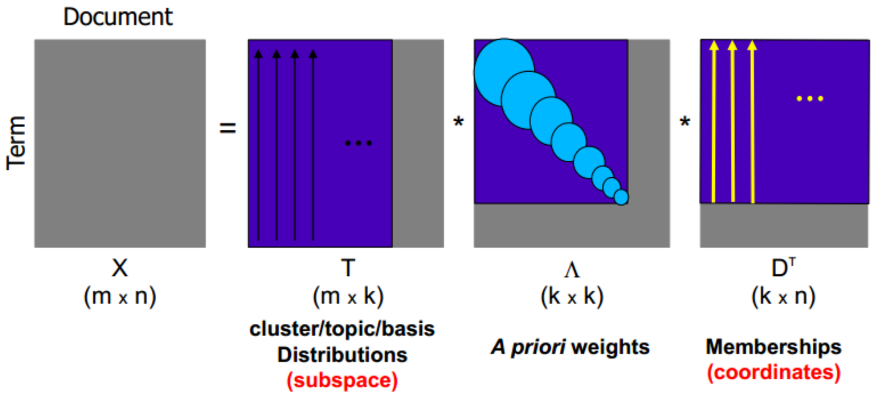


Figure 3: Matrix Decomposition of Subspace Analysis Methods

- Clustering: Binary Matrices for $D^\top$
- Latent Semantic Indexing: Arbitrary Matrices through Singular Value Decomposition
- Topic Models: Stochastic Matrices
- Sparse Coding: Sparse Arbitrary Matrices
- Deep Learning: Do the decomposition for multiple layers

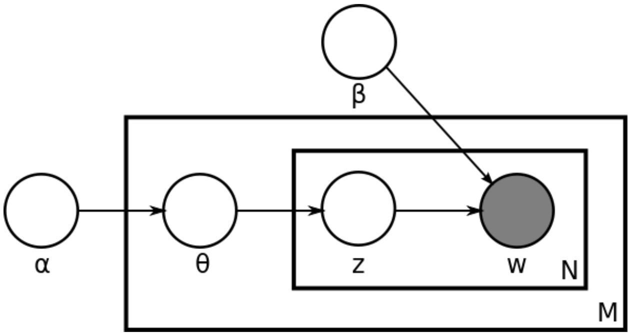


Figure 4: Plate Notation for LDA

4 Inference and Learning

The first task in inference is posterior inference, which we can compute through the joint distribution of a Bayesian network.

$$p(\beta, \theta, z, w) = \prod_{k=1}^{K} p(\beta_k|\eta) \prod_{d=1}^{D} p(\theta_d|\alpha) \prod_{n=1}^{N} p(z_{d,n}|\theta_d) p(w_{d,n}|z_{d,n}, \beta)$$

For posterior inference and learning questions, we may ask

- $p(\theta_n|D) =?$
- $p(z_{n,m}|D) =?$
- What to learn?
- What is the objective function in learning?

However, these tasks are intractable. For example,

$$p(\theta_n|D) = \frac{\sum_{z_{m,n}} \int \prod_{d=1}^{D} p(\theta_d|\alpha) \prod_{n=1}^{N} p(z_{d,n}|\theta_d) p(w_{d,n}|z_{d,n}, \beta) d\theta d\beta}{p(D)}$$

where

$$p(D) = \sum_{z_{m,n}} \int \cdots \int \prod_{d=1}^{D} p(\theta_d|\alpha) \prod_{n=1}^{N} p(z_{d,n}|\theta_d) p(w_{d,n}|z_{d,n}, \beta) d\theta_1 \cdots d\theta_N d\beta$$

As a result, we use approximate inference. We list some common approximation algorithms as the following. In this lecture, we only introduce the mean field approximation for topic models.

- Variational Inference:
  - Mean field approximation (Blei et al)
  - Expectation propagation (Minka et al)

  - Variational 2nd-order Taylor approximation (Ahmed & Xing, 2007)

- Markov Chain Monte Carlo:
  - Gibbs sampling (Griffiths & Steyvers, 2004)

Recall in the mean field approximation, we assume the variational distribution over the latent variables factorizes as

$$q(\beta, \theta, z) = \prod_k q(\beta_k) \prod_d q(\theta_d) \prod_n q(z_{d,n})$$

which means we assume the variational approximation q over β, θ, d are independent. Remember that mean-field family usually does NOT include the true posterior.

Then recall that in the mean field approximation, we intend to optimize the lower bound of the exact posterior:

$$\mathcal{L}(q(h)) = E_q[logp(w,h)] + \mathcal{H}(q(h))$$

where

$$h = \{\beta, \theta, z\}$$

$$q(\beta, \theta, z) = \prod_k q(\beta_k) \prod_d q(\theta_d) \prod_n q(z_{d,n})$$

Now we derive a coordinate ascent algorithm. Our objective function is

$$\mathcal{L}(q(h_i)) = \int q(h_i) E_{q_{-i}}[\log p(w,h)] dh_i + \mathcal{H}(q(h))$$

where h_i can be one of $\{\beta, \theta, z\}$, and $E_{q_{-i}}$ is the expectation over all other latent variables except for the j-th variable.

In Lecture 13, we know the optimal solution is

$$q(h_i) \propto \exp(E_{q_{-i}}[\log p(w,h)])$$

Now we have the following update rule for LDA,

$$q(\theta_d|\alpha) \propto \exp\left(\sum_{k=1}^{K} (\alpha_k - 1) \log \theta_{dk}\right)$$

$$q(z_{dn}|\theta_d) = \exp\left(\sum_{k=1}^{K} 1_{[z_{dn}=k]} \log \theta_{dk}\right)$$

$$q(\theta_d) = \exp\left(\sum_{k=1}^{K} \left(\sum_{n=1}^{N} q(z_{dn}=k)\right) \log \theta_{dk}\right)$$

And the algorithm is as follows

```
Initialize varientional topics $q(\beta_k)$;
while Lower bound $L(q)$ not converge do
    for each document $d \in \{1, 2, 3 \cdots D\}$ do
        Initialize varientional topic assignment $q(z_{dn})$;
        while Change of $q(\theta)$ is not small enough do
            Update varientional topic proportions $q(\theta_d)$;
            Update varientional topic assignments $q(z_{dn})$;
        end
        Update varientional topics $q(\beta_k)$;
    end
end
```

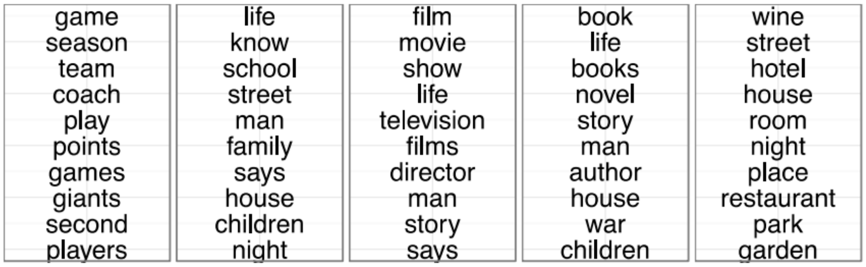
Algorithm 2: Coordinate ascent algorithm for LDA

However, mean-field algorithms could be very slow if we have millions of documents.

5 Evaluation

Despite that topic modeling is an unsupervised model, evaluation is very important. To evaluate the performance of a step, we need to fix the previous steps. For example, to evaluate a new inference method, we need to run both the new and old inference algorithms on identical models.

There are two ways of evaluating topic models inference. The empirical way is to visualize the results and judge the results by humans. The followings are the topic we discovered from New York Times using LDA.



game	life	film	book	wine
season	know	movie	life	street
team	school	show	books	hotel
coach	street	life	novel	house
play	man	television	story	room
points	family	films	man	night
games	says	director	author	place
giants	house	man	house	restaurant
second	children	story	war	park
players	night	says	children	garden

Figure 5: The 5 most frequent topics from the HDP on the New York Times

Another way to evaluate is to test on synthetic text where ground truth is known. Here, we show the comparison of Mean field approximation (BL) and Variational 2nd-order Taylor approximation (AX).

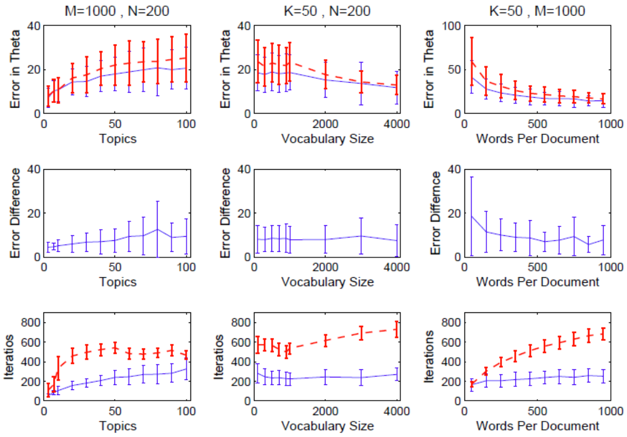


Figure 6: Inference On Simulated Data. Dotted and solid lines correspond to the BL and AX approaches respectively. Each column represents an experiment in which one dimension is varied. **Top row**: Average L2 error in topic vector estimation. **Middle row**: Error difference (L2(BL)-L2(AX)) in topic vector estimation on a per document level. **Bottom row**: Number of iterations needed by each approach to converge.

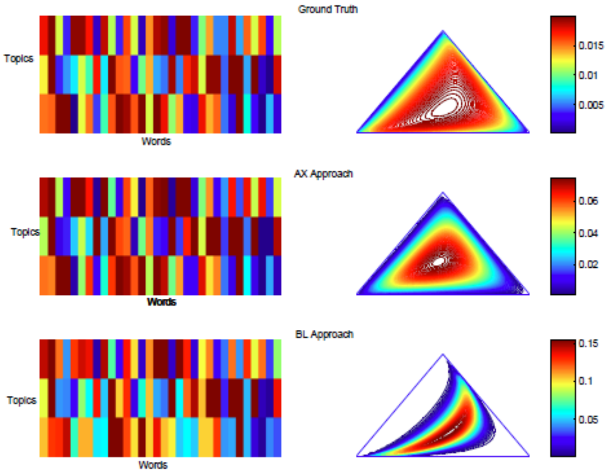


Figure 7: Parameter Estimation. Left panels represent topic distributions where each row is a topic, each column is a word, and colors correspond to probabilities. Right panels represent shapes of LN distribution over the sim- plex. Top row gives the ground truth model parameters, while middle and bottom rows give those estimated using the AX and BL approach respectively.

We also evaluate topic models on classification tasks. We use PNAS abstracts from 1997-2002 as a benchmark dataset, which contains 2500 documents with average of 170 words per document. We fitted 40-topics model using both approaches. We used topic model to generate low dimensional representation to predict the abstract category with SVM classifier.

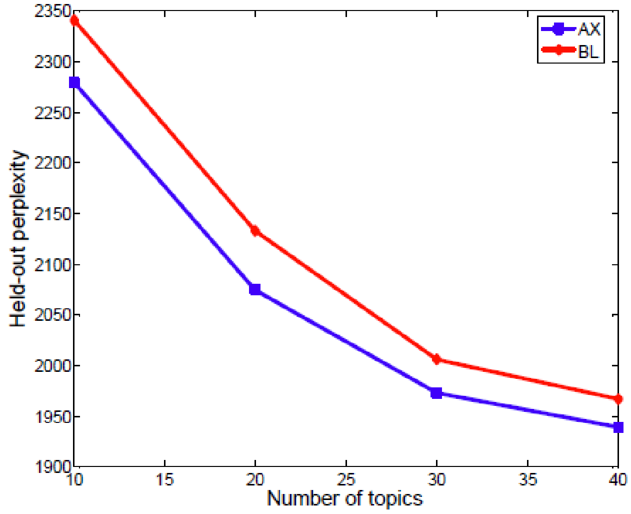
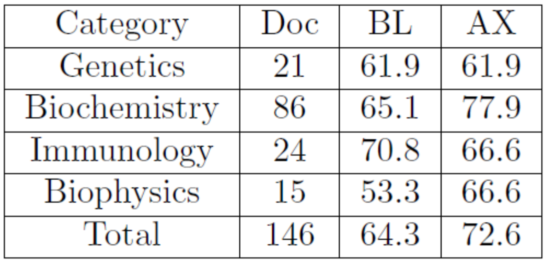


Figure 8: Test-set perplexities on the NIPS dataset.

Table 1: Document classification accuracies



Category	Doc	BL	AX
Genetics	21	61.9	61.9
Biochemistry	86	65.1	77.9
Immunology	24	70.8	66.6
Biophysics	15	53.3	66.6
Total	146	64.3	72.6

References

Ahmed, Amr and Xing, Eric P. Seeking the truly correlated topic posterior - on tight approximate inference of logistic-normal admixture model. 2:19–26, 2007. URL http://jmlr.csail.mit.edu/proceedings/papers/v2/ahmed07a/ahmed07a.pdf.

Blei, David M. Probabilistic topic models. *Commun. ACM*, 55(4):77–84, April 2012. ISSN 0001-0782. doi: 10.1145/2133806.2133826. URL http://doi.acm.org/10.1145/2133806.2133826.

Blei, David M and Lafferty, John D. Topic models. *Text mining: classification, clustering, and applications*, 10:71, 2009.

Blei, David M., Ng, Andrew Y., and Jordan, Michael I. Latent dirichlet allocation. *J. Mach. Learn. Res.*, 3: 993–1022, March 2003. ISSN 1532-4435. URL http://dl.acm.org/citation.cfm?id=944919.944937.

Griffiths, Thomas L. and Steyvers, Mark. Finding scientific topics. *Proceedings of the National Academy of Sciences*, 101(suppl 1):5228–5235, 2004.