Stochastic Neighbor Topic Model

Final Presentation for Advanced Introduction to Machine Learning, 2017

Wenchao. Du[1] Yue. Li[2]

[1]Language Technology Institute
Carnegie Mellon University

[2]Department of Statistics
Carnegie Mellon University

November 27, 2017

Outline

Problem Formulation

- **Dimensionality reduction of text**: In text mining, data representations of documents are very high-dimensional.
- **Topic modeling**: Standard dimensionality reduction technique for text data by introducing latent variables – *topics*. Each document is associated with a topic distribution.
- **Semi-supervised document clustering**: Given documents with meta-data such as tags, links, incorporate these information to improve document embedding quality.

Our Proposal

- **Assumption**: Documents that are in same category (or being linked) have similar topic distribution. Likewise, documents that are not in same category (or being linked) have dissimilar topics.
- **Previous work in literature**: Add additional regularization term on distances of document topic distributions to the objective of topic model. Distance measure includes Euclidean distance, KL divergence, or even hinge loss.
- **Our proposal**: We regularize distributional parameters using softmax of distances.
- **Novelty**:
  - A completely probabilistic approach that has potentially more discriminative power than previous methods
  - We developed an ADMM-based inference algorithm which provides better generality for this line of methods.

Outline

Stochastic Neighbor Topic Model(SNTM)

Generating Procedure

We formalize our stochastic neighbor topic model as follows:

1. For each document i, generate labels $c_i \sim \boldsymbol{\pi}$;
2. Generate $\boldsymbol{\theta}_{1:N}$ from $G(\boldsymbol{\theta}_{1:N}, c_{1:N})$;
3. For each word $w_{d,n}$:
   1. Draw topic assignment
   $$z_{d,n}|\boldsymbol{\theta}_d \sim \text{Multi}(\text{softmax}(\boldsymbol{\theta}_d))^1;$$
   2. Draw word
   $$w_{d,n}|z_{d,n}, \boldsymbol{\beta}_{1:K} \sim \text{Multi}(\boldsymbol{\beta}_{z_{d,n}}).$$

This is a hierarchical model assigning a latent 'topic' to each word $w_{d,n}$. We use softmax function to normalize $\boldsymbol{\theta}_d$, thus avoiding constraint optimization. And we can add dependence between $\boldsymbol{\theta}_d$'s by choosing proper form of prior distribution g.

[1] For $\boldsymbol{x} \in \mathbb{R}^K$, $\text{softmax}(\boldsymbol{x})_j = \frac{e^{x_j}}{\sum_{i=1}^{K} e^{x_k}}$ for $j = 1, 2, \ldots, K$.

Stochastic Neighbor Topic Model(SNTM)

Stochastic Neighbor Prior

We want a prior that pushes $\boldsymbol{\theta}_d$'s from different categories away, and promotes $\boldsymbol{\theta}_d$'s from same category to be closer.
Denote $g(\boldsymbol{\theta}_{1:N}, c_{1:N}) = -\log(G(\boldsymbol{\theta}_{1:N}, c_{1:N}))$. Inspired by SNE which defined a probability distribution over all potential neighbors of each document, we set

$$g(\boldsymbol{\theta}_{1:N}, c_{1:N}) = -\sum_{(i,j):i\sim j} \log p_{ij} + \lambda \sum_{d=1}^{D} \|\boldsymbol{\theta}_d\|^2, \tag{1}$$

where $i \sim j$ means that document i and j share the same label. Following SNE, we set

$$p_{ij} = \frac{\exp(-d(\boldsymbol{\theta}_i, \boldsymbol{\theta}_j)^2)}{\sum_{k\neq i} \exp(-d(\boldsymbol{\theta}_i, \boldsymbol{\theta}_k)^2)}. \tag{2}$$

The function $d(\cdot, \cdot)$ can be any distance function properly measure the similarity between $\boldsymbol{\theta}_d$'s.

Outline

Model Fitting

An Overview of Standard Methods

Sampling-based method:

- Slow convergence.
- Hard to diagnose.

Variational method:

- Making extra assumption that parameters have independent posterior distribution. Good posteriors are hard to find.
- Easy for exponential family distributions; complicated for others.
- Poor generality – variational updates have to be derived case by case.

ADMM is a **Splitting method** for optimization problem of form:

$$\boxed{\min_x f(x) + g(x)}$$

useful when $f + g$ is difficult to solve but f and g are easier to solve separately.

Model Fitting I

ADMM Framework

Write $\boldsymbol{\Theta} = [\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_D]$. With prior G, our objective function is

$$\arg\min_{\boldsymbol{\Theta}} -\sum_{\boldsymbol{z}} \log P(\boldsymbol{w}|\boldsymbol{z}) P(\boldsymbol{z}|\boldsymbol{\Theta}) + g(\boldsymbol{\Theta}). \tag{3}$$

We use an ADMM framework to solve this optimization problem. Denote the log-likelihood function by f, we have

Algorithm 1: Alternating Direction Method of Multipliers

Initialize $\boldsymbol{\Theta}_0$, $\boldsymbol{\Theta}'_0$, ρ and ν

repeat

$$\boldsymbol{\Theta}_{k+1} = \arg\min_{\boldsymbol{\Theta}} -f(\boldsymbol{z}, \boldsymbol{\Theta}) + \frac{\rho}{2}\|\boldsymbol{\Theta} - \boldsymbol{\Theta}'_k + \boldsymbol{u}_k\|^2$$

$$\boldsymbol{\Theta}'_{k+1} = \arg\min_{\boldsymbol{\Theta}'} g(\boldsymbol{\Theta}') + \frac{\rho}{2}\|\boldsymbol{\Theta}_{k+1} - \boldsymbol{\Theta}' + \boldsymbol{u}_k\|^2$$

$$\boldsymbol{u}_{k+1} = \nu(\boldsymbol{\Theta}_{k+1} - \boldsymbol{\Theta}'_{k+1}) + \boldsymbol{u}_k$$

until meet stopping criterion

Then we need to solve the two subproblems.

Model Fitting II

First Subproblem

The first subproblem is

$$\boxed{\arg\max_{\boldsymbol{\Theta}} f(\boldsymbol{z}, \boldsymbol{\Theta}) - \frac{\rho}{2}\|\boldsymbol{\Theta} - \boldsymbol{\Theta}_k^{'} + \boldsymbol{u}_k\|^2.} \tag{4}$$

This is logistic normal topic model where $\boldsymbol{\Theta} \sim \mathcal{N}(\boldsymbol{\Theta}_k^{'} - \boldsymbol{u}_k, \frac{1}{\rho}\boldsymbol{I})$,. We solve this subproblem using variational method with Laplace approximation, following existing literature [2].

[2] Wang, Chong, and David M. Blei. "Variational inference in nonconjugate models." Journal of Machine Learning Research 14.Apr (2013): 1005-1031.

Model Fitting III

Second Subproblem

The second subproblem is

$$\arg\min_{\mathbf{\Theta}'} - \sum_{(i,j):i\sim j} \log p_{ij} + \lambda \|\mathbf{\Theta}'\|^2 + \frac{\rho}{2}\|\mathbf{\Theta}_{k+1} - \mathbf{\Theta}' + \boldsymbol{u}_k\|^2. \tag{5}$$

We solve this subproblem using popular gradient descent methods such as Adagrad or accelerated gradient.

Model Fitting IV

Analysis

Advantages of ADMM:

- Now g can be solved exactly, taking full advantage of SOTA gradient descent methods.
- f is an existing topic model in literature of which inference methods are well studied.
- Better generality.

Outline

Experiment I: 20 News Group Data

Data Description

The 20 newsgroups corpus is a collection of approximately 20,000 newsgroup documents, partitioned almost evenly across 20 different newsgroups. Key features are:

- 20 document categories, single label;
- Use the version from R. F. Correas web-page which contains 8156 distinct words;

We did two experiments on the dataset. In all the experiments we use raw word count data, and do classification by SVM with linear kernel. The maximum iteration number of LDA is set to 30.

Experiment I: 20 News Group Data

Experiment I

We first evaluate the performance of our model on the full 20 classes. We randomly choose 100 documents from each category, and train our model. Denote the labeled data ratio by α, we present the results with $\alpha = 0.3, 0.4, 0.5$, and number of topics $K = 10, 20, 30$.

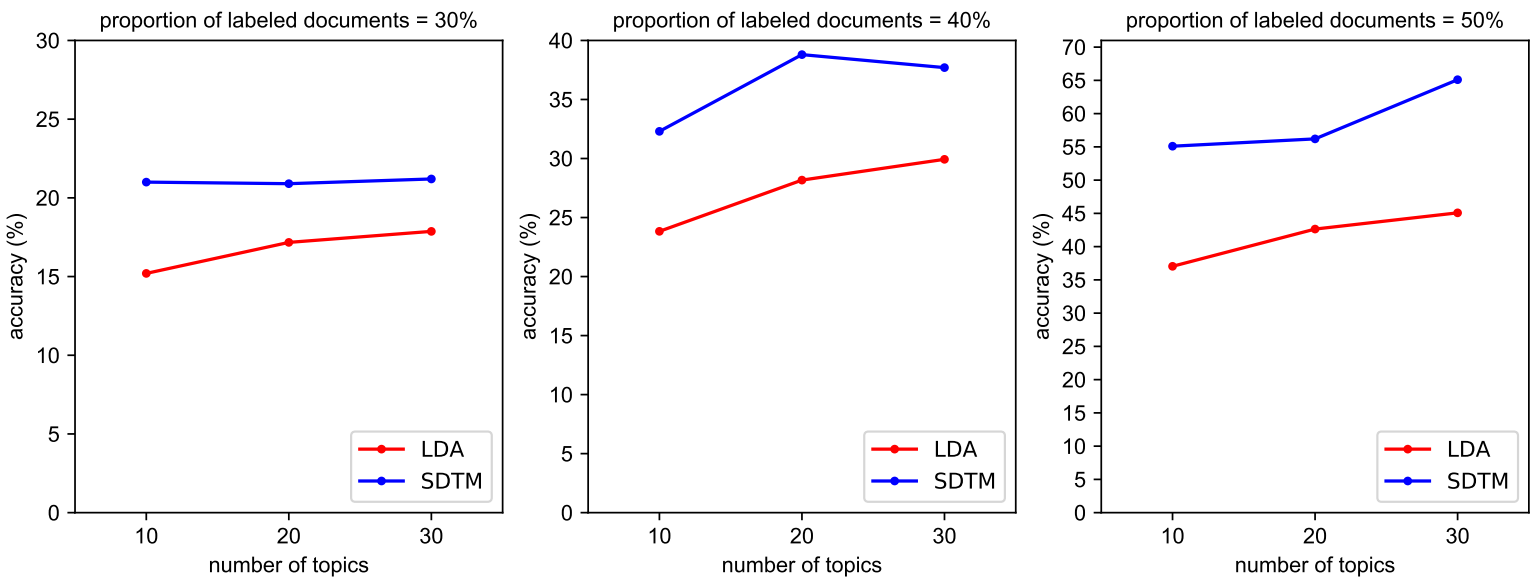

Figure: Clustering accuracy on unlabeled documents of different models

Experiment I: 20 News Group Data

Experiment II

Next we evaluate the performance of our model on 3 subclasses of the full dataset which have similar topics. Specifically we choose *alt.atheism*, *soc.religion.christian* and *talk.religion.misc*. Then we randomly choose 300 documents from each subclass. We also present the results with $\alpha = 1/3, 1/2$, and number of topics $K = 10, 20, 30$.

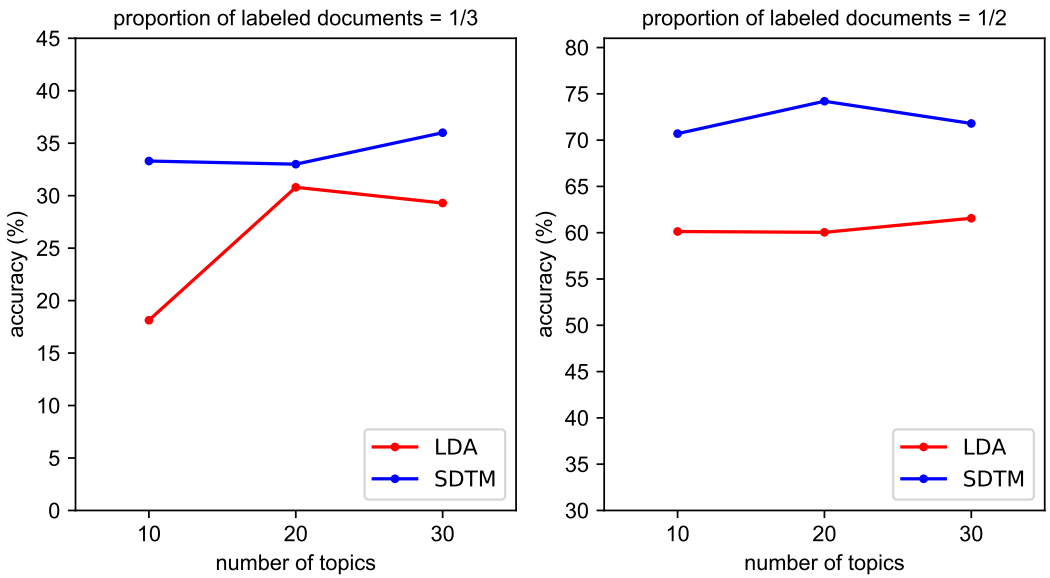

Figure: Clustering accuracy on unlabeled documents of different models

Summary

Next Step

- **Full Evaluation:** Compare with other approaches including graph-based and spectral methods, and evaluate topics qualitatively.
- **Application in Network Data:** Our method can be easily extended to directed network data as a joint modeling framework of community structure and topic distribution. We will try mixed-membership detection and link prediction with our method on citation network data.