TM-LDA: Efficient Online Modeling of Latent Topic Transitions in Social Media

Presentation for the Websci Seminar

Date: 10/10/2012

Presenter: Yitao Li

What will be presented:

- First, a brief overview of LDA
- How TM-LDA fits into the overall picture
- A few technical details of TM-LDA
- Some comment on the presentation of the paper, and more discussions, questions and anwsers about TM-LDA

An Overview of Latent Dirichlet Allocation

- Generative model for classification

  (i.e., an alternative to it would be using a discrminative model, such as logistic regression or SVM)

- Assumes all documents describe a fixed number of topics

- Each document is viewed as a mixture of topics

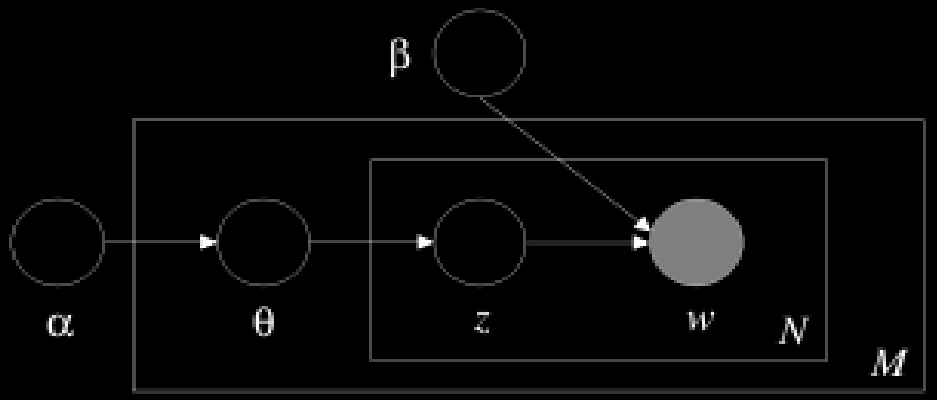
- Each topic has a probablistic distribution of words it generates

An Overview of Latent Dirichlet Allocation (Continued)

- **A few bits of notations:**
  - $V \in \mathbb{N}$: # of possible words, $K \in \mathbb{N}$: # of topics, $M \in \mathbb{N}$: # of documents
    $N \in \mathbb{N}$: # of words in a document
    (simplifying assumption: all documents have the same # of words)
  - $\theta_d \in \{[0,1]\}^K$ for $d \in [1, M] \cap \mathbb{N}$: distribution of topic in the d-th document (hence $\|\theta_d\|_1 = 1$), in addition, $\theta_d \sim \mathrm{Dir}(\alpha)$ for some $\alpha \in \mathbb{R}_+^K$
  - $\phi_k \in \{[0,1]\}^V$ for $k \in [1, K] \cap \mathbb{N}$: distribution of words in topic k (hence $\|\phi_k\|_1 = 1$), in addition, $\phi_k \sim \mathrm{Dir}(\beta)$ for some $\beta \in \mathbb{R}_+^V$
- **The generative process of LDA:**
  1. Choose $N \sim \mathrm{Poisson}(\xi)$.
  2. Choose $\theta \sim \mathrm{Dir}(\alpha)$.
  3. For each of the N words w_n:
     (a) Choose a topic $z_n \sim \mathrm{Multinomial}(\theta)$.
     (b) Choose a word w_n from $p(w_n \mid z_n, \beta)$, a multinomial probability conditioned on the topic z_n.
- **Reference:** D. M. Blei and J. D. Lafferty. Dynamic topic models. In *Proceedings of the 23rd International Conference on Machine learning (ICML)*, 2006.

An Overview of Latent Dirichlet Allocation (Continued)

- **LDA in plate notation:**



N: # of words
M: # of documents

- **The joint distributions:**

$$p(\theta, \mathbf{z}, \mathbf{w} \mid \alpha, \beta) = p(\theta \mid \alpha) \prod_{n=1}^{N} p(z_n \mid \theta) p(w_n \mid z_n, \beta),$$

$$p(\mathbf{D} \mid \alpha, \beta) = \prod_{d=1}^{M} \int p(\theta_d \mid \alpha) \left(\prod_{n=1}^{N_d} \sum_{z_{dn}} p(z_{dn} \mid \theta_d) p(w_{dn} \mid z_{dn}, \beta) \right) d\theta_d.$$

- **Reference:** D. M. Blei and J. D. Lafferty. Dynamic topic models. In *Proceedings of the 23rd International Conference on Machine learning (ICML)*, 2006.

An Overview of Latent Dirichlet Allocation (Continued)

- Objective:

  Estimate hidden variables after observing a document

  i.e., approximate the LHS of this equation:

$$p(\theta, \mathbf{z} \mid \mathbf{w}, \alpha, \beta) = \frac{p(\theta, \mathbf{z}, \mathbf{w} \mid \alpha, \beta)}{p(\mathbf{w} \mid \alpha, \beta)}$$

- Exact inference appears to be difficult

- Possible approaches:

  Variational inference, EM, Gibb sampling …

Beyond LDA: Where TM-LDA comes into the picture

- Objective of TM-LDA:

  Given a trained LDA model (which estimates the topic distribution of each historical document), find (approximately) the topic transition of a sequence of historical documents, and use it to *predict* the topic distribution of a future document

- As an extension of LDA:

  Of course the trained LDA itself can already accomplished this task to some extent with its estimations of all the parameters, but TM-LDA is able to incorporate the additional temporal information (namely, the “trend”) of the topics and perform better in some situations

Beyond LDA: Where TM-LDA comes into the picture (Continued)

- A few bits of notations

1. Topic distribution vector of a document

$$X = \{x \in \mathbb{R}_+^n : ||x||_1 = 1\}$$

Note: the topic distribution is represented as a row vector in the paper

2. Matrix determined by topic distribution vectors of document s through document t:

$$D^{(s,t)} = \begin{bmatrix} d_s \\ \vdots \\ d_t \end{bmatrix}$$

3. Moore-Penrose pseudoinverse: $\dagger$

Beyond LDA: Where TM-LDA comes into the picture (Continued)

- **Objective:**

  Given (m + 1) observed topic distributions $d_1, \ldots, d_{m+1}$ **(which are provided by the LDA)**

  Find the best transition matrix that minimize the prediction error

  More precisely: find $\arg\min_T \|LD^{(1,m)}T - D^{(2,m+1)}\|_F^2$

  where $D^{1,m}, D^{(2,m+1)} \in \mathbb{R}^{m \times n}$, $T \in \mathbb{R}^{n \times n}$, and L is the l^1-normalization matrix $(L = \mathtt{diag}(\|\mathtt{d_1T}\|_1^{-1}, \ldots, \|\mathtt{d_mT}\|_1^{-1})$

  Notation: $D^{(s,t)} = \begin{bmatrix} d_s \\ \vdots \\ d_t \end{bmatrix}$

Beyond LDA: Where TM-LDA comes into the picture (Continued)

- **Answer to** $\arg\min_T \|LD^{(1,m)}T - D^{(2,m+1)}\|_F^2$ **:**

$$T = D^{(1,m)\dagger} D^{(2,m+1)}$$

The derivation essentially follows from 3 mathematical facts:

#1. The Moore-Penrose pseudoinverse of a matrix exists and is unique.
#2. Square matrix A is a projection operator iff A is idempotent.
#3. If all entries of $D^{(1,m)}, D^{(2,m+1)} \in \mathbb{R}^{m \times n}$ are non-negative and each row of $D^{(1,m)}, D^{(2,m+1)}$ has l^1 norm 1, then the entries of each row of $D^{(1,m)} D^{(1,m)\dagger} D^{(2,m+1)}$ sums to 1.
(this is just stating "Theorem 1" of the paper in a more verbose manner)

Incremental Update of Transition Parameters

As value of m increases, the optimal solution

$$T = D^{(1,m)\dagger} D^{(2,m+1)}$$

can be updated via the Sherman-Morrison-Woodbury foruma:

$$\hat{A} = \begin{bmatrix} A \\ U_k \end{bmatrix}, \quad \hat{B} = \begin{bmatrix} B \\ V_k \end{bmatrix} \quad =>$$

$$\hat{T} = (\hat{A}'\hat{A})^{-1}(A'B + U_k' V_k)$$

$$= T + CV_k - C(I + U_k C)^{-1} C'(A'B + U_k' V_k)$$ where

$$C = (A'A)^{-1} U_k' \quad (' \text{ denotes matrix transpose})$$

However, when updating with SMW, one might worry about the condition number of the RHS and also the accumulation of numerical error resulted from successive updates

One can simplify the proof for "Theorem 1" into 2 lines:

Let $e = \begin{bmatrix} 1 & \dots & 1 \end{bmatrix} \in \mathbb{R}^{1 \times n}$, then $D^{(1,m)}e = D^{(2,m+1)}e = e \Rightarrow$

$$D^{(1,m)}D^{(1,m)^\dagger}D^{(2,m+1)}e = D^{(1,m)}D^{(1,m)^\dagger}(D^{(2,m+1)}e) = D^{(1,m)}D^{(1,m)^\dagger}e$$

$$= D^{(1,m)}D^{(1,m)^\dagger}(D^{(1,m+1)}e) = (D^{(1,m)}D^{(1,m)^\dagger}D^{(1,m+1)})e = D^{(1,m)}e = e$$

However, once that proof is out of the way, one will notice a slight mathematical problem immediately following it:

row sum is naturally 1. By adapting the result of Theorem 1 to TM-LDA, we obtain the following result:

$$D^{(1,m)}T^{(0)}\mathbf{e} = D^{(1,m)}D^{(1,m)\dagger}D^{(2,m+1)}\mathbf{e} = \mathbf{e}.$$

In other words, $\|d_i T^{(0)}\|_1 = 1$ for any $i \in \{1, 2, \dots, m\}$.

^ Notice the logic above will not follow through unless all entries of $D^{(1,m)}D^{(1,m)^\dagger}D^{(2,m+1)}$ are non-negative. The paper did not address this issue.

Evaluation of LM-TDA

- Emperically evaluated over Twitter posts
- Accuracy of the topic distribution prediction improved as by 11.4% compared to static LDA model (note that for the purpose of this comparison, static LDA model is given the distribution of words appearing in “future” document as input, while LM-TDA predicts without this data)

Evaluation of LM-TDA (Continued)

- However, in the event of prediction (based on past trend, which, recall, is represented by the topic transition matrix T) being far from actual outcome:
  one might infer some unusual event has altered the previous topic transitioning pattern

  1 question naturally arises: how does one decide whether a prediction failure of the algorithm is caused by unusual event or by something else?

  Answer seems to be subjective in some sense ...

"I am but mad north-north-west: when the wind is southerly I know a hawk from a handsaw."

– William Shakespeare, *Hamlet*, Act II, scene II

Questions?
Answers?

non-constructive criticism will be redirected to /dev/null

Thank you!