Undirected Topic Models

Kuhwan Jeong [1]

[1]Department of Statistics, Seoul National University, South Korea

December, 2016

Introduction

- There has been very little work on developing topic models using undirected graphical models.
- Several authors used RBMs in which word-count vectors are modeled as a Poisson distribution.
- They are unable to properly deal with documents of different lengths.
- Salakhutdinov and Hinton (2007) proposed a Constrained Poisson model that would ensure that the mean Poisson rates across all words sum up to the length of the document.
- The introduced model no longer defines a proper probability distribution over the word counts.

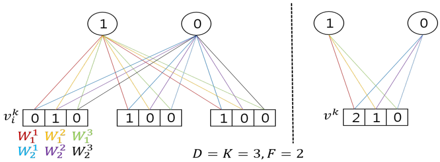
Replicated Softmax

(Hinton and Salakhutdinov, 2009)

- K is the dictionary size and D is the document length.
- Let $\mathbf{V}$ be a $D \times K$ observed binary matrix with $v_i^k = 1$ if i^{th} word takes on k^{th} value of the dictionary.
- Let $\mathbf{h} \in \{0, 1\}^F$ be binary hidden features.

Replicated Softmax

(Hinton and Salakhutdinov, 2009)



$$E(\mathbf{V}, \mathbf{h}) = -\sum_{i=1}^{D}\sum_{j=1}^{F}\sum_{k=1}^{K} W_j^k h_j v_i^k - \sum_{i=1}^{D}\sum_{k=1}^{K} v_i^k b^k - D\sum_{j=1}^{F} h_j a_j$$

$$= -\sum_{j=1}^{F}\sum_{k=1}^{K} W_j^k h_j v^k - \sum_{k=1}^{K} v^k b^k - D\sum_{j=1}^{F} h_j a_j,$$

where $v^k = \sum_{i=1}^{D} v_i^k$ denotes the count for the k^{th} word.

- Scaling up by D is crucial and allows hidden topic units to behave sensibly when dealing with documents of different lengths.

Replicated Softmax

(Hinton and Salakhutdinov, 2009)

- The marginal probability of $\mathbf{V}$ is

$$P(\mathbf{V}) = \frac{1}{Z}\sum_{\mathbf{h}} \exp\{-E(\mathbf{V},\mathbf{h})\} \text{ where } Z = \sum_{\mathbf{V},\mathbf{h}} \exp\{-E(\mathbf{V},\mathbf{h})\}.$$

- The conditional distributions are

$$P(h_j = 1|\mathbf{V}) = \sigma\Big(Da_j + \sum_{k=1}^{K} v^k W_j^k\Big),$$

$$v^1,\ldots,v^K|\mathbf{h} \sim \mathit{Multinomial}(D, p_1,\ldots,p_K)$$

where

$$p_k = \frac{\exp(b^k + \sum_{j=1}^{F} h_j W_j^k)}{\sum_{q=1}^{K}\exp(b^q + \sum_{j=1}^{F} h_j W_j^q)}.$$

Learning

- Given a collection of N documents $\{V_n\}_{n=1}^{N}$, the derivative of log-likelihood w.r.t. W_j^k takes the form:

$$\frac{1}{N}\sum_{n=1}^{N}\frac{\partial \log P(\mathbf{V}_n)}{\partial W_j^k} = \mathbb{E}_{P_{\text{data}}}[v^k h_j] - \mathbb{E}_{P_{\text{Model}}}[v^k h_j]$$

  where $P_{\text{data}}(\mathbf{h}, \mathbf{V}) = P(\mathbf{h}|\mathbf{V})\frac{1}{N}\sum_n \delta_{\mathbf{V}_n}(\mathbf{V})$.

- "Contrastive Divergence" method is used to estimate $\mathbb{E}_{P_{\text{Model}}}[v^k h_j]$.

Over-Replicated Softmax Model

(Srivastava, Salakhutdinov and Hinton, 2013)

- The Over-Replicated Softmax model is a family of two hidden layer Deep Boltzmann Machines (DBM).

- K is the dictionary size and D is the document size.

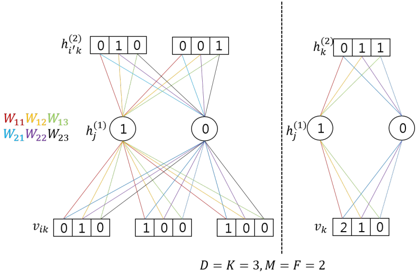
- Let $\mathbf{V}$ be a $D \times K$ observed binary matrix with $v_{ik} = 1$ if i^{th} word takes on k^{th} value of the dictionary.

- Let $\mathbf{h}^{(1)} \in \{0, 1\}^F$ be binary hidden features.

- Let $\mathbf{H}^{(2)}$ be a $M \times K$ observed binary matrix with $h_{mk}^{(2)} = 1$ if m^{th} hidden unit takes on k^{th} value of the dictionary.

Over-Replicated Softmax Model

(Srivastava, Salakhutdinov and Hinton, 2013)



$$E(\mathbf{V}, \mathbf{h}^{(1)}, \mathbf{H}^{(2)}) = -\sum_{j=1}^{F}\sum_{k=1}^{K} W_{jk} h_j^{(1)} \left(v_k + h_k^{(2)}\right) - \sum_{k=1}^{K}\left(v_k + h_k^{(2)}\right) b_k$$

$$-(M+D)\sum_{j=1}^{F} h_j^{(1)} a_j$$

where $v_k = \sum_{i=1}^{D} v_{ik}$ denotes the count for the k^{th} word in the input
and $h_k^{(2)} = \sum_{i'=1}^{M} h_{i'k}^{(2)}$ denotes the count for the k^{th} word in the second hidden layer.

Over-Replicated Softmax Model

(Srivastava, Salakhutdinov and Hinton, 2013)

- The marginal probability of $\mathbf{V}$ is

$$P(\mathbf{V}) = \frac{1}{Z} \sum_{\mathbf{h}^{(1)},\mathbf{H}^{(2)}} \exp\{-E(\mathbf{V}, \mathbf{h}^{(1)}, \mathbf{H}^{(2)})\}.$$

- Given a collection of N documents $\{V_n\}_{n=1}^{N}$, the derivative of log-likelihood w.r.t. W_{jk} takes the form:

$$\frac{1}{N} \sum_{n=1}^{N} \frac{\partial \log P(\mathbf{V}_n)}{\partial W_{jk}} = \mathbb{E}_{P_{\text{data}}}[(v_k+h_k^{(2)})h_j^{(1)}]-\mathbb{E}_{P_{\text{Model}}}[(v_k+h_k^{(2)})h_j^{(1)}]$$

where $P_{\text{data}}(\mathbf{h}^{(1)}, \mathbf{H}^{(2)}, \mathbf{V}) = P(\mathbf{h}^{(1)}, \mathbf{H}^{(2)}|\mathbf{V}) \frac{1}{N} \sum_n \delta_{\mathbf{V}_n}(\mathbf{V})$.

- Exact maximum likelihood learning is intractable.

Learning

- Consider any approximating distribution $Q(\mathbf{h}^{(1)}, \mathbf{H}^{(2)}|\boldsymbol{\mu})$, parameterized by $\boldsymbol{\mu}$, for the posterior $P(\mathbf{h}^{(1)}, \mathbf{H}^{(2)}|\mathbf{V})$.

- Then the log-likelihood has the following variational lower bound :

$$\log P(\mathbf{V}) \geq \sum_{\mathbf{h}^{(1)}, \mathbf{H}^{(2)}} Q(\mathbf{h}^{(1)}, \mathbf{H}^{(2)}|\boldsymbol{\mu}) \log P(\mathbf{h}^{(1)}, \mathbf{H}^{(2)}, \mathbf{V}) + \mathcal{H}(Q).$$

- We approximate $P(\mathbf{h}^{(1)}, \mathbf{H}^{(2)}|\mathbf{V})$ with a fully factorized distribution :

$$Q^{MF}(\mathbf{h}^{(1)}, \mathbf{H}^{(2)}|\boldsymbol{\mu}) = \prod_{j=1}^{F} q_1(h_j^{(1)}|\mu_j^{(1)}) \prod_{i=1}^{M} q_2(h_i^{(2)}|\mu_1^{(2)}, \ldots, \mu_K^{(2)})$$

where q_1 is a Bernoulli distribution and q_2 is a multinomial distribution with a single trial.

Learning

- In this case, the variational lower bound takes a simple form :

$$\log P(\mathbf{V}) \geq \left(\mathbf{v}^T + M\boldsymbol{\mu}^{(2)T}\right)\mathbf{W}\boldsymbol{\mu}^{(1)} - \log Z + \mathcal{H}(Q)$$

where $\mathbf{v} = (v_1, \ldots, v_K)^T$.

- For each training example, we maximize this lower bound w.r.t. $\boldsymbol{\mu}$ for fixed $\mathbf{W}$, which results in the fixed-point equations:

$$\mu_j^{(1)} \leftarrow \frac{\sum_{k=1}^{K} W_{jk}\left(v_k + M\mu_k^{(2)}\right)}{1 + \sum_{k=1}^{K} W_{jk}\left(v_k + M\mu_k^{(2)}\right)},$$

$$\mu_k^{(2)} \leftarrow \frac{\exp\left(\sum_{j=1}^{F} W_{jk}\mu_j^{(1)}\right)}{\sum_{q=1}^{K} \exp\left(\sum_{j=1}^{F} W_{jq}\mu_j^{(1)}\right)}.$$

Learning

1. Randomly initialize $\tilde{\mathbf{v}}^0, \tilde{\mathbf{h}}^{(1),0}, \tilde{\mathbf{H}}^{(2),0}$ and $\mathbf{W}^0$.

2. For $t = 0$ to τ(# of iterations)
   (a) For each training example $\mathbf{V}_n$, $n = 1$ to N
      - Randomly initialize $\boldsymbol{\mu}^{(1)}, \boldsymbol{\mu}^{(2)}$ and run mean-field updates until convergence.
      - Set $\boldsymbol{\mu}_n^{(1)} = \boldsymbol{\mu}^{(1)}$ and $\boldsymbol{\mu}_n^{(2)} = \boldsymbol{\mu}^{(2)}$.

   (b) Obtain a new state $\tilde{\mathbf{V}}^{t+1}, \tilde{\mathbf{h}}^{(1),t+1}, \tilde{\mathbf{H}}^{(2),t+1}$ by running a k-step Gibbs sampler, initialized at $\tilde{\mathbf{V}}^t, \tilde{\mathbf{h}}^{(1),t}, \tilde{\mathbf{H}}^{(2),t}$.

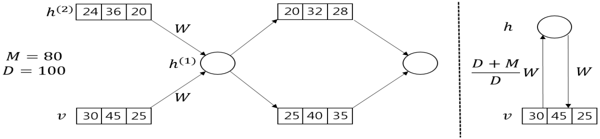
   (c) Update

$$\mathbf{W}^{t+1} = \mathbf{W}^t + \alpha_t \Bigg(\frac{1}{N} \sum_{n=1}^{N} \left(\mathbf{v}_n + \boldsymbol{\mu}_n^{(2)}\right)\left(\boldsymbol{\mu}_n^{(1)}\right)^T - \left(\tilde{\mathbf{v}}^{t+1} + \tilde{\mathbf{h}}^{(2),t+1}\right)\left(\tilde{\mathbf{h}}^{(1),t+1}\right)^T \Bigg)$$

   where $\mathbf{v} = (v_1, \ldots, v_k)$ and $\mathbf{h}^{(2)} = (h_1^{(2)}, \ldots, h_K^{(2)})$.

   (d) Decrease α_t.

Pretraining



- If we were given the initial state vector $\mathbf{h}^{(2)}$, we could train this DBM using one-step contrastive divergence with mean-field reconstructions of both $\mathbf{v}$ and $\mathbf{h}^{(2)}$.
- Since we are not given the initial state, one option is to set $\mathbf{h}^{(2)} = (M/D)\mathbf{v}$.
- Then the conditional distribution $P(h_j^{(1)} = 1|\mathbf{v}, \mathbf{h}^{(2)}) = \sigma\big(\mathbf{W}\big(\mathbf{v} + \mathbf{h}^{(2)}\big)\big)$ becomes $P(h_j^{(1)} = 1|\mathbf{v}) = \sigma\big(\frac{D+M}{D}\mathbf{W}\mathbf{v}\big)$.
- Mean-field reconstructions of $\mathbf{v}$ and $\mathbf{h}^{(2)}$ are

$$\mathbf{v} = (Dp_1, \ldots, Dp_K), \;\; \mathbf{h}^{(2)} = (Mp_1, \ldots, Mp_K),$$

  where $p_k = \exp(h^{(1)T}\mathbf{W}_{,k}) / \big(\sum_{q=1}^{K} \exp(h^{(1)T}\mathbf{W}_{,q})\big)$
- One-step contrastive divergence is exactly the same as training a RBM with the bottom-up weights scaled by a factor of $(D + M)/M$.

Experimental Results

- The average test perplexity per word was estimated as

$$\exp\left(-\frac{1}{N}\sum_{n=1}^{N}\frac{\log P(\mathbf{V}_n)}{D_n}\right).$$

- Replicated Softmax

Data set	Number of docs		K	$\bar{D}$	St. Dev.	Avg. Test perplexity per word (in nats)			
	Train	Test				LDA-50	LDA-200	R. Soft-50	Unigram
NIPS	1,690	50	13,649	98.0	245.3	3576	3391	3405	4385
20-news	11,314	7,531	2,000	51.8	70.8	1091	1058	953	1335
Reuters	794,414	10,000	10,000	94.6	69.3	1437	1142	988	2208

- Over-Replicated Softmax

Perplexities		
Unigram	1335	2208
Replicated Softmax	965	1081
Over-Rep. Softmax ($M = 50$)	961	1076
Over-Rep. Softmax ($M = 100$)	**958**	**1060**

  – All models use 128 topics.