Replicated Softmax:an Undirected Topic Model

Ruslan Salakhutdinov,Geoffrey Hinton-NIPS2009

Yunfei Wang

Department of Computer Science & Technology
Huazhong University of Science & Technology

April 22, 2013

Table of contents

1. Introduction

2. Replicated Softmax:A Generative Model of Word Counts

3. Evaluating Replicated Softmax as a Generative Model

Introduction

Advantages

- Capture the distributed representations provided by a whole set of active features.
- Deal with documents of different lengths.

Replicated Softmax I

Notation

K is dictionary size,D is document size;
observed matrix $V \in \{0,1\}^{K\times D}$,hidden topic features $h \in \{0,1\}^F$;
$v_i^k = 1$ if visible unit i takes on k^{th} value.

Energy of state $\{V, h\}$:

$$E(V,h) = -\sum_{i=1}^{D}\sum_{j=1}^{F}\sum_{k=1}^{K} W_{ij}^k v_i^k h_j - \sum_{i=1}^{D}\sum_{k=1}^{K} v_i^k b_i^k - \sum_{j=1}^{F} h_j a_j \tag{1}$$

where $\theta = \{W, a, b\}$ are the model parameters.

Replicated Softmax II

Probability of visible binary matrix V:

$$P(V) = \frac{1}{Z} \sum_{h} \exp(-E(V,h)) \tag{2}$$

where $Z = \sum_V \sum_h \exp(-E(V,h))$ is normalizing constant.
Conditional distributions are given by softmax and logistic functions:

$$p(v_i^k = 1|h) = \frac{\exp(b_i^k + \sum_{j=1}^{F} h_j W_{ij}^k)}{\sum q = 1^K \exp(b_i^q + \sum_{j=1}^{F} h_j W_{ij}^q)} \tag{3}$$

$$p(h_j = 1|V) = \sigma(a_j + \sum_{i=1}^{D} \sum k = 1^K v_i^k W_{ij}^k) \tag{4}$$

where $\sigma(x) = 1/(1 + \exp(-x))$ is logistic function.

Replicated Softmax III

Ignoring order of words,creating a separate RBM with D softmax units with same set of weights and biases for each document:

$$E(V,h) = -\sum_{j=1}^{F}\sum_{k=1}^{K} W_j^k h_j \hat{v}^k - \sum_{kl=1}^{K} \hat{v}^k b^k - D\sum_{j=1}^{F} h_j a_j \tag{5}$$

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

Bias terms of hidden units are scaled up by length of document,which makes hidden topic units behave sensibly when dealing with document with different sizes.

Replicated Softmax IV

Give a collection of N documents $\{V_n\}_{n=1}^{N}$,the derivation of log-likelihood with respect to W takes the form:

$$\frac{1}{N}\sum_{n=1}^{N}\frac{\partial \log P(V_n)}{\partial W_j^k} = E_{P_{data}}[\hat{v}^k h_j] - \underbrace{E_{P_{model}}[\hat{v}^k h_j]}_{untractable} \tag{6}$$

where $E_{P_{data}}[\cdot]$ and $E_{P_{model}}[\cdot]$denotes expectation with respect to data and model distribution respectively.

$$P_{data}(h,V) = p(h|V)P_{data}(V) \tag{7}$$

where $P_{data}(V) = \sum_n \sigma(V - V_n)/N$ representing the empirical distribution.

What's the challenge here?

Partition function:

$$Z = \sum_{V} \sum_{h} \exp(-E(V, h)) \tag{8}$$

Due to the presence of partition function,model selection,complexity control and exact maximum likelihood learning in RBM's are intractable.

Partition function can be eliminated through division operation when under same probability distributions.

But what can we do we encounter RBM's with different distributions?

Apparently,documents with different size(state of V changes) belong to different distributions.

Simple Importance Sampling

Two distributions $p_A(x) = p_A^*(x)/Z_Z$,$p_B(x) = p_B^*(x)/Z_B$,where $p_A(x)$ is distribution with known Z_A and $p_B(x)$ represents target distribution.

Using a simple importance sampling(IS) to estimate ratio of normalizing constants:

$$\frac{Z_B(x)}{Z_A(x)} = \frac{\int p_B^*(v)d_v}{Z_A} = \int \frac{p_B^*(v)}{p_A^*(v)} p_A(v)d_v = E_{P_A}\left[\frac{p_B^*(v)}{p_A^*(v)}\right] \qquad (9)$$

Drawing independent samples from p_A,the ration can be obtained by a simple Monte Carlo approximation:

$$\frac{Z_B(x)}{Z_A(x)} \approx \frac{1}{M}\sum_{i=1}^{M}\frac{p_B^*(v^i)}{p_A^*(v^i)} = \frac{1}{M}\sum_{i=1}^{M}\omega^i = \hat{r}_{IS} \qquad (10)$$

where $v^i \sim p_A$.Estimator $\hat{r}_{IS}$ will have large variance,unless p_A is near-perfect approximation to p_B.

Annealed Importance Sampling(AIS) I

Define a sequence of intermediate probability distributions:$\{p_0, \cdots, p_K\}$ with $p_0 = p_A$ and $p_K = p_B$.
For each $k = 0, \cdots, K-1$,we must be able to draw a sample v' from v using a Markov chain transition operation $T_k(v'; v)$ that leaves $p_k(v)$ invariant:

$$\int T_k(v'; v) p_k(v) d_v = p_k(v') \tag{11}$$

One general way to define this sequence is to set:

$$p_k(x) \propto p_A^*(x)^{1-\beta_k} p_B^*(x)^{\beta_k} \tag{12}$$

with $0 = \beta_0 < \beta_1 < \cdots < \beta_K = 1$ chosen by user.

Annealed Importance Sampling(AIS) II

```
Input: $p_A, p_B$;
Output: $\omega_{AIS}$;
Initialize $0 = \beta_0 < \beta_1 < \cdots < \beta_K = 1$;
// Generate $v_1, v_2, \cdots, v_K$ as follows;
Sample $v_1$ from $p_0 = p_A$;
for $k = 1$ to $K - 1$ do
    Sample $v_{k+1}$ given $v_k$ using $T_k(V_{k+1}; v_k)$;
end
return $\omega_{AIS} = \frac{p_1^*(v_1)}{p_0^*(v_1)} \frac{p_2^*(v_2)}{p_1^*(v_2)} \cdots \frac{p_{K-1}^*(v_{K-1})}{p_{K-2}^*(v_{K-1})} \frac{p_K^*(v_K)}{p_{K-1}^*(v_K)} = \prod_{k=1}^{K} \frac{p_k^*(v_k)}{p_{k-1}^*(v_k)}$;
```

Algorithm 1: Annealed Importance Sampling(AIS) run

There's no need to compute normalizing constants of any intermediate distributions.After performing M runs of AIS,the ratio of normalizing constants:

$$\frac{Z_B}{Z_A} \approx \frac{1}{M} \sum_{i=1}^{M} \omega_{AIS}^{i} = \hat{r}_{AIS} \tag{13}$$

Applying AIS into RBM's

For Replicated Softmax model with D words,joint distribution over $\{V, h\}$ is defined as:

$$p(V, h) = \frac{1}{Z} \exp(\sum_{j=1}^{F} \sum_{k=1}^{K} W_j^k h_j \hat{v}^k) \tag{14}$$

The sequence of intermediate distributions can be defined as follows:

$$p_k(V) = \frac{1}{Z_k} p^*(V) = \frac{1}{Z_k} \sum_h p_k^*(V, h) = \frac{1}{Z_k} \prod_{j=1}^{F} (1 + \exp(\beta_k \sum_{k=1}^{K} W_j^k \hat{v}^k)) \tag{15}$$