CS838-1 Advanced NLP: Latent Topic Models

Xiaojin Zhu

2007
Send comments to jerryzhu@cs.wisc.edu

Often a document can be thought of as a mixture of a small number of topics (e.g., a news article can be about 60% "finance", 30% "politics", and 10% "war".) We assume a document collection shares the same topics. And the job of latent topic models is to recover topics from document collections. This is interesting because

- The topics can be viewed as *latent semantic concepts*. Psychologists use latent topic models to explain the concept space we operate in, which is called the latent semantic space.
- Each document can be represented at the topics level instead of the word level. This allows two documents which share no common words (one says "buy", the other says "sell") to be measured as similar because they share the "finance" topic.

This is an example of unsupervised learning.

1 Naive Bayes

Recall one can model a document collection by K topics (clusters). Each topic is a multinomial over words, and each document is generated from a unique cluster.

$$p(w) = \sum_{k=1}^{K} p(z=k)p(w|z=k). \tag{1}$$

The parameters $p(z=k), p(w|z=k)$ can be learned with the EM algorithm. This is not a very flexible model, because it assumes "one document, one topic".

2 Probabilistic Latent Semantic Analysis (pLSA)

pLSA assumes *each* document d (with word vector w) to be generated from all topics, with document-specific topic weights. The generative process of pLSA is the following. Given a fixed document collection with n documents, we represent it as a $n \times V$ document-word matrix with entry $c(d, w)$, the count of word type w in document d. At each iteration, one picks a topic $z = 1 \ldots K \sim p(z)$, then picks a document and a word type independent of each other, but both depends on the topic: $d \sim p(d|z = k)$, $w \sim p(w|z = k)$. Generate (add one count of) word w to document d. Repeat until we generate the document-word matrix. Under this process, the probability of picking the cell (d, w) is

$$p(d, w) = \sum_{z=1}^{K} p(z)p(d|z)p(w|z). \tag{2}$$

The model parameters are $\Theta = \{p(z), p(d|z), p(w|z)\}$. We want to find the MLE to maximize the likelihood of the observed document-word matrix,

$$\begin{aligned} & \max\nolimits_\Theta \quad \sum_{d=1}^{n} \sum_{w=1}^{V} c(d, w) \log p(d, w) & (3) \\ = & \max\nolimits_\Theta \quad \sum_{d=1}^{n} \sum_{w=1}^{V} c(d, w) \log \left(\sum_{z=1}^{K} p(z)p(d|z)p(w|z) \right). & (4) \end{aligned}$$

Note z is the hidden variable, and note the sum inside log – these strongly hints EM... Indeed we can apply the EM algorithm here.

$$\begin{aligned} & \textstyle\sum_{d,w} c(d, w) \log p(d, w) & (5) \\ = & \textstyle\sum_{d,w} c(d, w) \log \left(\sum_{z=1}^{K} p(z)p(d|z)p(w|z) \right) & (6) \\ = & \textstyle\sum_{d,w} c(d, w) \log \left(\sum_{z=1}^{K} p(z|d, w, \Theta^t) \frac{p(z)p(d|z)p(w|z)}{p(z|d,w,\Theta^t)} \right) & (7) \\ \geq & \textstyle\sum_{d,w} c(d, w) \sum_{z=1}^{K} p(z|d, w, \Theta^t) \left(\log \frac{p(z)p(d|z)p(w|z)}{p(z|d,w,\Theta^t)} \right) & (8) \end{aligned}$$

Note Jensen's inequality involves $p(z|d, w, \Theta^t)$, which computes the probability of topics separately for each cell, under the current parameters Θ^t. This is exactly the E-step. They can be computed as

$$p(z|d, w, \Theta^t) \propto p(z|\Theta^t)p(d|z, \Theta^t)p(w|z, \Theta^t). \tag{9}$$

Maximizing (8) by setting the gradient to zero amounts to the M-step, which gives

$$\begin{aligned} p(z) & \propto \sum_{d} \sum_{w} c(d, w)p(z|d, w, \Theta^t) & (10) \\ p(d|z) & \propto \sum_{w} c(d, w)p(z|d, w, \Theta^t) & (11) \\ p(w|z) & \propto \sum_{d} c(d, w)p(z|d, w, \Theta^t). & (12) \end{aligned}$$

The E-step and M-step are repeated until convergence.

Once the model is trained, we can look at it in the following way:

- $p(w|z)$ are the topics. Each topic is defined by a word multinomial. Often people find that the topics seem to have distinct semantic meanings.
- From $p(d|z)$ and $p(z)$, we can compute $p(z|d) \propto p(d|z)p(z)$. $p(z|d)$ is the topic wights for document d.

One drawback of pLSA is that it is *transductive* in nature. That is, there is no easy way to handle a new document that is not already in the collection.

3 Latent Dirichlet Allocation (LDA)

LDA too assumes that each document is a mixture of multiple topics, and each document can have different topics weights. Unlike pLSA, LDA is a full generative model and readily generalizes to unseen documents. The LDA generative process is

1. Sample K multinomial distributions (each of size V) $\phi_{1:K}$ from a Dirichlet distribution $Dir(\beta)$, these are the topics. Note β is a parameter vector of length V.
2. For each document
   (a) Sample a topic multinomial (of size K) θ from a Dirichlet distribution $Dir(\alpha)$.
   (b) For each word position
      i. Sample topic index $z \sim \theta$
      ii. Sample a word from the topic $w \sim \phi_z$

The observation is the document collection $w_{1:n}$. The parameters are α and β. The other parameters z, ϕ and θ are hidden variables that will be marginalized out.

The probability of the a topic multinomial ϕ drawn from $Dir(\beta)$ is

$$p(\phi|\beta) = \frac{\Gamma(\sum_{i=1}^{V} \beta_i)}{\prod_{i=1}^{V} \Gamma(\beta_i)} \prod_{i=1}^{V} \phi_i^{\beta_i - 1}. \tag{13}$$

The probability of drawing the K topic multinomials are

$$p(\phi_{1:K}|\beta) = \prod_{j=1}^{K} p(\phi_j|\beta). \tag{14}$$

Similarly,

$$p(\theta|\alpha) = \frac{\Gamma(\sum_{i=1}^{K} \alpha_i)}{\prod_{i=1}^{K} \Gamma(\alpha_i)} \prod_{i=1}^{K} \theta_i^{\alpha_i - 1}. \tag{15}$$

Let $p(z|\theta)$ and $p(w|z,\phi) = p(w|\phi_z)$ be the corresponding multinomial probabilities. The probability of generating the words $w_{1:N}$ for a single document, given θ, ϕ is

$$\prod_{n=1}^{N} \sum_{z_n=1}^{K} p(z_n|\theta)p(w_n|z_n,\phi), \tag{16}$$

where we marginalize out z for each word position. Putting things together, the probability of a single document, after marginalizing out all hidden variables, is

$$p(w|\alpha,\beta) = \int_{\phi_{1:K}} \int_{\theta} p(\theta|\alpha) \left(\prod_{n=1}^{N} \sum_{z_n=1}^{K} p(z_n|\theta)p(w_n|z_n,\phi) \right) d\theta d\phi_{1:K}. \tag{17}$$

Finally, the probability of a document collection $w^{(1)}, \ldots, w^{(M)}$ is

$$p(w^{(1)}, \ldots, w^{(M)}|\alpha,\beta) = \prod_{d=1}^{M} p(w^d|\alpha,\beta). \tag{18}$$