Gibbs sampling in the generative model of Latent Dirichlet Allocation

Tom Griffiths
gruffydd@psych.stanford.edu

Our data consist of words $\mathbf{w} = \{w_1, \ldots, w_n\}$, where each w_i belongs to some document d_i, as in a word-document co-occurrence matrix. For each document we have a multinomial distribution over T topics, with parameters $\theta^{(d_i)}$, so for a word in document d_i, $P(z_i = j) = \theta_j^{(d_i)}$. The jth topic is represented by a multinomial distribution over the W words in the vocabulary, with parameters $\phi^{(j)}$, so $P(w_i|z_i = j) = \phi_{w_i}^{(j)}$. To make predictions about new documents, we → 向量 need to assume a prior distribution on the parameters $\theta^{(d_i)}$. The Dirichlet distribution is conjugate to the multinomial, so we take a Dirichlet prior on $\theta^{(d_i)}$. We then model the distribution over words in any one document as the mixture

$$P(w_i) = \sum_{j=1} P(w_i|z_i = j)P(z_i = j). \tag{1}$$

Blei, Ng and Jordan (2002) gave an algorithm for obtaining approximate maximum-likelihood estimates for $\phi^{(j)}$ and the hyperparameters of the prior on $\theta^{(d_i)}$, terming this procedure Latent Dirichlet Allocation (LDA). Here, we use a symmetric Dirichlet(α) prior on $\theta^{(d_i)}$ for all documents, a symmetric Dirichlet(β) prior on $\phi^{(j)}$ for all topics, and Markov chain Monte Carlo for inference. An advantage of this approach is that we do not need to explicitly represent the model parameters: we can integrate out θ and ϕ, defining model simply in terms of the assignments of words to topics indicated by the z_i. Since we are not performing inference in the Dirichlet hyperparameters, this approach is not necessarily going to lead to the same results as Latent Dirichlet Allocation. In particular, the symmetric prior on topics is likely to mean that there is likely to be little variation in the way topics are used, although the extent to which this is true will be influenced by the choice of α. An empirical Bayes procedure could be used to estimate asymmetric α parameters, resulting in an approach closer to Latent Dirichlet Allocation.

Markov chain Monte Carlo is a procedure for obtaining samples from complicated probability distributions, allowing a Markov chain to converge to the target distribution and then drawing samples from the Markov chain (see Gilks, Richardson & Spiegelhalter, 1996). Each state of the chain is an assignment of values to the variables being sampled, and transitions between states follow a simple rule. We use Gibbs sampling, where the next state is reached by sequentially sampling all variables from their distribution when conditioned on the current values of all other variables and the data. We will sample only the assignments of words to topics, z_i. Our complete probability model is

$$\begin{aligned} w_i|z_i, \phi^{(z_i)} &\sim \text{Discrete}(\phi^{(z_i)}) \\ \phi &\sim \text{Dirichlet}(\beta) \\ z_i|\theta^{(d_i)} &\sim \text{Discrete}(\theta^{(d_i)}) \\ \theta &\sim \text{Dirichlet}(\alpha) \end{aligned}$$

So the conditional posterior distribution for z_i is given by

$$P(z_i = j|\mathbf{z}_{-i}, \mathbf{w}) \propto P(w_i|z_i = j, \mathbf{z}_{-i}, \mathbf{w}_{-i})P(z_i = j|\mathbf{z}_{-i}), \tag{2}$$

where $\mathbf{z}_{-i}$ is the assignment of all z_k such that $k \neq i$. This is an application of Bayes' rule, where the first term on the right hand side is a likelihood, and the second a prior.

The parameters θ and ϕ do not appear in the above expression because we can obtain conditional probabilities for the z_i that depend only on $\mathbf{z}_{-i}$ and $\mathbf{w}$ by integrating over the parameter values that arise in each of the terms on the right hand side of the equation. For the first term, we have

$$P(w_i|z_i = j, \mathbf{z}_{-i}, \mathbf{w}_{-i}) = \int P(w_i|z_i = j, \phi^{(j)})P(\phi^{(j)}|\mathbf{z}_{-i}, \mathbf{w}_{-i})\ d\phi^{(j)}, \tag{3}$$

where $\phi^{(j)}$ is the multinomial distribution over words associated with topic j, and the integral is over all such distributions. We can obtain the rightmost term from Bayes' rule

$$P(\phi^{(j)}|\mathbf{z}_{-i}, \mathbf{w}_{-i}) \propto P(\mathbf{w}_{-i}|\phi^{(j)}, \mathbf{z}_{-i})P(\phi^{(j)}). \tag{4}$$

Since $P(\phi^{(j)})$ is Dirichlet(β) and conjugate to $P(\mathbf{w}_{-i}|\phi^{(j)}, \mathbf{z}_{-i})$, the posterior distribution $P(\phi^{(j)}|\mathbf{z}_{-i}, \mathbf{w}_{-i})$ will be Dirichlet($\beta + n_{-i,j}^{(w)}$), where $n_{-i,j}^{(w)}$ is the number of instances of word w assigned to topic j, not including the current word. The involvement of $\mathbf{z}_{-i}$ in this conditional probability is to partition the words into sets that are assigned to the different topics. Only the words assigned to topic j will influence the posterior distribution of $\phi^{(j)}$.

Since the first term on the right hand side of Equation 3 is just $\phi_{w_i}^{(j)}$, we can complete the integral to obtain

$$P(w_i|z_i = j, \mathbf{z}_{-i}, \mathbf{w}_{-i}) = \frac{n_{-i,j}^{(w_i)} + \beta}{n_{-i,j}^{(\cdot)} + W\beta}, \tag{5}$$

where $n_{-i,j}^{(\cdot)}$ is the total number of words assigned to topic j, not including the current one. This is the predictive distribution for a multinomial-Dirichlet model, and can only be obtained here because $\mathbf{z}$ is known. β is a hyperparameter that determines how heavily this empirical distribution is smoothed, and can be chosen to give the desired resolution in the resulting distribution.

We can find $P(z_i = j|\mathbf{z}_{-i})$ in the same way. Integrating over the multinomial distribution over topics for the document from which w_i is drawn, specified by $\theta^{(d_i)}$, we obtain

$$\begin{aligned} P(z_i = j|\mathbf{z}_{-i}) &= \int P(z_i = j|\theta^{(d_i)})P(\theta^{(d_i)}|\mathbf{z}_{-i})\ d\theta^{(d_i)} \\ &= \frac{n_{-i,j}^{(d_i)} + \alpha}{n_{-i,\cdot}^{(d_i)} + K\alpha}. \qquad (6) \end{aligned}$$

The result follows from the choice of a Dirichlet(α) prior for $\theta^{(d_i)}$. Here, $n_{-i,j}^{(d_i)}$ is the number of words from document d_i assigned to topic j, not including the current one, and $n_{-i,\cdot}^{(d_i)}$ is the total number of words in document d_i, not including the current one.

Putting together the results in Equations 5 and 6, we obtain the conditional probabilities

$$P(z_i = j|\mathbf{z}_{-i}, \mathbf{w}) \propto \frac{n_{-i,j}^{(w_i)} + \beta}{n_{-i,j}^{(\cdot)} + W\beta} \frac{n^{(d_i)} + \alpha}{n_{-i,\cdot}^{(d_i)} + T\alpha}. \qquad (7)$$

The Monte Carlo algorithm is then straightforward. The z_i are initialized to values between 1 and T, determining the initial state of the Markov chain. The chain is then run for a number of iterations, each time finding a new state by sampling each z_i from the distribution specified by Equation 7. After enough iterations for the chain to approach the target distribution, the current values of the z_i are recorded. Subsequent samples are taken after an appropriate lag, to ensure that their autocorrelation is low.

References

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2002). Latent Dirichlet allocation. In *Advances in Neural Information Processing Systems 14*.

Gilks, W., Richardson, S., & Spiegelhalter, D. J. (1996). *Markov Chain Monte Carlo in Practice.* Chapman & Hall, Suffolk.