Blei, Latent Dirichlet Allocation

Thomas W. Jones

6/10/2018

In 2003, David Blei, Andrew Ng, and Michael Jordan introduced *Latent Dirichlet Allocation* (LDA) as a method to model words in a corpus of documents. These words are sampled from latent variables called topics. In this paper, Blei et al. introduce their model as a generative model for documents, compare it to related models, describe an estimation method for LDA, and give examples of how LDA may be used for empirical tasks.

Background models

LDA has several predecessor models. The most famous (and still widely used) is called latent semantic analysis (LSA) or latent semantic indexing (LSI) (Deerwester et al. 1990). (This is the same model with two different names.) LSI performs a single-value decomposition on a TF-IDF-weighted document term matrix (DTM). This single-value decomposition finds a linear subspace of the original DTM. The features of this linear subspace may be interpreted as "topics" and the documents' coordinates within this subspace may be interpreted as the distribution of topics related to the documents. LSI is non-probabilistic and the relative magnitudes of its parameters are difficult to interpret and compare.

A single value decomposition decomposes a matrix, D (our DTM), into three components:

$$D = U\Sigma V \tag{1}$$

The rows of U indicate some distribution of topics related to documents. The rows of V indicate some distribution of words related to topics. The values that the entries take range from $-\infty$ to ∞. It's not obvious what negative values mean (negative correlation?) and one cannot compare the parameters of one model directly to the parameters of another. Probabilistic models have neither of these problems.

A significant step forward came from Hoffman (1999) with probabilistic latent semantic indexing (pLSI). pLSI models each word as being sampled from hierarchical multinomial distributions. pLSI gains an advantage in interpretability. The distribution of words within topics is interpretable as the probability of a word being sampled from a topic. The distribution of topics within documents is interpreted as the probability of a randomly-sampled word in a document having been sampled from a given topic. (More colloquially, the proportion of words in the document having come from each topic.)

pLSI decomposes D into two probability matrices.

$$D \sim \Theta\Phi \tag{2}$$

The rows of Θ, denoted θ_d are probability distributions relating topics to documents, as above. The rows of Phi are probability distributions relating words to topics, as above. Each of these rows sums to one. Their magnitudes are interpretable using probability theory. Not stated by Blei et al., but LDA has these same relationships. The method of estimation is different, however.

Blei et al. cite three limitations to pLSI. First, the number of parameters grows linearly with the number of documents. The model does not scale for large corpora. pLSI has $kV + kM$ parameters for V words and M documents. LDA has only $k + kV$ parameters. Second, it is unclear how to assign probabilities to a document outside of the training set. Third, empirically estimating pLSI with the EM algorithm often overfits training data. I am not sure that Blei et al. are right about the first issue. I am positive they are wrong about the second. But I'm not surprised that the third is true.

To the first point, it's not clear to me how LDA doesn't have to estimate k parameters for every document. Certainly when using it, you get $kM + kV$ parameters as output. To the second point, one must only use Bayes' rule to get predictions for held out documents. pLSI returns $P(word|topic)$ and $P(topic|document)$. Some algebra and Bayes' rule give us $P(topic|word)$, which can be multiplied to a normalized vector of word counts for each document, resulting in estimates of $P(topic|document)$ for new documents. Moreover, I have already implemented this method of prediction in `textmineR` as most packages for LDA don't even provide a good method for predicting on new documents, in spite of its theoretical ease.

Blei et al. also cite two simpler predecessor models. The simplest is the unigram model. Under the unigram model, words in every document are drawn independently from a single multinomial distribution. In this case, a whole corpus has only one topic. The mixture of unigrams model is similar to document clustering. Each document has only one topic from which its words are sampled, but there are multiple topics in the corpus.

Generative model

Blei et al. formulate LDA as a generative process. For a single process this is

1. Choose $N \sim Poisson(\zeta)$
2. Choose $\theta \sim Dirichlet(\alpha)$
3. For each of the N words,

a. Choose a topic $z_n \sim Multinomial(\theta)$
b. Choose a word w_n from $p(w_n|z_n, \beta)$ which is a multinomial distribution of words over topics.

Here, N is the total number of words in the document, θ is the proportion of topics in the document, α is the Dirichlet prior for topics over documents, and β relates words to topics but its meaning wasn't clear to me.

Note that their statement of the multinomial distribution is also confusing since the multinomial has two parameters, number of draws and probability of each category. They only state the probability, so I assume that they mean the categorical distribution, a multinomial of only a single draw.

In this formulation, Blei et al. do not formulate the distribution of words over topics in the way that has since become standard.

The formulation that has become standard is from Wikipedia, below:

1. Choose Φ such that each of its k rows is $\boldsymbol{\phi} \sim Dirichlet(\boldsymbol{\beta})$
2. Choose $\boldsymbol{\theta_d} \sim Dirichlet_k(\boldsymbol{\alpha})$
3. Choose $\mathbf{z_{d,n}} \sim Categorical(\boldsymbol{\theta_d})$
4. Choose $\mathbf{w_{d,n}} \sim Categorical(\boldsymbol{\phi_{z_{d,n}}})$

Estimation

Blei et al. use Bayesian variational expectation maximization to estimate LDA with data. As above, this is unconventional in comparison to what has become standard for LDA (Gibbs sampling) and what would be a more natural choice in Bayesian modeling more broadly (again, Gibbs sampling). But they do specify a formula for estimating LDA. Unfortunately, this is not the formula I need to implement Gibbs sampling.

Some additional observations

To understand the model, I need a clear formulation of the likelihood and priors. This paper does not state these explicitly. And a search through the web indicates that most papers don't either. Blei et al.'s paper also uses non-conventional notation and leaves some things poorly defined. For example, stating, as above, $p(w_n|z_n, \beta)$ represents a multinomial distribution without specifying its parameters is quite confusing.

I'm going to tweak my reading list following this to emphasize getting the model formulation right. From there, deriving the Gibbs formulas is (literally) algebraic. (Ok, I might have to take an integral too.)

Works cited

Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., & Harshman, R. (1990). Indexing by latent semantic analysis. Journal of the American society for information science, 41(6), 391.

Hofmann, T. (1999, July). Probabilistic latent semantic analysis. In Proceedings of the Fifteenth conference on Uncertainty in artificial intelligence (pp. 289-296). Morgan Kaufmann Publishers Inc..