A Hierarchical Topic Model

Jason D. M. Rennie
jrennie@gmail.com

September 28, 2005

Abstract

We describe a hierarchical topic model. We assume that there are various levels of specificity in a document collection. For example, a collection of mailing list posts can be organized according to sentence, paragraph, post and thread. We describe a model that encourages the largest changes in topics to occur at the highest levels of the hierarchy (e.g. post, thread).

1 The Basic Model

We assume that there is hierarchical structure to the document collection. We assume that there are a set of fundamental units which cannot be broken down any further, such as sentences. We represent each sentence as a bag-of-words, disgarding sequence information and simply storing its vector of word counts. We form the matrix Y out of the sentence word count vectors (one vector per row). We assume that each word count vector (row of Y) is generated by a multinomial model. We store the multinomial natural parameters in another matrix, X; each row of X contains the natural parameters for the corresponding row of Y. X and Y are the same size. Note that entries of X are real values; entries of Y are non-negative integers. Let X_{ij} (Y_{ij}) index the i^{th} row, j^{th} column entry of X (Y). The negative log-likelihood of the data is

$$-\log P(Y|X) = \sum_{i=1}^{n} \sum_{j=1}^{d} Y_{ij} \left[\log \left(\sum_{j'} \exp(X_{ij'}) \right) - X_{ij} \right] + g(Y), \qquad (1)$$

where $g(Y)$ is a term that only depends on the data and is unaffected by changes in parameter values.

2 The Hierarchy

The sentences are arranged in a hierarchy; there are $n_0 \equiv n$ sentences, which make up the bottom or 0^{th} level of the hierarchy. Each of the sentences has a "parent," a level-one node to which it is linked; there are $n_1 < n_0$ level-one nodes. If there is a second level, then each level-one node has a parent, a level-two node to which it is linked; there are $n_2 < n_1$ level-two nodes. The hierarchy may have any number of levels. Let $m+1$ be the number of "levels," including the sentence (0^{th}) level. We assume that there is only one level-m node, which we call the "root" node.

We define a "parent of" function, p, that returns the parent of each (non level-m) node in the hierarchy[1]. Furthermore, we use superscript notation, p^x, to indicate recursive application of p. If i is the index of a sentence, then $p(i)$, $p^2(i)$, and $p^3(i)$ return the indeices of its parent, grandparent, and great-grandparent respectively, where the level-k nodes are indexed $\{1, 2, \ldots, m_k\}$ according to their order in the document collection. Note that all sentences have the same level-m parent, $p^m(i) = 1\ \forall i$.

3 The Model

The rows of X, which are the multinomial natural parameters for the word counts (rows of Y), are not independent of each other. Here we describe how the rows of X are interdependent.

In earlier writing [1], we discussed an embedding of the multinomial parameter vector as a vector in Euclidean space, $\mathbb{R}^d$. To cause the rows of X to be tied, we embed the hierarchy as a set of Euclidean vectors and require that each row of X is a sum of vectors corresponding to its path through the hierarchy. Each node in the hierarchy has a corresponding vector. We arrange the vectors of level l as the rows of the matrix W^l. Matrix W^l has n_l rows. Each row of X is a sum of vectors from W,

$$X_{i,j} = W^0_{i,j} + \sum_{k=1}^{m} W^k_{p^k(i),j}, \tag{2}$$

where $W^k_{p^k(i),j}$ designates the entry of W^k in the $p^k(i)^{\text{th}}$ row and j^{th} column. In fact, each node in the hierarchy can be characterized by a point in Euclidean space which is the sum of the vectors corresponding to the node and all of its parents. If paragraphs are the level immediately above sentences in the hierarchy, then this point for each paragraph can be thought of as the "focus" for sentences contained in that paragraph. In fact, for each node in the hierachy, it's corresponding location is the location of its parent, plus its vector.

4 Regularization

If we simply minimize negative log-likliehood of the data, the hierarchical structure would have no impact on learning. To encourage the model to make use of the hierarchical structure, we add a penalty which is the trace norm of the

[1] Note that we use capital P for probability functions and lower-case p for the "parent of" function.

stacked W^i matrices. The trace norm is equal to the sum of the singular values of the stacked W^i matrices, or it is equal to the sum of the axis radii of the elipsoid that contains all of the W vectors. The penalty encourages vectors to be relatively short and to point in similar directions. Only when the data strongly shows evidence to the contrary should a vector be long or point in an unusual direction. Let W be the stacked W^i matrices,

$$W = \begin{bmatrix} M^0 \\ M^1 \\ \vdots \\ M^m \end{bmatrix}. \tag{3}$$

Then, our minimization objective is

$$J = \lambda \|W\|_{\text{tr}} + \sum_{i=1}^{n} \sum_{j=1}^{d} Y_{ij} \left[\log \left(\sum_{j'} \exp(X_{ij'}) \right) - X_{ij} \right], \tag{4}$$

where X is defined as above, $\|W\|_{\text{tr}}$ is the trace norm of W and λ is the regularization which trades-off between fit to the data and small trace norm.

References

[1] J. D. M. Rennie. Topics. http://people.csail.mit.edu/∼jrennie/writing, September 2005.