Topics

Jason D. M. Rennie
jrennie@gmail.com

September 21, 2005

Abstract

A "topic" is a notion that is commonly used in work on text classification and information retrieval. A topic defines the rate at which words from the vocabulary are expected to occur. A document can be defined in terms of its topics so that the rate of occurrence in the document is (approximately) equal to a convex combination of its "topics." In this work, we define two characteristics of a topic that allow better understanding of the nature of a topic. We then describe a new text dimensionality reduction technique that makes use of these new characterizations.

1 Introduction

Given a vocabulary of words indexed $\{1, 2, \ldots, n\}$, we define a topic as a distribution over those words. I.e. a topic is a non-negative, real-valued vector, $\vec{\mu}$, $\mu_i \geq 0 \ \forall i$, the elements of which sum to one, $\sum_i \mu_i = 1$. A topic specifies a rate of occurrence for each of the words in the vocabulary. Every non-empty document has a representation as a topic.

One methodology for a reduced dimensionality representation of a collection of documents is to find a set of topics (fewer than the number of documents) such that each document's topic representation is "close" to a convex combination of the chosen set of topics. This technique limits the number of topics that are allowed to represent the document collection.

2 Characterizing a Topic

Here, we define some ways to characterize a topic. A goal of these definitions is to give us the ability to characterize the *strength* of a set of topics, not just their number. First, we introduce an alternate parameterization of a topic.

The non-negative, sum-to-one topic parameterization ($\vec{\mu}$) we have introduced is known as the *mean* parameterization because the mean frequency of word i in any document generated from the topic is p is proportional to μ_i. This

representation provides good intuition, but due to the non-negativity and sum-to-one constraints, it is a somewhat unnatural representation with which to work.

All exponential family models have what is known as a natural parameterization in which the log-likelihood is a linear function of the data. We discuss this in §2 of [2]. In the case of a topic, which is the mean parameterization of a multinomial or unigram, there is a family of natural parameter vectors for each mean parameter vector. The degree of freedom is due to the sum-to-one constraint in the mean parameterization. Given a mean parameter vector, $\vec{\mu}$, a corresponding natural parameter vector is $\vec{\theta}$ such that $\theta_i = \log \mu_i + c$. The translation back to mean parameters is $\mu_i = \frac{\exp(\theta_i)}{\sum_i \exp(\theta_i)} = \exp(\theta_i - c)$.

For the purposes of characterizing a topic, we choose the natural parameterization with smallest Euclidean norm. It is easy to show that this corresponds to a choice of $c = -\frac{1}{n}\sum_i \log \mu_i$. In other words, we choose the corresponding natural parameter vector so that the parameter values sum to zero. Note that the uniform topic ($\mu_i = 1/n \; \forall i$) translates to a vector of all zeros in the natural parameter space. Furthermore, we can think of this as a natural embedding of a topic as a vector in Euclidean space. Thus, we can speak of a topic's *length* and *direction*. We define the length of a topic $\vec{\mu}$ to be

$$\text{length}(\vec{\mu}) = \sqrt{\sum_i \left(\log \mu_i - \frac{1}{n} \sum_{i'} \log \mu_{i'} \right)^2} \tag{1}$$

Let $f(\vec{\mu}) = \log \vec{\mu} - \frac{1}{n}\sum_i \log \mu_i$ (where log is applied element-wise to $\vec{\mu}$) be the embedding of the topic $\vec{\mu}$. Then, $\text{length}(\vec{\mu})$ is simply $\|f(\vec{\mu})\|_2$ (the L_2 norm of $f(\vec{\mu})$). Similarly, we define the direction of a topic $\vec{\mu}$ to be

$$\text{direction}(\vec{\mu}) = \frac{f(\vec{\mu})}{\text{length}(f(\vec{\mu}))}. \tag{2}$$

In other words, the direction of a topic is simply its Euclidean projection scaled to unit length.

These definitions suggest an alternate methodology for finding a reduced representation of a collection of documents. Instead of limiting the number of topics, we limit the sum of lengths of the topics. Then, a document is a linear combination of topics where the vector of linear weights form a unit-length vector. In other words, the goal is to find orthogonal matrices U and V and a diagonal matrix S so that USV^T approximates well natural parameter topic representations of the documents (one document per row). V is the topic matrix; each column corresponds to a topic direction. S is the length matrix, giving the lengths of each of the topics. U specifies the unit-length weights used to combine the topics.

What we have just described is, in fact, a trace norm constraint (see [1] for discussion). The topic lengths correspond to singular values. USV^T is a matrix of multinomial (natural) parameters where each row corresponds to a single document in the collection. I.e. each row of this matrix can be thought of as an

approximation to the natural parameter representation of the document topics. An advantage of using the trace norm as a constraint as opposed to simply limiting the number of topics is that it provides a continuum of constraints instead of a discrete set of constraints.

3 The Topic Length Model

Here we describe an instantiation of a topic model that makes use of a "sum of topic lengths" or trace norm constraint. We would like to find natural parameter topic representations of our documents that are somehow "close" to the true documents. We do this by treating each document as having been generated by a multinomial; we minimize negative log-likelihood. Instead of finding the best fit parameter matrix for various trace norm constraints, we minimize a weighted sum of trace norm and negative log-likelihood of the data. This allows us to explore the same set of solutions that we would have found had we explicitly imposed the trace norm constraint. Let Y be the matrix of term frequencies corresponding to our document collection, one row per document. Our minimization objective is

$$\min_X \lambda \|X\|_{\text{tr}} + \sum_{ij} X_{ij} Y_{ij}, \tag{3}$$

where X is the matrix of multinomial natural parameters that we learn, each row corresponding to a single document.

References

[1] J. D. M. Rennie. Equivalent ways of expressing the trace norm of a matrix. http://people.csail.mit.edu/∼jrennie/writing, September 2005.

[2] J. D. M. Rennie. Mixtures of multinomials. http://people.csail.mit.edu/∼jrennie/writing, September 2005.