Understanding clustering of textual documents

March 31, 2016

Contents

Abstract

1 Introduction

2 Gap statistics

3 Distribution for tf-idf vectors

To compute the gap statistic, we need a random sample from the distribution with no clusters. Now we would want this distibution to be similar of the distribution of tf-idf vectors. It has been argued[1] that the Dirichlet compound multinomial (DCM) distribution models the bag-of-words in textual documents. As the name suggests, DCM is a compound of Dirichlet and multinomial distributions. We will make this notion clear in the following. Before we delve into DCM, we will talk about multinomal and Dirichlet distributions.

3.1 Multinomial Model

Multinomial distribution is a multivariate generalization of the binomial distribution. It models the counts of each of the p outcomes in n trials. e.g. consider the experiment of throwing a dice n times, this has 6 possible outcomes. A sample drawn from $Mult(n, 6)$ is of the form $(x_1, x_2, \ldots, x_6)$ where each x_i is the count of number of times the outcome was i. As is pretty evident from the description that the count vectors for words in a document can be modelled by such a distribution. More specifically, a multinomial distribution can be used to model the probability of observing a given vector of word counts in a document. If θ_w is the probability of generating a word w, the the prob of generating a document x can be expressed as -

$$p(x|\theta) = \frac{n!}{\prod_{w=1}^{W} x_w!} \prod_{w=1}^{W} \theta_w^{x_w} \tag{1}$$

where $\theta = (\theta_1, \theta_2, \ldots, \theta_W)$, x_w is the frequency of word w in the document, W is the size of the vocabulary, and $n = \sum x_w$.
θ is the parameter of the distribution, which can be approximated as follows[1]:

$$\hat{\theta_w} = \frac{\sum_(d = 1)^D x_{dw}}{\sum_w = 1^W x_{dw}} \tag{2}$$

where d is the document number and D is the total number of documents in the corpus.

3.2 Dirichlet Model

Dirichlet distribution is a continous distribution which is the multivariate generalization of beta distribution. It is parameterized by a vector α containing real numbers. The Dirichlet distribution has the probability density function -

$$p(\theta|\alpha) = \frac{\Gamma(\sum_{w=1}^{W} \alpha_w)}{\prod_{w=1}^{W} \Gamma(\alpha_w)} \prod_{w=1}^{W} \theta_w^{\alpha_w - 1} \tag{3}$$

Dirichlet distribution can be used to model probabilities i.e. as a distributions over distributions. Thus θ is the above expression is a vector of probabilities.

3.3 Dirichlet Compound Multinomial (DCM) model

A DCM model of textual documents can be constructed hierarchically, exploiting the fact that Dirichlet distribution can describe probabilities. *In DCM model, the word-count vector for each document can be generated by a multinomial distribution whose parameters are generated by the Dirichlet distribution.* Thus to generate a sample from DCM,

1. Draw a sample from the Dirchlet distribution with parameter α. This gives θ.

2. With this θ as parameter, get a sample from multinomial distribution to get the count vector $x = (x_w)$.

This gives us word-count vectors, to generate sample representing tf-idf vectors, we apply appropriate transformation to the vectors thus generated.
We followed the above steps to generate samples of vectors which are well-suited for the cluster-analysis.
Mathematically, for DCM model

$$p(x|\alpha) = \int_{\theta} p(x|\theta)p(\theta|\alpha)d\theta \tag{4}$$

The only parameter of the DCM model is α which can be approximated as

$$\hat{\alpha_w} \simeq \frac{1}{D}\sum_{d=1}^{D} I(x_{dw} > 1) \tag{5}$$

Since $x_{dw} = 0$ for most of the words in the vocabulary, $\alpha_w << 1$ for most words. Experiments involving training on popular data sets show the value to be less that 0.1.

4 Computation of Gap statistics

5 Conlusion

References

[1] Rasmus E. Madsen, David Kauchak, Charles Elkan, Modeling Word Burstiness Using the Dirichlet Distribution, Proceedings of the 22 nd International Conference on Machine Learning, Bonn, Germany, 2005 (Online version)

[2] Charles Elkan, Deriving TF-IDF as a Fisher Kernel, M. Consens and G. Navarro (Eds.): SPIRE 2005, LNCS 3772, pp. 296301, 2005. (Online version)