On Finding the Natural Number of Topics with Latent Dirichlet Allocation: Some Observations

R. Arun, V. Suresh, C.E. Veni Madhavan, and M. Narasimha Murty

Department of Computer Science and Automation, Indian Institute of Science, Bangalore 560 012, India
{arun_r,vsuresh,cevm,mnm}@csa.iisc.ernet.in

Abstract. It is important to identify the "correct" number of topics in mechanisms like Latent Dirichlet Allocation(LDA) as they determine the quality of features that are presented as features for classifiers like SVM. In this work we propose a measure to identify the correct number of topics and offer empirical evidence in its favor in terms of classification accuracy and the number of topics that are naturally present in the corpus. We show the merit of the measure by applying it on real-world as well as synthetic data sets(both text and images). In proposing this measure, we view LDA as a matrix factorization mechanism, wherein a given corpus C is split into two matrix factors M_1 and M_2 as given by $C_{d*w} = M1_{d*t}$ x Q_{t*w}. Where d is the number of documents present in the corpus and w is the size of the vocabulary. The quality of the split depends on "t", the right number of topics chosen. The measure is computed in terms of symmetric KL-Divergence of salient distributions that are derived from these matrix factors. We observe that the divergence values are higher for non-optimal number of topics – this is shown by a 'dip' at the right value for 't'.

Keywords: LDA Topic SVD KL-Divergence.

1 Introduction

Topic Modelling is a widely used technique in information retrieval, data mining etc. The idea behind it is the fact that a small number of latent topics are enough to effectively represent a large corpus. As this is often the case with real world corpus such as text which have a large vocabulary, such models have proved to be very effective. However finding the right number of latent topics in a given corpus has remained an open ended question. Almost all previous methods including Latent Semantic Analysis [1], Probabilistic Latent Semantic Analysis [2], Latent Dirichlet Allocation [3], Non-Negative Matrix Factorization [4] which try to model the latent topics either as probability distributions or as a set of basis vectors in the topic space make the implicit assumption that the number of topics is known beforehand. While estimating the right number of topics for a small image or text corpus might seem easy, it becomes unreasonable to guess the same when the corpus size is huge. However the accuracy of all of the above mentioned methods is senstive to the number of topics.

M.J. Zaki et al. (Eds.): PAKDD 2010, Part I, LNAI 6118, pp. 391–402, 2010.

In this paper, we consider the Latent Dirichlet Allocation (LDA) [3] model as the basis for our work. We view LDA as a matrix factorization method which factorizes a document-word frequency matrix M into two matrices $M1$ and $M2$ of order $T*W$ and $D*T$ respectively where T is the number of topics and W is the size of the vocabulary of the corpus. We propose a new measure that computes the symmetric Kullback-Leibler divergence of the Singluar value distributions of matrix $M1$ and the distribution of the vector $L*M2$ where L is a $1*D$ vector containing the lengths of each document in the corpus. We show that under certain conditions these distributions are comparable and these conditions are expected to determine the 'right' number of topics. We also present empirical results that indicate that the proposed measure dips down and hits a low for the 'right' number of topics and increases again as the number of topics increase. The number of topics that is considered 'right' is any number in a small range that gives the best accuracy on a held out dataset.

This work is organized into the following sections: In Section 2, we review some related work in topic modelling and some methods proposed to choose the 'right' number of topics. In section 3, we motivate the rationale behind the measure proposed and explain how it is computed. In section 4 , we give experimental evidence to illustrate the robustness of the measure across text and image corpus. Finally we conclude in section 5 with a few points of discussion.

2 Background

2.1 Latent Dirichlet Allocation

LDA is a probabilistic generative model which assumes that every document is a distribution over topics and every topic is a distribution over words. Each word in a document is generated by first sampling a topic from the topic-distribution associated with the document and then sampling a word from the word distribution associated with the topic. Thus, given a corpus, LDA tries to find the right assignment of topic to every word such that the parameters of the generative model are maximized.

Topic Similarity. There have been a couple of approaches in the past which have tried to take advantage of the fact that the topics arising in read world data are correlated. Correlated Topic Models [12] is one such approach which tries to capture relation between topics using a covariance matrix. The Pachinko Allocation Model [14] on the other hand considers an acyclic graph where a topic is a node and is considered as a distribution over not only words but also other topics.

There have also been approaches like Hierarchical Dirichlet Process (HDP) [11] which try to find the right number of topics by assuming that the data has a hierarchical structure to it. Here, both HDP as well as LDA models for the same dataset are built and compared to find the right number of topics.

More recently [10] proposes a method to learn the right size of an ontology by measuring the change in the average cosine distance between topics found as