ABSTRACT...

In this paper, we empirically study standard algorithms for topic modelling using Latent Dirichlet Allocation. We also develop geometry based algorithms for learning the Latent Dirichlet Allocation (LDA) which are as effective as inference based methods while being more efficient. In particular, the most effective algorithms use a novel combination of dimension reduction, projection, k-means, and a scaling procedure.

We argue that the evaluation of topic modelling should include a prediction task that is closely aligned with the model. LDA, in particular, generates the words in a document. Thus, predicting missing words is appropriate for the evaluation of the LDA model, both when applied to a corpus and for genarated data. Moreover, this task is useful for recommendation systems, keyword suggestion, tag prediction, etc. Finally, for practitioners this evaluation task may be more indicative of the usefulness of various methods than internal measures such as perplexity or than indirect methods where other prediction algorithms are applied to learned topic models.

Finally, to empirically study algorithms, one needs to understand properties of datasets. We thus provide an analysis that predicts whether a corpus can be effectively modelled. For generated datasets, the prediction is based on the diversity of topics in a typical document, and the concentration of words in a topic.

INTRO

The Latent Dirichlet Allocation topic model has been tremendously influential in the development of methods for analyzing documents, and other types of data. Numerous algorithms for learning such models as well as variations of this model have been proposed; see [BleiCACM] for a recent survey. The bulk of these methods are based on inference techniques; they proceed by using learning the parameters using methods which include from Gibbs sampling, EM, variational methods, and modifying the model for easier computation. The evaluation of the methods have proceeded typically along two lines. One is to train on a set of documents and then to evaluate the how well the resulting model predicts a test set, typically using perplexity as a measure. Another is to use the model as a set of features in some classification task, say for example document classification, in some learning method, say for example, using a support vector machine.

In this paper, we suggest modifying both the inference approach to learning topic models and suggest a different evaluation method. We feel both methods are easier to understand and are more effective for yielding progress.

In terms of inference algorithms, we take the view that the model parameters are geometric objects. That is, topic centers are simply points in the word space when the data consist of documents. This view is quite a traditional view of data, e.g., that it is generated from points in space under some noise model. This view of data has long been associated with algorithms such as nearest

neigbors or k-means for classification or summarization. We take this view to develop an improved ...