A Sequential Algorithm for Fast Fitting of Dirichlet Process Mixture Models

Alec Greaves-Tunnell

Department of Statistics, University of Washington Seattle, WA, 98195, USA

Abstract

This paper discusses the motivation, implementation, and analysis of a sequential algorithm for approximate Bayesian inference in Dirichlet process mixture models introduced by Zhang et al. [2014]. Dirichlet process mixture models are a popular tool for clustering due to their flexibility, but standard sampling methods for posterior computation do not scale well to large data sets. As data continues to grow in many applications, there is increasing interest in computationally efficient methods for fast approximation of the posterior. The authors propose one such method, V-SUGS, which extends a simple sequential algorithm known as SUGS ("sequential update and greedy search") by improving the representation of uncertainty over cluster assignments in the posterior. Comparative analysis of the two algorithms shows that V-SUGS performs better on data whose clusterings are close or overlapping while preserving the attractive computational structure of the original algorithm. Further experiments demonstrate the speed and relative accuracy with which the model can be fit to two large biological data sets.

1 Introduction

Finite mixture modeling offers a useful method for the identification of latent classes or clusters in a given set of observations, but the requirement that the number of mixture components be specified ahead of time comes as a major limitation, particularly in applications involving large or complex data. The Bayesian nonparametric framework generalizes the class of finite mixture models such that this requirement is eliminated, as the number of mixture components and the parameters of each individual mixture component can be estimated simultaneously. Under this framework, a mixing distribution is specified over a

countably infinite number of elements, which results in a mixture model with an infinite number of components. However, since a training example is associated with only one mixture component, application of this model to any finite data set will result in a finite but variable number of active components. The Bayesian approach to this modeling task requires a prior over the mixing distribution; the most common prior to use is a Dirichlet process, and the resulting model is known as a Dirichlet process mixture model (DPMM).

Dirichlet process mixture models are valued for their capacity to grow with the complexity of the data to which they are applied, and interest in their application to clustering and density estimation problems has resulted in an extensive literature on methods for posterior computation. The bulk of this work has relied on Markov chain sampling methods, which promise exact sampling from the posterior after convergence of the Markov chain; a review of the major methods for sampling in DPMMs is given in Neal [2000]. While these methods collectively propose many elegant solutions for exact sampling from the posterior, they scale poorly in terms of computational cost to the analysis of very large data sets. Increasing interest in applying Bayesian nonparametric models such as DPMMs to large sets of data has thus led to emphasis on methods that can quickly compute some approximation to the posterior when sampling is inadvisable or downright impossible. Sequential algorithms, which process each observation individually in a single pass over the data, represent one promising category of methods that may be suited to this approximation task. Sequential algorithms for DPMMs recursively update an approximate representation of the posterior by incorporating each next observation into the existing model. This structure ensures that their computational complexity scales well with the total size of the data, and it makes them inherently suited for applications in which data arrive with some natural ordering.

This paper explores the implementation of SUGS and V-SUGS, two sequential algorithms for approximate posterior inference n DPMMs. SUGS, an acronym for "sequential update and greedy search," was introduced by Wang and Dunson [2011]. As its name suggests, it gives a simple recursive formula for updating an approximate posterior that greedily assigns each new observation to a mixture component. V-SUGS is an extension of this approach due to Zhang et al. [2014], who propose instead a probabilistic cluster assignment and subsequent variational update to the mixture component parameters. Comparative analysis of the two

algorithms shows that V-SUGS offers superior performance when clusters of data are close or overlapping, while both offer dramatic reductions in computational cost when compared to standard sampling methods. While relatively little comparison is made to other possible competitor algorithms, Zhang et al. [2014] demonstrate that their algorithm can quickly fit a DPMM on two large biological data sets, and they confirm that these approximations are reasonably accurate by comparison with the results from sampling. Unfortunately, little comparison is offered with other approximate methods, sequential or otherwise, beyond an initial evaluation of V-SUGS against SUGS on synthetic data.

2 Methods

3 Results

4 Discussion

References

Radford M Neal. Markov chain sampling methods for dirichlet process mixture models. *Journal of Computational and Graphical Statistics*, 9(2):249–265, 2000.

Lianming Wang and David B Dunson. Fast bayesian inference in dirichlet process mixture models. *Journal of Computational and Graphical Statistics*, 20(1):196–216, 2011.

Xiaole Zhang, David J Nott, Christopher Yau, and Ajay Jasra. A sequential algorithm for fast fitting of dirichlet process mixture models. *Journal of Computational and Graphical Statistics*, 23(4):1143–1162, 2014.