Journal of Machine Learning Research 12 (2011) 2383-2410 Submitted 10/09; Revised 12/10; Published 8/10

Distance Dependent Chinese Restaurant Processes

David M. Blei BLEI@CS.PRINCETON.EDU
Department of Computer Science
Princeton University
Princeton, NJ 08544, USA

Peter I. Frazier PF98@CORNELL.EDU
School of Operations Research and Information Engineering
Cornell University
Ithaca, NY 14853, USA

Editor: Carl Edward Rasmussen

Abstract

We develop the distance dependent Chinese restaurant process, a flexible class of distributions over partitions that allows for dependencies between the elements. This class can be used to model many kinds of dependencies between data in infinite clustering models, including dependencies arising from time, space, and network connectivity. We examine the properties of the distance dependent CRP, discuss its connections to Bayesian nonparametric mixture models, and derive a Gibbs sampler for both fully observed and latent mixture settings. We study its empirical performance with three text corpora. We show that relaxing the assumption of exchangeability with distance dependent CRPs can provide a better fit to sequential data and network data. We also show that the distance dependent CRP representation of the traditional CRP mixture leads to a faster-mixing Gibbs sampling algorithm than the one based on the original formulation.

Keywords: Chinese restaurant processes, Bayesian nonparametrics

1. Introduction

Dirichlet process (DP) mixture models provide a valuable suite of flexible clustering algorithms for high dimensional data analysis. Such models have been adapted to text modeling (Teh et al., 2006; Goldwater et al., 2006), computer vision (Sudderth et al., 2005), sequential models (Dunson, 2006; Fox et al., 2007), and computational biology (Xing et al., 2007). Moreover, recent years have seen significant advances in scalable approximate posterior inference methods for this class of models (Liang et al., 2007; Daume, 2007; Blei and Jordan, 2005). DP mixtures have become a valuable tool in modern machine learning.

DP mixtures can be described via the Chinese restaurant process (CRP), a distribution over partitions that embodies the assumed prior distribution over cluster structures (Pitman, 2002). The CRP is fancifully described by a sequence of customers sitting down at the tables of a Chinese restaurant. Each customer sits at a previously occupied table with probability proportional to the number of customers already sitting there, and at a new table with probability proportional to a concentration parameter. In a CRP mixture, customers are identified with data points, and data sitting at the same