Statistical Analysis and Data Mining: The ASA Data Science Journal Explore this journal >

View issue TOC
Volume 10, Issue 6
December 2017
Pages 393–409

ORIGINAL ARTICLE

A Bayesian mixture model for clustering and selection of feature occurrence rates under mean constraints

Qiwei Li, Michele Guindani, Brian J. Reich, Howard D. Bondell, Marina Vannucci

First published:
8 June 2017 Full publication history

DOI:
10.1002/sam.11350 View/save citation

Cited by (CrossRef):
0 articles Check for updates Citation tools

Abstract

In this paper, we consider the problem of modeling a matrix of count data, where multiple features are observed as counts over a number of samples. Due to the nature of the data generating mechanism, such data are often characterized by a high number of zeros and overdispersion. In order to take into account the skewness and heterogeneity of the data, some type of normalization and regularization is necessary for conducting inference on the occurrences of features across samples. We propose a zero-inflated Poisson mixture modeling framework that incorporates a model-based normalization through prior distributions with mean constraints, as well as a feature selection mechanism, which allows us to identify a parsimonious set of discriminatory features, and simultaneously cluster the samples into homogenous groups. We show how our approach improves on the accuracy of the clustering with respect to more standard approaches for the analysis of count data, by means of a simulation study and an application to a bag-of-words benchmark data set, where the features are represented by the frequencies of occurrence of each word.

Get access to the full text of this article

Article Information

Data Accessibility

Related content

Articles related to the one you are viewing

The articles below have been selected for you based on the article you are currently viewing.

Robustifying Bayesian nonparametric mixtures for count data

Antonio Canale, Igor Prünster

28 April 2016

Variable selection in covariate dependent random partition models: an application to urinary tract infection

William Barcella, Maria De Iorio, Gianluca Baio, James Malone-Lee

4 November 2015

The LZIP: A Bayesian latent factor model for correlated zero-inflated counts

Brian Neelon, Dongjun Chung

5 July 2016

Models for zero-inflated, correlated count data with extra heterogeneity: when is it too complex?

Sammy Chebon, Christel Faes, Frank Cools, Helena Geys

13 October 2016

A Nonparametric Model for Stationary Time Series

Isadora Antoniano-Villalobos, Stephen G. Walker

13 August 2015