Statistics Seminar

Department of Mathematical Sciences

DATE:	Thursday, Month 31, 2017
TIME:	1:15pm – 2:15pm
LOCATION:	Zoom meeting
SPEAKER:	Wenshu Dai, Binghamton University
TITLE:	Identifying the number of clusters in discrete mixture models

Abstract

Research on cluster analysis for categorical data continues to develop, with new clustering algorithms being proposed. However, in this context, the determination of the number of clusters is rarely addressed. In this paper, it proposes a new approach in which clustering of categorical data and the estimation of the number of clusters is carried out simultaneously. Assuming that the data originate from a finite mixture of multinomial distributions, it develops a method to select the number of mixture components based on a minimum message length (MML) criterion and implement a new expectation maximization(EM) algorithm to estimate all the model parameters. The proposed EM-MML approach, rather than selecting one among a set of pre-estimated candidate models (which requires running EM several times), seamlessly integrates estimation and model selection in a single algorithm. The performance of the proposed approach is compared with other well-known criteria (such as the Bayesian information criterion–BIC), resorting to synthetic data and to two real applications from the European Social Survey. The EM-MML computation time is a clear advantage of the proposed method. Also, the real data solutions are much more parsimonious than the solutions provided by competing methods, which reduces the risk of model order overestimation and increases interpretability.