Clusters and features from combinatorial stochastic processes

Tamara Broderick, University of California, Berkeley

In partitioning---a.k.a. clustering---data, we associate each data point with one and only one of some collection of groups called clusters or partition blocks. Here, we formally establish an analogous problem, called feature allocation, for associating data points with arbitrary non-negative integer numbers of groups, now called features or topics. Just as the exchangeable partition probability function (EPPF) can be used to describe the distribution of cluster membership under an exchangeable clustering model, we examine an analogous "exchangeable feature probability function" for certain types of feature models. Moreover, recalling Kingman's paintbox theorem as a characterization of the class of exchangeable clustering models, we develop a similar "feature paintbox" characterization of the class of exchangeable feature models. We examine models such as the Bayesian nonparametric Indian buffet process as examples within this broader class.

Authors: Tamara Broderick, Michael I. Jordan, Jim Pitman