Identifiability of nonparametric mixture models, clustering, and semi-supervised learning

Nikhyl Aragam

University of Chicago
E-mail: naragam@cs.cmu.edu

Abstract: Motivated by problems in data clustering and semi-supervised learning, we establish general conditions under which families of nonparametric mixture models are identifiable by introducing a novel framework for clustering overfitted parametric (i.e. misspecified) mixture models. These conditions generalize existing conditions in the literature, allowing for general nonparametric mixture components. Notably, our results avoid imposing assumptions on the mixture components, and instead impose regularity assumptions on the underlying mixing measure. After a discussion of some statistical aspects of this problem, we will discuss two applications of this framework. First, we extend classical model-based clustering to nonparametric settings and develop a practical algorithm for learning nonparametric mixtures. Second, we analyze the sample complexity of semi-supervised learning (SSL) and introduce new assumptions based on the mismatch between a mixture model learned from unlabeled data and the true mixture model induced by the (unknown) class conditional distributions. Under these assumptions, we establish an Omega(Klog K) labeled sample complexity bound without imposing parametric assumptions, where K is the number of classes. These results suggest that even in nonparametric settings it is possible to learn a near-optimal classifier using only a few labeled samples.