Adaptive Density-Based Clustering

Prof. Dr. Ingo Steinwart
(Universität Stuttgart)

November 8, 2017

A central task in nonparametric statistics is cluster analysis, where the goal is to find clusters in unlabeled data. One widely accepted definition of clusters has its roots in a paper by Carmichael et al., where clusters are described to be densely populated areas in the input space that are separated by less populated areas. The non-parametric mathematical translation of this idea usually assumes that the data is generated by some unknown probability measure that has a density with respect to the Lebesgue measure. Given a threshold level, the clusters are then defined to be the connected components of the density level set. Here, the choice of the threshold, which is left to the user, is a notoriously difficult problem, typically only addressed by heuristics.
In this talk, we show how a simple algorithm based on a density estimator can find the smallest level for which there are more than one connected component in the level set. For some classical density estimators we further establish rates of convergence and present a simple approach for selecting the width parameter. It turns out that in many cases this approach is adaptive, i.e. it achieves the previously established rates of convergence without knowing specifics about the distribution. Finally, we discuss some practical aspects of the algorithm.