

Abstract Submitted
for the MAR17 Meeting of
The American Physical Society

An information theoretic approach to geometric clustering DJ STROUSE, Princeton University, DAVID SCHWAB, Northwestern University — Clustering is a basic task in data analysis for both understanding and pre-processing data. Classic clustering methods, such as k-means or EM fitting of a gaussian mixture model, are based on geometry. These geometric clustering methods group data points together based on their Euclidean distance from one another; roughly speaking, points within a cluster have smaller distances to one another than to points in other clusters. More recently, however, distributional clustering methods, such as the information bottleneck (IB) and deterministic information bottleneck (DIB), have been introduced that group data points based upon their conditional distributions over a target variable. Here, points within a cluster provide similar information about the target variable. Are distributional and geometric clustering related, and if so, how? Can we blend these two approaches? Here we first describe a method to incorporate geometric information into the (D)IB clustering algorithm, where the target variable our clustering should be informative about is the spatial location of the contained data points. This enables us to derive a novel set of geometric clustering algorithms, which we then compare to the classic methods mentioned above. Finally, we compare both approaches on data.

DJ Strouse
Princeton University

Date submitted: 11 Nov 2016

Electronic form version 1.4