STATISTICAL ANALYSIS AND DATA MINING

Research Article

On the limits of clustering in high dimensions via cost functions

Hoyt A. Koepke, Bertrand S. Clarke

First published: 16 November 2010
https://doi.org/10.1002/sam.10095
Cited by: 3

Abstract

This paper establishes a negative result for clustering: above a certain ratio of random noise to nonrandom information, it is impossible for a large class of cost functions to distinguish between two partitions of a data set. In particular, it is shown that as the dimension increases, the ability to distinguish an accurate partitioning from an inaccurate one is lost unless the informative components are both sufficiently numerous and sufficiently informative. We examine squared error cost functions in detail. More generally, it is seen that the VC - dimension is an essential hypothesis for the class of cost functions to satisfy for an impossibility proof to be feasible. Separately, we provide bounds on the probabilistic behavior of cost functions that show how rapidly the ability to distinguish two clusterings decays. In two examples, one simulated and one with genomic data, bounds on the ability of squared - error and other cost functions to distinguish between two partitions are computed. Thus, one should not rely on clustering results alone for high dimensional low sample size data and one should do feature selection. Copyright © 2010 Wiley Periodicals, Inc. Statistical Analysis and Data Mining 4: 30–53 2011