Mathematical Statistics
Stockholm University
Master Thesis **2022:3**
http://www.math.su.se

Statistical generalization and applications of a robust, fast and fully automated density based clustering method for big data

Anton Holm*

June 2022

Abstract

In recent years, machine learning has taken a larger place in data analysis. The aim is often to predict some response variables based on other explanatory variables. When only the explanatory variables are available, it is still possible to do inference on the data. One way to do so is by performing clustering. Essentially, clustering is a method where data points with similar characteristics are grouped while keeping data points with dissimilar characteristics far from each other. There are currently several different branches of clustering and this thesis will be focusing on density-based clustering. A comparison between a kernel density estimator and a k-nearest neighbor (kNN) density estimator is performed, showing the strength and robustness of using a kNN approach. The main goal of this thesis is the construction of a fully automated density-based clustering method. The method is statistically robust to clusters with varying shape and density, works fast on large data sets, is easy to understand and interpret even for non-statisticians, and only relies on a single parameter. The method is tested on a generated data set showing promising results. Lastly, future improvements are discussed, suggesting the use of fuzzy clustering and substitution of Euclidean distance by graph-based distance in efficiently identifying clusters with non-linear shape.

*Postal address: Mathematical Statistics, Stockholm University, SE-106 91, Sweden. E-mail: anton.holm.klang@gmail.com. Supervisor: Chun-Biu Li.