Mathematical Statistics
Stockholm University
Bachelor Thesis **2023:1**
http://www.math.su.se

Initialization of the k-means algorithm
A comparison of three methods

Simon Jorstedt*

June 2023

Abstract

k-means is a simple and flexible clustering algorithm that has remained in common use for 50+ years. In this thesis, we discuss the algorithm in general, its advantages, weaknesses and how its ability to locate clusters can be enhanced with a suitable initialization method. We formulate appropriate requirements for the (batched) UnifRandom, k-means++ and Kaufman initialization methods and compare their performance on real and generated data through simulations. We find that all three methods (followed by the k-means procedure) are able to accurately locate at least up to nine well-separated clusters, but the appropriately batched UnifRandom and the Kaufman methods are both significantly more computationally expensive than the k-means++ method already for $K = 5$ clusters in a dataset of $N = 1000$ points.

*Postal address: Mathematical Statistics, Stockholm University, SE-106 91, Sweden. E-mail: jorstedtsimon@gmail.com. Supervisor: Taras Bodnar, Lina Palmborg, Dongni Zhang.