On the choice of initial seed values for the Lloyd's k-means algorithm

Kalev Pärna
University of Tartu, Tartu, Estonia – kalev.parna@ut.ee

This paper is about efficient ways of converting continuous data into a compact discrete form. In information theory such a conversion is called quantization and it is used for transmission of data through a discrete channel which is capable of admitting a finite number (k) of values, only. In statistics and data mining the same method is called k-means clustering and its aim is to partition a data set into k non-overlapping clusters by minimizing the within-sum of squares of deviations from their respective cluster centers (k-means). Efficient calculation of k-means, especially in multivariate setting, is still a problem which needs further research. Well-known Lloyd's iterative method for calculation of k-means is sensitive with respect to initial values and, therefore, much research has been focused on the choice of initial seed values of k-means. In this paper we propose to use certain theoretical results about asymptotic density of points of k-means in the process where the number k grows infinitely. Namely, it is known that for large values of k the optimal k points are distributed in accordance with density $f^*(x)$ which is a power function of the initial data density $f(x)$ (for example, in one-dimensional case $f^*(x)$ is proportional to $f(x)^{1/3}$). Our main idea is to use this asymptotic distribution for placement of initial seeds of k-means. In order to benefit from the asymptotic theory, we propose a 3-steps method for calculation of k-means, consisting of 1) estimation of $f^*(x)$ from the data, 2) generation of k points from $f^*(x)$, and, 3) using these points as initial values in the Lloyd's iterative algorithm.

Keywords: quantization; asymptotic distribution of k-means, Lloyd's algorithm.