Asymptotically optimal placement of initial centers in k-means clustering

Kalev Pärna

University of Tartu, Estonia, kalev.parna@ut.ee

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

We are discussing efficient ways of converting data into a compact discrete form – a problem that arises in many areas. In information theory such a conversion is called quantization and it is used to transmit the data through a discrete channel which allows k different values only. In statistics and data mining, k-means clustering aims at partitioning the 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. Lloyd's iterative method [2] – a standard procedure for calculation of k-means – is sensitive with respect to initial centers and, therefore, different methods have been proposed for the choice of initial seed values of k-means [1].

In this paper we focus on how to make use of certain theoretical results about asymptotic behavior of optimal centers if k tends to infinity. It is known that, for large k, the optimal centers are distributed in accordance with the density $f^\star(x)$, which is a power function of the initial data density $f(x)$ [3]. In 1-dimensional case, for example, the asymptotic density of k-means is proportional to $[f(x)]^{1/3}$. In order to benefit from this asymptotics, we propose to use the density $f^\star(x)$ for placement of initial seeds in the Lloyd's iterative algorithm. Our method consists of 1) estimation of $f^\star(x)$ from the data, 2) using $f^\star(x)$ for reweighting initial data (change of measure), 3) sampling k points from reweighted data, and 4) using the points sampled as initial values for further iterations in Lloyd's algorithm.

References

[1] Arthur, D. and Vassilvitskii, S. (2007). k-means++: the advantages of careful seeding. *Proceedings of The 18-th Annual ACM-SIAM Symposium on Discrete Algorithms*. SIAM, 1027–1035.

[2] Lloyd, S.P. (1982). Least squares quantization in PCM. *IEEE Transactions on Information Theory* **IT-28**, 129–137.

[3] Zador, P. (1982). Asymptotic quantization error of continuous signals and the quantization dimension. *IEEE Transactions on Information Theory* **IT-28**, 139–149.