1 Basic Concepts about Clustering

Let d be a positive integer and $\mathbb{R}$ the field of real numbers. For a set S of n points $\vec{p_i} \in \mathbb{R}^d$, we denote by $|S|$ the number of points of S. We consider the problem that we will call "k-means globally optimum clustering".

Definition 1. *The "k-means globally optimum clustering" is to split $S \subset \mathbb{R}^d$ of n points $\vec{p_i}$, $i = 1, \dots, n$ into k disjoint nonempty subsets $S_1, \dots, S_k$ called clusters in such a way that the following expression is minimized:*

$$f_{S_1,\dots,S_k}(S) = \sum_{j=1}^{k} \sum_{\vec{p} \in S_j} \|\vec{p} - \vec{q}_j\|^2, \quad \textit{where } \vec{q}_j = \frac{\sum_{\vec{p} \in S_j} \vec{p}}{|S_j|}.$$

$S_1, \dots, S_k$ is called an optimal partition of S.

It is well known that, given S, there always exists $\vec{q_1}, \dots, \vec{q_k}$ such that the partition defined as,

$$S_j = \bigcap_{l=1}^{k} \{\vec{p} \in S \ : \ \|\vec{p} - \vec{q_j}\|^2 \leq \|\vec{p} - \vec{q_l}\|^2\},$$

is an optimal partition.[1] Indeed, the common approach to attack this problem is to use *Lloyd's heuristic* [2], which was first used in [3] and, under minor modifications, performs quite well in practice, see [1, 4].

References

[1] David Arthur and Sergei Vassilvitskii. k-means++: the advantages of careful seeding. In *Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms*, SODA '07, pages 1027–1035, Philadelphia, PA, USA, 2007. Society for Industrial and Applied Mathematics.

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

[3] James B. MacQueen. Some methods for classification and analysis of multivariate observations. In *Proceedings of the 5th Berkeley Symposium on Mathematical Statistics and Probability*, volume 1, pages 281–297. University of California Press, 1967.

[4] Chen Zhang and Shixiong Xia. K-means clustering algorithm with improved initial center. In *Knowledge Discovery and Data Mining, 2009. WKDD 2009. Second International Workshop on*, pages 790 –792, jan. 2009.

[1] Using this definition it could be that one point belong to more than one clusters. Fortunately, it is always possible to solve the ties in a reasonable manner