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,

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

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].

We will need the following concepts from topology:

- A set contained in $\mathbb{R}^d$ is *convex* if for any pair of points within the set, every point in the straight line segment that joins them is also within the object.
- Given a set of points $S \subset \mathbb{R}^d$, the convex hull of S is the smallest set of $\mathbb{R}^d$ which contains S.
- Given $\vec{a} \in \mathbb{R}^d - \{\vec{0}\}$ and $b \in \mathbb{R}$, the set $\mathcal{H} = \{\vec{x} \in \mathbb{R}^d : (\vec{a})^{\mathbf{T}}\vec{x} = b\}$ is called a hyperplane.
- A point $\vec{p} \in \mathbb{R}^d$ lies in the *left side* of hyperplane $\mathcal{H}$ if $(\vec{a})^{\mathbf{T}}\vec{p} > b$. If $(\vec{a})^{\mathbf{T}}\vec{p} < b$, the point $\vec{p}$ lies in the *right side* of hyperplane $\mathcal{H}$.
- An hyperplane $\mathcal{H}$ *separates* two sets $S,\ S' \subset \mathbb{R}^d$ if all the points in S lies in the left side of $\mathcal{H}$ and all the points in S' lies in the right side of $\mathcal{H}$.

We cite here the maximum separation hyperplane.

Lemma 1. *For any two convex sets $S,\ S' \subset \mathbb{R}^d$ such that $S \cap S' = \emptyset$, there exists an hyperplane $\mathcal{H}$ that separates S and S'.*

[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

As it was stated before, it is known that one optimal partition is defined using k centroids. Partitions defined by centroid have a very interesting property.

Lemma 2. *Given a set of point $S \subset \mathbb{R}^d$ and centroids $\vec{q_1}, \ldots, \vec{q_k} \in \mathbb{R}^d$, the partition $\S_1, \ldots, \S_k$ defined as*

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

for $j = 1, \ldots, k$ satisfies:

- *the intersection of the convex hull of any two different clusters S_i, S_j is empty,*
- *for each pair S_i, S_j exists an hyperplane $\mathcal{H}$ that separates S_i and S_j.*

Proof. The first assertion of the lemma is proved by induction. For $k = 2$, it is trivial. The general case is done noting that the intersection of two convex sets is a convex set. So, the convex hull of

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

is just the intersection of the convex hulls of

$$\{\vec{p} \in S \ : \ \|\vec{p} - \vec{q_j}\|^2 \leq \|\vec{p} - \vec{q_l}\|^2\},$$

for $l \neq j$, which are disjoint by induction.

The second assertion is a direct application of Lemma 1 and that S_i, S_j are convex sets. $\square$

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.