Two Examples to Show How k-Means Reaches Richness and Consistency

Jia-heng WEI*

Institute of Mathematics and Statistics, Xi’an Jiaotong University, Xi’an, Shaanxi, China
*Corresponding author

Keywords: k-Means clustering, Kleinberg axioms, Richness, Consistency.

Abstract. In [1] three axioms of clustering algorithms were introduced. In this paper, we demonstrated how k-Means clustering method reaches two of the three axioms: richness and consistency using two well-designed examples. For richness, a metric of form $d_q(q \geq 1)$ doesn’t satisfy this property, but a semi-metric $d_q(0 < q < 1)$ satisfies this property. For consistency, a counter-example is provided; however, this property is achieved in a more general sense when it comes to the Lloyd’s algorithm [6], an approximation method for k-Means.

Introduction

Centroid-based clustering methods are a group of popular methods that minimize the sum of the cost of each point, where the cost is based on the distance between the point and the centroid. Let $P$ be the set of all points, and $C$ be the set of $k$ centroids. Let $d$ be the distance function, where $d(p, C) = \min_{c \in C} d(p, c)$. Let $g: R^+ \rightarrow R^-$ be a non-decreasing function with $g(0) = 0$. Then we say the $(k, g)$-centroid clustering method solves the following optimization problem [3]:

$$\min_{C} F(C) = \sum_{p \in P} g(d(p, C)) \tag{1}$$

where popular $g(x)$ can be $\sqrt{x}$, $x$, $x^2$, etc. When $g(x) = x^2$, the method is called k-Means, which is NP-hard (even in planar situations [5]), and only local optimum can be effectively calculated by Lloyd’s algorithm [6].

Kleinberg [1] introduced three axioms for clustering methods:

Scale Invariance: the result doesn't change when all distances are multiplied by the same constant;

Richness: for any partitioning of $P$, there exists some distance that yields the desired result;

Consistency: reducing distances between points within the same cluster while increasing distances among different clusters doesn't change the result.

However, the impossibility theorem [1,2] stated that no algorithm is able to satisfy all three axioms. In [4], a number of clustering methods including Single-Linkage, MST cuts family, Min-SUM k-clustering and Constant partitioning were examined, and whether the properties are satisfied were indicated.

For k-Means, a direct deduction implies that scale invariance is satisfied. In this paper, we analyzed the rest two axioms with toy examples and acquired a few results. For richness, it cannot be reached if $d$ is a metric, but if we don’t require the triangle inequality (i.e., $d$ is a semi-metric, as stated in the theorem), richness can be reached. For consistency, it’s not satisfied by k-Means, but can be satisfied by the Lloyd algorithm [6], an approximate method to solve k-Means.

How k-Means Reaches Richness

When an algorithm reaches richness, according to the definition, for any partition of the dataset, there exists some distance function $d$ that eventually yields the desired output. One common way to define the distance function $d$ is given by the follow:

$$d_q(x, y) = \|y - x\|_q, \text{ where } \|v\|_q = (\sum_{i=1}^{n} |v_i|^q)^{\frac{1}{q}} \tag{2}$$
When \( q \geq 1 \), \( d_q \) is a metric since it satisfies the triangle inequality; when \( 0 < q < 1 \), \( d_q \) is a semi-metric. Figure 1 shows the unit circles in 2-d space with distances under different \( q \)'s.

![Figure 1. Unit circles in 2-d space with distances \( d_q \)'s for \( q=0 \text{ to } 5 \) (from the inner side to the outer side).](image)

Now, we consider the following example with only four points in the 2-d space. Notice that in some occasions a set of 2-d points is not linearly separable, but most clustering algorithms especially k-Means will produce partitions with linear boundaries. From this point of view, we are able to create a set of points that is not the output of k-Means with a metric as a distance function. The coordinates of those four points are \( p_1 = (0,1) \), \( p_2 = (0, -(1+2b)) \), \( p_3 = (-1,0) \), \( p_4 = (1+2a, 0) \).

Let \( r_1 = (0, -b) \) be the centroid of \( (p_1, p_2) \) and \( r_2 = (a, 0) \) be the centroid of \( (p_3, p_4) \). Recall that richness implies that any partition of \( P \) is the result of the clustering algorithm under certain distance; we assign \( (p_1, p_2) \) of class 1 and \( (p_3, p_4) \) of class 2. Figure 2 demonstrates all the details.

![Figure 2. The example used in this section, where orange points refer to class 1 and blue points refer to class 2, and the triangle points refer to the centroid of points with the same color.](image)

We now assume that for some \( q \) the distance function \( d_q \) yields such result. This means that

\[
\begin{align*}
    &\{ d_p(p_1, r_1) < d_p(p_1, r_2) \\
    &d_p(p_3, r_2) < d_p(p_3, r_1) \}
\end{align*}
\]

Or equivalently,

\[
\begin{align*}
    &\{ (1+b)^q < 1+a^q \\
    &(1+a)^q < 1+b^q \}
\end{align*}
\]

Since

\[
\frac{d}{dq} ((1+a)^q - 1 - a^q) = (a+1)^q \log(a+1) - a^q \log(a) > 0
\]

and when \( q = 1, (1+a)^q - 1 - a^q = 0 \), we know that for \( q > 1 \),
where yields a contradiction. Consequently, any metric \( d_q \) won't yield the desired output. However, for \( 0 < q < 1 \), we have \((1 + a)^q < 1 + a^q\). As a result, the desired output is archived if we assign \( a = b \). If \( a \) and \( b \) are arbitrarily given, the two inequalities are still satisfied as \( q \) goes to 0, since the left side goes to 1 while the right side converges to 2.

From this example, we see that when we want to make k-Means to output any desired result, we probably need to use a semi-metric as the distance function, which is not common in real examples. On the other hand, when we apply k-Means to real world datasets, we do not expect the clusters are integrated with each other, which conforms with the intuition.

### How k-means Reaches Consistency

In [3] an example was given to point out that consistency is not satisfied. That is, as the distance of points within the same cluster decreases and that of points between clusters increases (the consistency condition), the result might change. This seems ridiculous; when we assume that the clustering centers stay still, the result will not change at all. However, the clustering centers may move; it's even possible that they are far away from the original ones. Consequently, the new clusters may also change greatly.

The example is shown in Figure 3. In the first situation, let \( b_{00} \) be the centroid of the set composed of blue and orange points \( \{b_1, ..., b_m, o_1, ..., o_m\} \), which is the first class; let \( g_0 \) be the centroid of the set of green points \( \{g_1, ..., g_m\} \), which is the second class. The distances are given as follows:

\[
d(b_i, b_{00}) = d(o_i, b_{00}) = d(b_i, o_i) = 1, \ d(b_i, g_i) = d(o_i, g_i) = 1 + \delta, \ d(g_i, g_0) = \gamma < 1
\]  
(7)

In the second situation, we reduce the distance between points in the first class. Let \( b_0 \) be the centroid of \( \{b_1, ..., b_m\} \), and \( o_0 \) be the centroid of \( \{o_1, ..., o_m\} \). Let

\[
d(b_i, b_0) = d(o_i, o_0) = \frac{1}{2}
\]  
while remaining \( d(b_i, o_i) = 1 \) and other distances.

![Figure 3. The example used in the section.](image)

Now we compute the loss function \( F(C) \) in two situations. In the first one,

\[
F(\{b_{00}, g_0\}) = 2m \times 1^2 + m\gamma^2 = m(2 + \gamma^2)
\]  
(9)

In the second situation,

\[
F(\{b_0, g_0\}) = m \times \left(\frac{1}{2}\right)^2 + m(1 + \delta)^2 = m\left(\frac{1}{2} + (1 + \delta)^2\right)
\]  
(10)

When we choose some appropriate \( \gamma \) and \( \delta \), we have \((2 + \gamma^2) > \left(\frac{1}{2} + (1 + \delta)^2\right)\), and thus \( F(\{b_0, g_0\}) < F(\{b_{00}, g_0\}) \), which changes the clustering result. From this example, we see that the core thing that leads to this change is that the clustering centers have changed.

On the other side, if we use the Lloyd's algorithm [6], which yields a local optimum, instead of looking for the global optimum, it is reasonable to say that consistency is achieved as well. To prove
this, it is enough to argue that $\{b_{00}, g_0\}$ still produces a local optimum. Given these two centroids, every $g_i$ is assigned to class 2 since

$$d(g_i, g_0) = \gamma < 1 < 1 + \delta = d(g_i, b_{00})$$

(11)
every $b_i$ is assigned to $b_{00}$ because

$$d(b_i, b_{00}) \leq \max\{\max_i d(b_i, b_j), \max_j d(b_i, o_j)\} = 1 < 1 + \delta = d(g_i, b_{00})$$

(12)
and same for $o_i$. Thus, the algorithm terminates and $\{b_{00}, g_0\}$ is the output of $C$. From this, we see that a relaxed consistency is satisfied by the Lloyd’s algorithm, an approximation algorithm to $k$-Means.

In real world situations when the Lloyd’s algorithm is used to cluster the datasets, we still expect the same result when we change the distance function that follow the consistency condition; when we remove some points that are close to points in different clusters, it is unlikely that other points are assigned to different clusters.

**Conclusion and Future Work**

In this paper, we demonstrated two examples to show how $k$-Means reaches two of the three Kleinberg axioms: richness and consistency. In the first example, we provided a well-designed set of four points to show that for a certain partition of the set, a metric $d_q(q \geq 1)$ cannot yield the desired clustering result, but a semi-metric $d_q(1 > q > 0)$ is able to. Additionally, for any $a$ and $b$ in the context, there exists some $q$ that yields the desired result. In the second example, we examined a set of points which doesn't satisfy consistency with $k$-Means algorithm. However, when it comes to the Lloyd's algorithm, it is reasonable to say that a more general kind of consistency is achieved, because it might output the same result.

In real world situations, on the other hand, such extreme situations as in this paper are still not likely to occur. As discussed in sections 2 and 3, when we use a metric of $d_q$ form as the distance function and Lloyd's algorithm to find a local optimum, the result tends to conform with the intuition: clusters that integrate with each other are not likely to be the result, and change of distances according to the consistency condition does not yield different results.

In the future, we plan to do several improvements to our results. First, we will make theoretical analysis to a more general case including all centroid-based clustering methods. Second, we plan to explore more examples and find a more general pattern of cases when a metric cannot yield certain partition of the dataset. Third, we also plan to generalize properties of the datasets which don't satisfy the consistency.

**References**


Reference to a book:


