1 Framework

1.1 Context

We first recall the overall setting in which we are working in. Let $\{x_i\}_{i=1}^n \subset \mathbb{R}^D$ be a set of D-dimensional data points, or observations. We wish to assign to these data points $y_1, ..., y_k$ labels, such that some loss function is minimized across all potential assignments.

In the k-means context, our loss function was realized as the L^2 norm or Euclidean norm. If we defined a cluster as $C_i = \{x_j : y_j = i\}$, then our loss function here is simply:

$$F(C_1, ..., C_k) = \sum_{j=1}^{k} \left(\sum_{x_i \in C_j} \|x_i - \mu_j\|_2^2 \right) = \sum_{j=1}^{k} \sum_{x_i \in C_j} \left\| x_i - \frac{1}{|C_j|} \sum_{x_l \in C_j} x_l \right\|_2^2$$

where we use the distance between data points in the cluster to the mean μ_j of the cluster C_j.

Although there is strong geometric interpretation here, we notice that from a statistical standpoint, it is well known that the mean is not resistant to outliers. Indeed, consider an outlier point, that is, suppose there exists $x^o = x_j$ such that $\|x_i - x^o\|_2^2 > \delta$ for all $x_i \neq x_j$. In particular, since we have a finite set of points, it must be bounded. Let M be the upper bound for the norms other than the outlier, $\{\|x_i\| : i \neq j\}$, and choose $\delta > Mn$ for example. Then, we must have that this outlier is a point on its own, as including it in any cluster with other points will incur a penalty of at least Mn, and the other points may only be at most M away from the origin.

Although this is not necessarily detrimental to the algorithm, as some outliers such as these should be removed in a data cleaning step, in other cases of sampling data from distributions with long tails, this may result in incorrect clustering. Analogously to usual statistical methods then, we introduce the median as an alternative choice of centroid, more resistant to outliers, that can be used with such noisy distributions and sampling.

1.2 Definitions

Firstly, we will define the median that we will be using. Let $\{x_i\}_{i=1}^n \subset \mathbb{R}^D$ be a set of points, and define the median as:

$$m = \arg\min \sum_{i=1}^{n} |x_i - y| \implies m_j = \arg\min \sum_{i=1}^{n} |(x_i)_j - y_j|$$

that is, the component-wise median of the points in the cluster. We may recognize this also as the L^1 norm.

In a similar fashion then, we may define our loss functional as being:

$$F(C_1, ..., C_k) = \sum_{j=1}^{k} \left(\sum_{x_i \in C_j} \|x_i - m_j\|_2^2 \right)$$

where, as last time, m_j corresponds to the median of the cluster C_j.

It may be useful to have an understanding of why this corresponds to a median, and how this relates to the intuitive definition of a median being the "middle value". Recall that geometrically, $|x - y|$ is symmetric around $x = y$, where the value of that function is 0. Thus, the minimum value of $\sum_{i=1}^n \sum_{j=1}^D |(x_i)_j - y_j|$ must be y_j such that the $(x_i)_j$ values are "balanced" in a way around y_j; that is, we must have that $\sum_{(x_i)_j > y_j} |(x_i)_j - y_j| = \sum_{(x_i)_j < y_j} |(x_i)_j - y_j|$ for y_j achieving the minimum.

From here, we proceed in the same way as the normal k-means algorithm. We choose arbitrary starting cluster centroids, with some heuristic for reasonableness. Then, we iteratively update the cluster points by adding data points to each cluster such that the overall functional is minimized.

1.3 Choosing the value of k

TBD

2 Bibliography

Cardot, Hervé, Peggy Cénac, and Jean-Marie Monnez. "A fast and recursive algorithm for clustering large datasets with k-medians." Computational Statistics & Data Analysis 56.6 (2012): 1434-1449.

Fischer, Aurélie. "On the number of groups in clustering." Statistics & Probability Letters 81.12 (2011): 1771-1781.

Godichon-Baggioni, Antoine, and Sobihan Surendran. "A penalized criterion for selecting the number of clusters for K-medians." arXiv preprint arXiv:2209.03597 (2022).