Algorithm (Pseudocode)

Notations

Let $X = \{x_1, ... x_n\}$ be the set of points in d-dimensional Euclidean space, and let k be a positive integer specifying the number of clusters. Let $||x_i - x_j||$ denote the Euclidean distance between x_i and x_j. For a point x and a subset $Y \subseteq X$ of points, the distance is defined as $d(x, Y) = min_{y \in Y} ||x - y||$. For a subset $Y \subseteq X$ of points, let its *centroid* be given by

$$\text{centroid}(Y) = \frac{1}{|Y|} \sum_{y \in Y} x$$

Let $C = \{c_1, ... c_k\}$ be the ser of points and let $Y \subseteq X$. We define the *cost* of Y with respect to C as

$$\phi_Y(C) = \sum_{y \in Y} d^2(y, C) = \sum_{y \in Y} \min_{i=1,\ldots,k} ||y - c_i||^2$$

Algorithm

k-means++(k) initialization

1. $C \leftarrow$ sample a point uniformlt at random from X
2. **while** $|C| < k$ **do**
   - Sample $x \in X$ with probability $\frac{d^2(x,C)}{\phi_X(C)}$
   - $C \leftarrow C \cup \{x\}$
3. **end while**

k-means$||$(k, l) initialization

1. $C \leftarrow$ sample a point uniformlt at random from X
2. $\psi \leftarrow \phi_X(C)$

3. **for** $O(\log\psi)$ times **do**
    - $C' \leftarrow$ sample each point $x \in X$ independently with probability $p_x = \frac{l \cdot d^2(x,C)}{\phi_X(C)}$
    - $C \leftarrow C \cup C'$
4. **end for**
5. For $x \in C$, set w_x to be the number of points in X closer to x than any point in C
6. Recluster the weighted points in C into k clusters