K-means Clustering

Input: integer k>0, set S of points in the euclidean space
Output: A (partitional) clustering of S

1. Select k points in S as the initial centroids
2. Repeat until the centroids do not change
   Form k clusters by assigning points to the closest centroids
   For each cluster recompute its centroid

- Initial centroids are often chosen randomly.
- Centroids are often the mean of the points in the cluster.
- 'Closeness' is measured by Euclidean distance, cosine similarity, correlation, etc.

k-means++

Algorithm 1 k-means++(k) initialization.

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

where:

$$d(x,\mathcal{C}) = \min_{c \in \mathcal{C}} d(x,c), \quad \Phi_X(\mathcal{C}) = \sum_{x \in X} d^2(x,\mathcal{C})$$

$d^2(x,C)$ measures how "good" is the clustering for point x.
Points that are *relatively* far away from "their" centroids will be selected with higher probability.