Indian Institute of Technology, Tirupati
March 7th, 2023

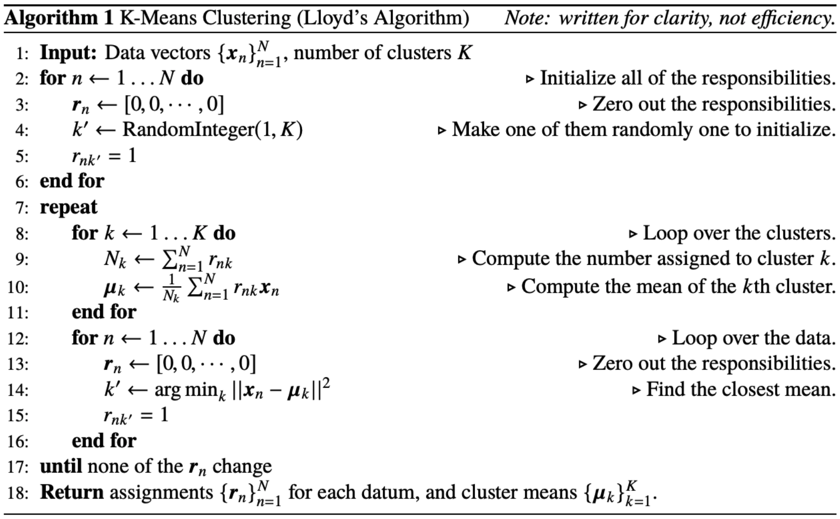
CS204P : Data Structures & Algorithms Lab

For a given set of data points in $\mathbf{R}^2$ space find K-Means Clustering of the points using a Red-Black Tree.

Input : $[set\, of\, data\, points : \{(x_i, y_i) \,|\, i = 0, 1, \ldots, n\},\ K : number\, of\, clusters]$

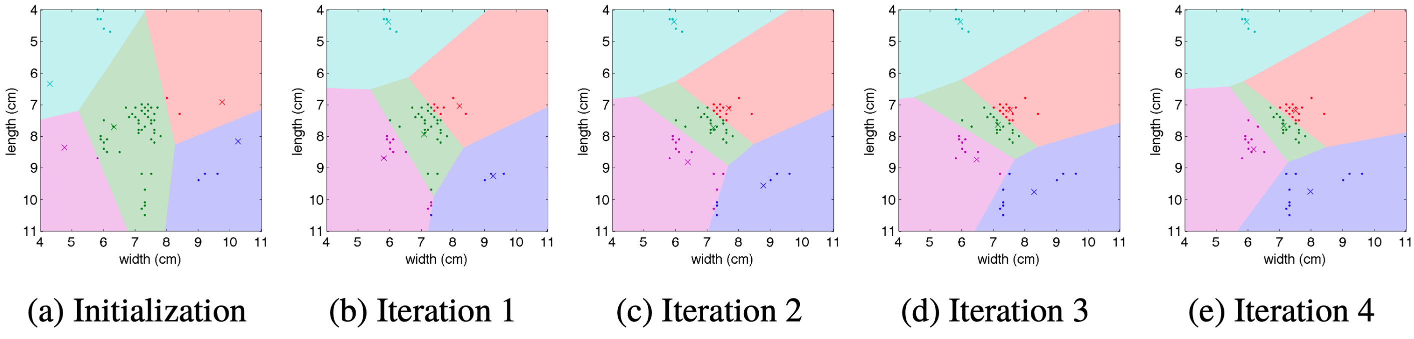
Output : $[set\, of\, data\, points\, with\, their\, associated\, cluster : \{(x_i, y_i, k) \,|\, i = 0, 1, \ldots, n\,;\ k \in K\}]$

K-means Clustering Algorithm



```
Algorithm 1 K-Means Clustering (Lloyd's Algorithm)        Note: written for clarity, not efficiency.
Input: Data vectors $\{\boldsymbol{x}_n\}_{n=1}^N$, number of clusters $K$
for $n \leftarrow 1 \ldots N$ do                           ▷ Initialize all of the responsibilities.
    $\boldsymbol{r}_n \leftarrow [0, 0, \cdots, 0]$                           ▷ Zero out the responsibilities.
    $k' \leftarrow$ RandomInteger$(1, K)$                  ▷ Make one of them randomly one to initialize.
    $r_{nk'} = 1$
end for
repeat
    for $k \leftarrow 1 \ldots K$ do                                      ▷ Loop over the clusters.
        $N_k \leftarrow \sum_{n=1}^N r_{nk}$                 ▷ Compute the number assigned to cluster $k$.
        $\boldsymbol{\mu}_k \leftarrow \frac{1}{N_k} \sum_{n=1}^N r_{nk} \boldsymbol{x}_n$                ▷ Compute the mean of the $k$th cluster.
    end for
    for $n \leftarrow 1 \ldots N$ do                                          ▷ Loop over the data.
        $\boldsymbol{r}_n \leftarrow [0, 0, \cdots, 0]$                       ▷ Zero out the responsibilities.
        $k' \leftarrow \arg\min_k ||\boldsymbol{x}_n - \boldsymbol{\mu}_k||^2$                     ▷ Find the closest mean.
        $r_{nk'} = 1$
    end for
until none of the $\boldsymbol{r}_n$ change
Return assignments $\{\boldsymbol{r}_n\}_{n=1}^N$ for each datum, and cluster means $\{\boldsymbol{\mu}_k\}_{k=1}^K$.
```

Figure 1: K-Means Clustering



(a) Initialization (b) Iteration 1 (c) Iteration 2 (d) Iteration 3 (e) Iteration 4

Figure 2: Clustering Convergence