K-means Clustering

Initialization:

$X = (x_1, x_2, \ldots x_n)$ set of n data points

k be no. of clusters

- Centroids are initialized as $C = \{c_1, c_2, \ldots c_k\}$

Distance Calculation

Euclidean Distance:

$$d(x_i, c_j) = \sqrt{\sum_{m=1}^{d} (x_{i,m} - c_{j,m})^2}$$

- $x_{i,m}$ is m-th coordinate of x_i

$c_{j,m}$ is m-th " " c_j

Cluster Assignment

$$\text{Cluster}(x_i) = \arg\min_j d(x_i, c_j)$$

Update Centroids

$$c_j = \frac{1}{n_j} \sum_{x_i \in \text{cluster}_j} x_i$$

$\sum_{x_i \in \text{cluster}_j}$ is sum of all points in cluster.

$$SSE = \sum_{j=1}^{k} \sum_{x_i \in \text{cluster}_j} \|x_i - c_j\|^2$$