2/4/25 Lab-10

Date / /
Page

Build a K-means clustering algorithm to cluster a set of data stored in a .csv

input

- Dataset $D = \{x_1, x_2, \ldots x_N\}$ for each $x_i \in \mathbb{R}^d$, d is dimensional factor
- no. of clusters: K

Output

- K cluster centroids and cluster assignments for each data point

algorithm

1. Initialize centroid
   - Randomly select k data points from D as initial centroids $\{\mu_1, \mu_2, \ldots \mu_K\}$
   → choose 1st centroid from D
   → from remaining centroid select a point with probability

2. Iterate until convergence / max itr
   - assign clusters
   - for each data point x_i:

Page

Compute the Euclidean distance

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

3. update centroid

- for each cluster $j = 1, 2, \ldots k$:
  → compute new centroid μ_j,

$$\mu_j = \frac{1}{|C_j|} \sum_{x_i \in C_j} x_i$$

- check convergence
  - if centroids move less than a small threshold, stop
  - other repeat the "assign cluster" and "update centroid type"

3. output

Return the final centroids $\{\mu_1, \mu_2, \ldots \mu_K\}$ and cluster assignments for each data point