Machine Learning
K-Means Clustering

DR. BHARGAVI R

SCOPE

VIT CHENNAI

Clustering -Introduction

- Unsupervised learning
- Clustering : Set of techniques for finding homogeneous subgroups/clusters, in a data set.
- Good clustering technique results in clusters with:
  - within each group are quite similar to each other (small within-cluster variation)
  - Observations in different groups are quite different from each other.

Applications

- Market segmentation
- Color compression
- Recommendation systems
- Document Analysis etc.

K – Means Clustering (Partition based)

- Well known Partition based clustering.
- Partition the observations into a pre-specified number of non-overlapping clusters (K number of clusters).
- Objective – Find K clusters such that the total within-cluster variation, summed over all K clusters, is as small as possible.

$$\min_{C_1, C_2, \ldots, C_K} \left\{ \sum_{k=1}^{K} \frac{1}{|C_k|} \sum_{i, i' \in C_k} \sum_{j=1}^{m} \left(x_{ij} - x_{i'j} \right)^2 \right\}$$

where $|C_k|$ denotes the number of observations in the k^{th} cluster

- This is difficult to solve since there are K^n ways to partition n observations into K clusters (unless n and K are small).

K – Means Clustering (cont…)

$$\frac{1}{|C_k|}\sum_{i,i' \in C_k}\sum_{j=1}^{m}\left(x_{ij}-x_{i'j}\right)^2 = 2\sum_{i, \in C_k}\sum_{j=1}^{m}\left(x_{ij}-\bar{x}_{kj}\right)^2$$ where

$\bar{x}_{kj} = \frac{1}{|C_k|}\sum_{i \in C_k} x_{ij}$ is the mean for feature j in cluster k

Algorithm

1. **Initialize:** Randomly assign each observation any cluster number from 1 to K. These serve as initial cluster assignments for the observations.
2. **Iterate:** Until the cluster assignments stop changing:
   a. For each of the K clusters, compute the cluster centroid. The k^{th} cluster centroid is the vector of the m feature means for the observations in the k^{th} cluster.
   b. Assign each observation to the cluster whose centroid is closest.

Example

Consider 8 instances of two dimensional data as (2,10), (2,5), (8,4), (5,8), (7,5), (6,4), (1,2), (4,9).

Let the initial centroids be first three instances. i.e C1(2,10), C2(2,5), and C3(8,4)

X_1	X_2	Distance C1	Distance C2	Distance C3	Cluster Assignment
2	10	0	5	8.485	C1
2	5	5	0	6.082	C2
8	4	8.485	6.082	0	C3
5	8	3.605	4.242	5	C1
7	5	7.071	5	1.414	C3
6	4	7.211	4.123	2	C3
1	2	8.062	3.162	7.28	C2
4	9	2.236	4.472	6.40	C1

Example (cont…)

- Now, update the centroids of the 3 clusters with updated instances.
  - Centroid of C1 (2, 10), (5, 8), (4, 9)= (3.66, 9)
  - Centroid of C2 (2, 5), (1, 2)= (1.5, 3.5)
  - Centroid of C3 (8, 4), (7, 5), (6, 4)= (7, 4.33)
- Repeat the distance computations and reassignment of clusters with new centroids till there is no change in cluster assignments or max iterations.

Discussion

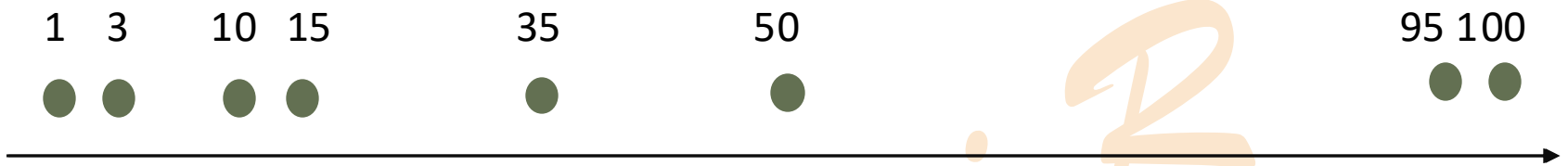
- Does not guarantee global optimum solution.
- Final clusters obtained will depend on the initial (random) cluster assignment.
  - Run the algorithm multiple times from different random initial configurations.
  - Then select the best solution for which the objective function has smallest value.

K-Means++

- K-Means with smarter initialization of the centroids to achieve quality clusters.
- Steps involved in initialization of K-Means ++
  - Randomly select the first centroid from the data points.
  - For each of the remaining data points compute its distance from the nearest, previously chosen centroid.
  - Select the next centroid from the data points such that the probability of choosing a point as centroid is directly proportional to its distance from the nearest, previously chosen centroid.
  - Repeat steps 2 and 3 until k centroids have been selected

Example

Consider the one dimensional dataset ------ 1, 3, 10, 15, 35, 50, 95, 100



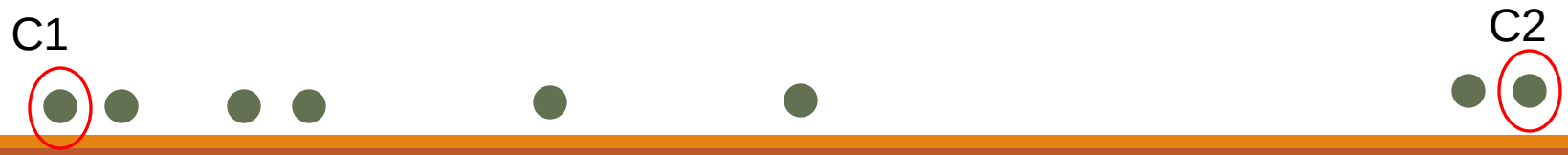
Let the initial cluster chosen be 1

To choose the next centroid we follow the steps defined in the k-means ++

Example (cont…)

Centroids existing	Data	Distance from {1}
{1}	3	2
	10	9
	15	14
	35	34
	50	49
	95	94
	100	99

Max. distance/ farthest point. Hence choose this data point



Example (cont…)

Centroids existing	Data	Distance from {1}	Distance from {100}	Distance from nearest existing centroids
{1, 100}	3	2	97	2
	10	9	90	9
	15	14	85	14
	35	34	65	34
	50	49	50	49
	95	94	5	5

Max. distance/ farthest point. Hence choose this data point