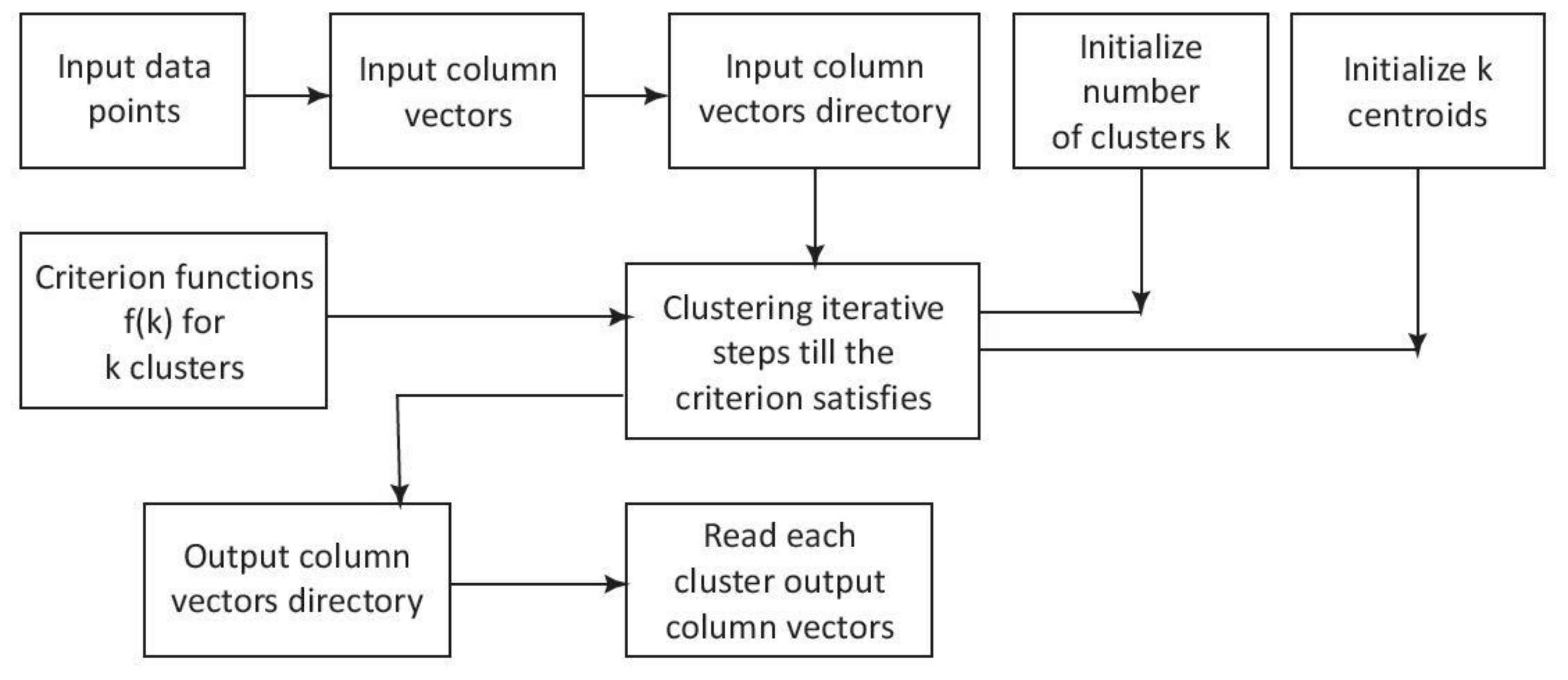
Lesson 11
Clustering Algorithms

Partitions/Centroid based Clustering Algorithms

- K-means, K-medoids, Fuzzy k-means, Mean-shift clustering and other related methods K-Means Clustering

Figure 6.10: K-means Clustering Algorithm



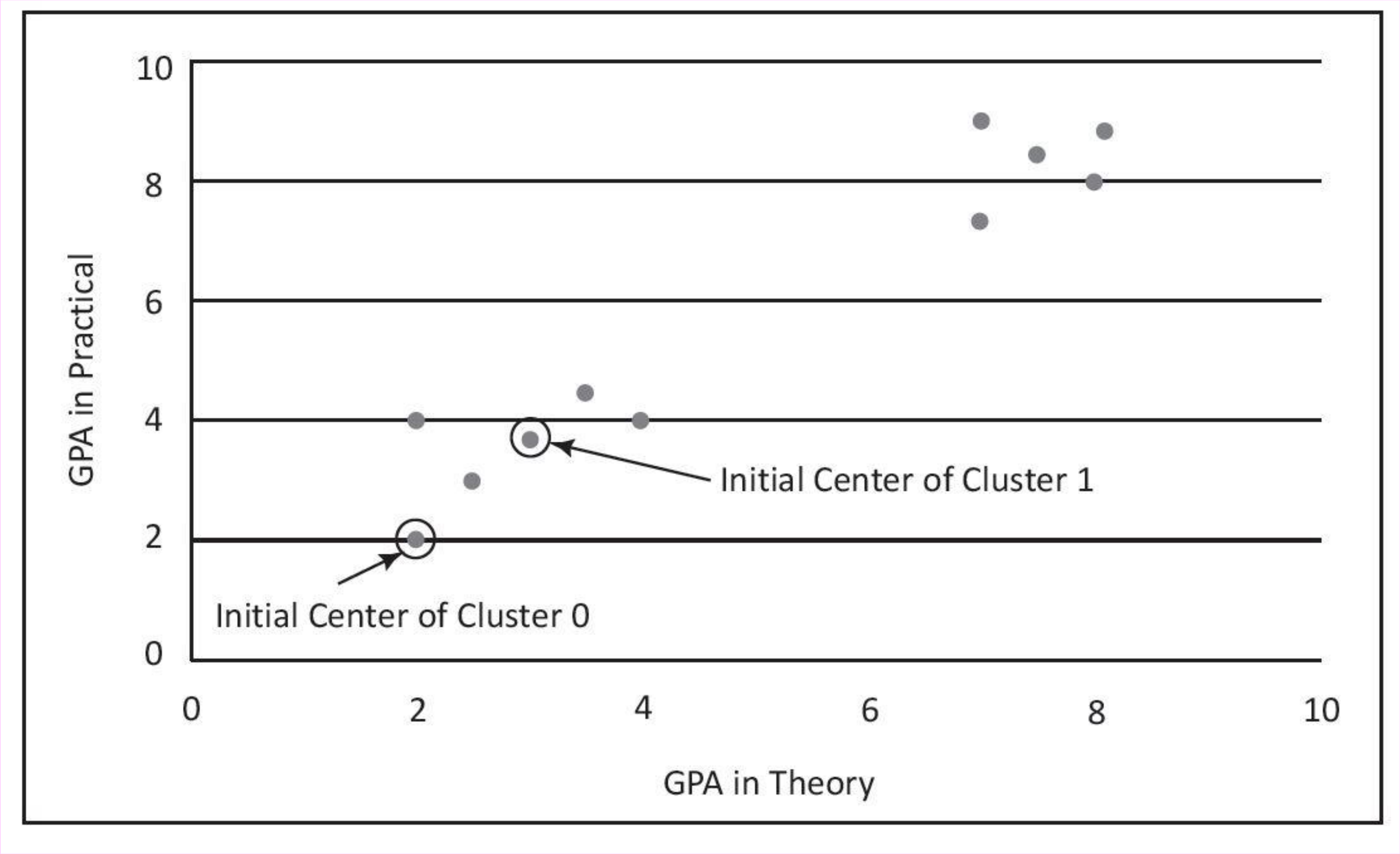
K-Means algorithm

- MacQueen (1967)
- Simplest unsupervised learning algorithms for clustering
- Groups the objects based on the attributes (features) into k number of groups where k is a positive integer number

K-Means Step 1

1. Randomly initialize the k cluster centroid points (= **C1**, **C**2, ..., **Ck**) as k-partition centers which mean partitions with these cluster centroids

Figure 6.11 Initialization



K-Means Step 2

2. Go through each of the data points and assign points to a cluster where the distance from a centroid is minimum.

K-Means Step 3

3. Identify the centroid of the new cluster formed, Centroid is average of all the data points in a cluster

. Algorithm calculates the average of all the points in a cluster, and moves the centroid to that average location

K-Means Step 4

4. Repeat until no change in the clusters takes place (or possibly until some other stopping condition is met, more than Threshold % points already in the cluster)

K-Means Algorithm Inputs

(1) Input: N (objects) from inputs column vectors directory, and initialized value of k (the number of clusters)

K-Means Algorithm Outputs

(2) Output: in Output column vectors directory for a set of k clusters, on using the criterion-functions f(k)

K-Means Algorithm Centroid Output

- The centroid C_k for each cluster computes after minimizing the sum of squared distances (Euclidean distances from the) $\sum D_{Eu}^2$.
- D_{Eu} is between data point in a cluster from C_k

Algorithm: Iterative relocation algorithm

- (i) Initialize k centroids as the initial solution.
- (ii) (Re) compute memberships for the objects using the current cluster centroids

Properties of K-means

1. Number of clusters which form are always k clusters, where k is the number of partition centers cluster.
2. Each cluster consists of at least one object in each cluster

Properties of K-means

3. The clusters are flat (non-hierarchical) and they do not overlap
4. Every member object of a cluster is closer to its cluster than any other cluster.

Function O (f (n))

- Refers to efficiency of algorithm in terms of function f (n) [O is called big O notation.]
- The n = Number of data points taken in the algorithm

Function O (f (n))

- If f (n) = n^2, then the requirement of the algorithm is proportional to n^2
- Space requirement O ($n \times d$) means that memory taken by the algorithm is proportional to $n \times d$.

O (f (n)) for K-means

- Advantage: Computing the distances between points and group centres has linear complexity O (*n)*.

Disadvantages of K-Means

(i) need to choose k,

(ii) need to start and randomly choose the cluster centres, the results may be choice dependent.

(iii) Less consistency of the results compared to other methods.

K-Means Use Cases to solve a number of real-life situations

- Identifying abnormal data items in a very large dataset
- For example, identifying potentially fraudulent credit card transactions, risky loan applications and medical claim fraud detection.

Use Cases

- An image retrieval system using similarities
- Finding diabetic/non-diabetic or hypertension/non-hypertension group structure from the input value
- Finding segment of customers and customer category using the spending behavior characteristic

K-Medoids

- Similar to a mean or centroid, but restricts to members of the dataset
- A dataset may have more than one medoid

K-Medoids algorithm

- Initializes k data points as exemplars (centers), which shift iteratively for minimizing dissimilarities
- The algorithm K-medoids does clustering using an algorithm, which has flavours of k-means algorithm and medoid-shift algorithm.

Algorithm Steps

1. Step 1: Choose a set of medoids for the data-points.
2. Step 2: Compute distances from each medoid to other data-points.

Algorithm Steps

3. Step 3: Cluster the data points according to their similarities with the medoid.
4. Step 4: Optimize the set of medoids using iterative process.

Iterations

- The sum of pair of dissimilarities minimizes in K-medoids compared to minimizing the sum of squared Euclidean distances in K-means

Advantage of K-Medoids Method

- Uses in graphs and other non-metric spaces
- Average of dissimilarities minimizes taking all cluster members (objects)..

Connectivity and spectrum based Clustering Algorithm

- Hierarchical clustering, when closeness relates to connectivity then spectral clustering

Figure 6.12 Original object points and one hierarchical cluster representation of those object points

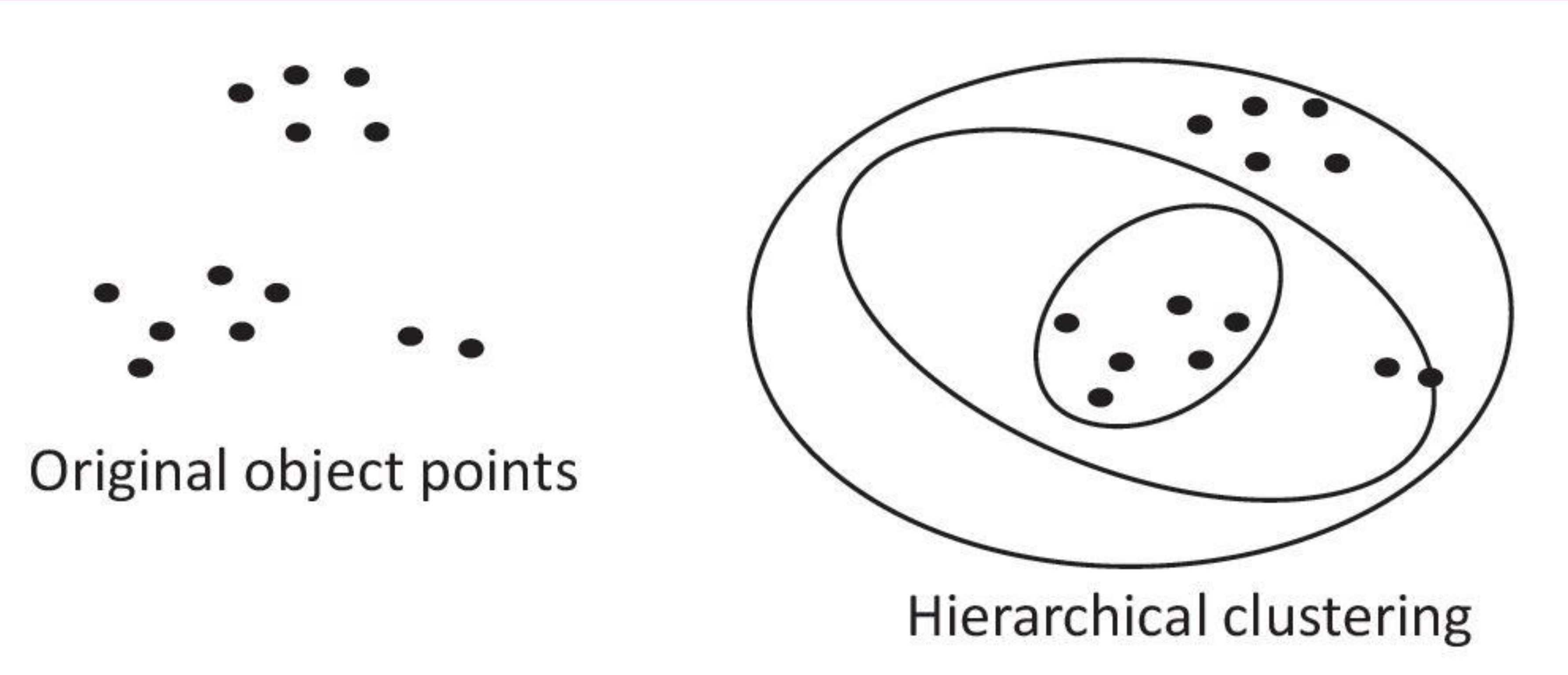
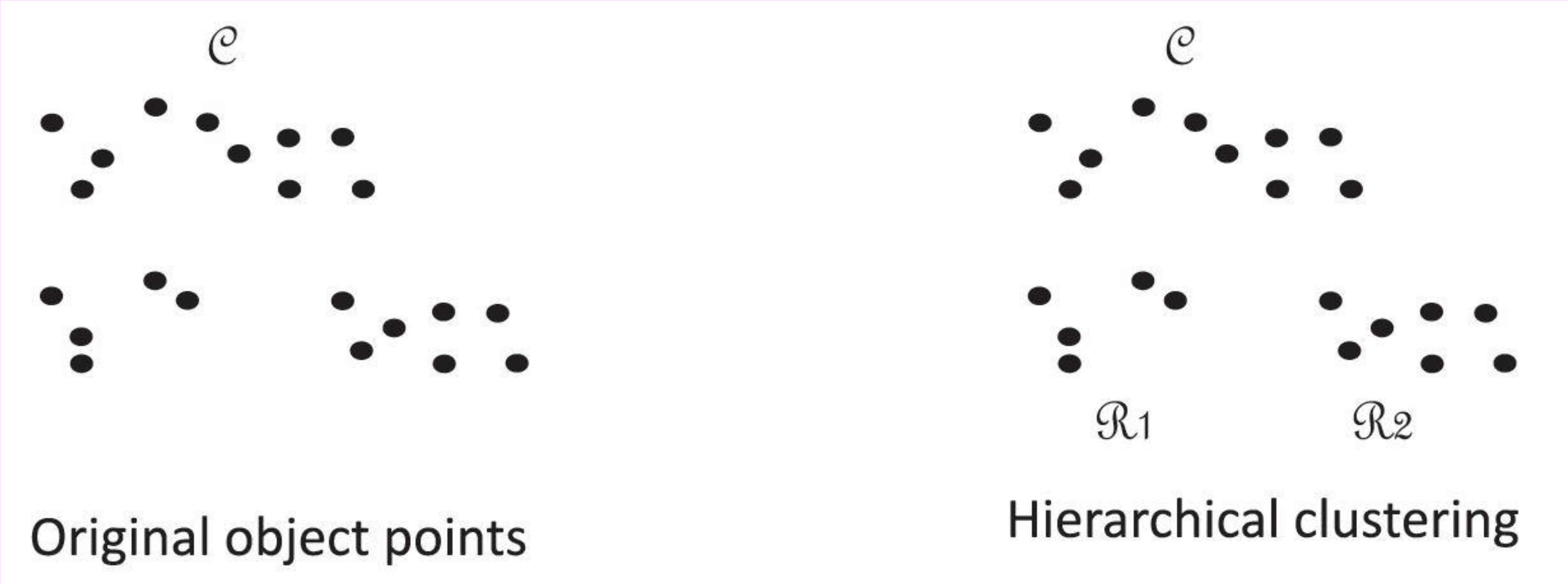


Figure 6.13 Hierarchical clustering of (i) original object points in city C showing high total sales per day, (ii) clusters of j set of regions R1 and R2 showing high total sales per day.



Probabilistic distribution based

- Latent-Dirichlet-Allocation (LDA) (Section 6.9), Gaussian Mixture Model (GMM), Expectation Maximization (EM) clustering and others, [Expectation Maximization (EM) algorithm uses a set of parameters that maximize the probability of the chosen PDF for data as a metric.]

Expectation Maximization

- Algorithm uses a set of parameters that maximize the probability of the chosen PDF for the data as a metric

PCA Clustering Algorithm

- Dimensionality reduction based Principal Component Analysis (PCA)

DBSCAN Clustering Algorithm

- Density based Density-Based Spatial Clustering of Applications with Noise (DBSCAN)

Neural Networks/Deep Learning Clustering Algorithms

- — Auto-encoders, self-organizing maps

Summary

We learnt:

- Clustering Algorithms
- K-Means Algorithm
- K-Medoids Algorithm
- Hierarchical Clustering
- Other Methods

End of Lesson 11 on Clustering Algorithms