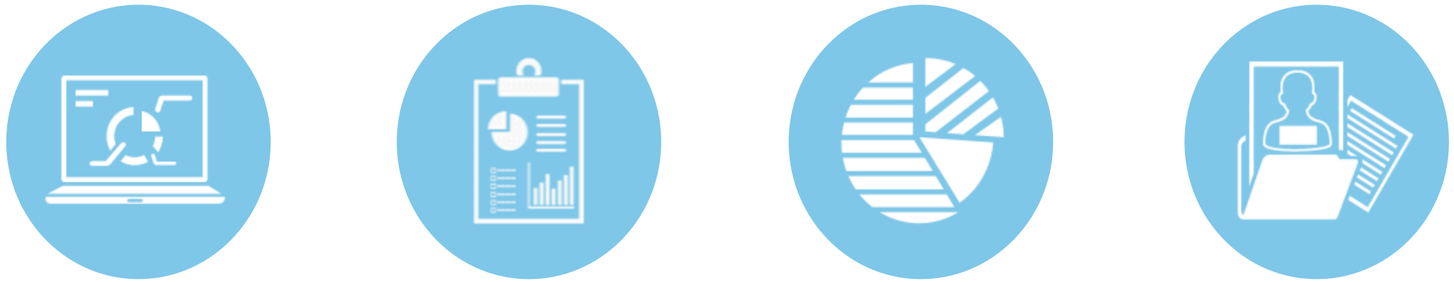
Data Science with R

Lesson 9— Clustering



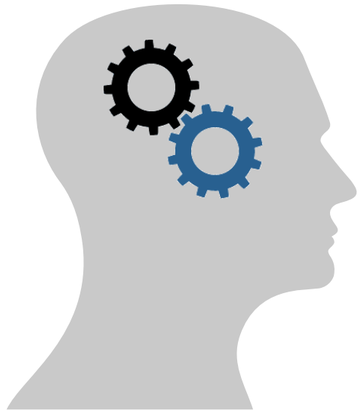
Learning Objectives

- Define clustering
- List clustering methods

Clustering

Topic 1—Introduction to Clustering

Introduction

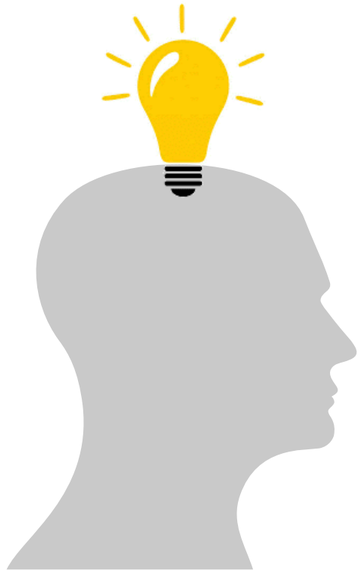


Consider a scenario where a pizza store wants to open a new center.

To choose the best location, it should analyze factors such as distance, accessibility, ease of delivery, population, etc.

Keeping all these factors in mind, how can the best location be predicted?

Introduction



The team should conduct a thorough analysis that would help in understanding how the delivery locations can be grouped, hence reducing the average distance for both people and delivery executives.

This can be done using clustering algorithms.

What Is Clustering?

Cluster analysis or clustering is the most commonly used technique of unsupervised learning used to find data clusters so that each cluster has the most closely matched data.

Unsupervised Learning is a subset of Machine Learning used to extract inferences from datasets that consist of input data without labeled responses.

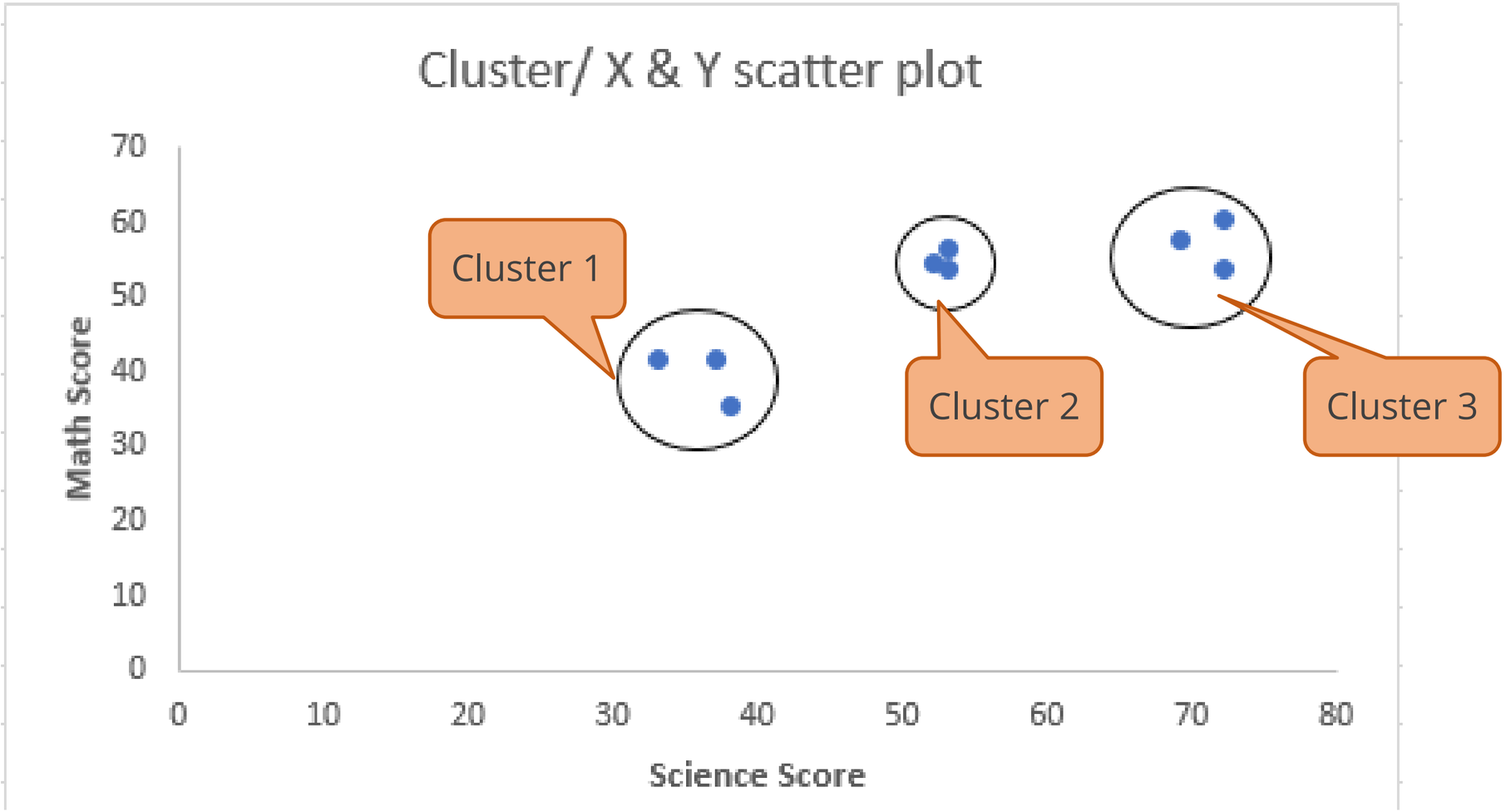
Clustering: Example

Consider a scenario where you need to create a cluster/group of students who are of similar aptitude using clustering. The following data is available.

ID	Math	Science
1	37	42
2	33	42
3	38	36
4	53	54
5	52	55
6	53	57
7	69	58
8	72	54
9	72	61

Clustering: Example

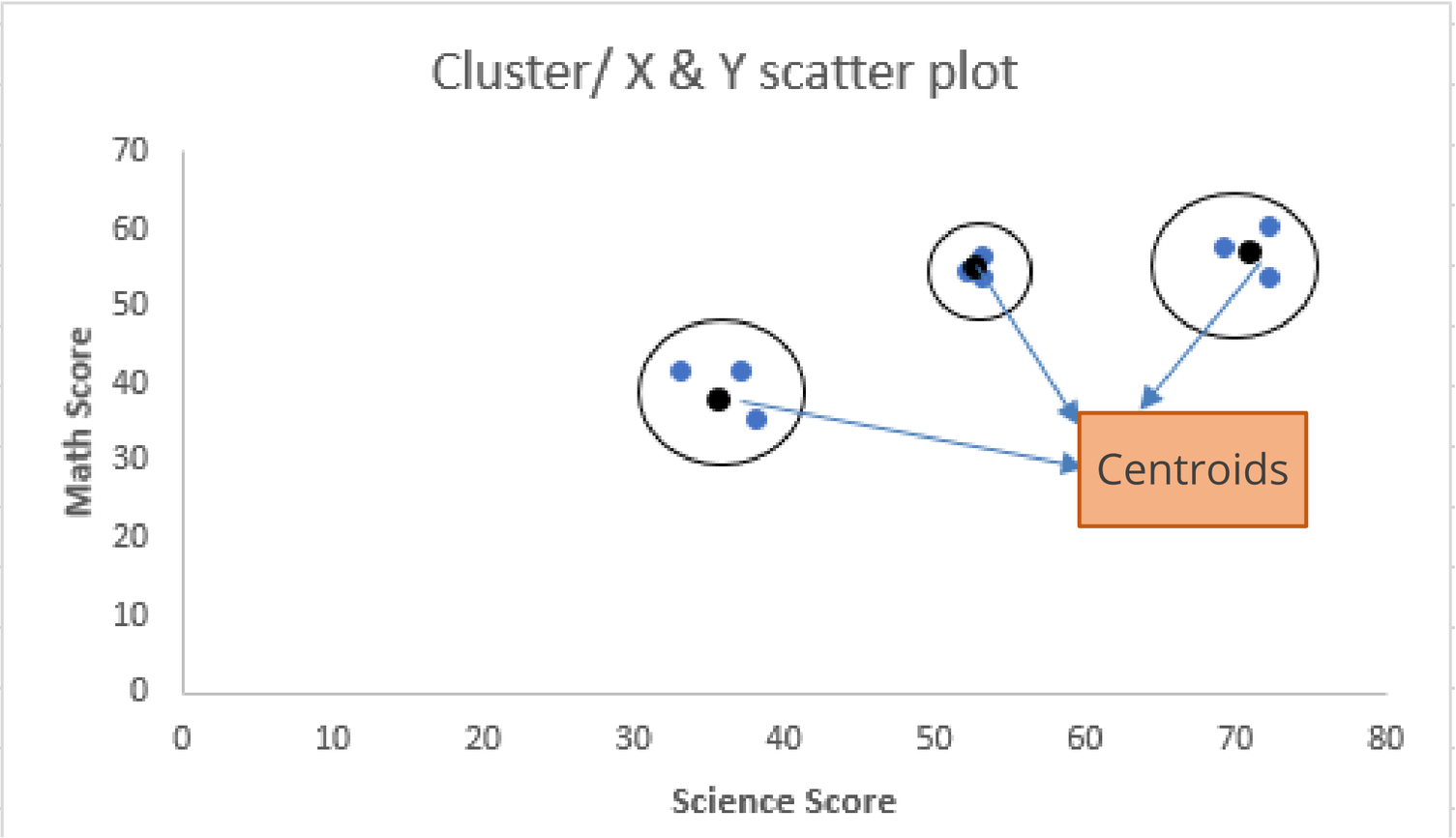
PLOTTING THE OBSERVATION



Clustering: Example

CENTROIDS

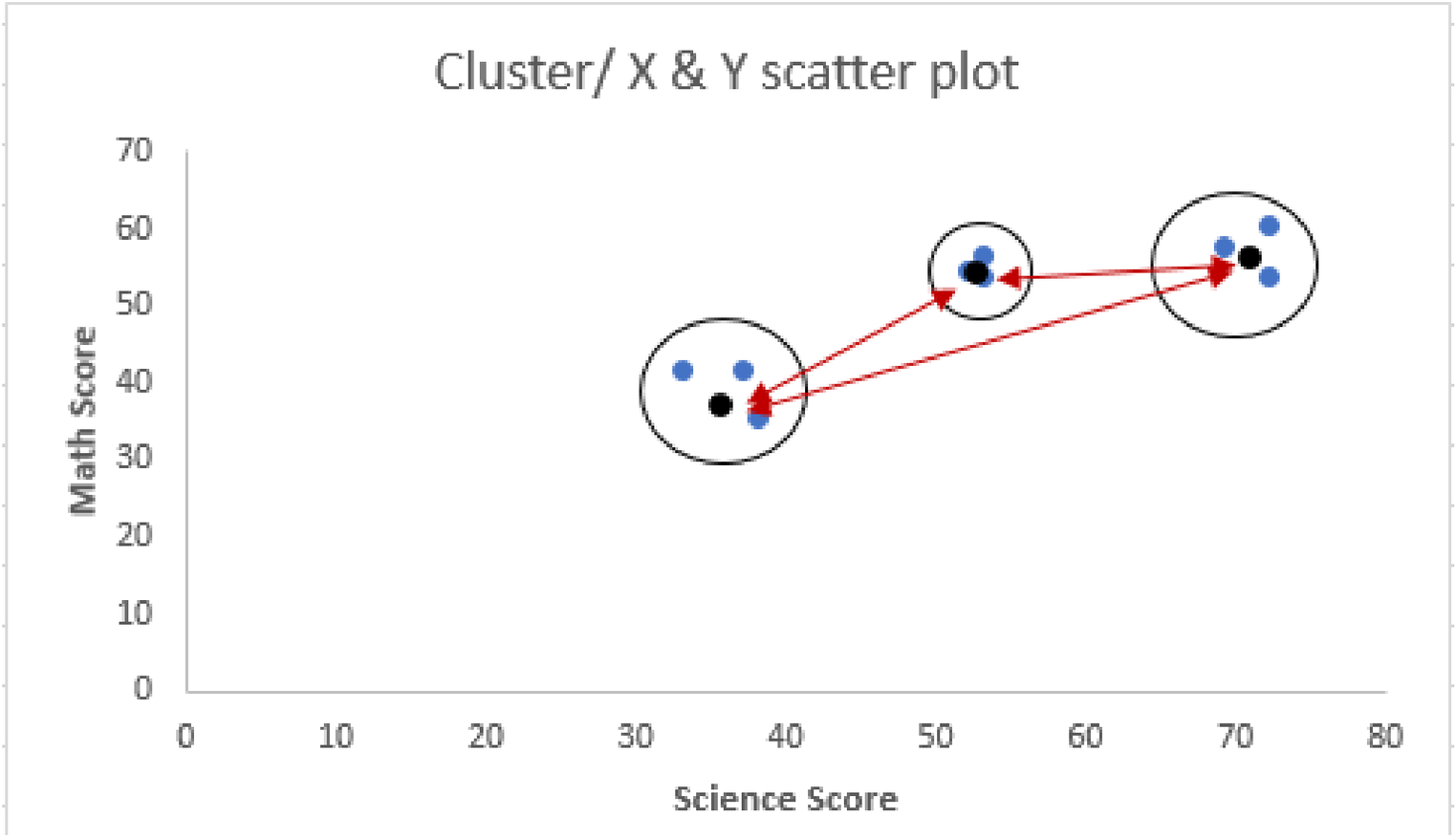
Each of these clusters has center points, called centroids.



Clustering: Example

DISTANCE BETWEEN CLUSTERS

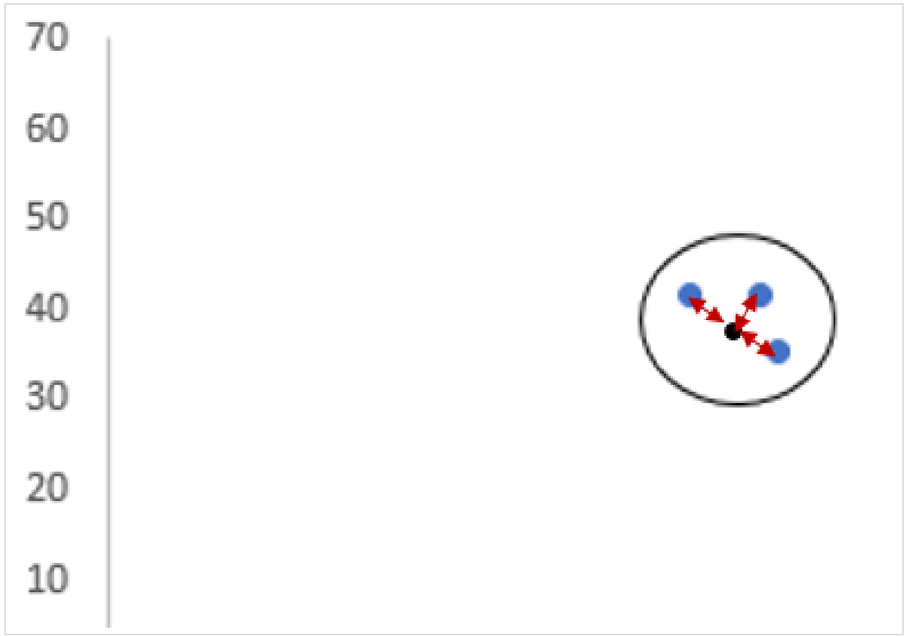
Distance between the cluster centroids is termed distance between clusters.



Clustering: Example

DISTANCE WITHIN CLUSTERS

Average distance of observation in a cluster from its cluster centroid is called distance within cluster.



Other Examples of Clustering

- Grouping the content of a website or product in a retail business
- Segmenting customers or users into different groups on the basis of their metadata and behavioral characteristics
- Segmenting communities in ecology
- Finding clusters of similar genes in DNA analysis
- Creating image segments to be used in image analysis applications

All of this is done using various clustering methods.

Clustering

Topic 2—Clustering Methods

Clustering Methods

Prototype-based Clustering

Hierarchical Clustering

Density-based Clustering (DBSCAN)

Clustering Methods

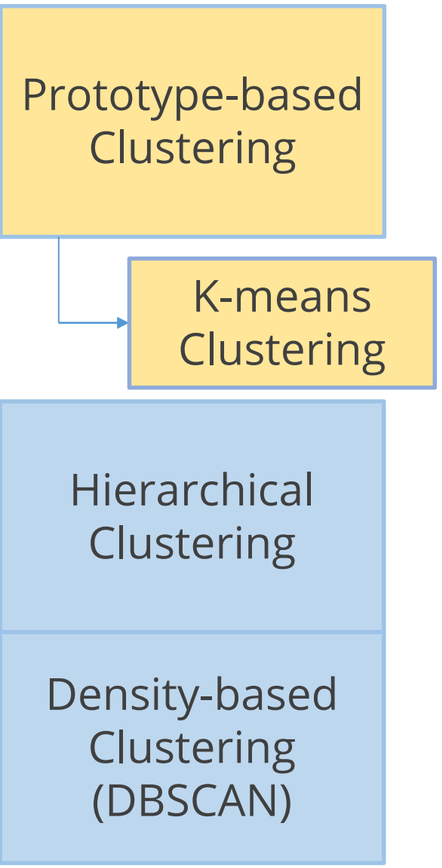
- Prototype-based Clustering
- Hierarchical Clustering
- Density-based Clustering (DBSCAN)

Prototype-based clustering assumes that most of the data is located near prototypes (element of data space representing a group of elements).

Example: centroid (average) or medoid (most frequently occurring point)

It is widely used in banking and sports stat predictions to provide robustifying efforts based on statistics.

Clustering Methods



K-means, a Prototype-based method, is the most popular method for clustering. It involves:

- Assigning training data to matching cluster based on similarity.
- Iterative process to get data points in the best clusters possible

Clustering Methods

K-MEANS CLUSTERING: EXAMPLE

- Prototype-based Clustering
  - K-means Clustering
- Hierarchical Clustering
- Density-based Clustering (DBSCAN)

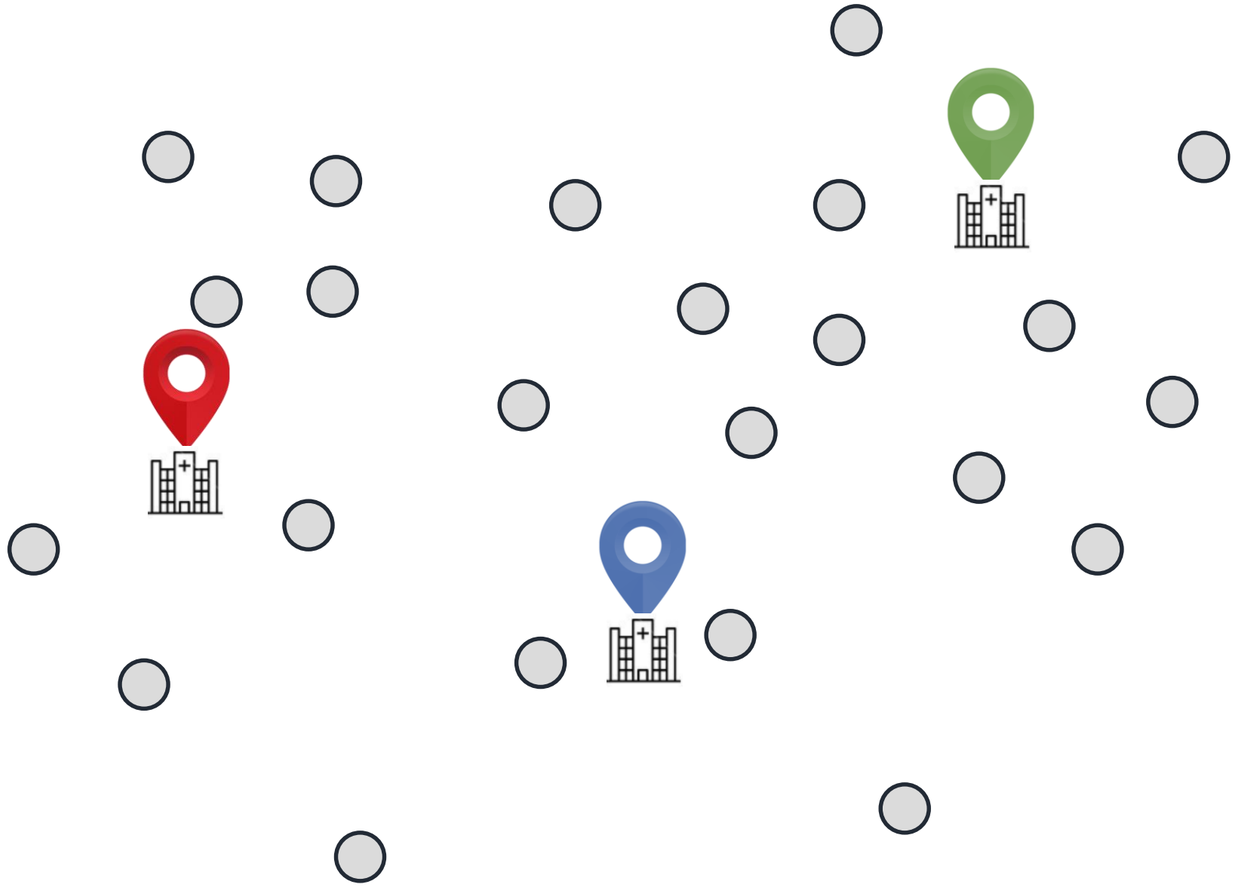
The government of California wants to identify high density clusters to build hospitals. (No other ground truth or features are provided apart from the population data). How can the clusters be identified?

Clustering Methods

K-MEANS CLUSTERING: EXAMPLE

- Prototype-based Clustering
  - K-means Clustering
- Hierarchical Clustering
- Density-based Clustering (DBSCAN)

Start by picking k random centroids. Assume, k = 3.

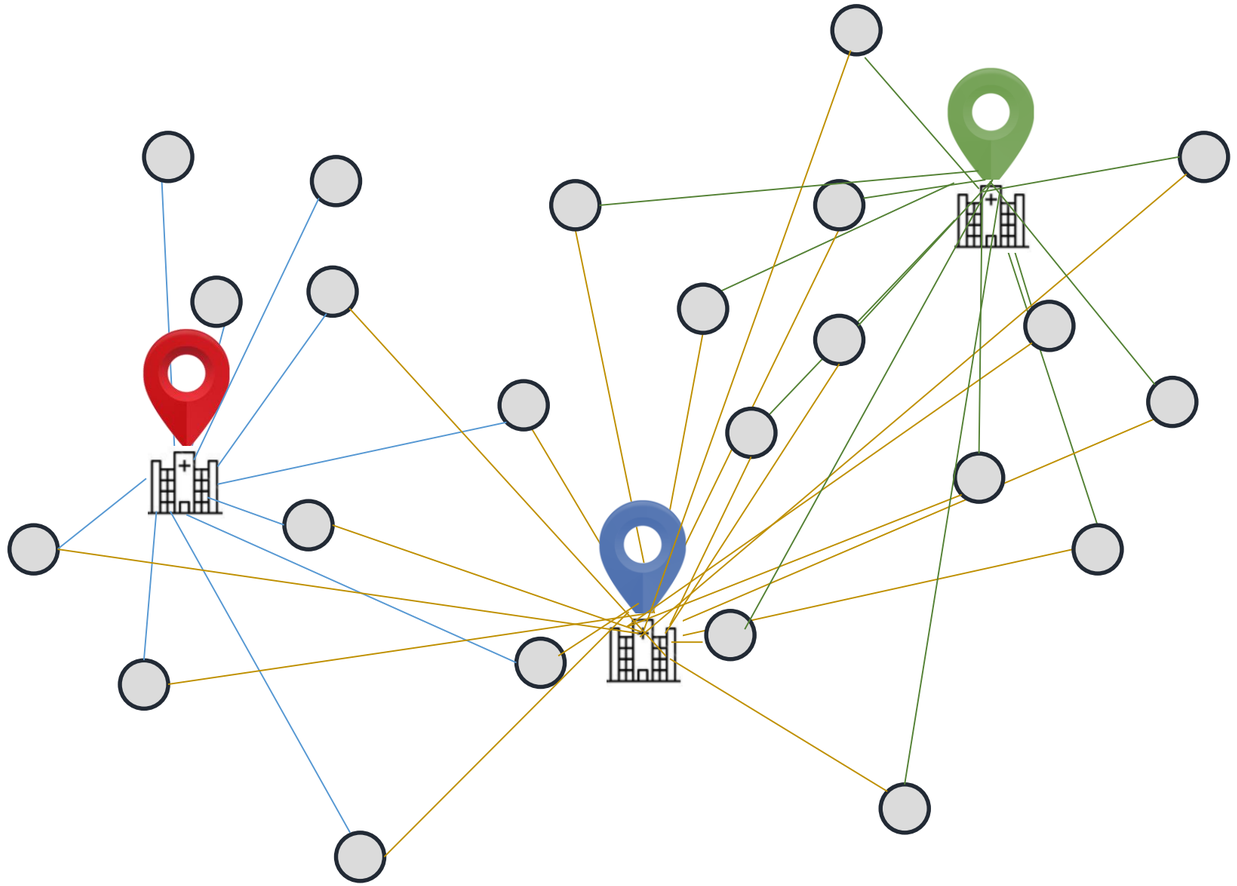


Clustering Methods

K-MEANS CLUSTERING: EXAMPLE

- Prototype-based Clustering
  - K-means Clustering
- Hierarchical Clustering
- Density-based Clustering (DBSCAN)

Assign each point to the nearest centroid.

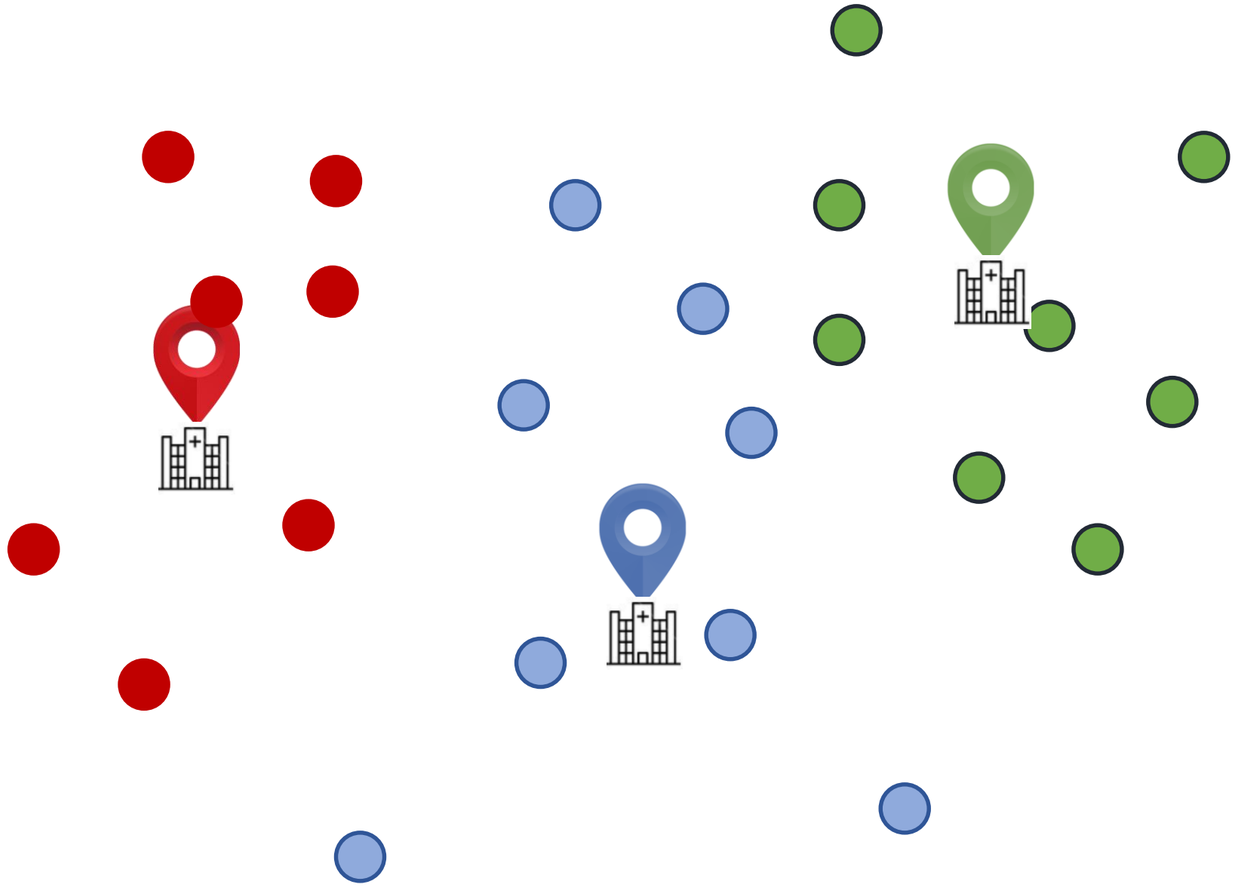


Clustering Methods

K-MEANS CLUSTERING: EXAMPLE

- Prototype-based Clustering
  - K-means Clustering
- Hierarchical Clustering
- Density-based Clustering (DBSCAN)

Move each centroid to the center of the respective cluster.

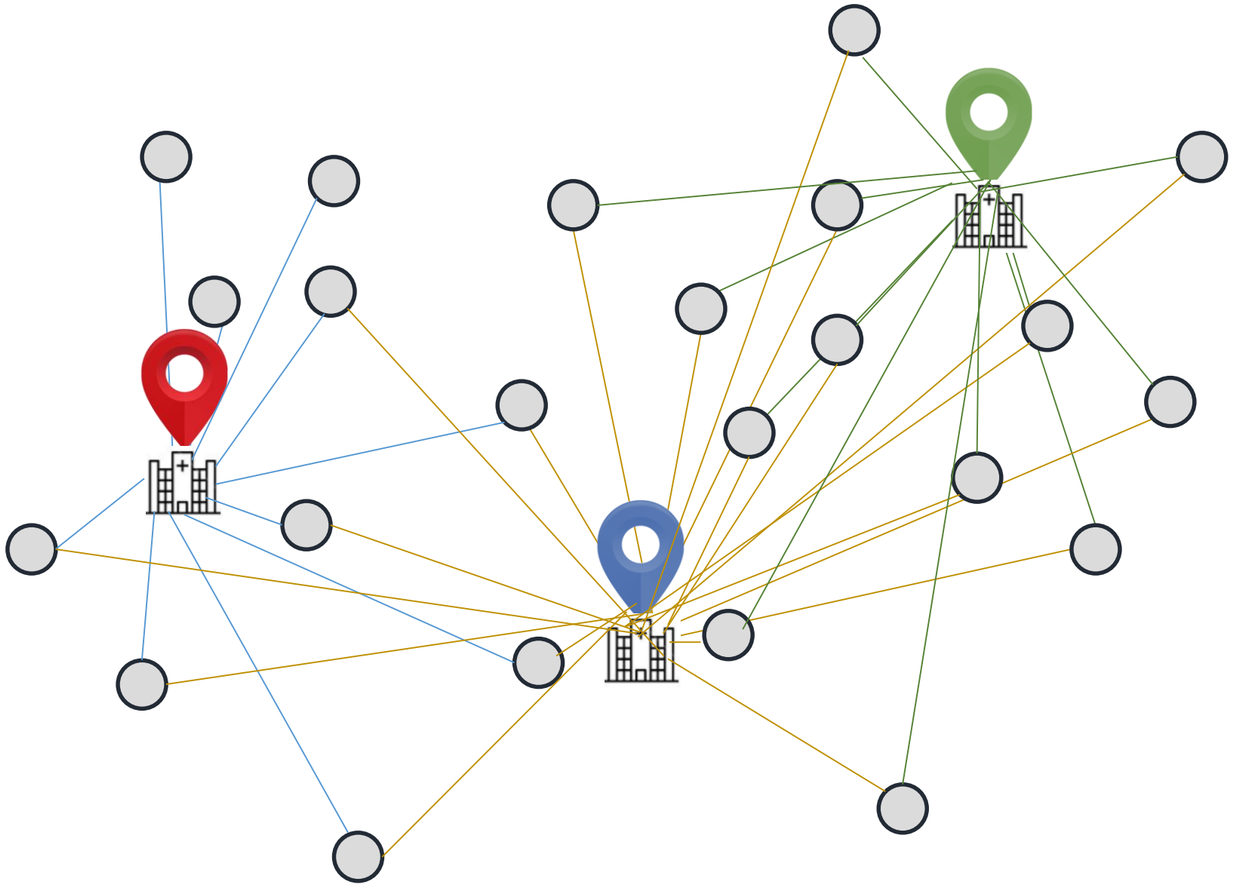


Clustering Methods

K-MEANS CLUSTERING: EXAMPLE

- Prototype-based Clustering
  - K-means Clustering
- Hierarchical Clustering
- Density-based Clustering (DBSCAN)

Calculate the distance of the centroids from each point again.

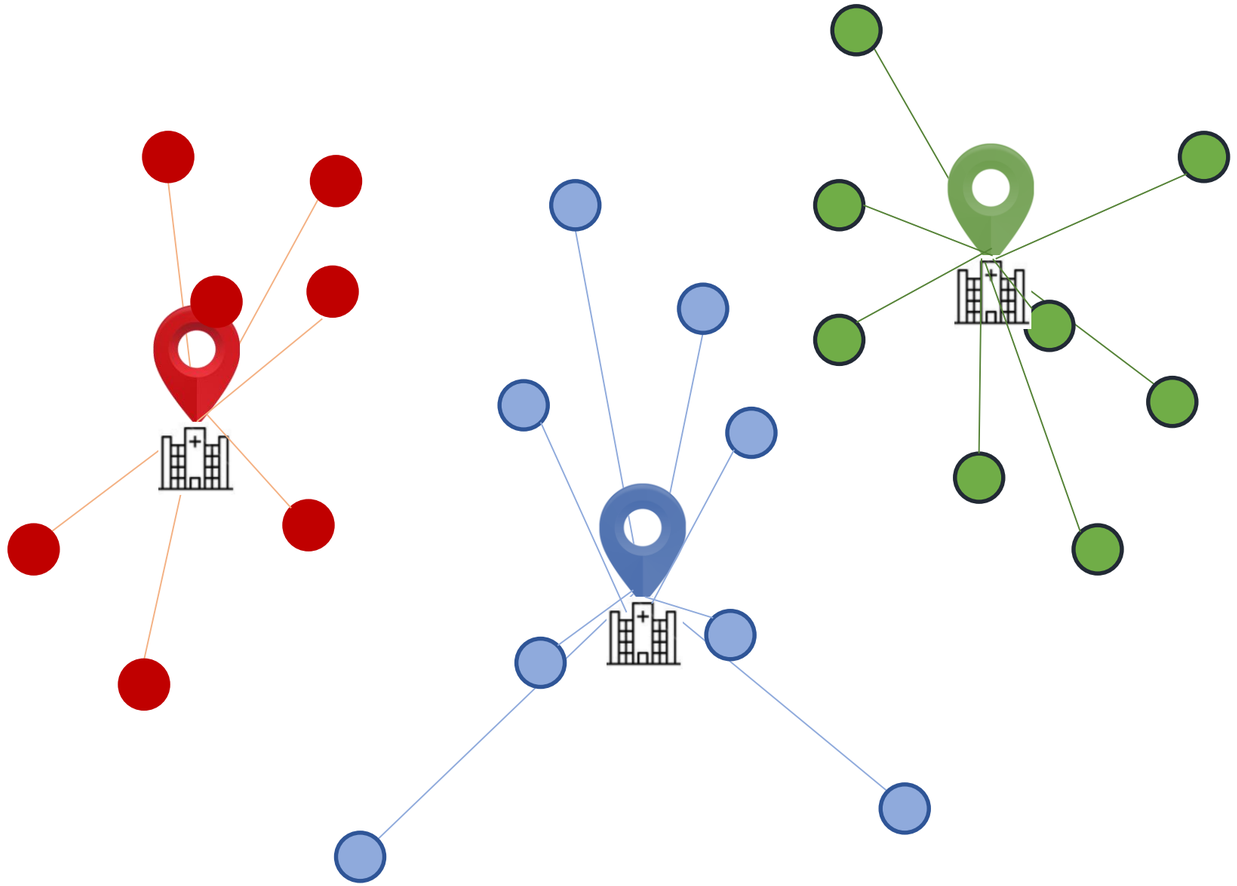


Clustering Methods

K-MEANS CLUSTERING: EXAMPLE

- Prototype-based Clustering
  - K-means Clustering
- Hierarchical Clustering
- Density-based Clustering (DBSCAN)

Move points across clusters and re-calculate the distance from the centroid.

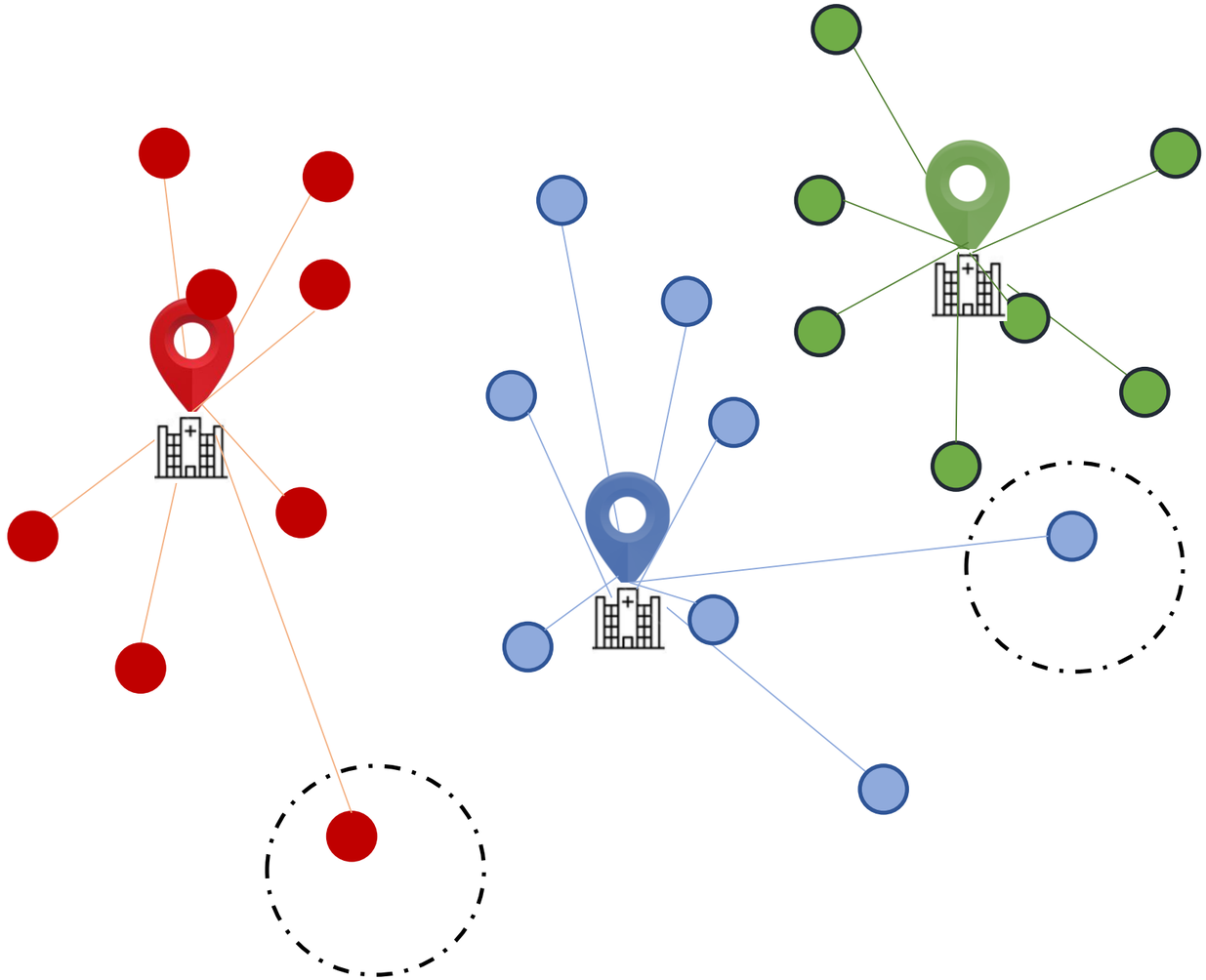


Clustering Methods

K-MEANS CLUSTERING: EXAMPLE

- Prototype-based Clustering
  - K-means Clustering
- Hierarchical Clustering
- Density-based Clustering (DBSCAN)

Keep moving the points across clusters until the distance from the center is minimized.

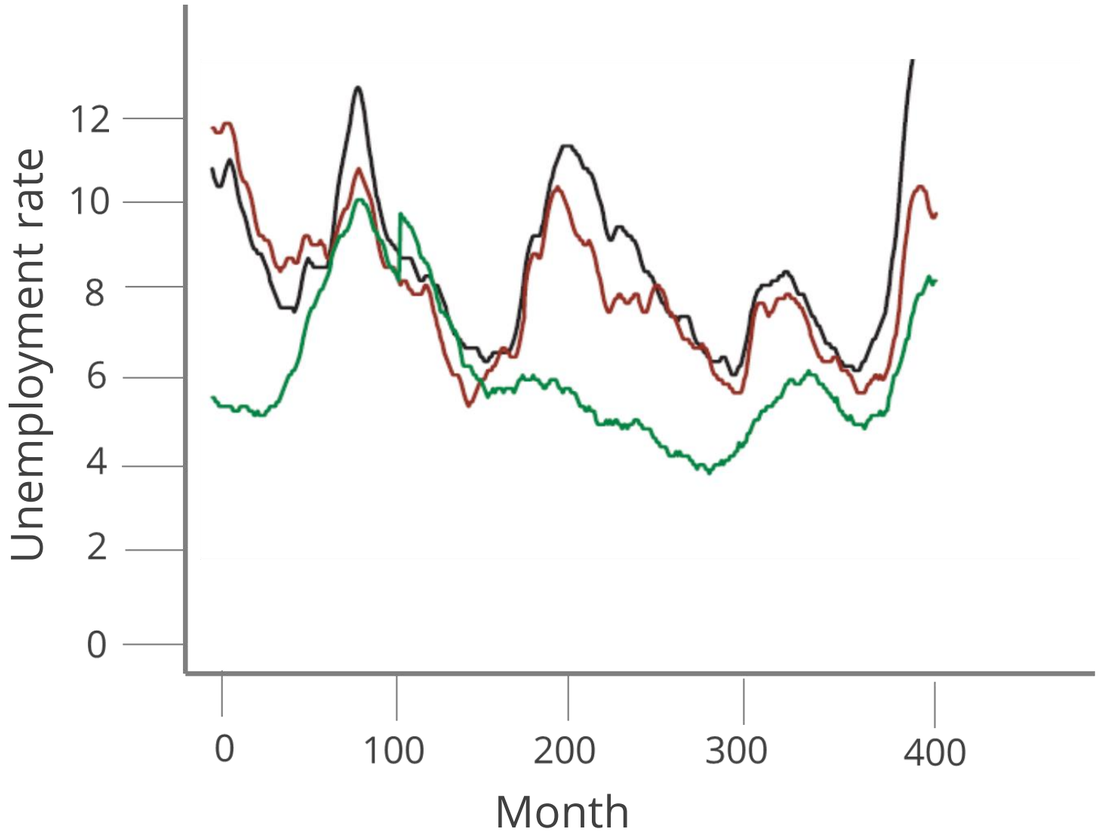


Clustering Methods

K-MEANS CLUSTERING USING R: CASE STUDY

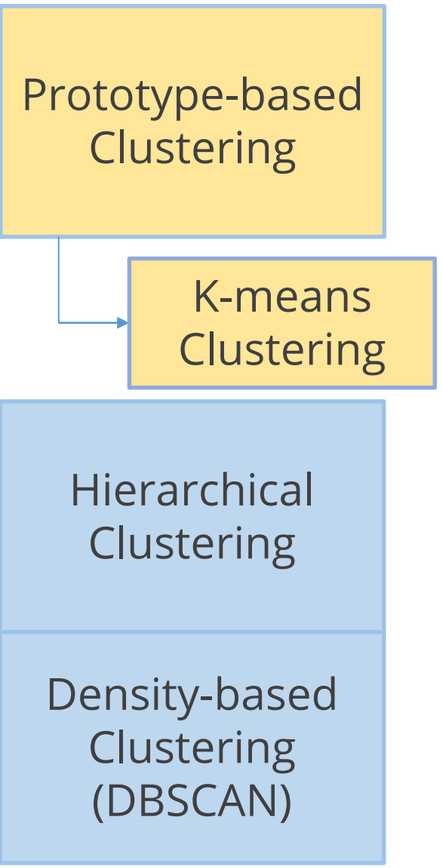
- Prototype-based Clustering
  - K-means Clustering
- Hierarchical Clustering
- Density-based Clustering (DBSCAN)

The monthly and seasonal adjusted unemployment rates, from January 1976 to August 2010, for 50 U.S. states were captured. The graph below shows the time series plots of three states: Iowa (green), New York (red), and California (black). Cluster states group wise.



Clustering Methods

K-MEANS CLUSTERING USING R: CASE STUDY



Assume:

- Each state is characterized by a feature vector, with p = 416.
- New York and California form a cluster.

Calculate the 416 monthly averages with two observations each.

Clustering Methods

K-MEANS CLUSTERING USING R: CASE STUDY

- Prototype-based Clustering
  - K-means Clustering
- Hierarchical Clustering
- Density-based Clustering (DBSCAN)

```
## read the data; series are stored column-wise with labels in
first ## row
raw <- read.csv("C:/DataMining/Data/unempstates.csv")
## transpose the data then we have 50 rows (states) and 416
columns (time periods)
rawt=matrix(nrow=50,ncol=416)
rawt=t(raw)
## k-means clustering in 416 dimensions
set.seed(1)
grpunemp2 <- kmeans(rawt, centers=2, nstart=10)
sort(grpunemp2$cluster)
grpunemp3 <- kmeans(rawt, centers=3, nstart=10)
sort(grpunemp3$cluster)
grpunemp4 <- kmeans(rawt, centers=4, nstart=10)
sort(grpunemp4$cluster)
grpunemp5 <- kmeans(rawt, centers=5, nstart=10)
sort(grpunemp5$cluster)
```

Clustering Methods

- Prototype-based Clustering
- Hierarchical Clustering
- Density-based Clustering (DBSCAN)

It clusters n units/objects, each with p features, into smaller groups and creates a hierarchy of clusters as a dendrogram.

Dendrograms are units in the same cluster joined by a horizontal line. They provide a visual representation of clusters.

Clustering Methods

- Prototype-based Clustering
- Hierarchical Clustering
- Density-based Clustering (DBSCAN)

They are of two types of Hierarchical clustering:

Type	Method	Approach
Agglomerative clustering	Starts at the individual leaves and successively merges clusters together	Bottom-up
Divisive clustering	Starts at the root and recursively splits the clusters	Top-down

Clustering Methods

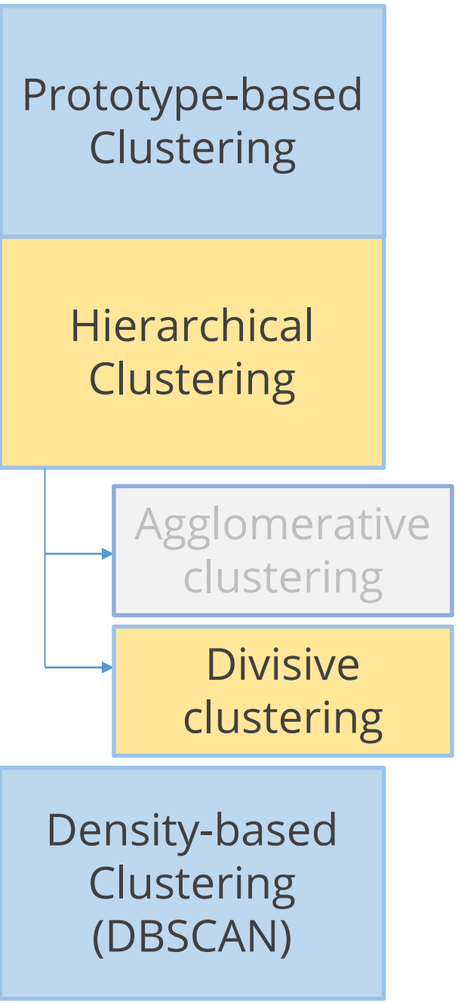
- Prototype-based Clustering
- Hierarchical Clustering
  - Agglomerative clustering
  - Divisive clustering
- Density-based Clustering (DBSCAN)

Agglomerative clustering is a process where:

- An n × n distance matrix is considered, where the number in the i^{th} row and j^{th} column is the distance between the i^{th} and j^{th} units.
- The distance matrix is symmetric with zeros in the diagonal.
- Rows and columns are merged as clusters and the distances between them are updated.

For R package cluster, use the **agnes** function.
For stats package, use the **hclust** function.

Clustering Methods



Divisive clustering is a “top down” clustering approach (all observations start in one cluster, and splits are performed recursively as one moves down the hierarchy) used in practical applications of image retrieval.

wikipedia

Clustering Methods

CASE – STUDY

- Prototype-based Clustering
- Hierarchical Clustering
- Density-based Clustering (DBSCAN)

The protein intakes in 25 European countries were captured across 9 food sources, as given in the table below:

Country	Red Meat	White Meat	Eggs	Milk	Fish	Cereals	Starch	Nuts	Fr&Veg
Albania	10.1	1.4	0.5	8.9	0.2	42.3	0.6	5.5	1.7
Austria	8.9	14	4.3	19.9	2.1	28	3.6	1.3	4.3
Belgium	13.5	9.3	4.1	17.5	4.5	26.6	5.7	2.1	4
Bulgaria	7.8	6	1.6	8.3	1.2	56.7	1.1	3.7	4.2
Czechoslovakia	9.7	11.4	2.8	12.5	2	34.3	5	1.1	4
Denmark	10.6	10.8	3.7	25	9.9	21.9	4.8	0.7	2.4
E Germany	8.4	11.6	3.7	11.1	5.4	24.6	6.5	0.8	3.6
Finland	9.5	4.9	2.7	33.7	5.8	26.3	5.1	1	1.4
France	18	9.9	3.3	19.5	5.7	28.1	4.8	2.4	6.5
Greece	10.2	3	2.8	17.6	5.9	41.7	2.2	7.8	6.5
Hungary	5.3	12.4	2.9	9.7	0.3	40.1	4	5.4	4.2
Ireland	13.9	10	4.7	25.8	2.2	24	6.2	1.6	2.9
Italy	9	5.1	2.9	13.7	3.4	36.8	2.1	4.3	6.7
Netherlands	9.5	13.6	3.6	23.4	2.5	22.4	4.2	1.8	3.7
Norway	9.4	4.7	2.7	23.3	9.7	23	4.6	1.6	2.7
Poland	6.9	10.2	2.7	19.3	3	36.1	5.9	2	6.6
Portugal	6.2	3.7	1.1	4.9	14.2	27	5.9	4.7	7.9
Romania	6.2	6.3	1.5	11.1	1	49.6	3.1	5.3	2.8
Spain	7.1	3.4	3.1	8.6	7	29.2	5.7	5.9	7.2
Sweden	9.9	7.8	3.5	24.7	7.5	19.5	3.7	1.4	2
Switzerland	13.1	10.1	3.1	23.8	2.3	25.6	2.8	2.4	4.9
United Kingdom	17.4	5.7	4.7	20.6	4.3	24.3	4.7	3.4	3.3
USSR	9.3	4.6	2.1	16.6	3	43.6	6.4	3.4	2.9
W Germany	11.4	12.5	4.1	18.8	3.4	18.6	5.2	1.5	3.8
Yugoslavia	4.4	5	1.2	9.5	0.6	55.9	3	5.7	3.2

Determine whether the listed 25 countries can be separated into a smaller number of clusters.

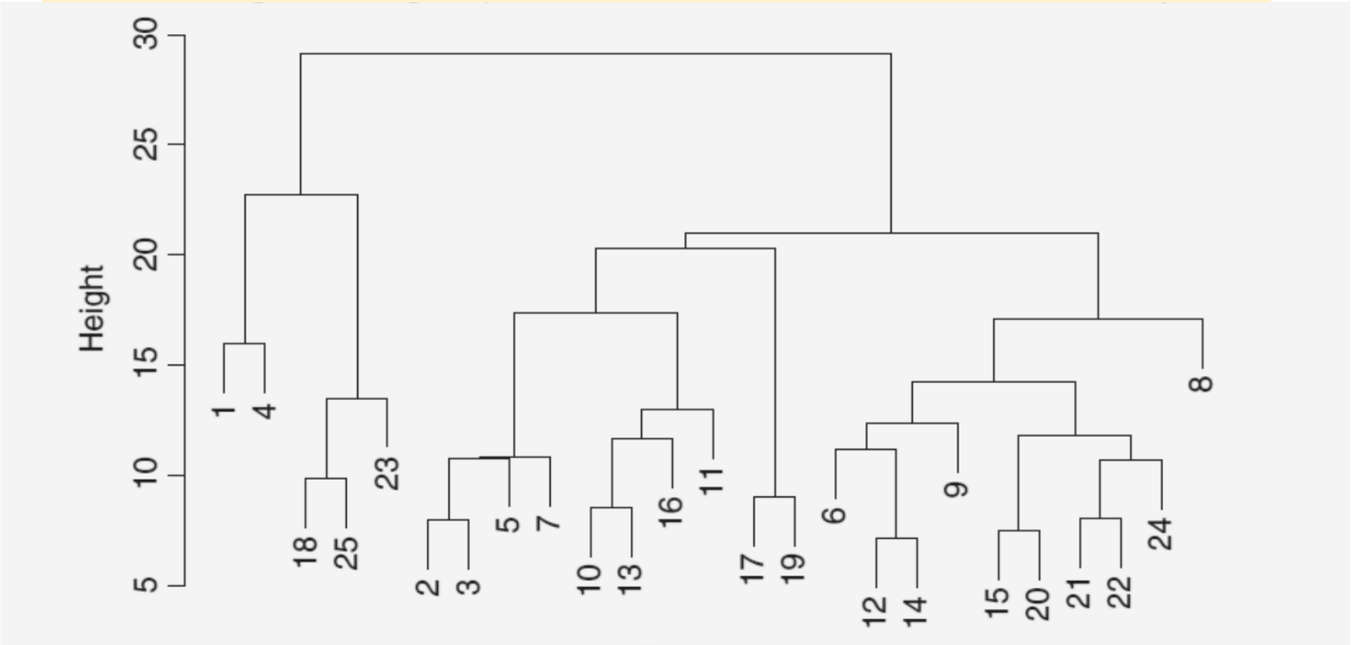
Clustering Methods

CASE STUDY

- Prototype-based Clustering
- Hierarchical Clustering
- Density-based Clustering (DBSCAN)

```
library(cluster)
food <- read.csv("C:/DataMining/Data/protein.csv")
foodagg=agnes(food,diss=FALSE,metric="euclidian")
plot(foodagg) ## dendrogram
```

Dendogram of agnes(x = food, diss = FALSE, metric = “Euclidian”



Food
Agglomerative coefficient = 0.64

Clustering Methods

- Prototype-based Clustering
- Hierarchical Clustering
- Density-based Clustering (DBSCAN)

DBSCAN (Density-Based Spatial Clustering and Application with Noise) is used to identify clusters of any shape in a data set containing noise and outliers.

DBSCAN algorithms are used to find associations and structures in data and predict trends.

Key Takeaways

- Cluster analysis or clustering is the most commonly used technique of unsupervised learning to find data clusters such that each cluster has most closely matched data.
- Prototype-based clustering assumes that most of the data is located near prototypes (element of data space representing a group of elements).
- K-means is a Prototype-based method for clustering that involves assigning training data to matching cluster based on similarity and using an Iterative process to get data points in the best clusters possible.
- Hierarchical Clustering clusters n units/objects, each with p features, into smaller groups and creates a hierarchy of clusters as a dendrogram.
- DBSCAN (Density-Based Spatial Clustering and Application with Noise) is used to identify clusters of any shape in a dataset containing noise and outliers.