UNSUPERVISED LEARNING

By Alpha Sow

UNSUPERVISED LEARNING

PCA

Intuition,
theories,

Examples

Clustering

K-means

Affinity
Propagation

Spectral Clustering

Agglomerative
Clustering

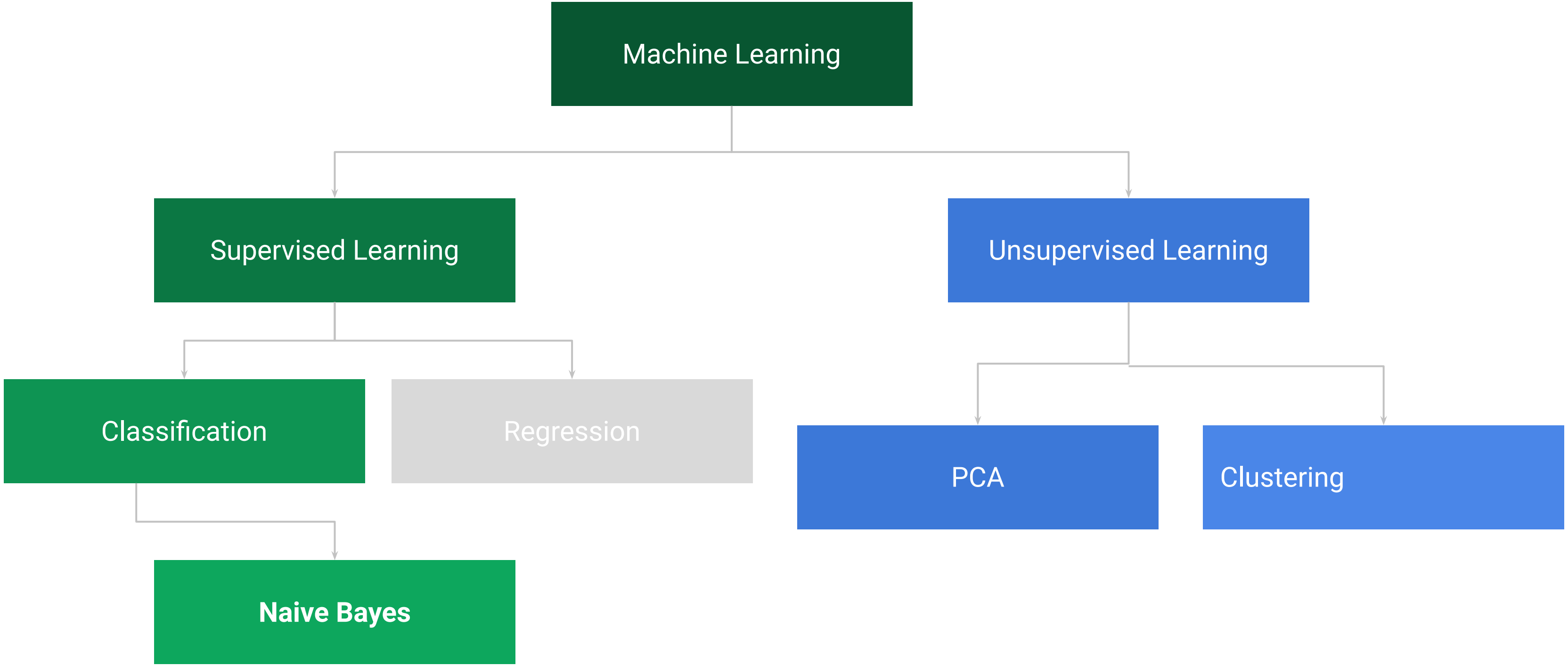
Accuracy Metrics

Industry Application

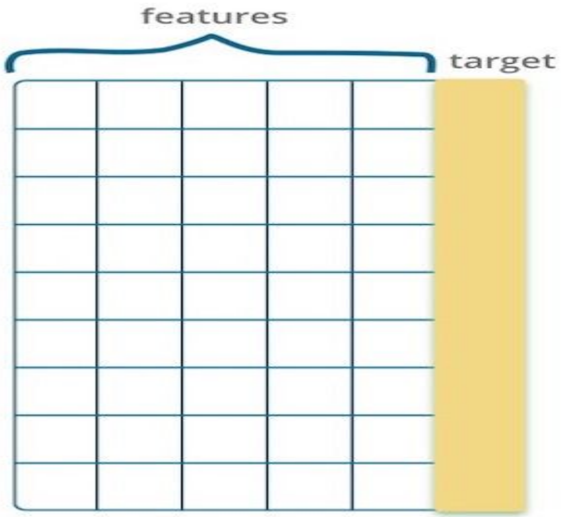
Oil and Gas

Biotech

Economics

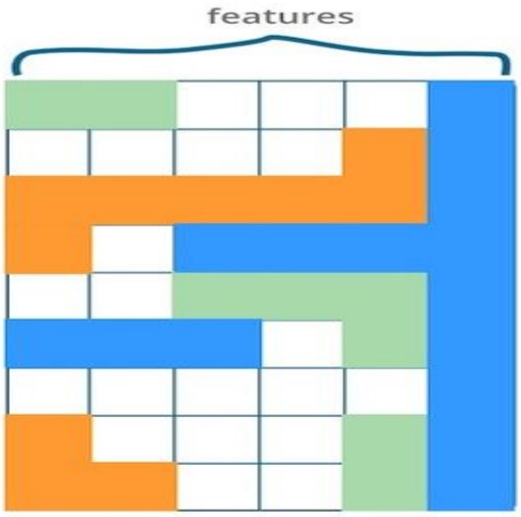
Machine Learning
Supervised Learning
Unsupervised Learning
Classification
Regression
PCA
Clustering
Naive Bayes

Supervised vs Unsupervised Learning



Supervised Learning

Goal: predict a target



Unsupervised Learning

Goal: find patterns

UNSUPERVISED LEARNING

The main feature of unsupervised learning algorithms, when compared to classification and regression methods, is that input data are unlabeled (i.e. no labels or classes given) and that the algorithm learns the structure of the data without any assistance. This creates two main differences.

- First, it allows us to process large amounts of data because the data does not need to be manually labeled.
- Second, it is difficult to evaluate the quality of an unsupervised algorithm due to the absence of an explicit goodness metric as used in supervised learning.

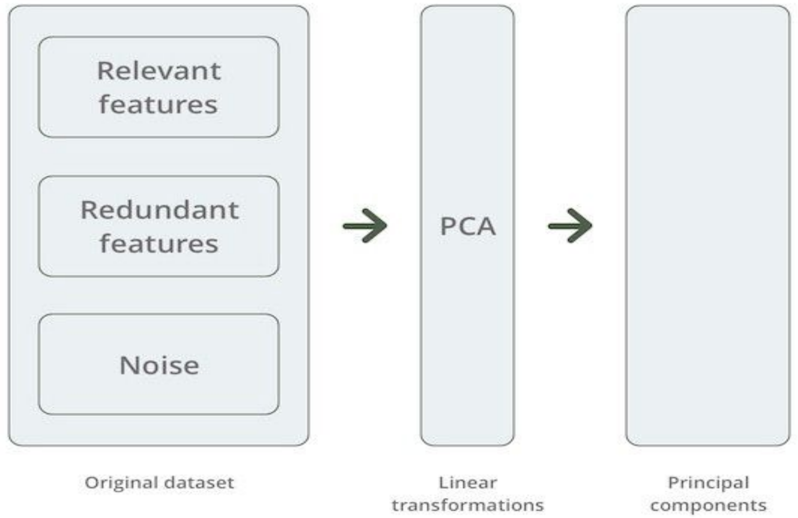
Principal Component Analysis (PCA)

PCA is fundamentally a dimensionality reduction algorithm, but it can also be useful as a tool for :

- visualization,
- noise filtering,
- feature extraction and engineering,
- and much more.

After a brief conceptual discussion of the PCA algorithm, we will see a couple examples of these further applications. We begin with the standard imports:

How PCA works

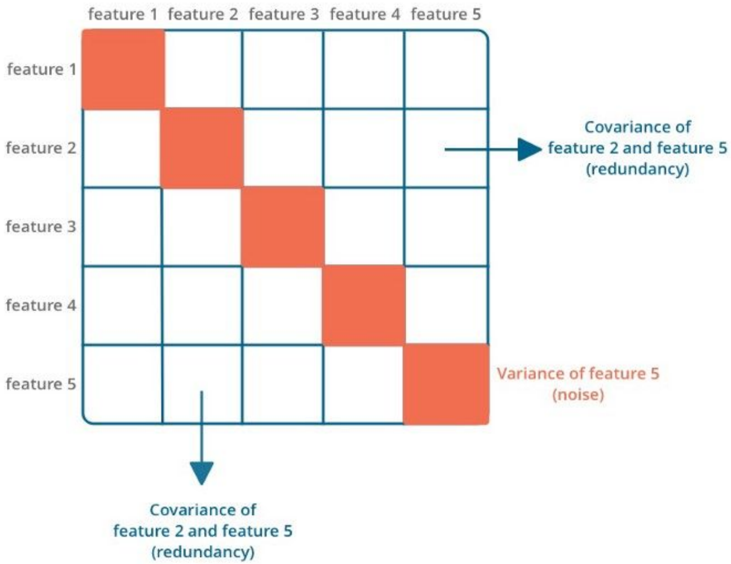


Set of Principal Components,

Ranked in descending order of how much they contribute to describing patterns in the data.

In statistical parlance- according to how much variance they explain.

Covariance Matrix



feature 1 feature 2 feature 3 feature 4 feature 5

feature 1

feature 2

feature 3

feature 4

feature 5

Covariance of feature 2 and feature 5 (redundancy)

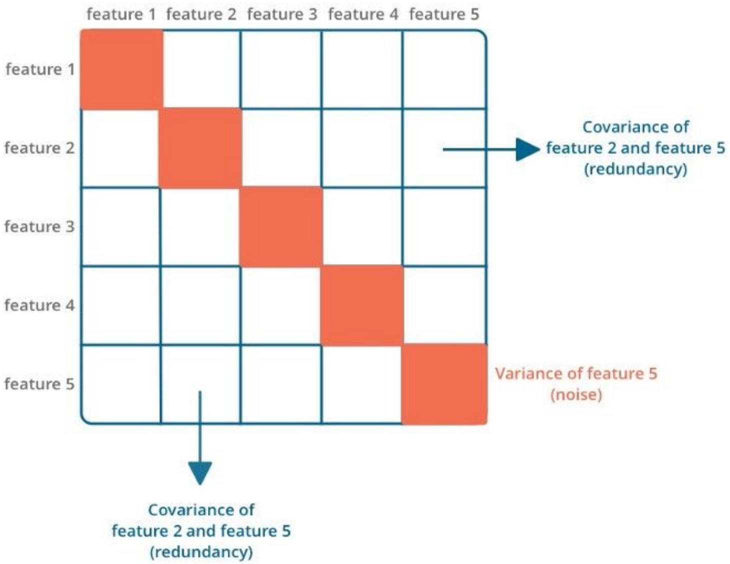
Variance of feature 5 (noise)

Covariance of feature 2 and feature 5 (redundancy)

- Variance of each feature (relevant or pure noise/ diagonal cells.
- Strength of linear relationship between pairs of features (redundant features, all non-diagonal values.

$$cov(X_i, X_j) = E[(X_i - \mu_i)(X_j - \mu_j)] = E[X_i X_j] - \mu_i \mu_j,$$

Covariance Matrix



Covariance of feature 2 and feature 5 (redundancy)

Variance of feature 5 (noise)

Covariance of feature 2 and feature 5 (redundancy)

The covariance matrix highlights the things that get in the way of seeing patterns in the data.

- Reduce noise, by maximizing feature variance.
- Reduce redundancy, by minimizing the covariance between pairs of features.

The basis of PCA is the covariance matrix

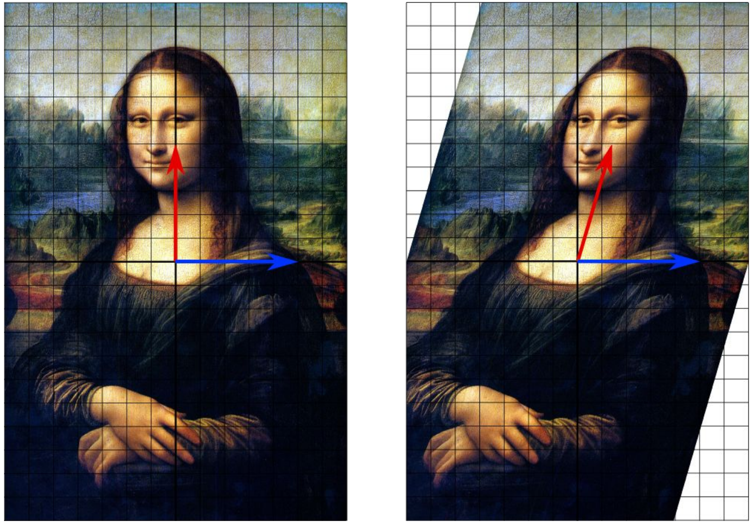
$$\Sigma = E[(\mathbf{X} - E[\mathbf{X}])(\mathbf{X} - E[\mathbf{X}])^T]$$

Approach to identify Principal Components

Eigenvectors of the Covariance Matrix

Singular Value Decomposition

Eigenvectors of the Covariance Matrix



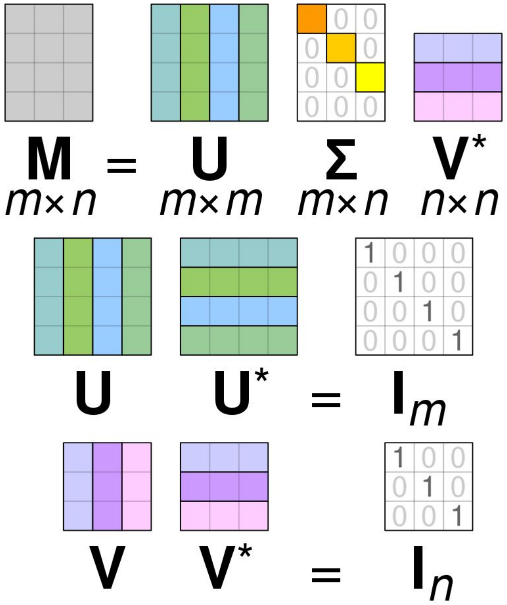
Linear transformation is a nonzero vector that changes by a scalar factor when that linear transformation is applied to it

Matrix as linear Operators have

- Eigenvalue
- Eigenvectors

$$Mw_i = \lambda_i w_i$$

Singular Value Decomposition



In linear algebra, the singular value decomposition (SVD) is a factorization of a real or complex matrix that generalizes the eigendecomposition of a square normal matrix to any m x n matrix via an extension of the polar decomposition.

CLUSTERING

Clustering

Basically, we say to ourselves, "I have these points here, and I can see that they organize into groups. It would be nice to describe these things more concretely, and, when a new point comes in, assign it to the correct group."

5000
Ravensburger

Clustering

“The validation of clustering structures is the most difficult and frustrating part of cluster analysis. Without a strong effort in this direction, cluster analysis will remain a black art accessible only to those true believers who have experience and great courage.”

– *Algorithms for Clustering Data*, Jain and Dubes

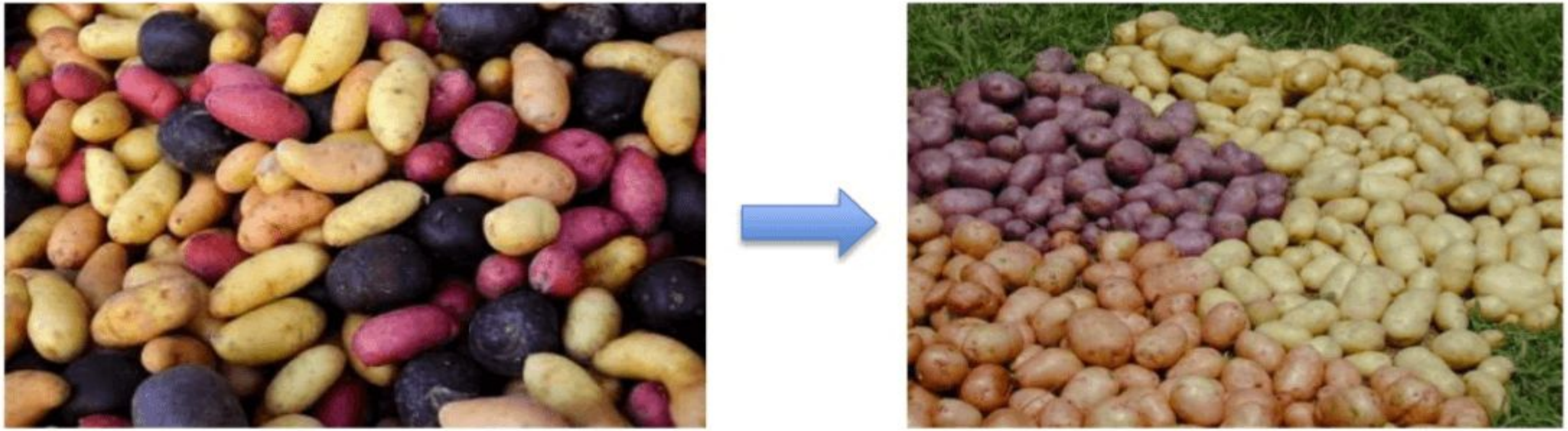
Clustering Analysis

Clustering Analysis is finding groups of objects such that the objects in a group will be similar (or related) to one another and different from (or unrelated to) the objects in other groups such that

- Intra-cluster distances are minimized
- Inter-cluster distances are maximized

Clustering can be classified to **partitional clustering** and **hierarchical clustering**. Partitional clustering creates a division data objects into non-overlapping subsets (clusters) such that each data object is in exactly one subset; Hierarchical clustering creates a set of nested clusters organized as a hierarchical tree

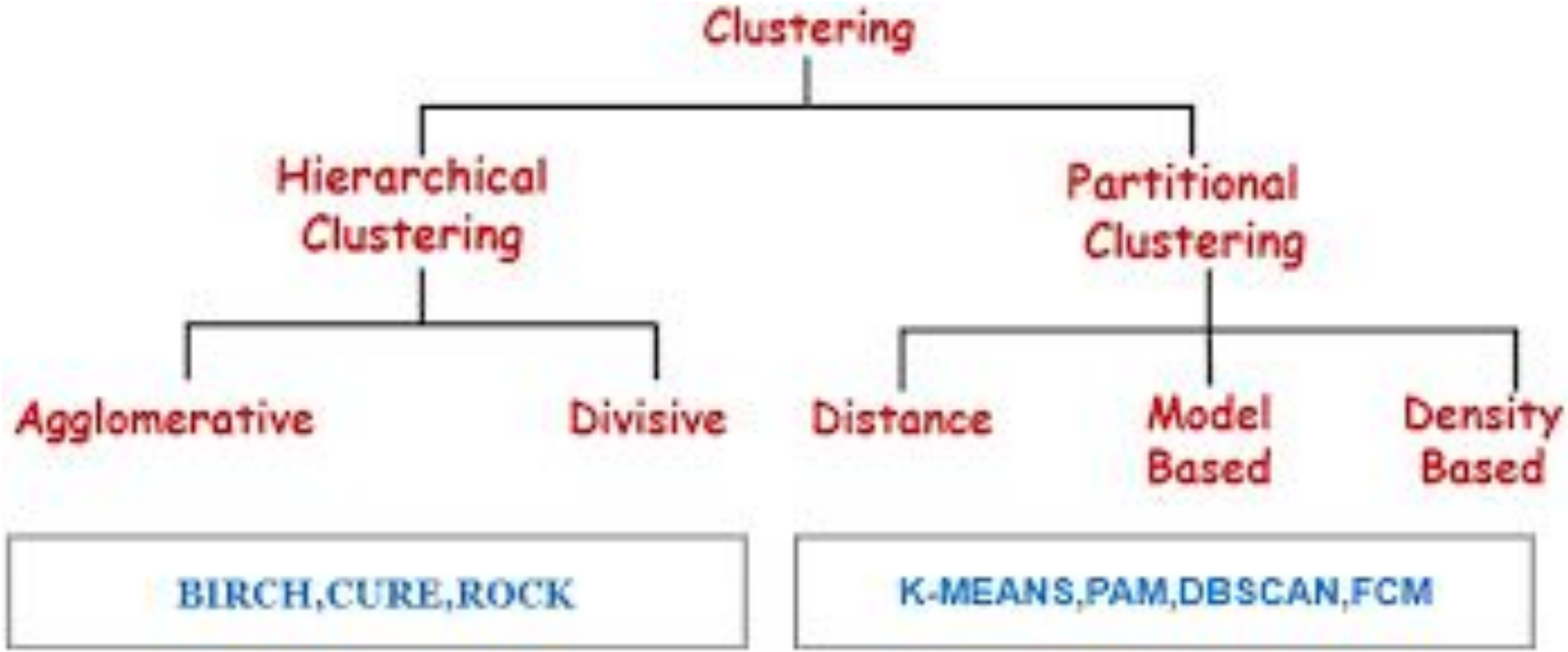
Clustering



sample

Cluster/group

Family of clusters



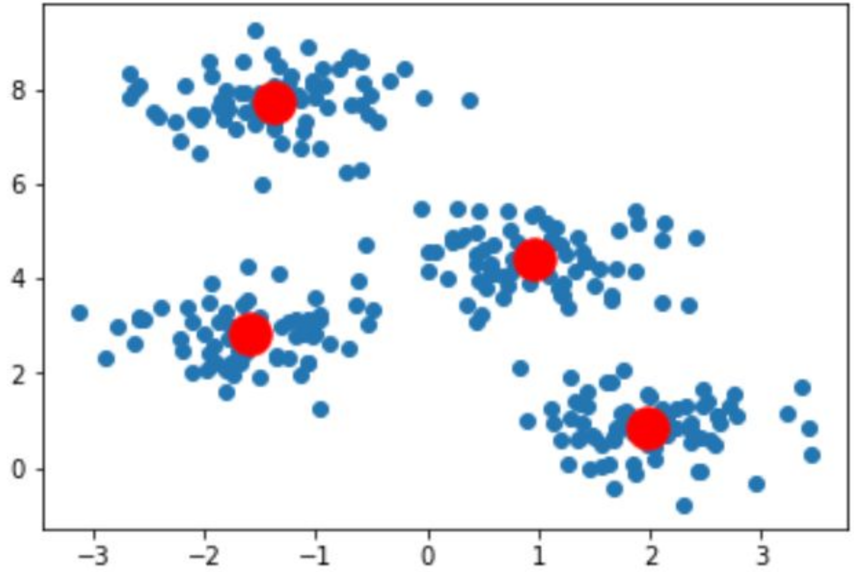
Partitional vs. Hierarchical

1. **Partitional Algorithms (k means)**
   Partition the data space
   Finds all clusters simultaneously

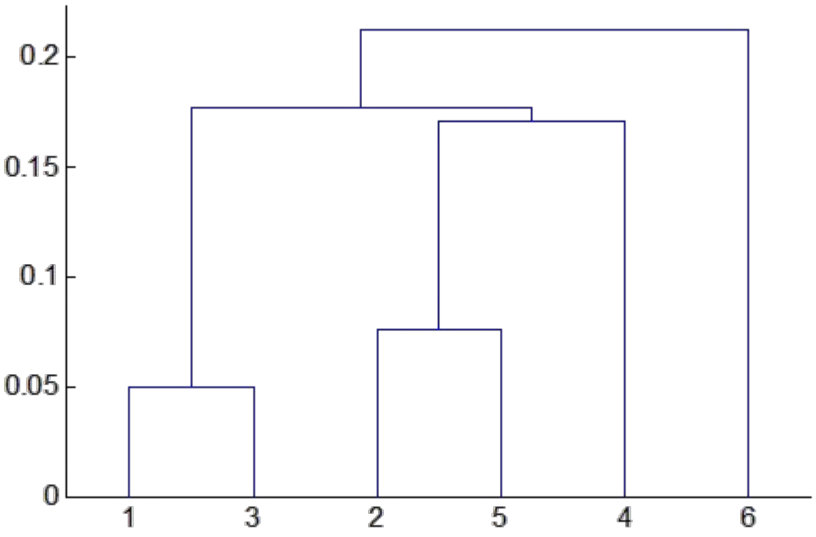
Hierarchical algorithms

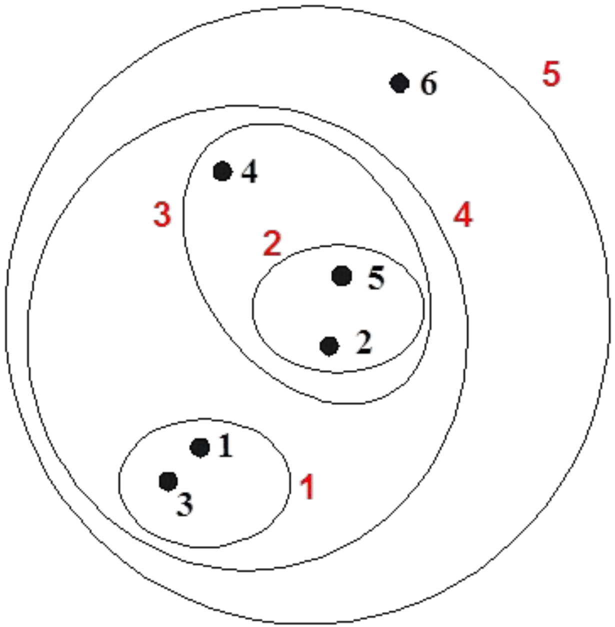
Generate nested cluster hierarchy
Agglomerative (bottom-up)
Divisive(top down)
Distance between clusters:
Single-linkage, complete linkage, average-linkage)

Partitional Clustering



Hierarchical Clustering





Types of Clusters

1 **Well-separated clusters**: A cluster is a set of points such that any point in a cluster is closer (or more similar) to **every** other point in the cluster than to any point not in the cluster.

2 **Center-based clusters**: A cluster is a set of objects such that an object in a cluster is closer (more similar) to the “center” of a cluster, than to the center of any other cluster

Types of Clusters

3 **Contiguous clusters** (Nearest neighbor or Transitive): A cluster is a set of points such that a point in a cluster is closer (or more similar) to **one or more** other points in the cluster than to any point not in the cluster.

4 **Density-based clusters**: A cluster is a dense region of points, which is separated by low-density regions, from other regions of high density. Used when the clusters are irregular or intertwined, and when noise and outliers are present.

Types of Clusters

5 **Property or Conceptual**: Finds clusters that share some common property or represent a particular concept.

6 **Described by an Objective Function**: Finds clusters that minimize or maximize an objective function. It maps the clustering problem to a different domain and solve a related problem in that domain. Therefore **clustering is equivalent to breaking the graph into connected components, one for each cluster**, and we want to minimize the edge weight between clusters and maximize the edge weight within clusters

K-means & its variants

K-means is a partitional clustering approach where number of clusters, K must be specified. Each cluster is associated with a centroid (center point) and each point is assigned to the cluster with the closest centroid.

01 | Initial centroids are often chosen randomly

02 | The centroid is typically the mean of the points in the cluster

03 | Except for number of clusters, the **measure of "Closeness"** should also be specified. (Usually measured by *Euclidean distance*, *cosine similarity*, or *correlation*).

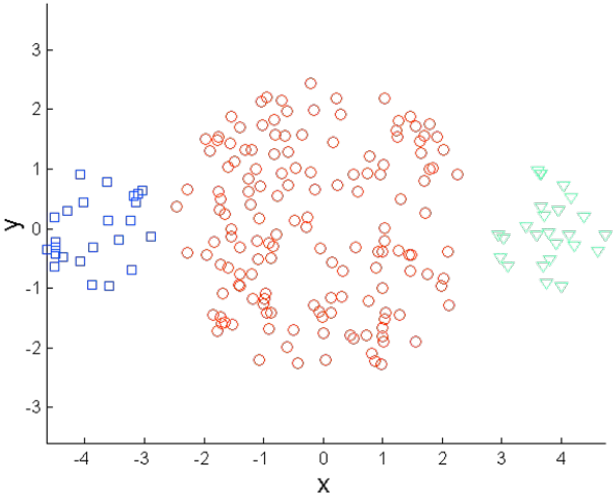
K-means & its limitation

K-means has problems when clusters are of

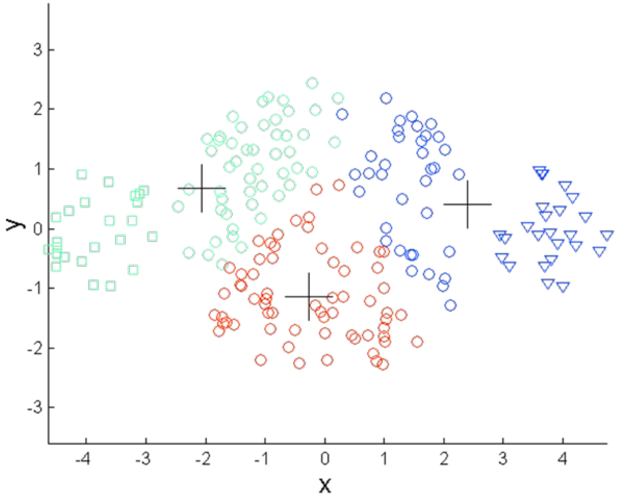
- differing Sizes
- differing Densities
- Non-globular shapes

K-means also has problems when the data contains outliers.

K-means & its limitation(size)

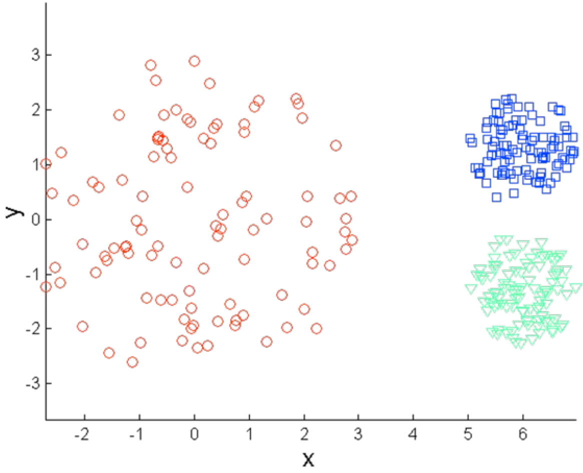


Original Points

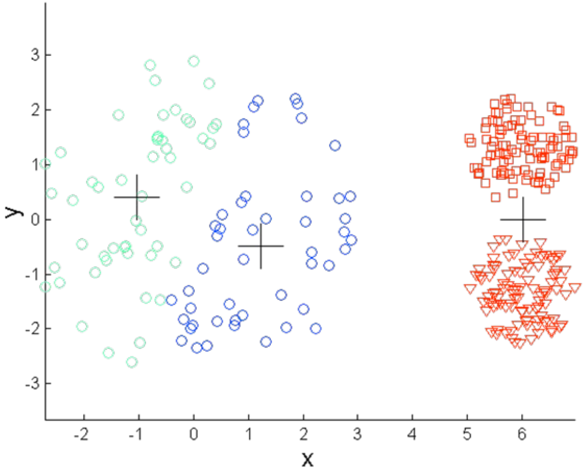


K-means (3 Clusters)

K-means & its limitation(Density)

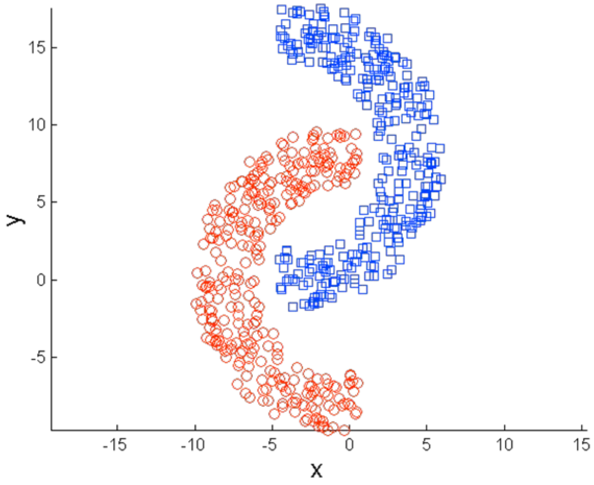


Original Points

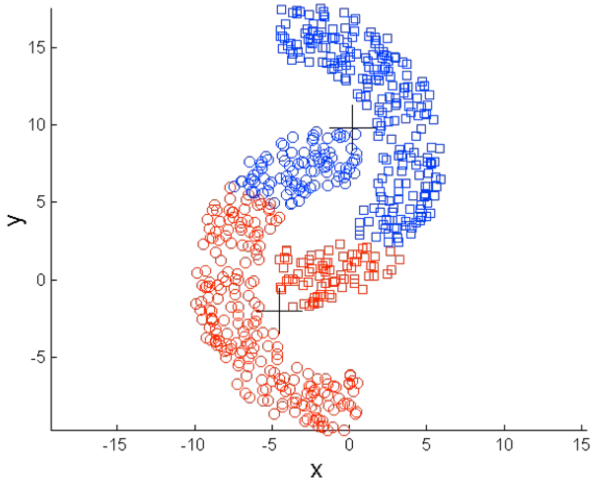


K-means (3 Clusters)

K-means & its limitation(Non Globular Shape)



Original Points



K-means (2 Clusters)

Hierarchical Clustering

In data mining and statistics, hierarchical clustering (also called hierarchical cluster analysis or HCA) is a method of cluster analysis which seeks to build a hierarchy of clusters. Strategies for hierarchical clustering generally fall into two types:

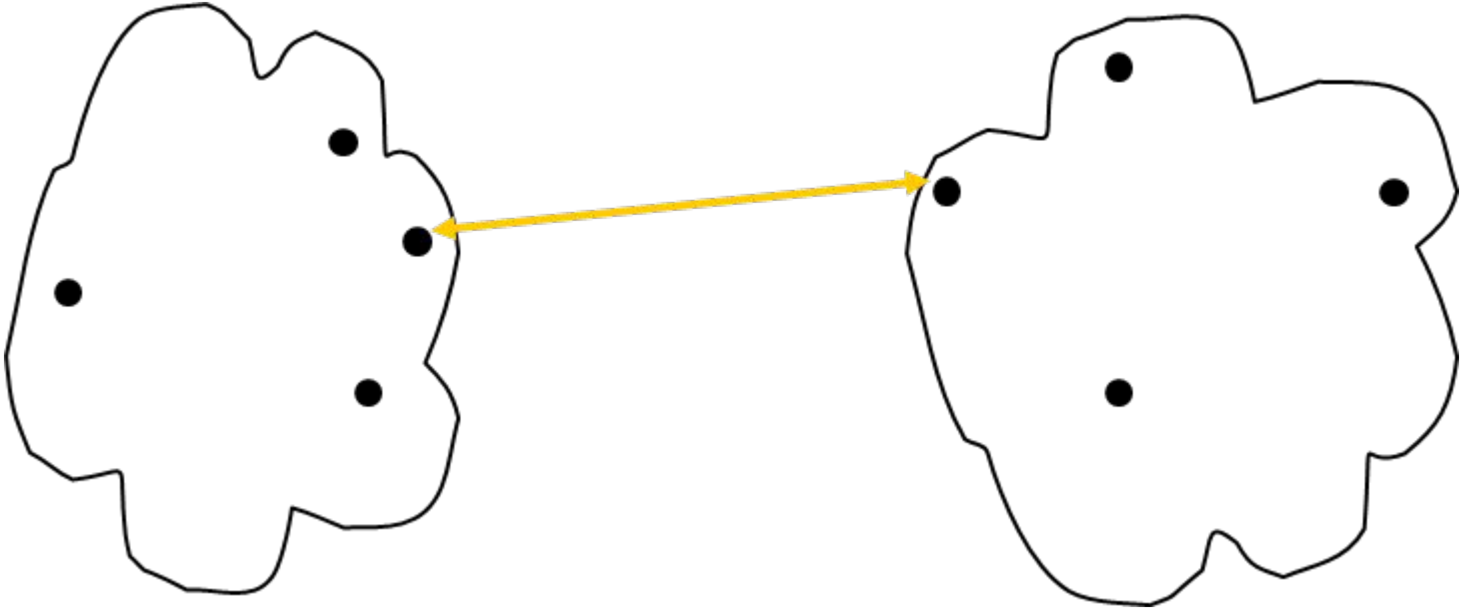
- Agglomerative: This is a "bottom-up" approach: each observation starts in its own cluster, and pairs of clusters are merged as one moves up the hierarchy.
  Start with the points as individual clusters.
  At each step, it merges the closest pair of clusters until only one cluster (or k clusters) left.

Hierarchical Clustering

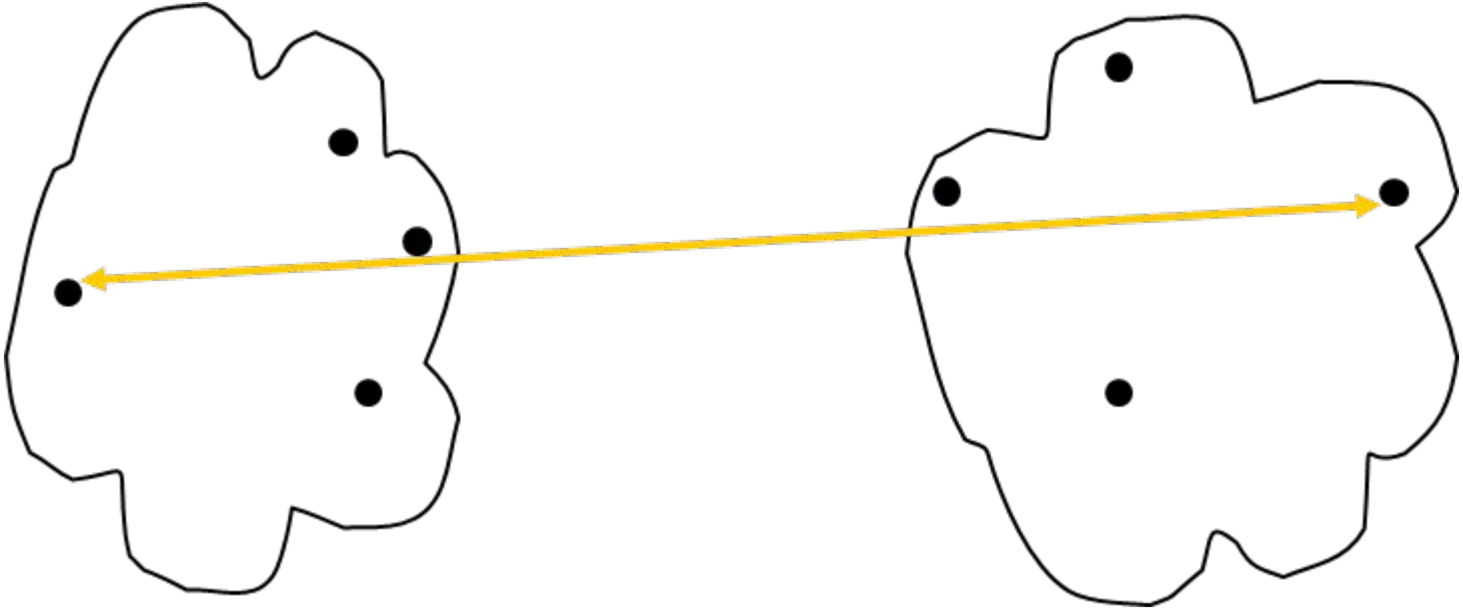
In data mining and statistics, hierarchical clustering (also called hierarchical cluster analysis or HCA) is a method of cluster analysis which seeks to build a hierarchy of clusters. Strategies for hierarchical clustering generally fall into two types:

- Divisive: This is a "top-down" approach: all observations start in one cluster, and splits are performed recursively as one moves down the hierarchy.
  Starts with one, all-inclusive cluster.
  At each step, it splits a cluster until each cluster contains a point (or there are k clusters).

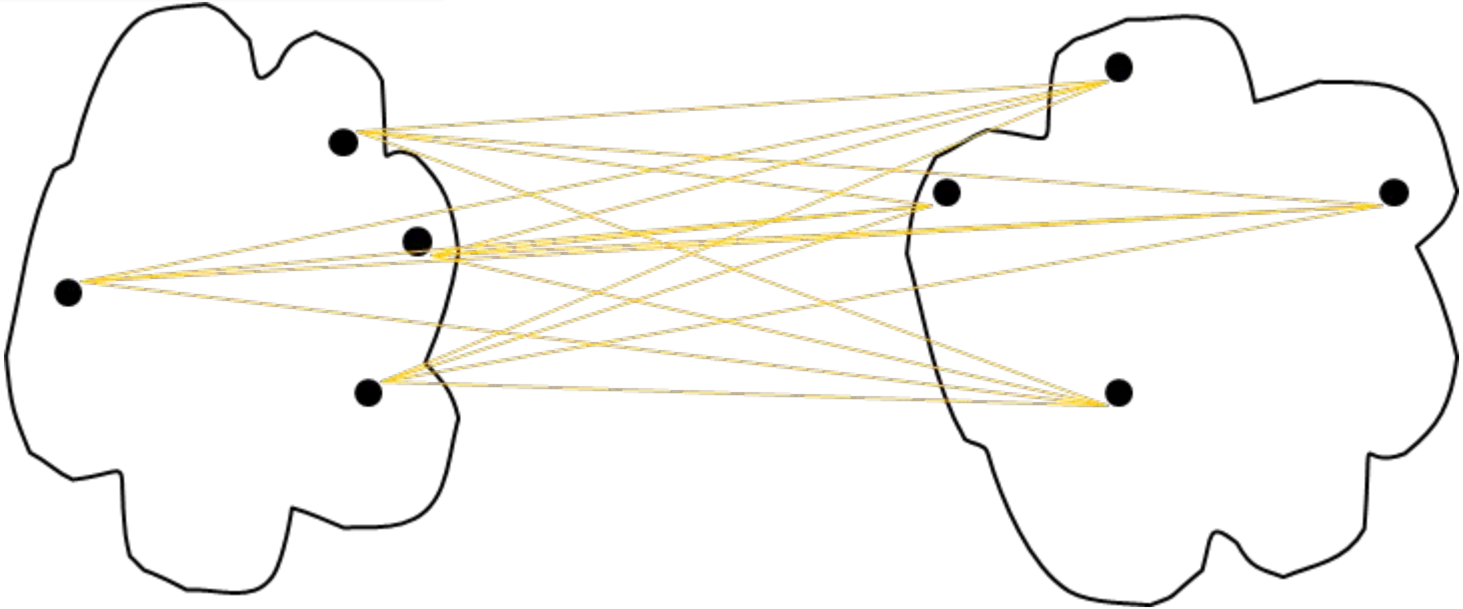
Single Affinity(min)



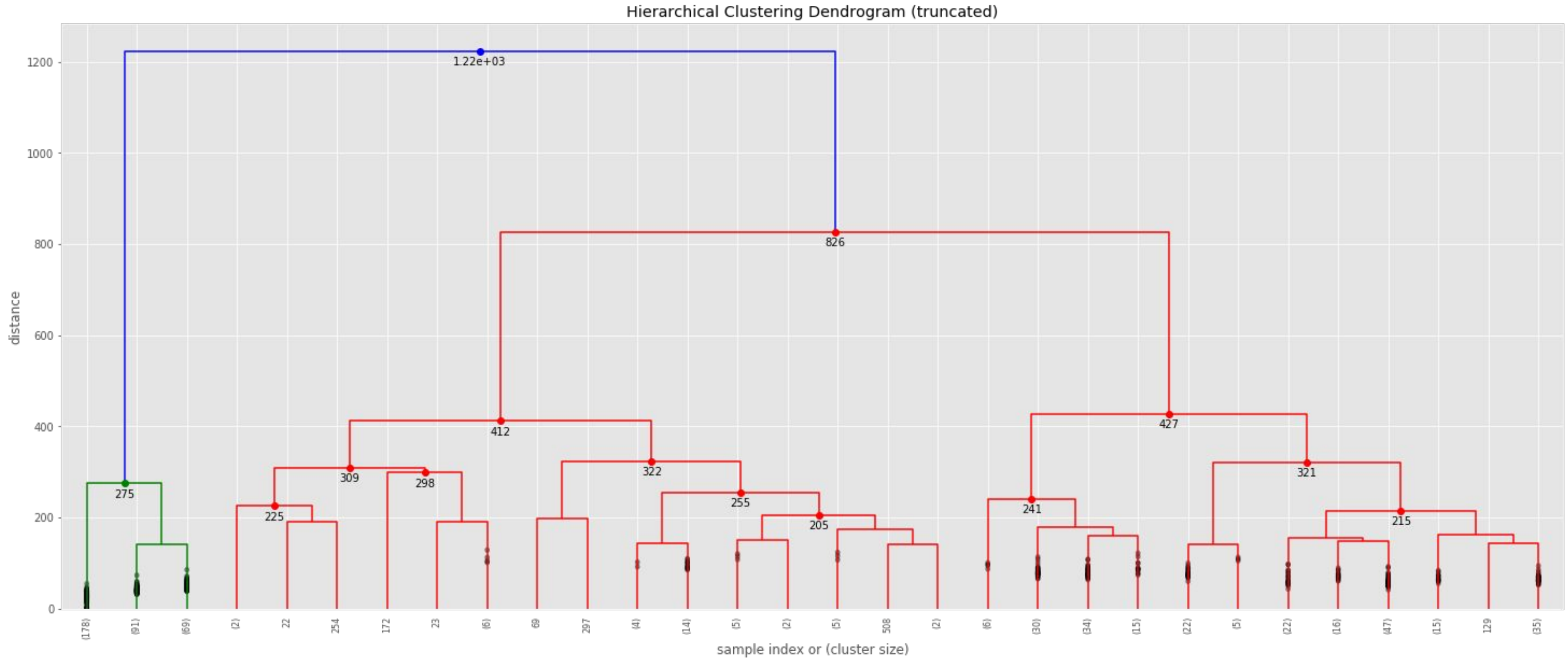
Complete Affinity



Average Affinity



Dendrogram


Hierarchical Clustering Dendrogram (truncated)
distance
0
200
400
600
800
1000
1200
1.22e+03
826
412
427
309
298
322
321
275
225
255
205
241
215
(178)
(91)
(69)
(2)
22
254
172
23
(6)
69
297
(4)
(14)
(5)
(2)
(5)
508
(2)
(6)
(30)
(34)
(15)
(22)
(5)
(22)
(16)
(47)
(15)
129
(35)
sample index or (cluster size)

Resources

- “Introduction to Data Mining,” by P.-N. Tan, M. Steinbach, V. Kumar, Addison-Wesley.
- Bradley, P. S., Fayyad, U., & Reina, C. (1998). Scaling EM (expectation-maximization) clustering to large databases.
- Gower, J. C., & Ross, G. J. (1969). Minimum spanning trees and single linkage cluster analysis. Applied statistics, 54-64.
- Jana, P. K., & Naik, A. (2009, December). An efficient minimum spanning tree based clustering algorithm. In Methods and Models in Computer Science, 2009. ICM2CS 2009. Proceeding of International Conference on (pp. 1-5). IEEE.