Cluster analysis

Louis Jachiet

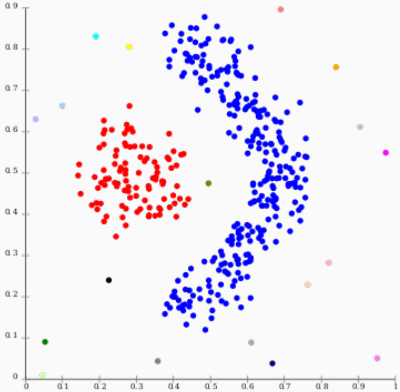
What is clustering?

Wikipedia definition

Cluster analysis or clustering is the task of grouping a set of objects in such a way that objects in the same group (called a cluster) are more similar (in some sense) to each other than to those in other groups (clusters).

Clustering

Infer structure from data.



Examples

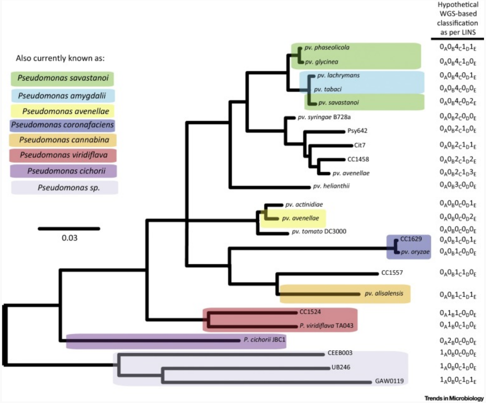


Figure 1: Cluster of species (from *cell.com*)

Examples

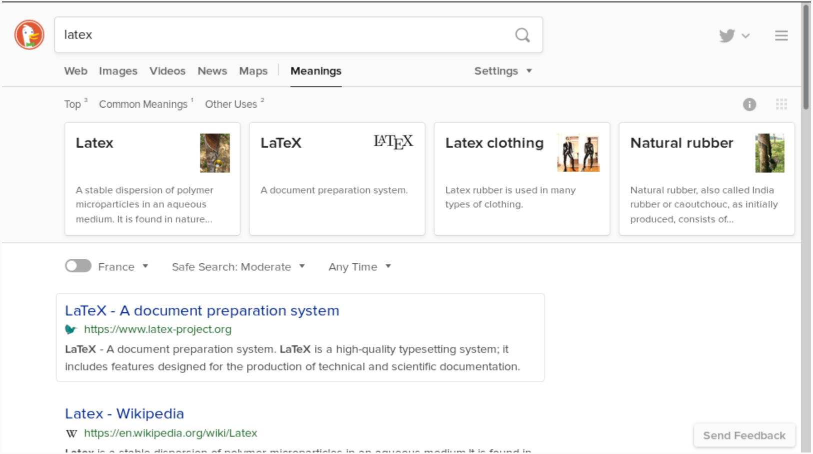
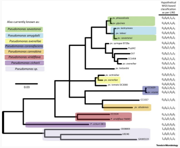
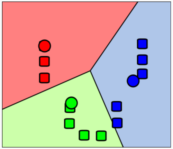


Figure 2: Screenshot of Duckduckgo

Clustering clustering algorithms

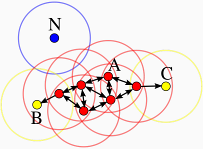
Hierarchical





Partitional

Density based



Clustering goal

Minimize some *error function* on clusters!

Clustering goal

Minimize some *error function* on clusters!

Classical error functions

- Closest link
- Farthest link
- Average distance
- Squared Centroid distance
- ...

Hierarchical clustering

Agglomerative Hierarchical Clustering

- Each item starts in its own cluster

Agglomerative Hierarchical Clustering

- Each item starts in its own cluster
- While we have more than one cluster

Agglomerative Hierarchical Clustering

- Each item starts in its own cluster
- While we have more than one cluster
    - Merge the two “closest” clusters

Agglomerative Hierarchical Clustering

- Each item starts in its own cluster
- While we have more than one cluster
  - Merge the two “closest” clusters

$\Rightarrow$ What defines cluster proximity?

Agglomerative Hierarchical Clustering

- Each item starts in its own cluster
- While we have more than one cluster
  - Merge the two “closest” clusters

$\Rightarrow$ What defines cluster proximity?

MAX?

Agglomerative Hierarchical Clustering

- Each item starts in its own cluster
- While we have more than one cluster
  - Merge the two “closest” clusters

$\Rightarrow$ What defines cluster proximity?
MAX? MIN?

Agglomerative Hierarchical Clustering

- Each item starts in its own cluster
- While we have more than one cluster
  - Merge the two "closest" clusters

$\Rightarrow$ What defines cluster proximity?

MAX? MIN? Average?

Agglomerative Hierarchical Clustering

- Each item starts in its own cluster
- While we have more than one cluster
  - Merge the two "closest" clusters

$\Rightarrow$ What defines cluster proximity?

MAX? MIN? Average? Ward?

BIRCH CF-tree

Summarize N points $v_1 \dots v_N$ with $(N, \vec{LS}, SS)$ where

- $\vec{LS} = \sum v_i$
- $SS = \sum v_i^2$

Computing centroid:

$$\vec{C} = \frac{\vec{LS}}{N}$$

Computing radius:

$$R^2 = \frac{SS}{N} - \vec{C}^2$$

Combine clusters:

$(N_1, \vec{LS}_1, SS_1) + (N_2, \vec{LS}_2, SS_2) = (N_1 + N_2, \vec{LS}_1 + \vec{LS}_2, SS_1 + SS_2)$

BIRCH algorithm

Parameters

- B branching factor
- T threshold

BIRCH algorithm

Parameters

- B branching factor
- T threshold

Adding point in the CF-tree

- Try to find the appropriate leaf
- When it doesn't exist, create one
- Explode node if it is too big

BIRCH algorithm

Parameters

- B branching factor
- T threshold

Adding point in the CF-tree

- Try to find the appropriate leaf
- When it doesn't exist, create one
- Explode node if it is too big

Algorithm

- Initialize empty CF-tree
- Add each point to the CF-tree
- Compute clusters over the data summarized by the CF-tree

Clu-Stream

- Uses micro-clusters to store statistics on-line
  - Clustering Features $CF = (N, LS, SS, LT, ST)$
    - N: numer of data points
    - LS: linear sum of the N data points
    - SS: square sum of the N data points
    - LT: linear sum of the time stamps
    - ST: square sum of the time stamps
- Uses pyramidal time frame

Partitional clustering

K-Means

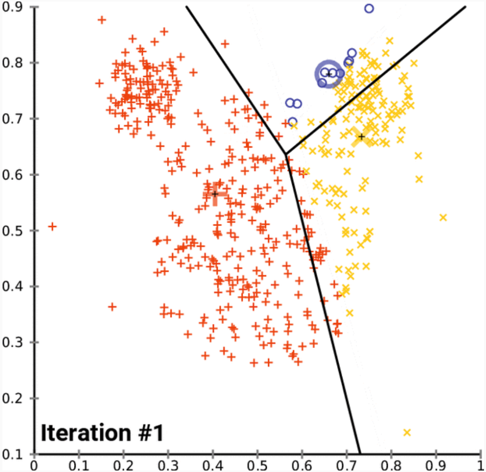
Input

- K a number
- P a set of points

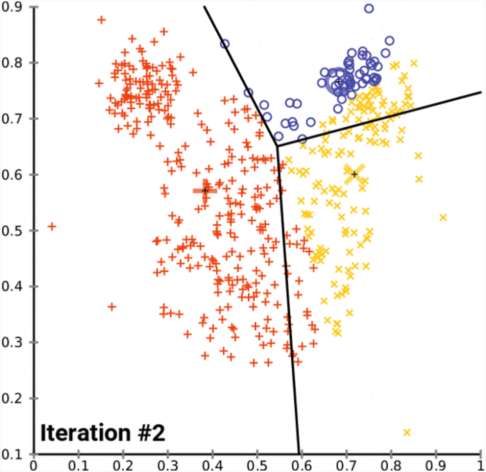
Algorithm

- Select K points $C_1 \ldots C_K$ (Forgy or Random)
- Iterate
  - Partition P into $P_1 \ldots P_K$ ($p \in P$ goes into P_i when C_i is the closest to p among the $C_1 \ldots C_K$)
  - Set $C_i = center(P_i)$

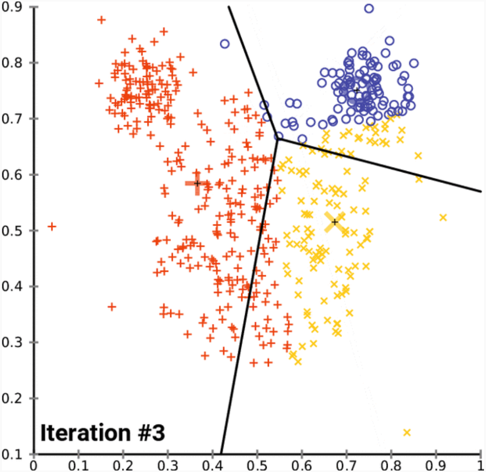
K-Means



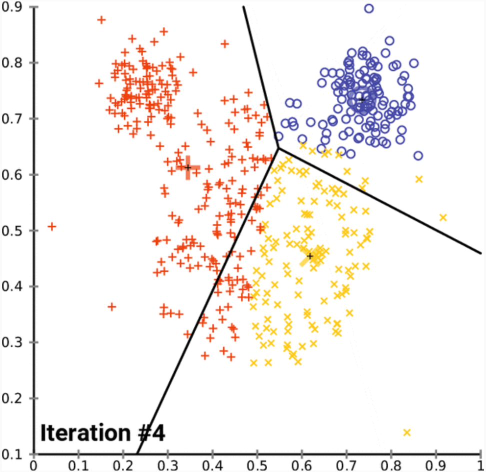
K-Means



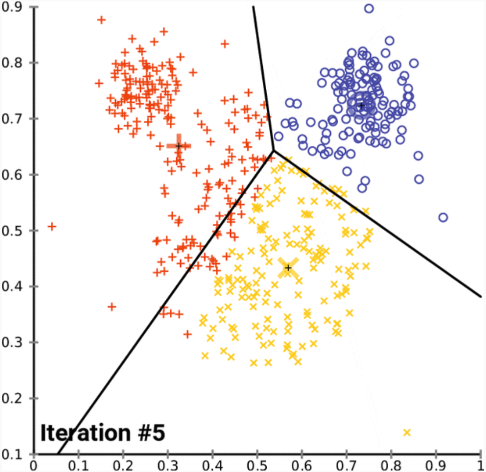
K-Means



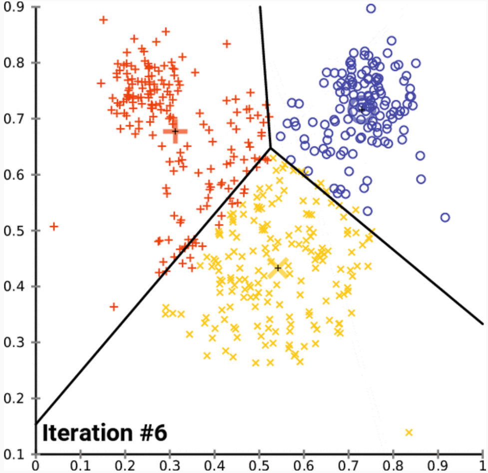
K-Means



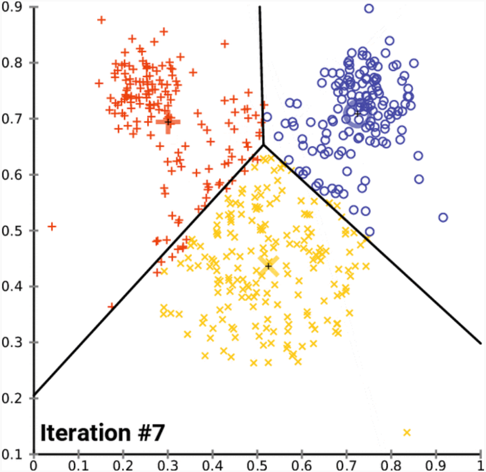
K-Means



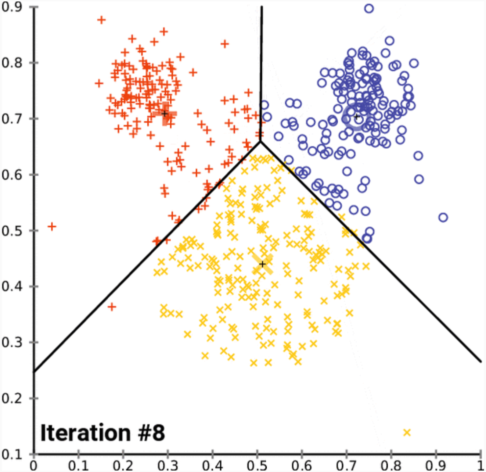
K-Means



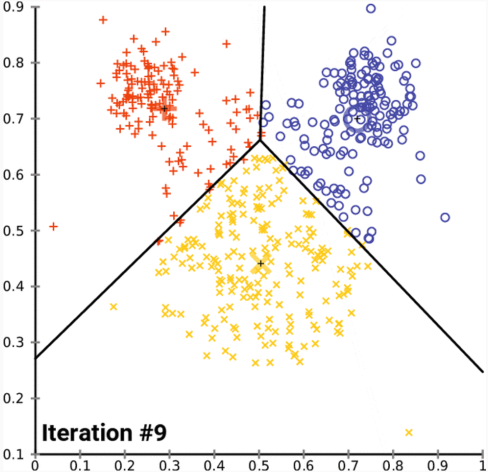
K-Means



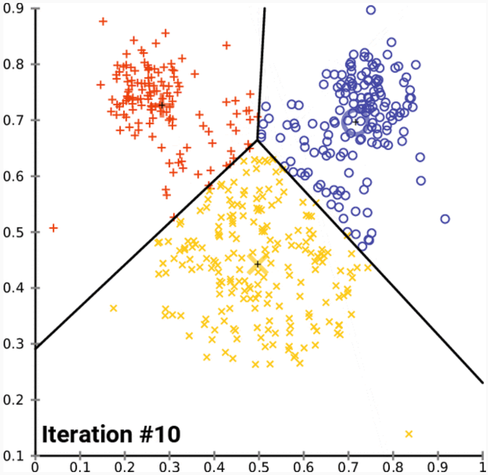
K-Means



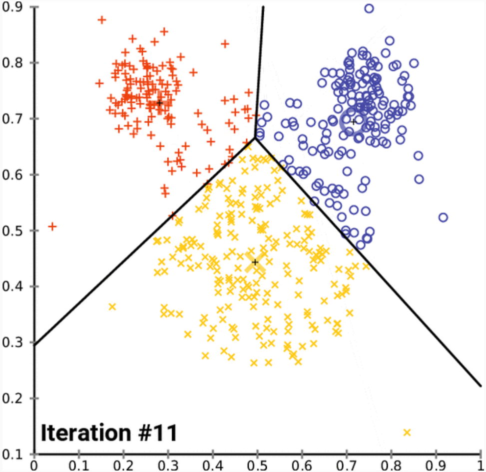
K-Means



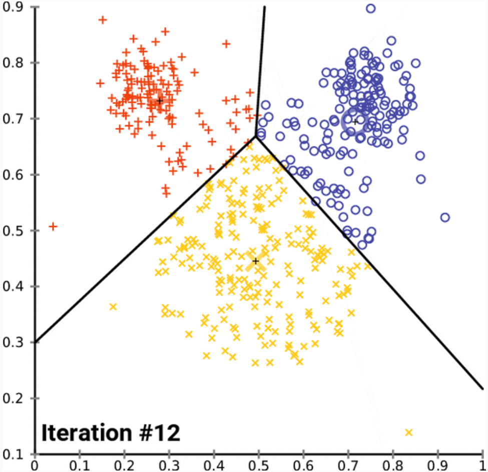
K-Means



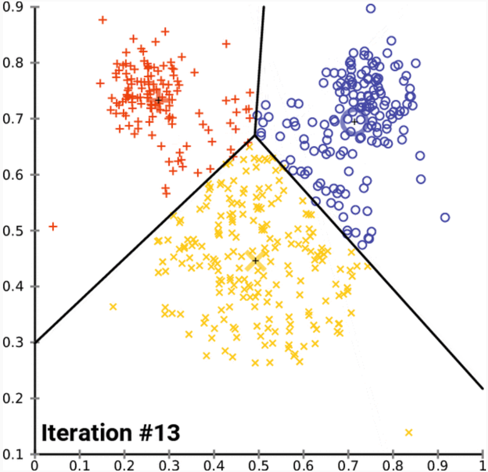
K-Means



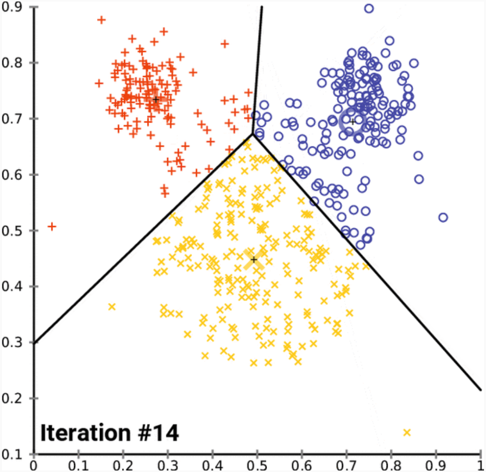
K-Means



K-Means



K-Means



K-Means

K-Means can be slow . . . and does not necessarily converges to an optimal solution!

K-Means++

Improve initialization

- First point is selected randomly
- Each new point is choosen randomly but depending on the distance from selected points

Finding K?

- Test several K and plot (elbow method)
- Silhouette method
- Maximizing Bayesian Information Criterion
- ...

Extension for streaming data

StreamKM++

- Creates a *coreset* of points (inspired by K-means++)
- Runs K-means on the coreset

Density

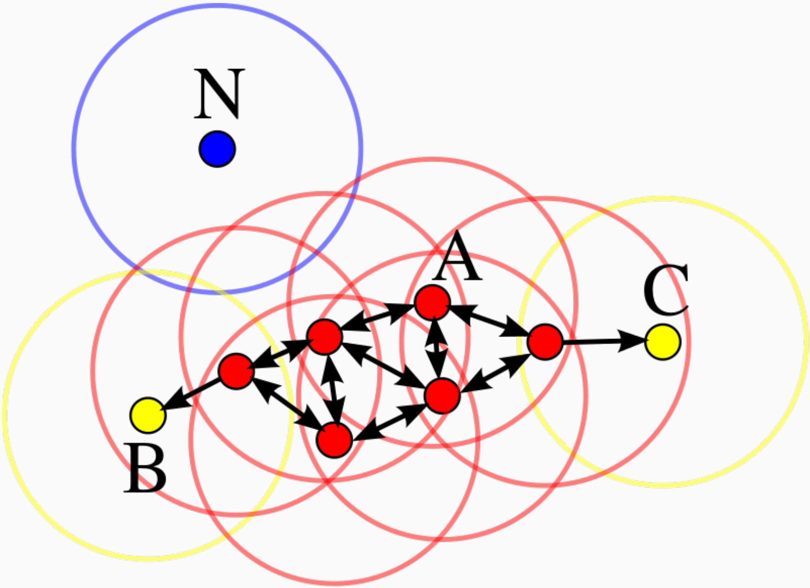
DBSCAN(ϵ, μ)

- ϵ-neighborhood(p): set of points that are at a distance of *p* less or equal to ϵ
- Core object: object whose ϵ-neighborhood has an overall weight at least μ
- A point *p* is *directly density-reachable* from *q* if
  - *p* is in ϵ-neighborhood(q)
  - *q* is a core object
- A point *p* is *density-reachable* from *q* if
  - there is a chain of points $p_1, \ldots, p_n$ such that p_{i+1} is directly density-reachable from p_i
- A point *p* is *density-connected* from *q* if
  - there is point *o* such that *p* and *q* are density-reachable from *o*

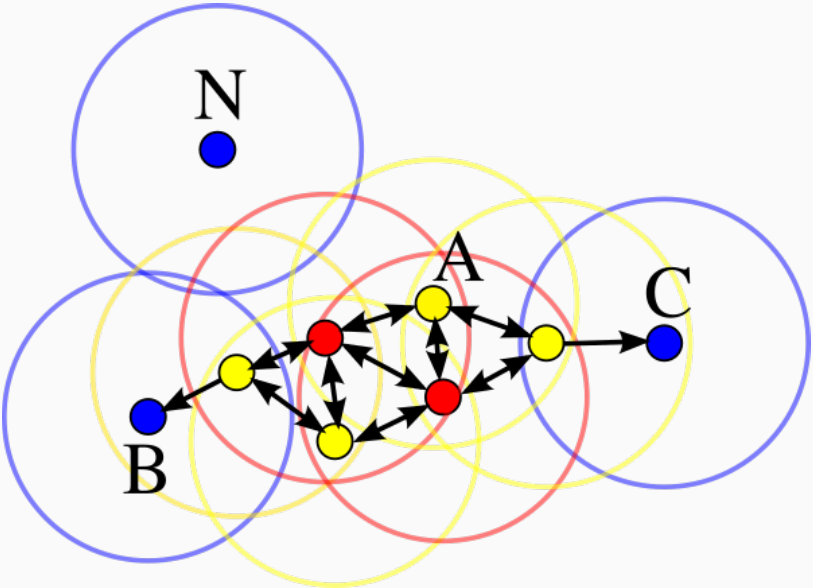
DBSCAN

- A *cluster* C of points satisfies
  - if $p \in C$ and q is density-reachable from p, then $q \in C$
  - all points $p, q \in C$ are density-connected
- A *cluster* is uniquely determined by any of its core points
- A *cluster* can be obtained
  - choosing an arbitrary core point as a seed
  - retrieve all points that are density-reachable from the seed

DBSCAN($\epsilon, 3$)



DBSCAN($\epsilon, 4$)



DBSCAN

Algorithm

- select an arbitrary non treated point p
- retrieve $N = \epsilon$-neighborhood(p)
- if $|N| \geq \mu$
  - set $T = N \setminus \{p\}$
  - While $T \neq \emptyset$
    - Set $(p', T) = T$
    - Mark p' as cluster p
    - Set $N' = \epsilon$-neighborhood(p')
    - If $|N'| \geq \mu$ then $T = T \cup N'$
- Continue the process until all of the points have been processed