

GL Applied Data Science Program

Unsupervised Learning - Clustering

This file is meant for personal use by hhung.inbox@gmail.com only.
Sharing or publishing the contents in part or full is liable for legal action.

Overview

Overview of this week / module:

- Data collection and visualization for exploratory data analysis
- Network analysis
- Unsupervised learning - clustering

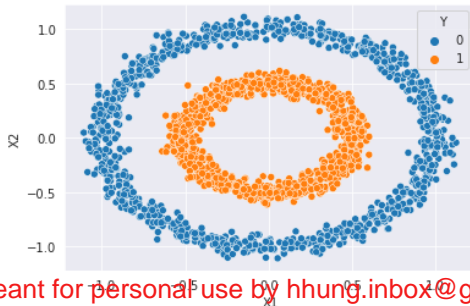
Overview of this lecture:

- Clustering methods
- Community detection in networks

This file is meant for personal use by hhung.inbox@gmail.com only.
Sharing or publishing the contents in part or full is liable for legal action.

Case study: clustering

- Find groups, so that elements within cluster are very similar and elements between clusters are very different
- **Examples:**
 - Find customer groups to adjust advertisement
 - Find subtypes of diseases to fine-tune treatment
- Our eye is very good at identifying cluster



This file is meant for personal use by hhung.inbox@gmail.com only.
Sharing or publishing the contents in part or full is liable for legal action.

Clustering

This file is meant for personal use by hhung.inbox@gmail.com only.
Sharing or publishing the contents in part or full is liable for legal action.

Clustering

N samples, k clusters: k^N possible assignments

- E.g., $N = 100$, $k = 3$: $3^{100} = 5 * 10^{47}$!!
⇒ impossible to search through all assignments

We will discuss:

- k -means clustering
- Gaussian mixture models
- Hierarchical clustering
- DBSCAN

This file is meant for personal use by hhung.inbox@gmail.com only.
Sharing or publishing the contents in part or full is liable for legal action.

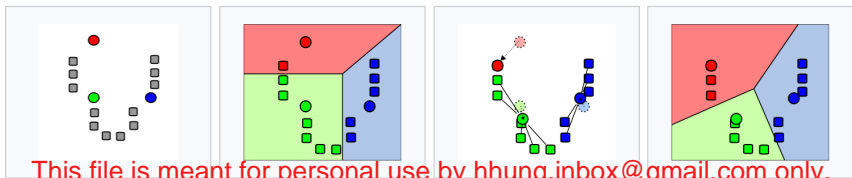
K -means clustering

- Choose the number of clusters K
- Minimize the sum of the pairwise distances between samples within each cluster (also known as the *Within-Groups Sum of Squares*)

This file is meant for personal use by hhung.inbox@gmail.com only.
Sharing or publishing the contents in part or full is liable for legal action.

K-means clustering

- Choose the number of clusters K
- Minimize the sum of the pairwise distances between samples within each cluster (also known as the *Within-Groups Sum of Squares*)
 - One can show that this is equivalent to minimizing the sum of the distances to the cluster means
- Exact solution of minimization problem is computationally infeasible
 - Use greedy algorithm with random restarts to avoid local optima
- Leads to spherical shaped clusters of similar radii



This file is meant for personal use by hhung.inbox@gmail.com only.

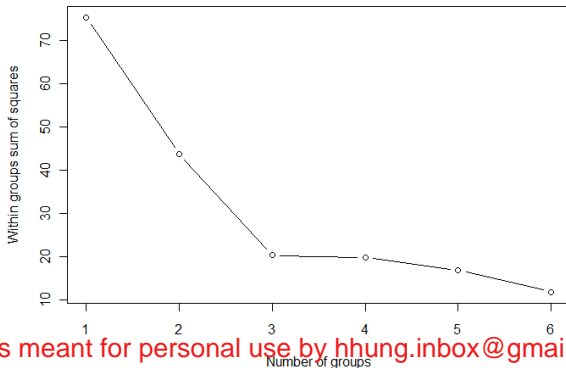
Sharing or publishing the contents in part or full is liable for legal action.

Image source: Wikipedia

Choosing the number of clusters

- Run K -means clustering for several number of groups K
- Plot Within-Groups Sum of Squares versus the number of groups
- Choose number of groups after the last big drop of the curve

Example:



This file is meant for personal use by hhung.inbox@gmail.com only.
Sharing or publishing the contents in part or full is liable for legal action.

Partitioning around medoids (PAM)

- *K*-Means: Cluster centers can be arbitrary points in space
⇒ very sensitive to outliers!

This file is meant for personal use by hhung.inbox@gmail.com only.
Sharing or publishing the contents in part or full is liable for legal action.

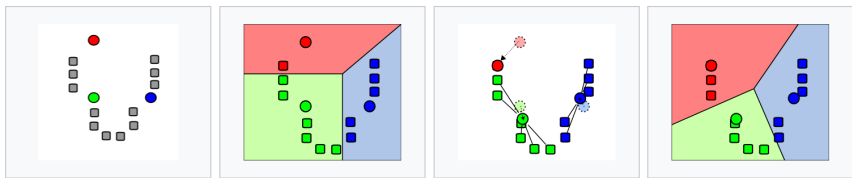
Partitioning around medoids (PAM)

- *K*-Means: Cluster centers can be arbitrary points in space
⇒ very sensitive to outliers!
- Robust alternative: Partitioning around medoids (PAM)
 - Cluster center must be an observation (“medoid”)
 - More robust against outliers
 - Also gives a representative object for each cluster (e.g., for easy interpretation)

This file is meant for personal use by hhung.inbox@gmail.com only.
Sharing or publishing the contents in part or full is liable for legal action.

Gaussian mixture model

- Soft version of K -means clustering, where each sample is attributed to a cluster with a certain probability
 - This allows for points between clusters to belong to multiple clusters
- Distribution of samples within each cluster is modeled by a Gaussian
 - This allows for ellipsoidal shaped clusters and the number of clusters can be determined in a statistically sound way (for example using the so-called *Bayesian Information Criterion*)

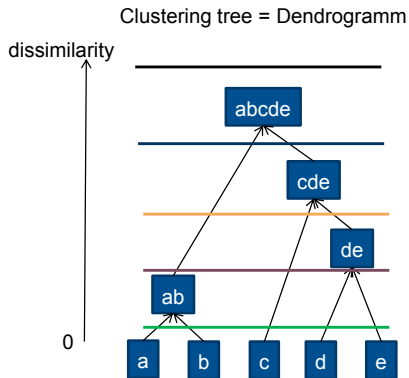


This file is meant for personal use by hhung.inbox@gmail.com only.
Sharing or publishing the contents in part or full is liable for legal action.

Image source: Wikipedia

Hierarchical clustering

- **Agglomerative clustering:**
Build up clusters from individual observations
- **Divisive clustering:** Start with whole group of observations and split off clusters



Advantage of hierarchical clustering:

- Solve clustering for all possible numbers of cluster $1, 2, \dots, n$ at once
- Choose desired number of clusters later

This file is meant for personal use by hhung.inbox@gmail.com only.

Sharing or publishing the contents in part or full is liable for legal action.

Distance measures between clusters

How we define distances between clusters can have a huge effect on what kinds of clusters we obtain!

- **single linkage** (i.e., minimum distance)
- **complete linkage** (i.e., maximum distance)
- **average linkage** (i.e., average distance)

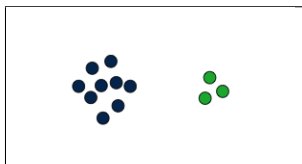
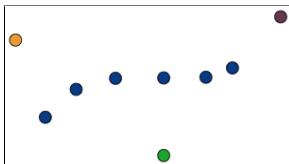
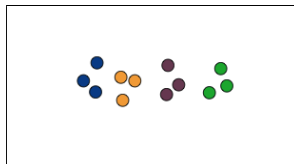
This file is meant for personal use by hhung.inbox@gmail.com only.
Sharing or publishing the contents in part or full is liable for legal action.

Distance measures between clusters

How we define distances between clusters can have a huge effect on what kinds of clusters we obtain!

- **single linkage** (i.e., minimum distance)
- **complete linkage** (i.e., maximum distance)
- **average linkage** (i.e., average distance)

Which clustering output belongs to which choice of linkage?



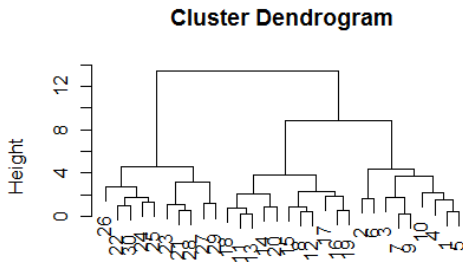
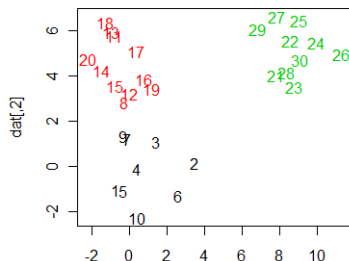
This file is meant for personal use by hhung.inbox@gmail.com only.

Sharing or publishing the contents in part or full is liable for legal action.

Choosing the number of clusters

- No strict rule
- Find the largest vertical “drop” in the tree

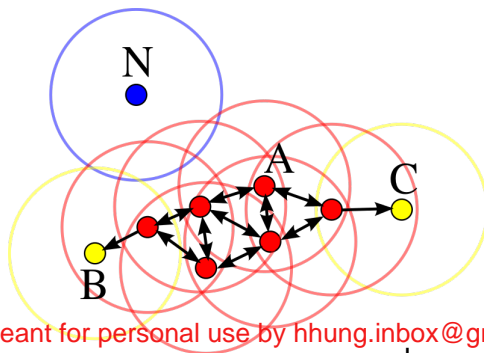
Example:



This file is meant for personal use by hhung.inbox@gmail.com only.
Sharing or publishing the contents in part or full is liable for legal action.

DBSCAN

- Uses 2 parameters: minPts (minimum number of points) and ϵ (radius of neighborhood)
- Core points have at least minPts within distance ϵ
- Clusters are defined by looking at all points reachable from a core point



This file is meant for personal use by hhung.inbox@gmail.com only.
Sharing or publishing the contents in part or full is liable for legal action.

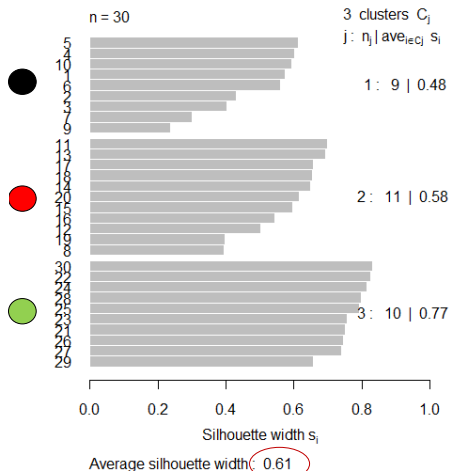
Image source: Wikipedia

Quality of clustering: Silhouette plot

Compute for each sample $x^{(i)}$:

- $a(x^{(i)})$ = average distance between $x^{(i)}$ and all other points in its cluster
- $b(x^{(i)})$ = average distance between $x^{(i)}$ and the closest cluster it does not belong to
- $S(x^{(i)}) \in [-1, 1]$ with

$$S(x^{(i)}) = \frac{(b(x^{(i)}) - a(x^{(i)}))}{\max(a(x^{(i)}), b(x^{(i)}))}$$



This file is meant for personal use by hhung.inbox@gmail.com only.

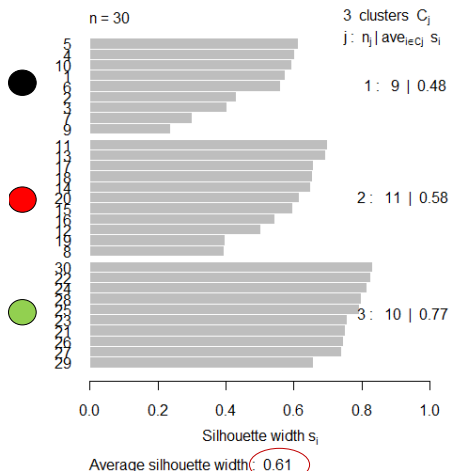
Sharing or publishing the contents in part or full is liable for legal action.

Quality of clustering: Silhouette plot

Compute for each sample $x^{(i)}$:

- $a(x^{(i)})$ = average distance between $x^{(i)}$ and all other points in its cluster
- $b(x^{(i)})$ = average distance between $x^{(i)}$ and the closest cluster it does not belong to
- $S(x^{(i)}) \in [-1, 1]$ with

$$S(x^{(i)}) = \frac{(b(x^{(i)}) - a(x^{(i)}))}{\max(a(x^{(i)}), b(x^{(i)}))}$$

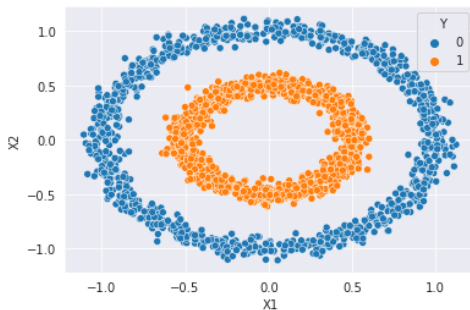


Note: $S(x^{(i)})$ large (0.5 is often used as cut-off): well clustered; $S(x^{(i)})$ small: badly clustered, $S(x^{(i)}) < 0$: assigned to wrong cluster.

Sharing or publishing the contents in part or full is liable for legal action.

Case study: clustering

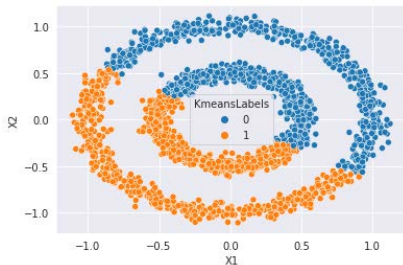
Which clustering methods are able to identify the two clusters?



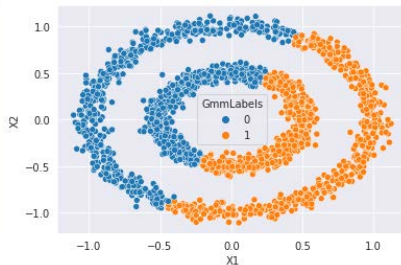
This file is meant for personal use by hhung.inbox@gmail.com only.
Sharing or publishing the contents in part or full is liable for legal action.

Case study: clustering

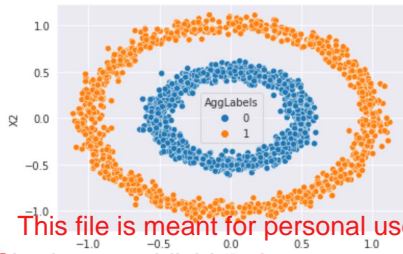
k-means clustering



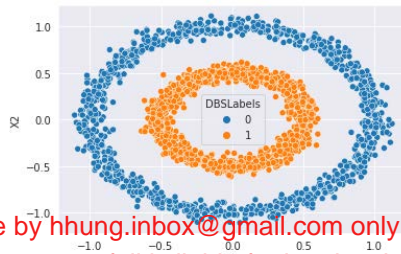
Gaussian mixture model



Hierarchical clustering (single linkage)



DBSCAN



This file is meant for personal use by hhung.inbox@gmail.com only.
Sharing or publishing the contents in part or full is liable for legal action.

Community detection

Community detection:

- detect subsets of nodes that are more densely connected between each other in the network than outside the community

Clustering

- determine subsets of points that are 'close' to each other given a pairwise distance or similarity measure defined by the network
- examples for vertex similarity measures: hop distance, number of different neighbors, correlation between adjacency matrix columns,...
- can use clustering methods discussed so far based on these similarity measures to identify communities in a network

This file is meant for personal use by hhung.inbox@gmail.com only.

Sharing or publishing the contents in part or full is liable for legal action.

Other methods: Divisive and agglomerative algorithms

- **Algorithm of Girvan and Newman** (2002): iteratively remove edges with highest **betweenness centrality**
 - Intuition: intercommunity edges have a large value of edge betweenness, because many shortest paths connecting vertices of different communities will pass through them

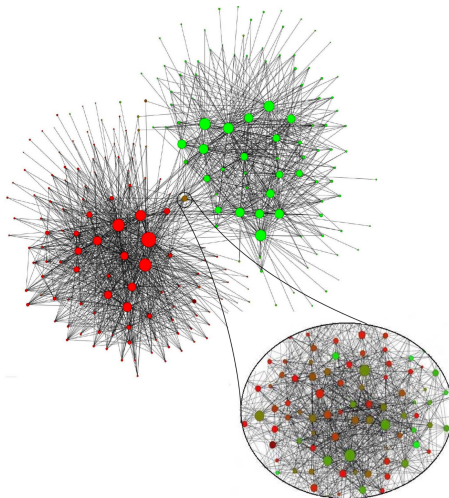
This file is meant for personal use by hhung.inbox@gmail.com only.
Sharing or publishing the contents in part or full is liable for legal action.

Other methods: Divisive and agglomerative algorithms

- **Algorithm of Girvan and Newman** (2002): iteratively remove edges with highest **betweenness centrality**
 - Intuition: intercommunity edges have a large value of edge betweenness, because many shortest paths connecting vertices of different communities will pass through them
- **Louvain method** by Blondel et al. (2008): iteratively merge pairs of clusters that are connected by more edges than expected if the edges were randomly distributed (also known as **modularity score**)
 - Is extremely fast and provides decomposition of network into communities for different levels of organization

This file is meant for personal use by hhung.inbox@gmail.com only.
Sharing or publishing the contents in part or full is liable for legal action.

Louvain method



Belgian mobile phone network with 2M customers (red: French-speaking, green: Dutch-speaking).

This file is meant for personal use by thung.inbox@gmail.com only.
Sharing or publishing the contents in part or full is liable for legal action.

References

- For clustering
 - Chapter 14 in
T. Hastie, R. Tibshirani, & J. Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer, 2009.
- For community detection in networks:
 - V. D. Blondel, et al. *Fast unfolding of communities in large networks*. Journal of Statistical Mechanics: Theory and Experiment 10, 2008.
 - S. Fortunato. *Community detection in graphs*. Physics Reports 486, 2010.
 - Lecture notes on Laplacian and spectral clustering (prominent method not discussed in this module) by T. Roughgarden & G. Valiant:
<http://web.stanford.edu/class/cs168/1/111.pdf>

This file is meant for personal use by thung.inbox@gmail.com only.

Sharing or publishing the contents in part or full is liable for legal action.