Practical AI: clustering practicum

Stanislav Protasov for
Harbour.Space University

Reading

https://scikit-learn.org/stable/modules/clustering.html

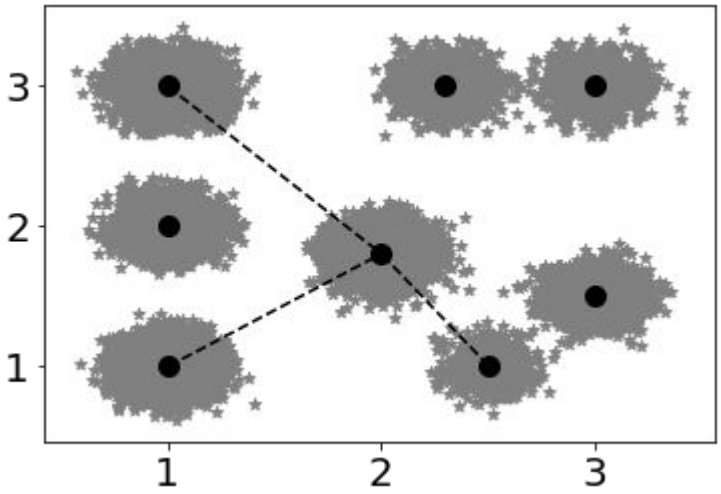
https://louvain-igraph.readthedocs.io/en/latest/intro.html

Agenda

- Problem statement
- How we measure quality
- Couple of algorithms
  - K-means
  - Louvain modularity
  - Hierarchical clustering
  - DBScan

Why do we cluster?

- And **modelling** is done to **simplify data**
- We simplify because we cannot **make decisions** based on millions of numbers
  - E.g. Linear regression brings **few numbers** to describe a domain instead of holding samples
  - “Terminator and similar” is a good way to describe customer’s preferences
- Clustering is a way to bring **limited number of entities** (clusters or representatives) while **preserving** general **idea** about the structure.



Clustering - what is this?

Set partitioning - grouping of the set's elements into non-empty subsets, in such a way that every element is included in one and only one of the subsets.

Number of partitions - **Bells number** (~ e^x) $B_{n+1} = \sum_{k=0}^{n} \binom{n}{k} B_k$

Number of non-empty partitions of size k- **Stirling number of second kind** $\left\{ {n \atop k} \right\} = \frac{1}{k!} \sum_{i=0}^{k} (-1)^i \binom{k}{i} (k-i)^n.$

NB: for any metric introduced, we cannot solve a problem with brute force

Clustering - what to do then?

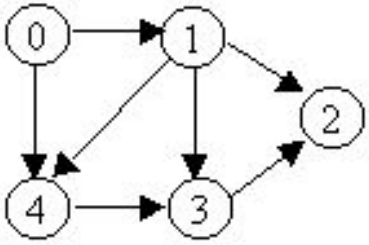
Thus we put **limitations**:

- Pre-define number of clusters
- Implement iterative approaches
- Rely on distance to avoid considering obviously bad case

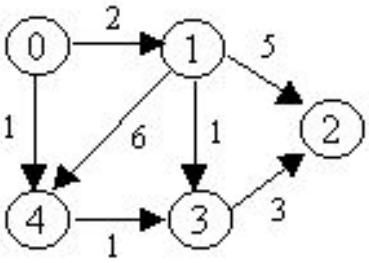
But even then clustering is usually slow.

Clustering - what is the object?

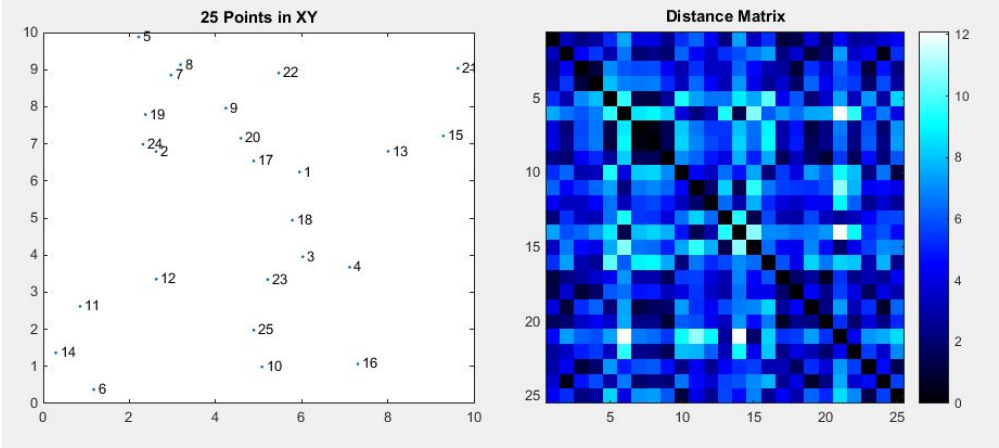
As we don’t have any idea about cluster form, we will rely on distance. There are 2 major approaches to define distance



$$A = \begin{pmatrix} 0 & 1 & 0 & 0 & 1 \\ 0 & 0 & 1 & 1 & 1 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \end{pmatrix}$$



$$A = \begin{pmatrix} \infty & 2 & \infty & \infty & 1 \\ \infty & \infty & 5 & 1 & 6 \\ \infty & \infty & \infty & \infty & \infty \\ \infty & \infty & 3 & \infty & \infty \\ \infty & \infty & \infty & 1 & \infty \end{pmatrix}$$



Clustering - how to understand success?

General idea: ... include groups with **small distances between cluster members**, dense areas of the data space ...

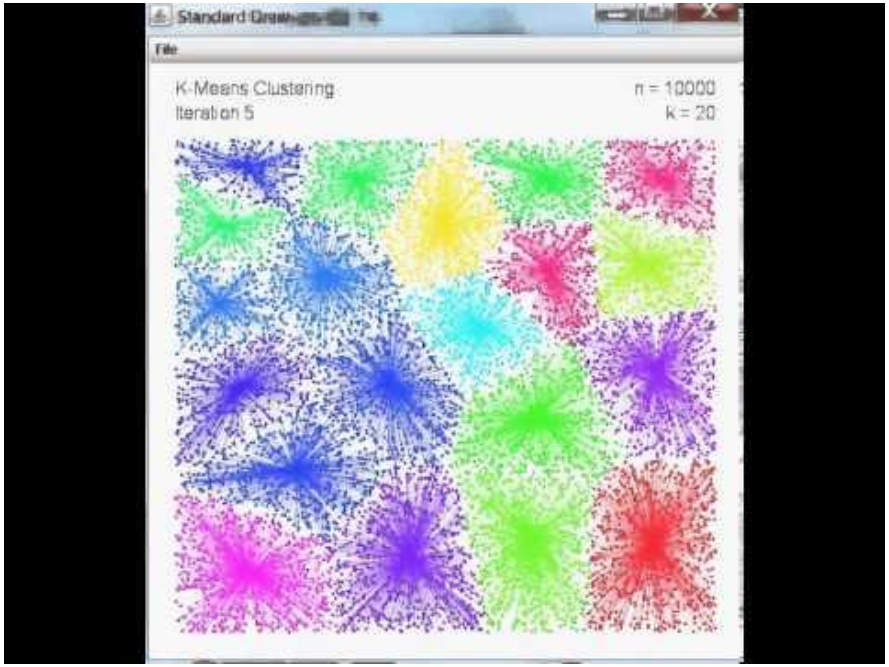
Also: maximize between-cluster variance, minimize within-class variance.

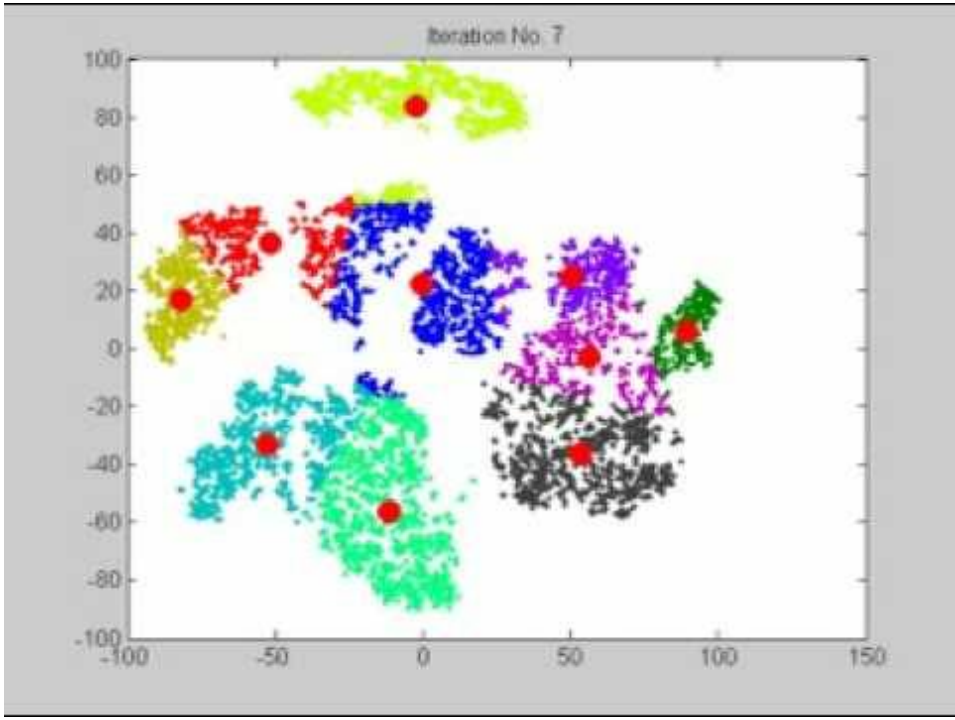
Internal evaluation (on the training data).

- Davies–Bouldin index $DB = \frac{1}{n}\sum_{i=1}^{n} \max_{j \neq i} \left(\frac{\sigma_i + \sigma_j}{d(c_i, c_j)} \right)$
- Dunn index
- Silhouette Coefficient $D = \frac{\min_{1 \le i < j \le n} d(i,j)}{\max_{1 \le k \le n} d'(k)},$

Purity, coverage, Differential edit distance - rely on pre-defined clusters (compare with validation set)

K-Means


File
K-Means Clustering
Iteration 5
n = 10000
k = 20


Iteration No. 7
100
80
60
40
20
0
-20
-40
-60
-80
-100
-100
-50
0
50
100
150

Lab #1 Clustering with kNN

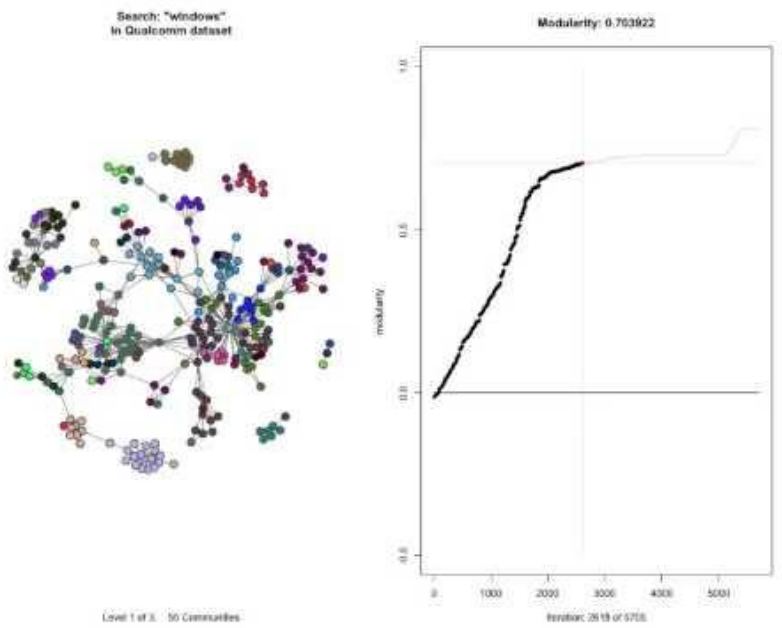
- Consider clustering example.
- Run.
- What is silhouette score for k={2, 3}?
- Why?

Louvain modularity

$$Q = \frac{1}{2m} \sum_{ij} \left[A_{ij} - \frac{k_i k_j}{2m} \right] \delta(c_i, c_j),$$

- A_{ij} represents the edge weight between nodes i and j;
- k_i and k_j are the sum of the weights of the edges attached to nodes i and j, respectively;
- $2m$ is the sum of all of the edge weights in the graph;
- c_i and c_j are the communities of the nodes; and
- δ is a simple delta function.

- Graph-based
- Considers only existing edges (no cetroids)
- Starts with community number == number of nodes.
- Searches for communities. Change element assignment if this improves **modularity**



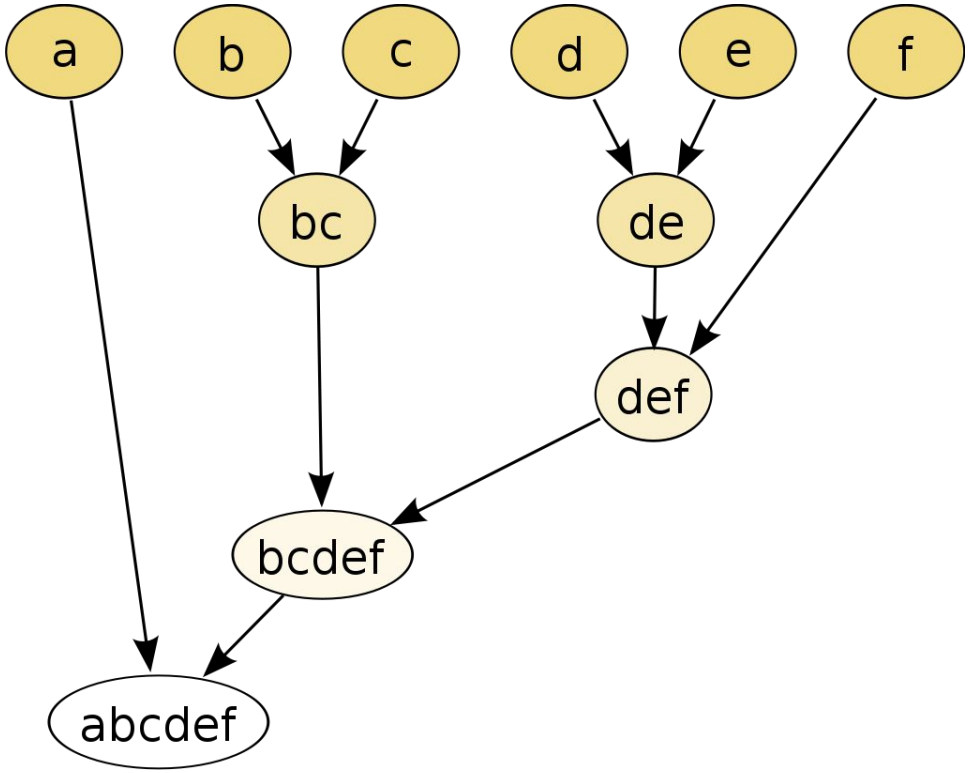
Hierarchical Clustering

Bottom-up (merge smaller clusters to improve metric)

Top-down (divide clusters to improve)

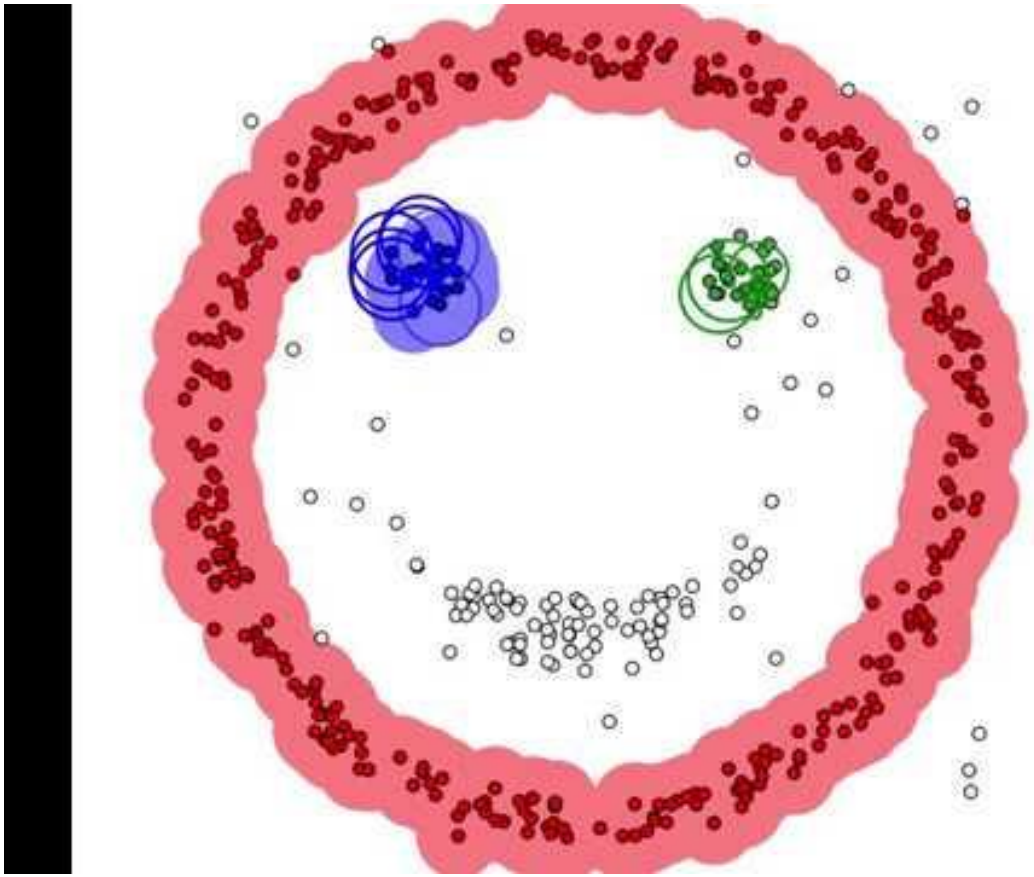
Rely on:

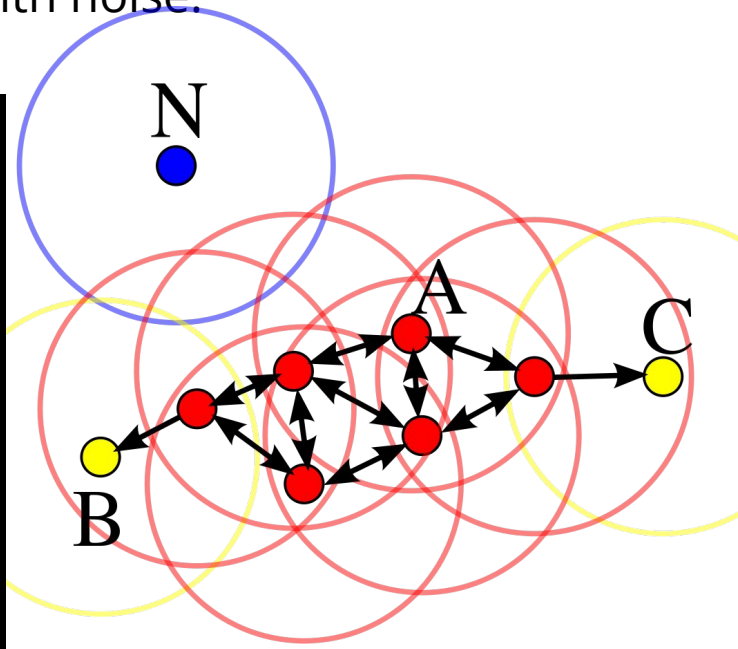
- Single links
- All within-cluster links
- Centroids



DBScan

Density-based spatial clustering of applications with noise.





Lab #2

Run different clustering algorithms on the same data.

Estimate quality.

Visualize your data. Please, refer to https://nikkimarinsek.com/blog/7-ways-to-label-a-cluster-plot-python

Homework

User has multiple subscriptions. Too many to show a news feed. We need to bring this number to 10 without losing quality.

https://github.com/hsu-ai-course/hsu.ai/tree/master/homeworks/13