Fundamentals of machine learning

Course 5: Clustering

Titouan Vayer & Mathurin Massias
email: titouan.vayer@inria.fr, mathurin.massias@inria.fr

February 13, 2025

ENS DE LYON

Table of contents

What is clustering and density estimation ?

K-means clustering
- The principle
- The algorithm
- Some failures of K-means

Spectral clustering

Hierarchical Clustering Analysis

Table of contents

What is clustering and density estimation ?

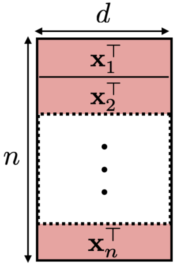
K-means clustering
- The principle
- The algorithm
- Some failures of K-means

Spectral clustering

Hierarchical Clustering Analysis

Unsupervised dataset

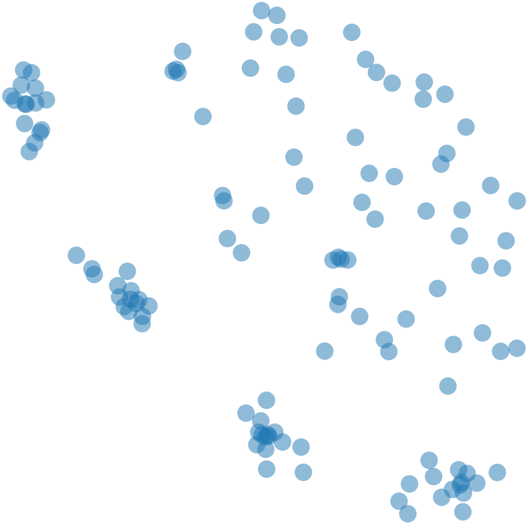
$$\mathbf{X} = \begin{bmatrix} \mathbf{x}_1^\top \\ \mathbf{x}_2^\top \\ \vdots \\ \mathbf{x}_n^\top \end{bmatrix} = \begin{bmatrix} x_{11} & x_{12} & \dots & x_{1d} \\ \vdots & \vdots & \vdots & \vdots \\ x_{n1} & x_{n2} & \dots & x_{nd} \end{bmatrix}$$



Unsupervised learning

- The dataset contains the samples $(\mathbf{x}_i)_{i=1}^n$ where n is the number of samples of size d.
- d and n define the dimensionality of the learning problem.
- Data stored as a matrix $\mathbf{X} \in \mathbb{R}^{n \times d}$ that contains the training samples as rows.

Motivating example



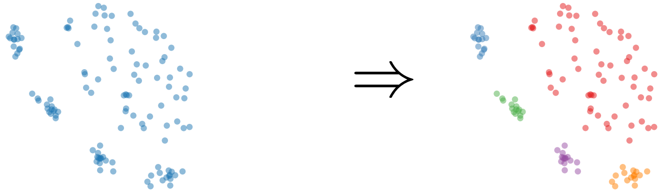
Understanding the data

- The samples come from a certain distribution $p(\mathbf{x})$ ($\mathbf{x}_i \sim p(\mathbf{x})$).
- $p(\mathbf{x})$ is unknown !
- Density estimation: find $\hat{p}(\mathbf{x}) \approx p(\mathbf{x})$.
- One can generate new samples from this approximate distribution.

Understanding the data

- The samples come from a certain distribution $p(\mathbf{x})$ ($\mathbf{x}_i \sim p(\mathbf{x})$).
- $p(\mathbf{x})$ is unknown !
- Density estimation: find $\hat{p}(\mathbf{x}) \approx p(\mathbf{x})$.
- One can generate new samples from this approximate distribution.
- $p(\mathbf{x})$ is usually complicated: find a understandable/compact representation of it.
- Clustering: group points together.
- Find most "representative" points of $p(\mathbf{x})$.

Clustering



Objective

$$\{\mathbf{x}_i\}_{i=1}^{n} \quad \Rightarrow \quad \{\hat{y}_i\}_{i=1}^{n}$$

- Organize training examples in groups/clusters
- Find the labels $\hat{y}_i \in \mathcal{Y} = \{1, \ldots, K\}$.

Parameters

- K number of clusters.
- Similarity measure between samples.

Methods

- K-means.
- Gaussian mixtures.
- Spectral clustering.
- Hierarchical clustering.

Table of contents

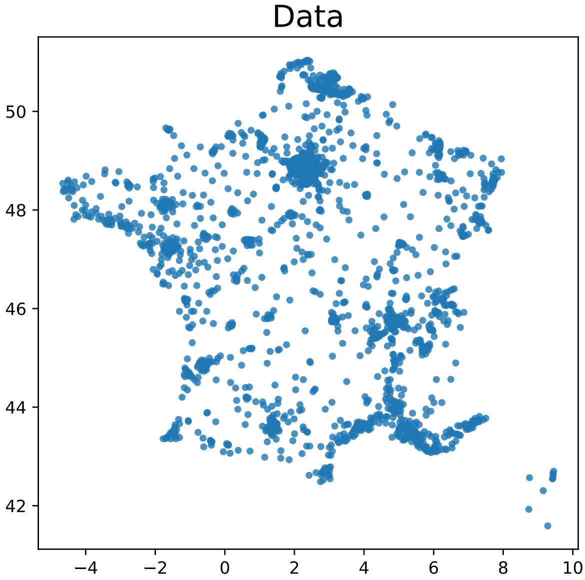
What is clustering and density estimation ?

K-means clustering
- The principle
- The algorithm
- Some failures of K-means

Spectral clustering

Hierarchical Clustering Analysis

Motivating example



K-means clustering

- Find K centroids $\mathbf{c}_1, \cdots, \mathbf{c}_K \in \mathbb{R}^d$ that optimize MacQueen 1967:

$$\min_{\mathbf{c}_1, \cdots, \mathbf{c}_K} \sum_{i=1}^{n} \min_{k \in [\![K]\!]} \|\mathbf{x}_i - \mathbf{c}_k\|_2^2$$

- $\min_{k \in [\![K]\!]} \|\mathbf{x}_i - \mathbf{c}_k\|_2^2$: squared distance between $\mathbf{x}_i$ and its *closest* centroid.

K-means clustering

- Find K centroids $\mathbf{c}_1, \cdots, \mathbf{c}_K \in \mathbb{R}^d$ that optimize MacQueen 1967:

$$\min_{\mathbf{c}_1, \cdots, \mathbf{c}_K} \sum_{i=1}^{n} \min_{k \in [\![K]\!]} \|\mathbf{x}_i - \mathbf{c}_k\|_2^2$$

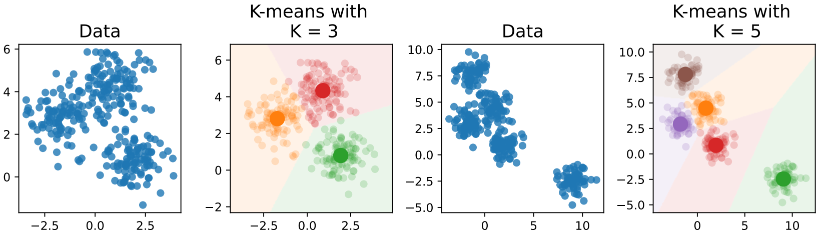
- $\min_{k \in [\![K]\!]} \|\mathbf{x}_i - \mathbf{c}_k\|_2^2$: squared distance between $\mathbf{x}_i$ and its *closest* centroid.
- Group according to the closest centroids $\approx$ most "representative point".
- On average the dist. between $\mathbf{x}_i$ and its closest centroid is minimum.

K-means clustering

- Find K centroids $\mathbf{c}_1, \cdots, \mathbf{c}_K \in \mathbb{R}^d$ that optimize MacQueen 1967:

$$\min_{\mathbf{c}_1, \cdots, \mathbf{c}_K} \sum_{i=1}^{n} \min_{k \in [\![K]\!]} \|\mathbf{x}_i - \mathbf{c}_k\|_2^2$$

- $\min_{k \in [\![K]\!]} \|\mathbf{x}_i - \mathbf{c}_k\|_2^2$: squared distance between $\mathbf{x}_i$ and its *closest* centroid.
- Group according to the closest centroids $\approx$ most "representative point".
- On average the dist. between $\mathbf{x}_i$ and its closest centroid is minimum.



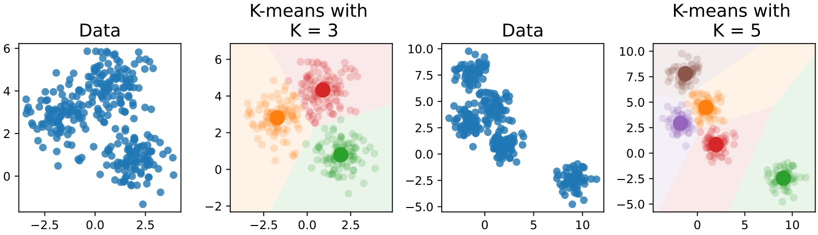
K-means clustering:

Equivalent formulation

- Find K centroids $\mathbf{c}_1, \cdots, \mathbf{c}_K \in \mathbb{R}^d$ that optimize MacQueen 1967:

$$\min_{\substack{\mathbf{c}_1,\cdots,\mathbf{c}_K \\ \mathbf{P}\in\{0,1\}^{n\times K}, \mathbf{P}\mathbf{1}_K=\mathbf{1}_n}} \quad \sum_{i=1,k=1}^{n,K} P_{ik}\|\mathbf{x}_i - \mathbf{c}_k\|_2^2$$

- $P_{ik} \in \{0,1\}, \mathbf{P}\mathbf{1}_K = \mathbf{1}_n$: assign point $\mathbf{x}_i$ to centroid k ($\mathbf{x}_i$ to cluster k).
- Binary assignment problem.
- K-means is NP-Hard (even for $k = 2$ Drineas et al. 2004).



The algorithm

Lloyd's algorithm Lloyd 1982

Lloyd's alternating scheme

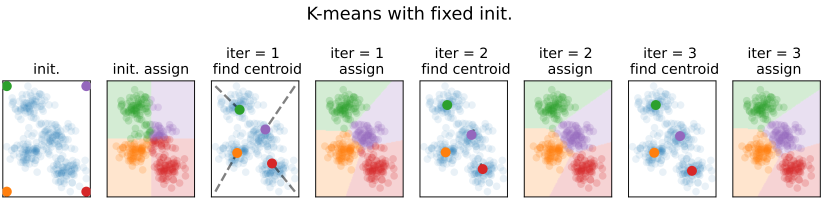
Init. centroids $\mathbf{c}_1, \cdots, \mathbf{c}_K$
while Not converged **do**
    Assign each $\mathbf{x}_i$ to its closest centroid $\mathbf{c}_{k_i}$.
    Update each $\mathbf{c}_k$ as the mean of the points in each cluster: $\mathbf{c}_k \leftarrow \frac{1}{|C_k|} \sum_{i \in C_k} \mathbf{x}_i$
end while

The algorithm

Lloyd's algorithm Lloyd 1982

Lloyd's alternating scheme

Init. centroids $\mathbf{c}_1, \cdots, \mathbf{c}_K$
while Not converged **do**
    Assign each $\mathbf{x}_i$ to its closest centroid $\mathbf{c}_{k_i}$.
    Update each $\mathbf{c}_k$ as the mean of the points in each cluster: $\mathbf{c}_k \leftarrow \frac{1}{|C_k|}\sum_{i \in C_k} \mathbf{x}_i$
end while



The algorithm

Lloyd's algorithm Lloyd 1982

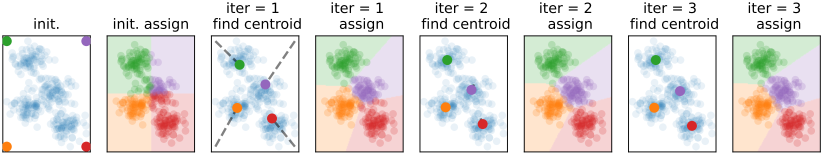
Lloyd's alternating scheme

Init. centroids $\mathbf{c}_1, \cdots, \mathbf{c}_K$
while Not converged **do**
  Assign each $\mathbf{x}_i$ to its closest centroid $\mathbf{c}_{k_i}$.
  Update each $\mathbf{c}_k$ as the mean of the points in each cluster: $\mathbf{c}_k \leftarrow \frac{1}{|C_k|}\sum_{i \in C_k} \mathbf{x}_i$
end while

Remarks

- Heuristic algorithm but very efficient and simple !
- Alternating minimization scheme (BCD on $\mathbf{P}, \mathbf{c}_1, \cdots, \mathbf{c}_k$).
- Time complexity $\mathcal{O}(ndk \times n_{iter})$.
- Provably near-optimal clustering solutions when applied to well-clusterable data Ostrovsky et al. 2013.

K-means with fixed init.



The algorithm

Lloyd's algorithm Lloyd 1982

Lloyd's alternating scheme

1: Init. centroids $\mathbf{c}_1, \cdots, \mathbf{c}_K$
2: **while** Not converged **do**
3: Assign each $\mathbf{x}_i$ to its closest centroid $\mathbf{c}_{k_i}$.
4: Update each $\mathbf{c}_k$ as the mean of the points in each cluster: $\mathbf{c}_k \leftarrow \frac{1}{|C_k|} \sum_{i \in C_k} \mathbf{x}_i$
5: **end while**

The python code

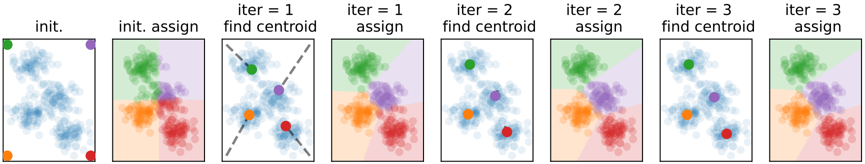
```python
from sklearn.cluster import KMeans
# K-means with K=2
clf = KMeans(2)

# fit the model and predict classes
y = clf.fit_predict(X)

# distance from samples to clusters
dist = clf.transform(X)

# get the centroids
C = clf.cluster_centers_
```

K-means with fixed init.



Voronoi diagram

A partition of the space Voronoi 1908

- $S = \{\mathbf{c}_1, \cdots, \mathbf{c}_K\}$ a finite set of points.
- Voronoï cell of $\mathbf{c}_k$ (ideas trace back to Descartes!):

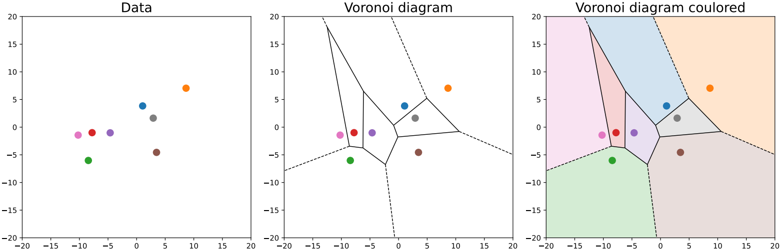
$$\mathrm{Vor}_S(\mathbf{c}_k) = \{\mathbf{x} \in \mathbb{R}^d : \forall \mathbf{c}_j \in S \setminus \{\mathbf{c}_k\}, \|\mathbf{x} - \mathbf{c}_k\| \leq \|\mathbf{x} - \mathbf{c}_j\|\}$$

Voronoi diagram

A partition of the space Voronoi 1908

- $S = \{\mathbf{c}_1, \cdots, \mathbf{c}_K\}$ a finite set of points.
- Voronoï cell of $\mathbf{c}_k$ (ideas trace back to Descartes!):

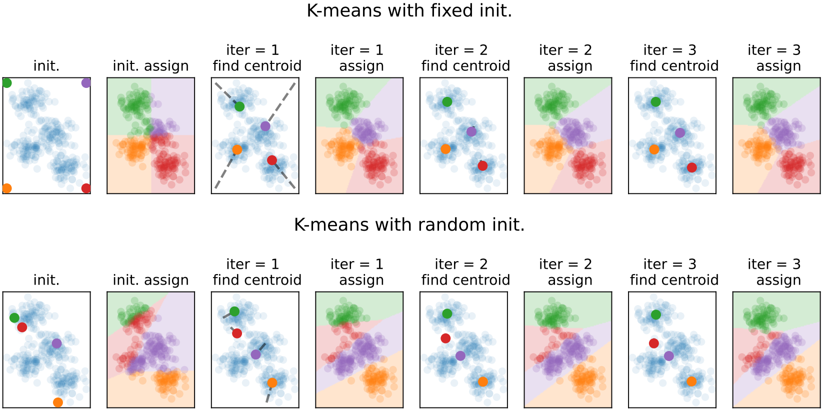
$$\mathrm{Vor}_S(\mathbf{c}_k) = \{\mathbf{x} \in \mathbb{R}^d : \forall \mathbf{c}_j \in S \setminus \{\mathbf{c}_k\}, \|\mathbf{x} - \mathbf{c}_k\| \leq \|\mathbf{x} - \mathbf{c}_j\|\}$$



- Each “assign step” of the K-means finds Voronoï cells of the centroids.

Problem of initialisation

Lloyd's algorithm highly depends on the init. (bad local minima).



The algorithm

Finding a good initialisation.

- "Naive" Lloyd's algorithm may find sub-optimal clustering.
- Spreading out the K initial cluster centers is important.

The algorithm

Finding a good initialisation.

- "Naive" Lloyd's algorithm may find sub-optimal clustering.
- Spreading out the K initial cluster centers is important.

K-means ++ Arthur and Vassilvitskii 2006.

- Choose one center uniformly at random.
- $D(\mathbf{x}_i)$: distance between $\mathbf{x}_i$ and its nearest already selected centroid.
- Choose one new cluster amoung points with prob. $\propto D(\mathbf{x}_i)^2$.
- Repeat until K centroids are chosen.

The algorithm

Finding a good initialisation.

- "Naive" Lloyd's algorithm may find sub-optimal clustering.
- Spreading out the K initial cluster centers is important.

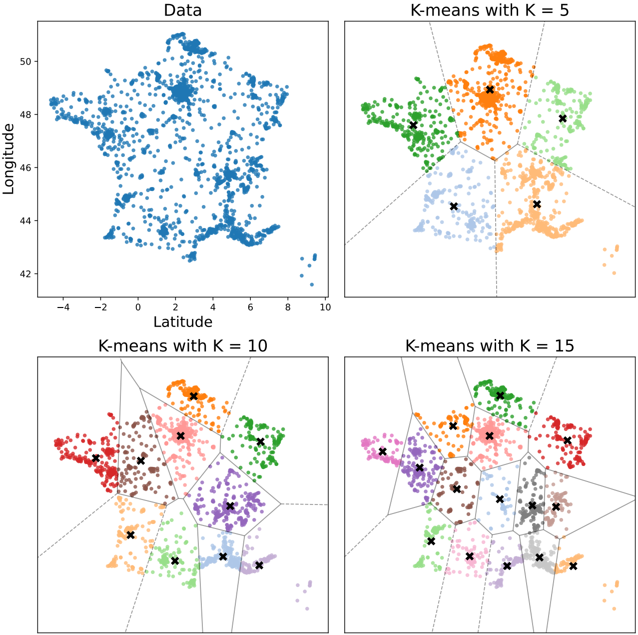
K-means ++ Arthur and Vassilvitskii 2006.

- Choose one center uniformly at random.
- $D(\mathbf{x}_i)$: distance between $\mathbf{x}_i$ and its nearest already selected centroid.
- Choose one new cluster amoung points with prob. $\propto D(\mathbf{x}_i)^2$.
- Repeat until K centroids are chosen.

In practice

- Run the K-means algorithm with different init.
- Choose the best configuration in the end (with lowest clustering error).

Illustration with map of France



Some statistical guarantees of k-means ++

- On expectation it leads to a solution close to the optimum.
- The quantification error:

$$Q_n(\mathbf{c}_1, \cdots, \mathbf{c}_K) = \sum_{i=1}^{n} \min_{k \in [\![K]\!]} \|\mathbf{x}_i - \mathbf{c}_k\|_2^2 \,.$$

Some statistical guarantees of k-means ++

- On expectation it leads to a solution close to the optimum.
- The quantification error:

$$Q_n(\mathbf{c}_1, \cdots, \mathbf{c}_K) = \sum_{i=1}^{n} \min_{k \in [\![K]\!]} \|\mathbf{x}_i - \mathbf{c}_k\|_2^2 .$$

Arthur and Vassilvitskii 2006

If $\mathbf{c}_1, \cdots, \mathbf{c}_K$ are centroids obtained by k-means++, then

$$\mathbb{E}[Q_n(\mathbf{c}_1, \cdots, \mathbf{c}_K)] \leq 8(\log K + 2) \min_{\mathbf{c}'_1, \cdots, \mathbf{c}'_K} Q_n(\mathbf{c}'_1, \cdots, \mathbf{c}'_K)$$

where the expectation is take with respect to the random choice of initial centroids.

K-means variants

K-medoids Maranzana 1963

- Choose centroids amoung points:

$$\mathbf{c}_1, \cdots, \mathbf{c}_K \in \{\mathbf{x}_1, \cdots, \mathbf{x}_n\}.$$

- Partitioning Around Medoids (PAM) algorithm (on the board).

K-means variants

K-medoids Maranzana 1963

- Choose centroids amoung points:

$$\mathbf{c}_1, \cdots, \mathbf{c}_K \in \{\mathbf{x}_1, \cdots, \mathbf{x}_n\}.$$

- Partitioning Around Medoids (PAM) algorithm (on the board).

Change the metric

- Solve for some "distance function" d:

$$\min_{\mathbf{c}_1, \cdots, \mathbf{c}_K} \sum_{i=1}^{n} \min_{k \in [\![K]\!]} d(\mathbf{x}_i, \mathbf{c}_k)$$

- *e.g.* robust k-means with $d = \|\cdot\|_1$ (K-median Bradley, Mangasarian, and Street 1996), or even non-Euclidean data !

K-means variants

K-medoids Maranzana 1963

- Choose centroids amoung points:

$$\mathbf{c}_1, \cdots, \mathbf{c}_K \in \{\mathbf{x}_1, \cdots, \mathbf{x}_n\}.$$

- Partitioning Around Medoids (PAM) algorithm (on the board).

Change the metric

- Solve for some "distance function" d:

$$\min_{\mathbf{c}_1, \cdots, \mathbf{c}_K} \sum_{i=1}^{n} \min_{k \in [\![K]\!]} d(\mathbf{x}_i, \mathbf{c}_k)$$

- *e.g.* robust k-means with $d = \|\cdot\|_1$ (K-median Bradley, Mangasarian, and Street 1996), or even non-Euclidean data !

Large-scale dataset

- For very large n: Minibatch-Kmeans Sculley 2010 or Stochastic Gradient Descent (SGD) Bottou and Bengio 1994.

Illustration with map of France

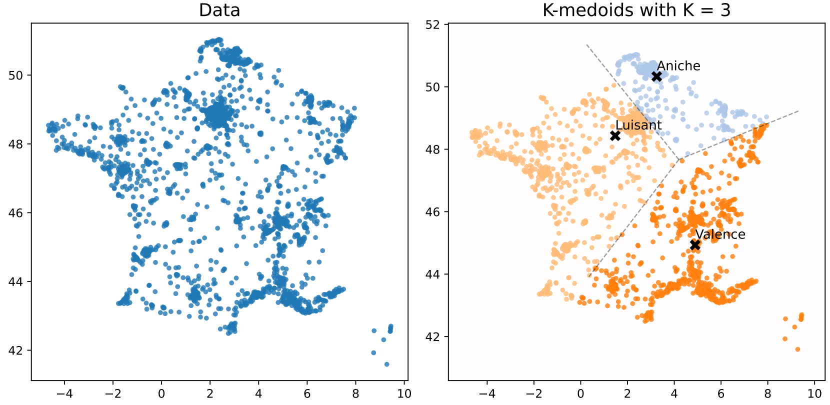


Illustration with map of France

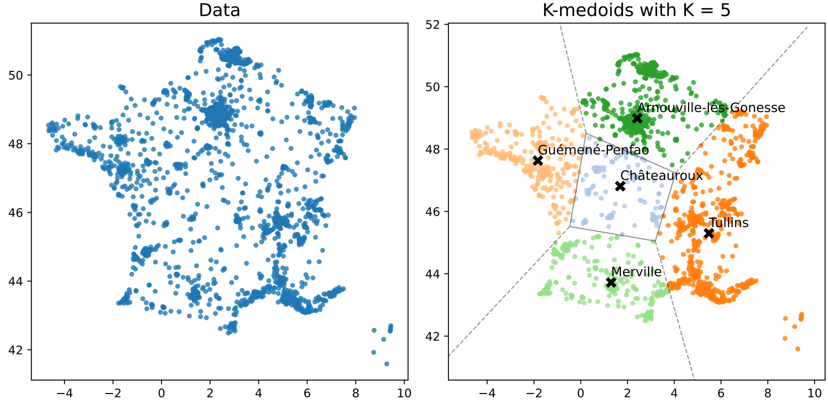


Illustration with map of France

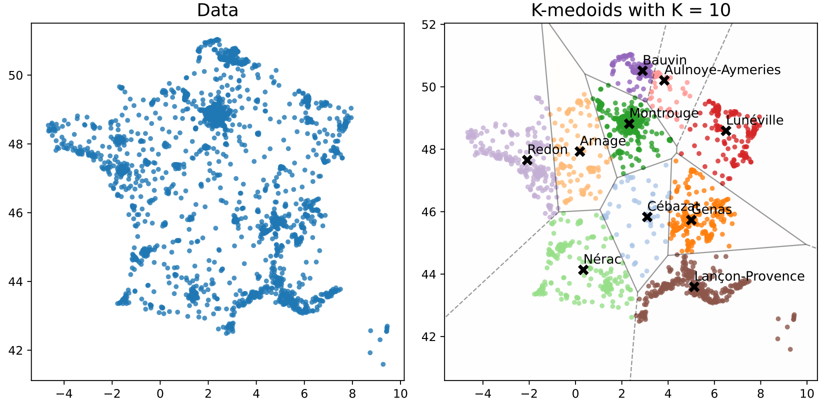


Illustration with map of France

With more point **and** with $d(\mathbf{x}, \mathbf{y})$ is the shortest-path distance between the points $\mathbf{x}, \mathbf{y}$ `https://github.com/tvayer/Kmeanscountry`.

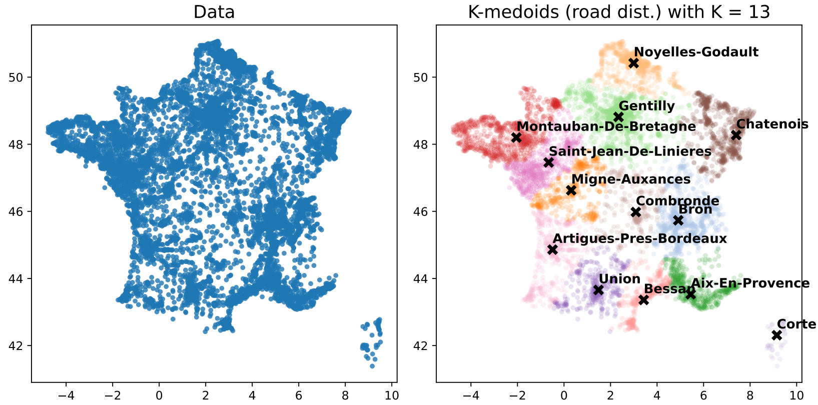


Illustration on time series

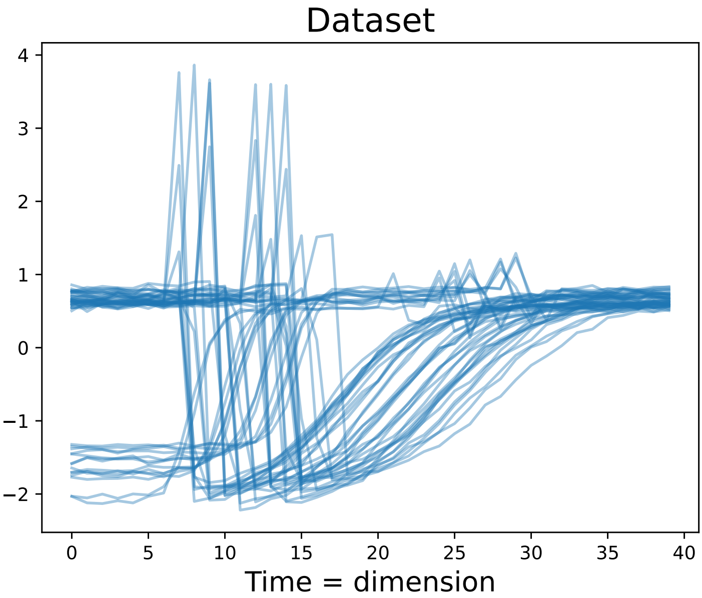
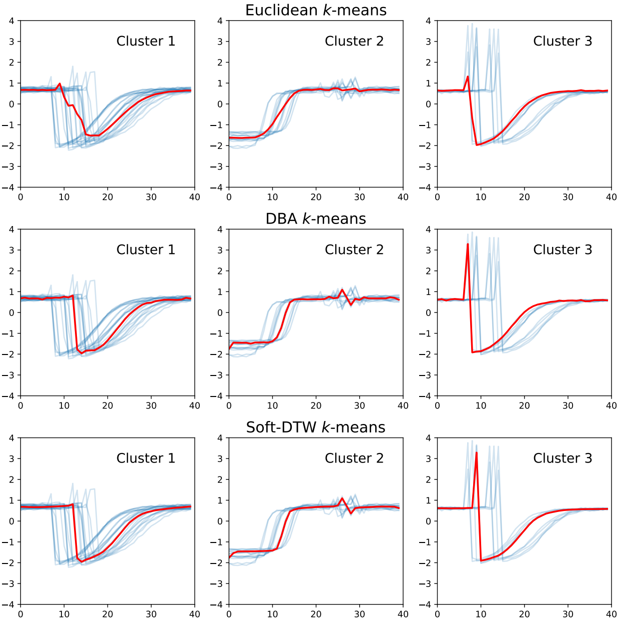
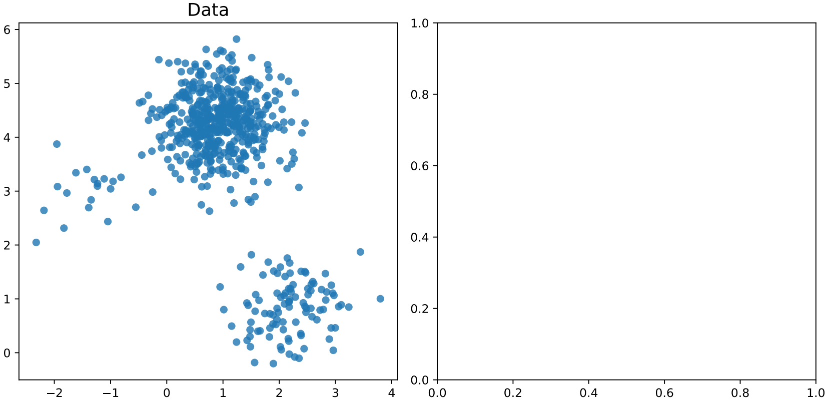


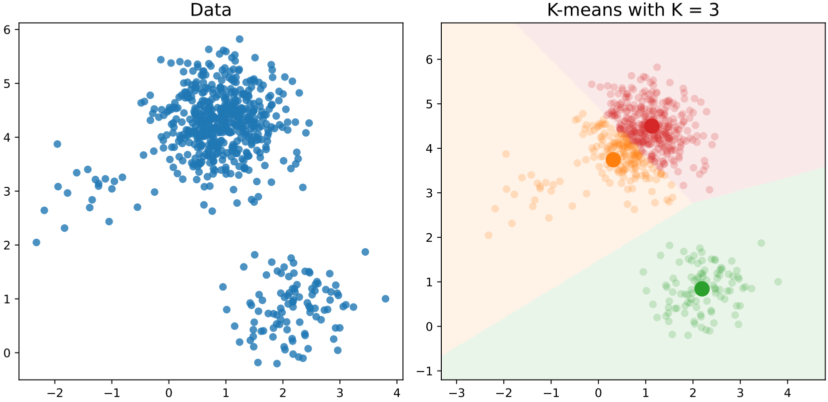
Illustration on time series



Some failures of K-means



Some failures of K-means



Some failures of K-means

- Density of each cluster is important in K-means !
- In the classical formulation each point has the same importance/mass:

$$\min_{\mathbf{c}_1,\cdots,\mathbf{c}_K} \sum_{i=1}^{n} \frac{1}{n} \min_{k\in[\![K]\!]} \|\mathbf{x}_i - \mathbf{c}_k\|_2^2$$

- One "remedy" is weighted K-means

$$\min_{\mathbf{c}_1,\cdots,\mathbf{c}_K} \sum_{i=1}^{n} w_i \min_{k\in[\![K]\!]} \|\mathbf{x}_i - \mathbf{c}_k\|_2^2 \text{ with } w_1,\cdots,w_n \geq 0 \text{ s.t. } \sum_{i=1}^{n} w_i = 1\,.$$

Some failures of K-means

- Density of each cluster is important in K-means !
- In the classical formulation each point has the same importance/mass:

$$\min_{\mathbf{c}_1,\cdots,\mathbf{c}_K} \sum_{i=1}^{n} \frac{1}{n} \min_{k\in[\![K]\!]} \|\mathbf{x}_i - \mathbf{c}_k\|_2^2$$

- One "remedy" is weighted K-means

$$\min_{\mathbf{c}_1,\cdots,\mathbf{c}_K} \sum_{i=1}^{n} w_i \min_{k\in[\![K]\!]} \|\mathbf{x}_i - \mathbf{c}_k\|_2^2 \text{ with } w_1,\cdots,w_n \geq 0 \text{ s.t. } \sum_{i=1}^{n} w_i = 1\,.$$

- E.g. $w_i = \frac{1}{K}\frac{1}{|C_k|}$ if $i \in C_k$: points in small clusters more important (unknown).

Some failures of K-means

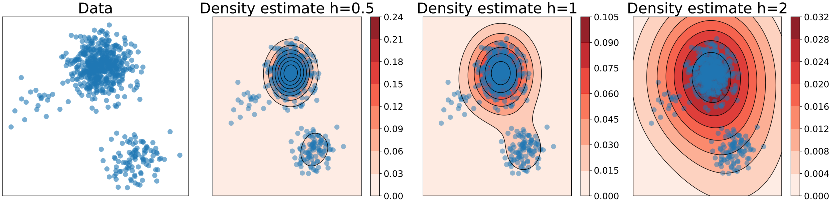
- Density of each cluster is important in K-means !
- In the classical formulation each point has the same importance/mass:

$$\min_{\mathbf{c}_1,\cdots,\mathbf{c}_K} \sum_{i=1}^{n} \frac{1}{n} \min_{k\in[\![K]\!]} \|\mathbf{x}_i - \mathbf{c}_k\|_2^2$$

- One "remedy" is weighted K-means

$$\min_{\mathbf{c}_1,\cdots,\mathbf{c}_K} \sum_{i=1}^{n} w_i \min_{k\in[\![K]\!]} \|\mathbf{x}_i - \mathbf{c}_k\|_2^2 \text{ with } w_1,\cdots,w_n \geq 0 \text{ s.t. } \sum_{i=1}^{n} w_i = 1\,.$$

- E.g. $w_i = \frac{1}{K}\frac{1}{|C_k|}$ if $i \in C_k$: points in small clusters more important (unknown).
- E.g. $w_i \propto 1/\text{local density around the point}$.



Some failures of K-means

With weights (unknown):

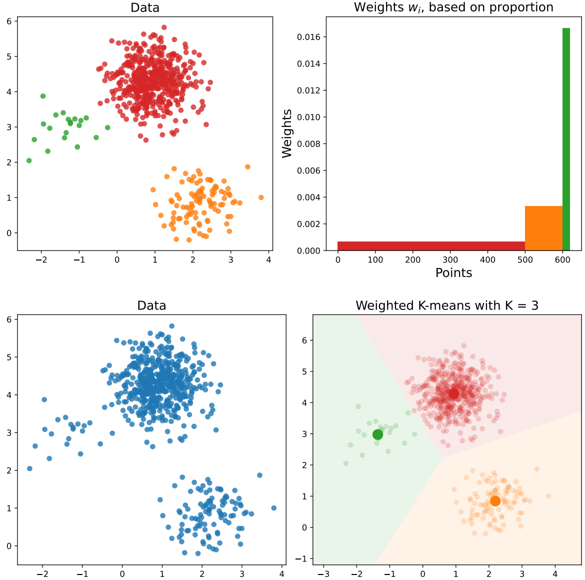


Table of contents

What is clustering and density estimation ?

K-means clustering

- The principle
- The algorithm
- Some failures of K-means

Spectral clustering

Hierarchical Clustering Analysis

Spectral clustering

The principle

- Divide data points into K groups *s.t.* points in the same group $=$ similar, points in different groups $=$ dissimilar.

Spectral clustering

The principle

- Divide data points into K groups *s.t.* points in the same group $=$ similar, points in different groups $=$ dissimilar.

Spectral clustering Von Luxburg 2007

- Let $\mathbf{A} \in \mathbb{R}^{n\times n}$ be a similarity matrix between the points $(\mathbf{x}_i, \mathbf{x}_j)$.
- *e.g.* Gaussian similarity $A_{ij} = \exp(-\gamma\|\mathbf{x}_i - \mathbf{x}_j\|_2^2)$,
- More generally weighted adjacency matrix of a graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$.

Spectral clustering

The principle

- Divide data points into K groups *s.t.* points in the same group $=$ similar, points in different groups $=$ dissimilar.

Spectral clustering Von Luxburg 2007

- Let $\mathbf{A} \in \mathbb{R}^{n\times n}$ be a similarity matrix between the points $(\mathbf{x}_i, \mathbf{x}_j)$.
- *e.g.* Gaussian similarity $A_{ij} = \exp(-\gamma\|\mathbf{x}_i - \mathbf{x}_j\|_2^2)$,
- More generally weighted adjacency matrix of a graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$.
- **Step 1:** eigenvalue decomposition of $\mathbf{L} \succeq 0$, smallest K eigenvalues.
- $\mathbf{L} = \mathbf{D} - \mathbf{A}$ the graph Laplacian matrix, $\mathbf{D} = \text{diag}(\mathbf{A1})$ degree matrix.
- Eigenvalues decomposition solves (Ky-Fan theorem):

$$\min_{\substack{\mathbf{U}\in\mathbb{R}^{n\times K}\\ \mathbf{U}^\top\mathbf{U}=\mathbf{I}_K}} \text{tr}(\mathbf{U}^\top\mathbf{L}\mathbf{U}) \qquad .$$

Spectral clustering

The principle

- Divide data points into K groups *s.t.* points in the same group $=$ similar, points in different groups $=$ dissimilar.

Spectral clustering Von Luxburg 2007

- Let $\mathbf{A} \in \mathbb{R}^{n\times n}$ be a similarity matrix between the points $(\mathbf{x}_i, \mathbf{x}_j)$.
- *e.g.* Gaussian similarity $A_{ij} = \exp(-\gamma\|\mathbf{x}_i - \mathbf{x}_j\|_2^2)$,
- More generally weighted adjacency matrix of a graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$.
- **Step 1:** eigenvalue decomposition of $\mathbf{L} \succeq 0$, smallest K eigenvalues.
- $\mathbf{L} = \mathbf{D} - \mathbf{A}$ the graph Laplacian matrix, $\mathbf{D} = \text{diag}(\mathbf{A1})$ degree matrix.
- Eigenvalues decomposition solves (Ky-Fan theorem):

$$\min_{\substack{\mathbf{U}\in\mathbb{R}^{n\times K}\\ \mathbf{U}^\top\mathbf{U}=\mathbf{I}_K}} \text{tr}(\mathbf{U}^\top\mathbf{L}\mathbf{U}) = \frac{1}{2}\sum_{i,j=1}^{n} A_{ij}\|\mathbf{u}_i - \mathbf{u}_j\|_2^2, \text{ where } \mathbf{U} = \begin{pmatrix}\mathbf{u}_1^\top\\ \vdots \\ \mathbf{u}_n^\top\end{pmatrix}.$$

- Interpretation: $\mathbf{u}_i$ closed to $\mathbf{u}_j$ when A_{ij} is big.

Spectral clustering

Spectral clustering Von Luxburg 2007

- **Step 1:** eigenvalue decomposition of $\mathbf{L} \succeq 0$, smallest K eigenvalues.
- Interpretation: gives $\mathbf{u}_i$ closed to $\mathbf{u}_j$ when A_{ij} is big.

Spectral clustering

Spectral clustering Von Luxburg 2007

- **Step 1:** eigenvalue decomposition of $\mathbf{L} \succeq 0$, smallest K eigenvalues.
- Interpretation: gives $\mathbf{u}_i$ closed to $\mathbf{u}_j$ when A_{ij} is big.
- **Step 2:** Identifiy $\mathbf{x}_i$ as $\propto \mathbf{u}_i$.

Spectral clustering

Spectral clustering Von Luxburg 2007

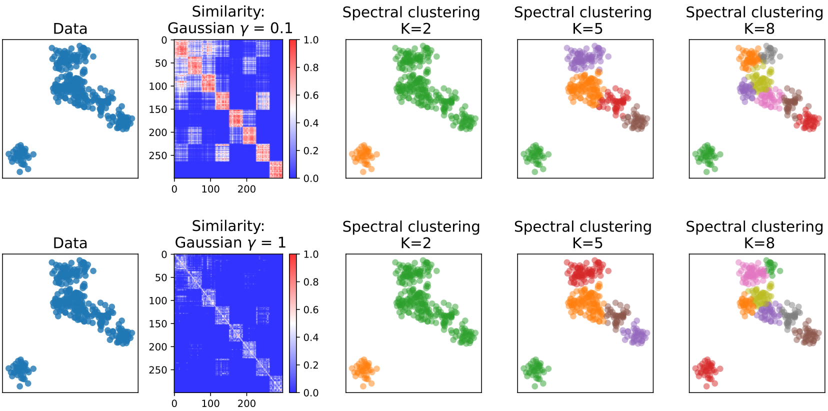
- **Step 1:** eigenvalue decomposition of $\mathbf{L} \succeq 0$, <u>smallest</u> K eigenvalues.
- Interpretation: gives $\mathbf{u}_i$ closed to $\mathbf{u}_j$ when A_{ij} is big.
- **Step 2:** Identifiy $\mathbf{x}_i$ as $\propto \mathbf{u}_i$.
- **Step 3:** Run K-means on $\mathbf{u}_1, \cdots, \mathbf{u}_n$: find $\mathbf{c}_1, \cdots, \mathbf{c}_K \in \mathbb{R}^K$ centroids.
- Cluster $\mathbf{x}_1, \cdots, \mathbf{x}_n$ accordingly.

Spectral clustering

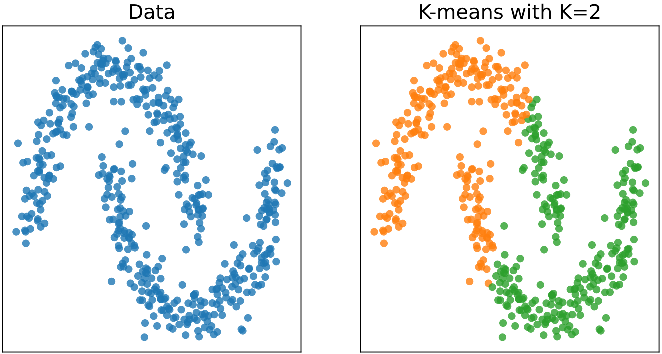
Spectral clustering Von Luxburg 2007

- **Step 1:** eigenvalue decomposition of $\mathbf{L} \succeq 0$, <u>smallest</u> K eigenvalues.
- Interpretation: gives $\mathbf{u}_i$ closed to $\mathbf{u}_j$ when A_{ij} is big.
- **Step 2:** Identifiy $\mathbf{x}_i$ as $\propto \mathbf{u}_i$.
- **Step 3:** Run K-means on $\mathbf{u}_1, \cdots, \mathbf{u}_n$: find $\mathbf{c}_1, \cdots, \mathbf{c}_K \in \mathbb{R}^K$ centroids.
- Cluster $\mathbf{x}_1, \cdots, \mathbf{x}_n$ accordingly.
- Variants with normalized graph Laplacian $\mathbf{L} = \mathbf{I} - \mathbf{D}^{-1/2}\mathbf{A}\mathbf{D}^{-1/2}$.
- Strong connections with graph partitioning and dimension reduction (on the board).

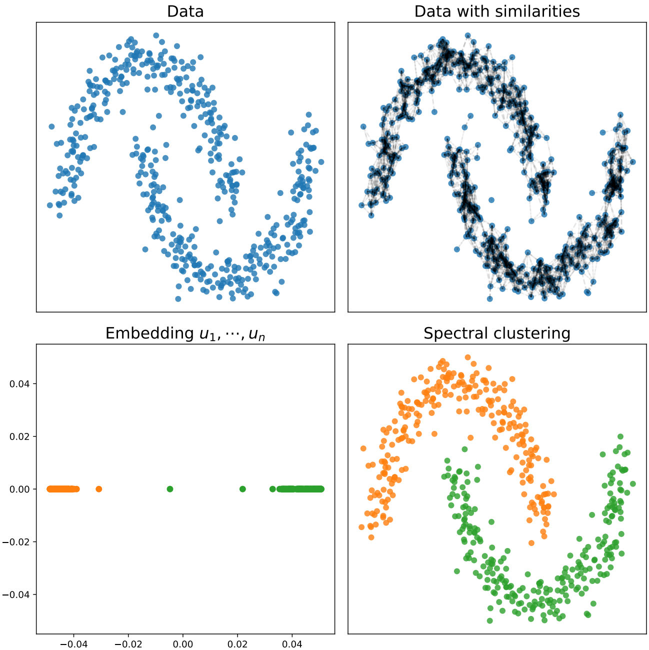
Examples


Data
Similarity:
Gaussian γ = 0.1
Spectral clustering
K=2
Spectral clustering
K=5
Spectral clustering
K=8
Data
Similarity:
Gaussian γ = 1
Spectral clustering
K=2
Spectral clustering
K=5
Spectral clustering
K=8

Examples



Examples



Data
Data with similarities
Embedding $u_1, \cdots, u_n$
Spectral clustering
0.04
0.02
0.00
−0.02
−0.04
−0.04
−0.02
0.00
0.02
0.04

Table of contents

What is clustering and density estimation ?

K-means clustering

The principle

The algorithm

Some failures of K-means

Spectral clustering

Hierarchical Clustering Analysis

Hierarchical Clustering Analysis (HCA)

Agglomerative HCA

Algorithm

Init. clusters $C_1, \cdots, C_n$ (one per sample)
while $|\{C_i\}_i| > 1$ **do**
  Find pair C_i, C_j minimizing $\Delta(C_i, C_j)$ amoung all pairs.
  Merge C_i, C_j
end while

Hierarchical Clustering Analysis (HCA)

Agglomerative HCA

Algorithm

```
Init. clusters $C_1, \cdots, C_n$ (one per sample)
while $|\{C_i\}_i| > 1$ do
    Find pair $C_i, C_j$ minimizing $\Delta(C_i, C_j)$ amoung all pairs.
    Merge $C_i, C_j$
end while
```

Tutorial Nielsen and Nielsen 2016

- Find clusters recursively through Agglomeration
- The linkage function $\Delta(C_i, C_j)$ is a measure of “distance” between two clusters.
- Final clustering with a fixed K nb. of clusters.
- The tree visualization is called *dendrogram*.

Hierarchical Clustering Analysis (HCA)

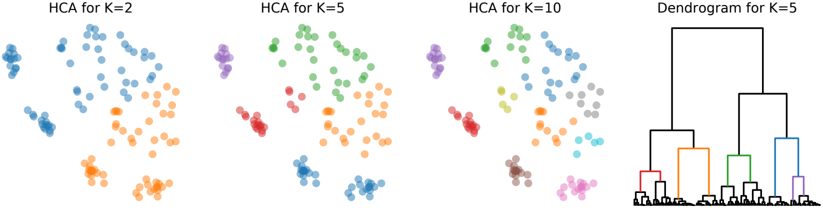
Agglomerative HCA

Algorithm

Init. clusters $C_1, \cdots, C_n$ (one per sample)
while $|\{C_i\}_i| > 1$ **do**
  Find pair C_i, C_j minimizing $\Delta(C_i, C_j)$ amoung all pairs.
  Merge C_i, C_j
end while

Tutorial Nielsen and Nielsen 2016

- Find clusters recursively through Agglomeration
- The linkage function $\Delta(C_i, C_j)$ is a measure of "distance" between two clusters.
- Final clustering with a fixed K nb. of clusters.
- The tree visualization is called *dendrogram*.



References I

Arthur, David and Sergei Vassilvitskii (2006). *k-means++: The advantages of careful seeding*. Tech. rep. Stanford.

Bottou, Leon and Yoshua Bengio (1994). “Convergence properties of the k-means algorithms”. In: *Advances in neural information processing systems* 7.

Bradley, Paul, Olvi Mangasarian, and W Street (1996). “Clustering via concave minimization”. In: *Advances in neural information processing systems* 9.

Drineas, Petros et al. (2004). “Clustering large graphs via the singular value decomposition”. In: *Machine learning* 56, pp. 9–33.

Lloyd, Stuart (1982). “Least squares quantization in PCM”. In: *IEEE transactions on information theory* 28.2, pp. 129–137.

MacQueen, J (1967). “Classification and analysis of multivariate observations”. In: *5th Berkeley Symp. Math. Statist. Probability*. University of California Los Angeles LA USA, pp. 281–297.

Maranzana, Francesco E. (1963). “On the location of supply points to minimize transportation costs”. In: *IBM Systems Journal* 2.2, pp. 129–135.

References II

Nielsen, Frank and Frank Nielsen (2016). "Hierarchical clustering". In: *Introduction to HPC with MPI for Data Science*, pp. 195–211.

Ostrovsky, Rafail et al. (2013). "The effectiveness of Lloyd-type methods for the k-means problem". In: *Journal of the ACM (JACM)* 59.6, pp. 1–22.

Sculley, David (2010). "Web-scale k-means clustering". In: *Proceedings of the 19th international conference on World wide web*, pp. 1177–1178.

Von Luxburg, Ulrike (2007). "A tutorial on spectral clustering". In: *Statistics and computing* 17, pp. 395–416.

Voronoi, Georges (1908). "Nouvelles applications des paramètres continus à la théorie des formes quadratiques. Premier mémoire. Sur quelques propriétés des formes quadratiques positives parfaites.". In: *Journal für die reine und angewandte Mathematik (Crelles Journal)* 1908.133, pp. 97–102.