Big Data Analytics

Jour 3 — Algorithmes non supervisés

François-Marie Giraud

https://www.orsys.fr/

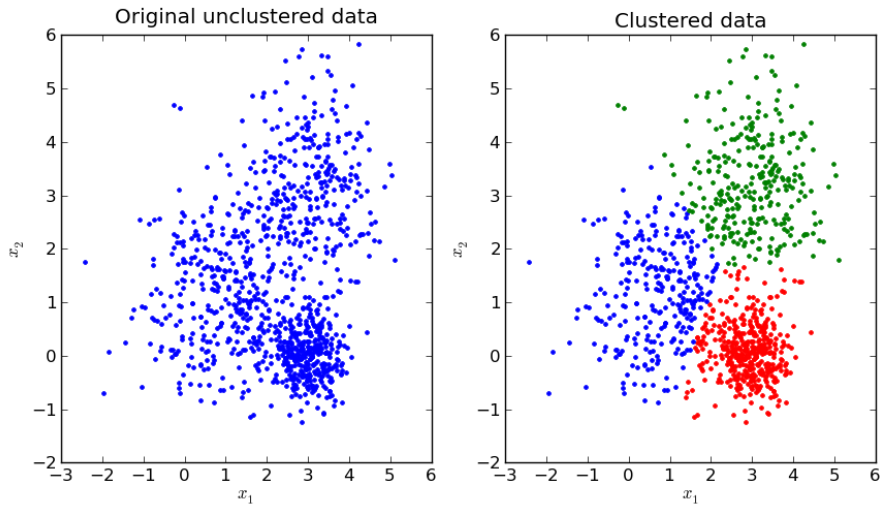
Apprentissage non-supervisé

Apprentissage non-supervisé

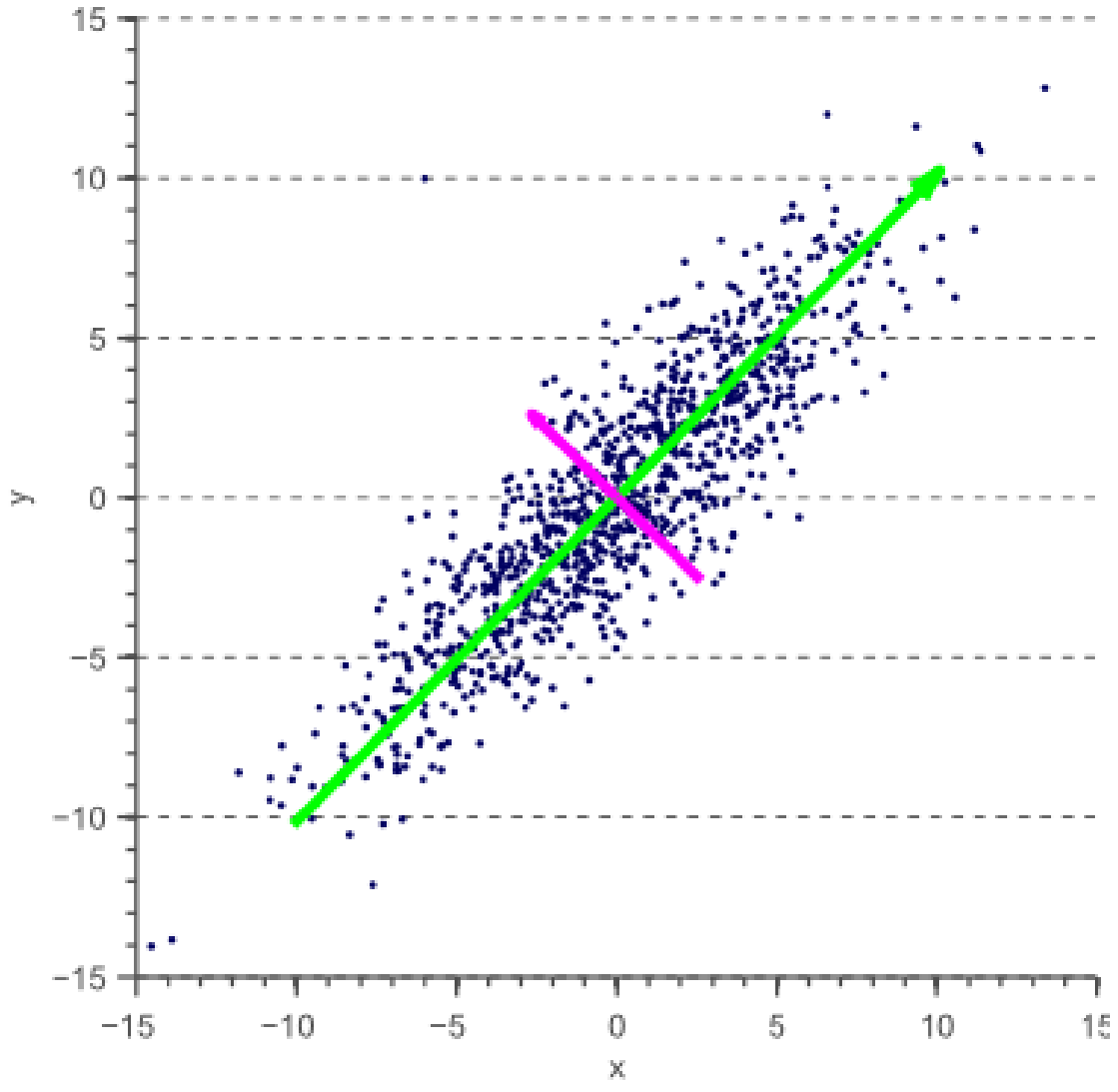
Introduction

But

Détection de variables cachées, de « structures » cachées (clusters, variété topologique (manifold), ...)



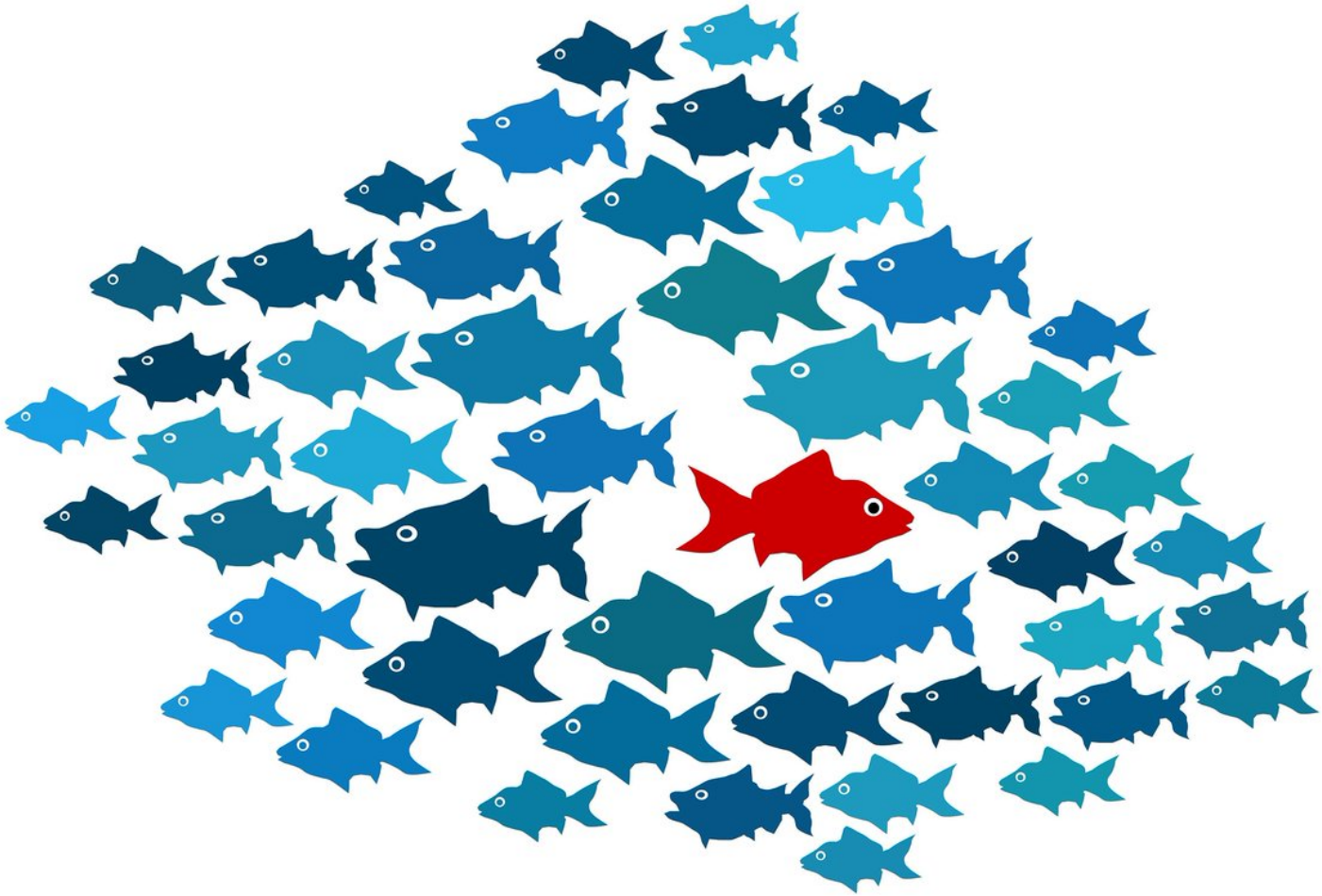
Réduction de Dimensionnalité



Clustering

Par exemple : Création de segments de clientèle

Détection d'anomalie



Apprentissage non-supervisé

Réduction de la dimensionnalité

Problématique

Comment appréhender des données en grande dimension ?

$$X = \begin{bmatrix} X_{1,1} & X_{1,2} & \dots & X_{1,D} \\ X_{2,1} & X_{2,2} & \dots & X_{2,D} \\ \vdots & \vdots & \ddots & \vdots \\ X_{N,1} & X_{N,2} & \dots & X_{N,D} \end{bmatrix}$$

La malédiction des grandes dimensions

Nombre d'extrémités dans une espace de dimension :

Dimensions	Points d'un hypercube
1	2
2	4
3	8
4	16
5	32

Besoin en données

Corrélé exponentiellement à la taille des espaces.

Approches

- Séléction de dimensions
- Projections linéaires (ACP, LDA, ...)
- Projections non-linéaires (t-SNE, UMAP, kernels, embeddings, ...)

Projections linéaires

Deux méthodes principales :

- Principal Component Analysis (Non-supervisée)
- Linear Discriminant Analysis (Supervisée)

PCA — Caractéristiques

- Crée de nouvelles features (combinaisons linéaires des originales)
- Ces features sont décorrélées
- Les premières sont les plus informatives
- Pour réduire la dimensionnalité : ne conserver que les k premières

PCA — Caractéristiques

- Crée de nouvelles features (combinaisons linéaires des originales)
- Ces features sont décorrélées
- Les premières sont les plus informatives
- Pour réduire la dimensionnalité : ne conserver que les k premières

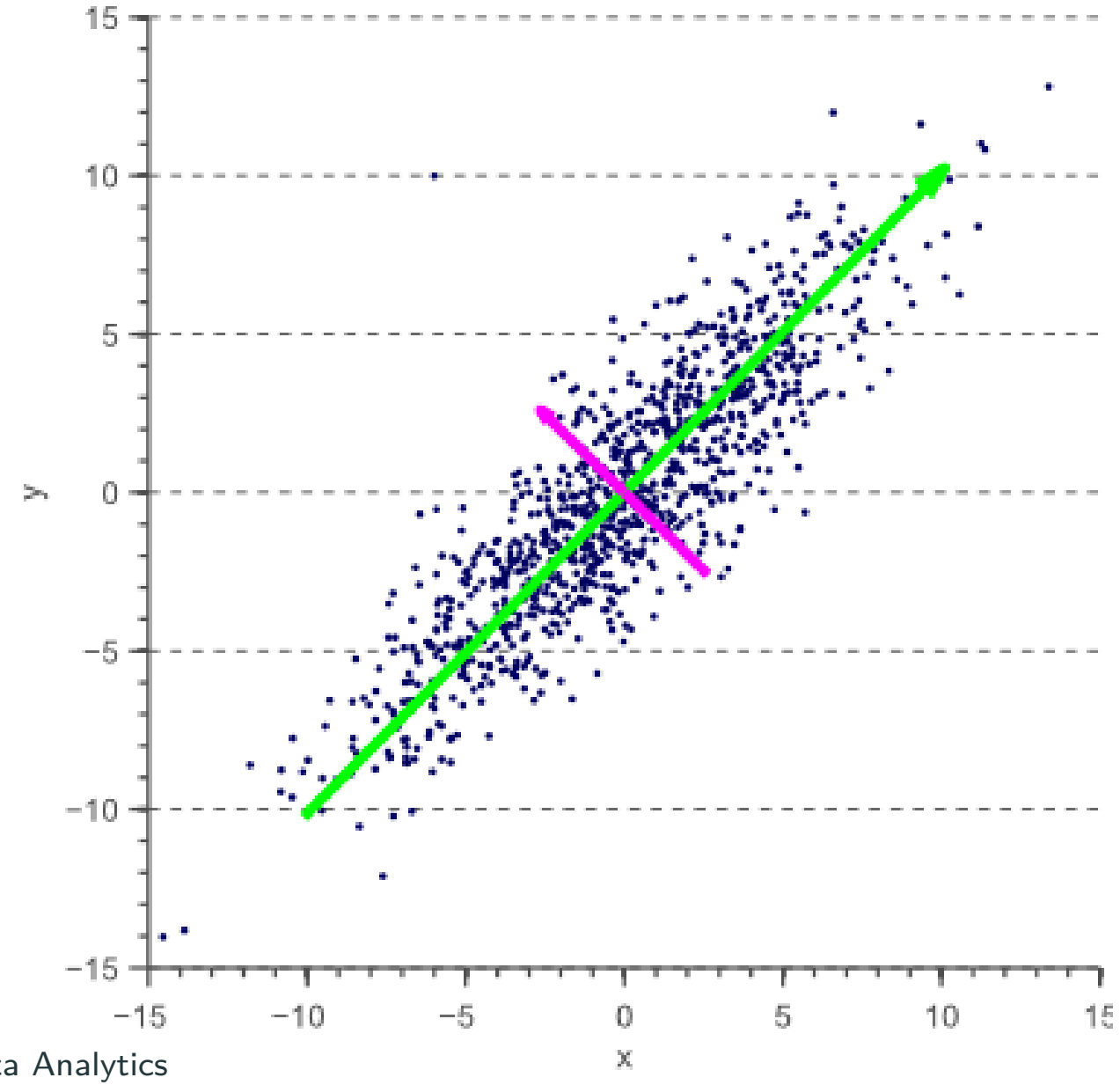
PCA — Caractéristiques

- Crée de nouvelles features (combinaisons linéaires des originales)
- Ces features sont décorrélées
- Les premières sont les plus informatives
- Pour réduire la dimensionnalité : ne conserver que les k premières

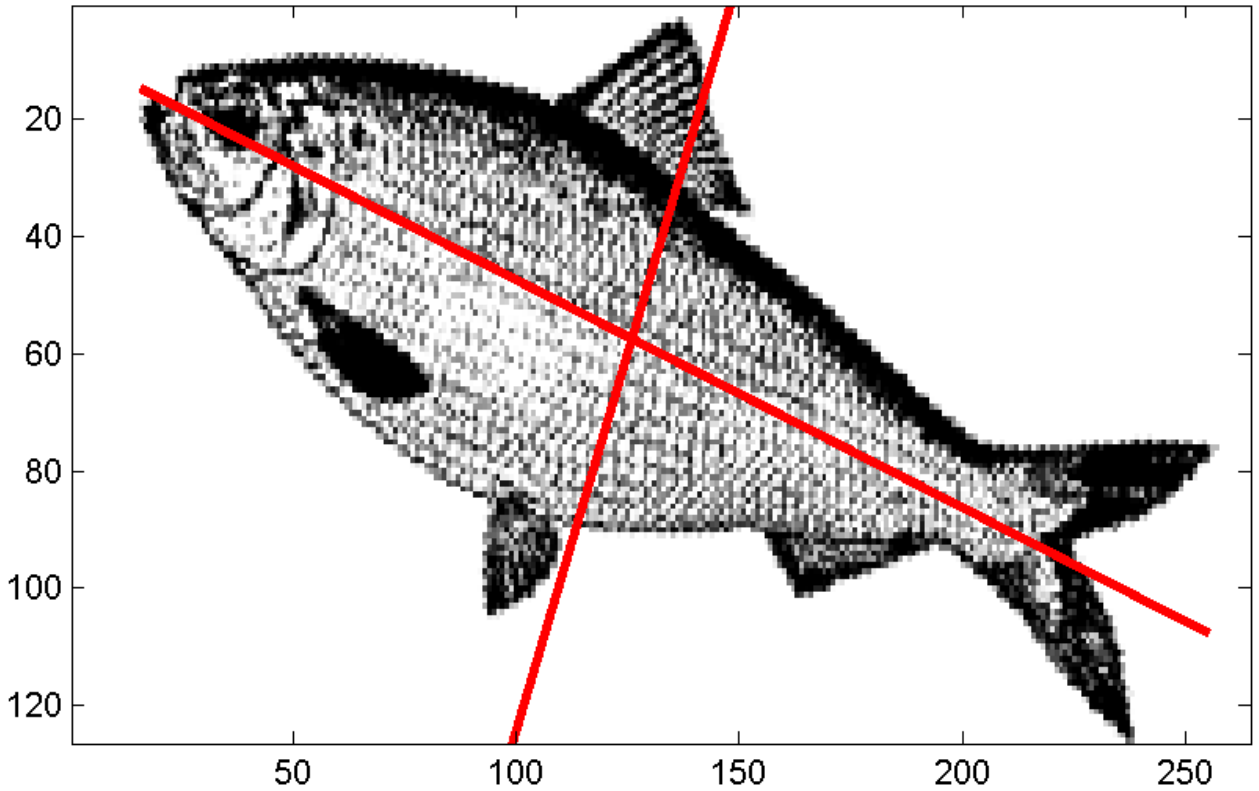
PCA — Caractéristiques

- Crée de nouvelles features (combinaisons linéaires des originales)
- Ces features sont décorrélées
- Les premières sont les plus informatives
- Pour réduire la dimensionnalité : ne conserver que les k premières

PCA — Exemple 1



PCA — Exemple 2



PCA — Point de départ

$$X = \begin{bmatrix} X_{1,1} & \cdots & X_{1,D} \\ \vdots & \ddots & \vdots \\ X_{N,1} & \cdots & X_{N,D} \end{bmatrix}$$

PCA — Procédé

Chaque dimension est centrée :

$$\overline{X} = \begin{bmatrix} X_{1,1} - \overline{X_1} & \dots & X_{1,D} - \overline{X_D} \\ \vdots & \ddots & \vdots \\ X_{N,1} - \overline{X_1} & \dots & X_{N,D} - \overline{X_D} \end{bmatrix}$$

puis optionnellement réduite :

$$\widetilde{X} = \begin{bmatrix} \frac{X_{1,1} - \overline{X_1}}{\sigma(X_1)} & \dots & \frac{X_{1,D} - \overline{X_D}}{\sigma(X_D)} \\ \vdots & \ddots & \vdots \\ \frac{X_{N,1} - \overline{X_1}}{\sigma(X_1)} & \dots & \frac{X_{N,D} - \overline{X_D}}{\sigma(X_D)} \end{bmatrix}$$

PCA — Procédé (suite)

- Calcul de la matrice de covariance (resp corrélation) :

$$\frac{1}{N} \times \overline{X^T} \times \overline{X} \quad (\frac{1}{N} \times \widetilde{X^T} \times \widetilde{X})$$

- Calcul des vecteurs propres de cette matrice : ce sont les combinaisons linéaires des nouvelles features
- Tri des vecteurs par valeur propre décroissante
- Réduction de dimension : on ne conserve que les k premiers vecteurs propres

PCA — Procédé (suite)

- Calcul de la matrice de covariance (resp corrélation) :

$$\frac{1}{N} \times \overline{X^T} \times \overline{X} \quad (\frac{1}{N} \times \widetilde{X^T} \times \widetilde{X})$$

- Calcul des vecteurs propres de cette matrice : ce sont les combinaisons linéaires des nouvelles features
- Tri des vecteurs par valeur propre décroissante
- Réduction de dimension : on ne conserve que les k premiers vecteurs propres

PCA — Procédé (suite)

- Calcul de la matrice de covariance (resp corrélation) :

$$\frac{1}{N} \times \overline{X^T} \times \overline{X} \quad (\frac{1}{N} \times \widetilde{X^T} \times \widetilde{X})$$

- Calcul des vecteurs propres de cette matrice : ce sont les combinaisons linéaires des nouvelles features
- Tri des vecteurs par valeur propre décroissante
- Réduction de dimension : on ne conserve que les k premiers vecteurs propres

PCA — Procédé (suite)

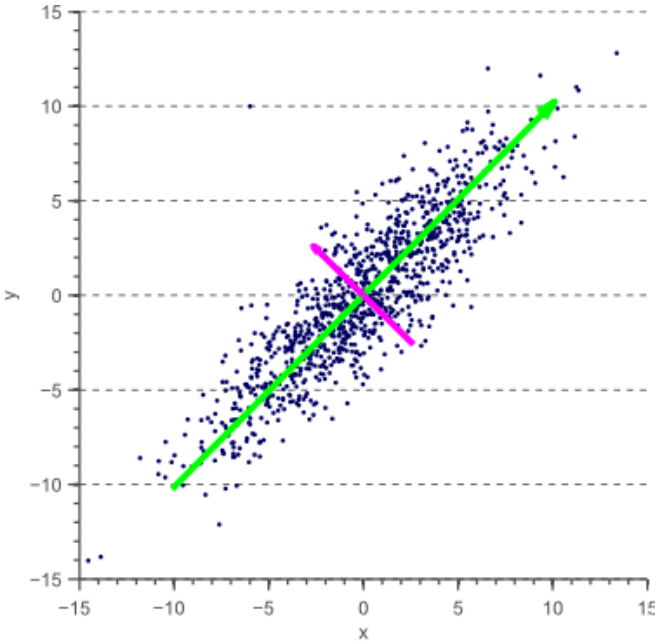
- Calcul de la matrice de covariance (resp corrélation) :

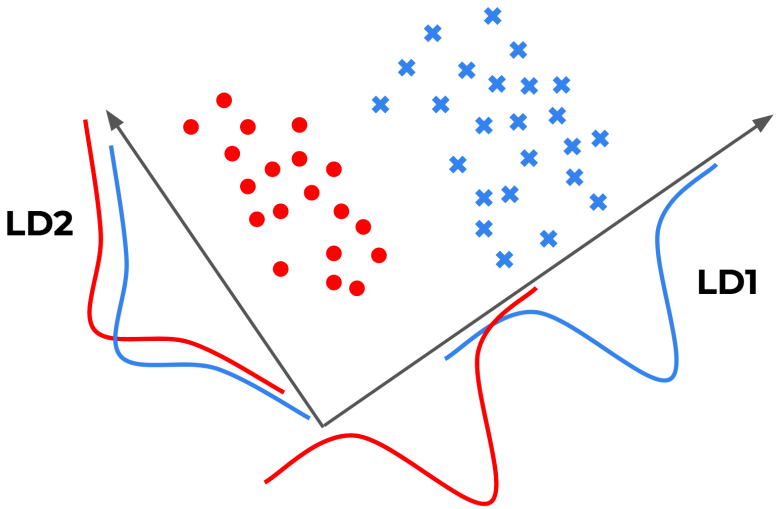
$$\frac{1}{N} \times \overline{X^T} \times \overline{X} \quad (\frac{1}{N} \times \widetilde{X^T} \times \widetilde{X})$$

- Calcul des vecteurs propres de cette matrice : ce sont les combinaisons linéaires des nouvelles features
- Tri des vecteurs par valeur propre décroissante
- Réduction de dimension : on ne conserve que les k premiers vecteurs propres

PCA — Linear Discriminant Analysis

Maximisation de la variance inter-classes.





Analyse des Correspondances Multiples (ACM)

ACP sur des données qualitatives (Ex : enquètes d'opinions avec QCM)
Chaque variable qualitative est transformé en vecteur sparse.
On obtient une matrice binaire sur laquelle on procède à l'ACP.

Analyse Factorielle pour données mixtes (AFDM)

Quand on a des variables qualitative ET quantitatives pour décrires nos échantillons, on discrétise chaque variable quantitative. On peut ainsi procéder à l'Analyse en Composantes Multiples

Analyse Factorielle des Correspondances (AFC)

Méthode sur un tableau de contingence :

Yaourts	Nantes	Bordeaux	Limoges	Tours	Poitiers	TOTAL
Ananas	14	15	9	20	20	78
Banane	15	10	14	20	21	80
Fraise	16	16	26	8	22	88
Framboise	18	14	24	20	17	93
Abricot	17	18	20	22	16	93
TOTAL	80	73	93	90	96	432

On procède alors à une double ACP (une sur le profil ligne, l'autre sur le profil colonne) en utilisant une métrique particulière : le χ^2

Projections non linéaires

Plus intéressant pour découvrir des patterns en haute dimension :

- t-SNE
- UMAP

Projections non linéaires

Plus intéressant pour découvrir des patterns en haute dimension :

- t-SNE
- UMAP

Méthode

Algorithmes de disposition de graphe.

Minimisation de la différence topologique entre deux graphes :

- Graphe construit en haute dimension
- Graphe construit dans la dimension de projection

Méthode

Algorithmes de disposition de graphe.

Minimisation de la différence topologique entre deux graphes :

- Graphe construit en haute dimension
- Graphe construit dans la dimension de projection

Paramètres

Dans les deux algorithmes, on peut contrôler le trade-off structure local / structure globale :

- La perplexité de t-SNE $\approx$ nombre de voisins considérés comme connectés
- UMAP utilise explicitement un argument de nombre de voisins connectés

Paramètres

Dans les deux algorithmes, on peut contrôler le trade-off structure local / structure globale :

- La perplexité de t-SNE $\approx$ nombre de voisins considérés comme connectés
- UMAP utilise explicitement un argument de nombre de voisins connectés

t-SNE — Bonnes pratiques & Analyse

Article t-SNE

UMAP — Bonnes pratiques & Analyse

Article UMAP

Support TP : PCA/LDA

- PCA sur le dataset Iris — Tutoriel
- PCA & LDA sur le dataset Iris — Tutoriel

Avez-vous des questions ?

Apprentissage non-supervisé

Clustering

But

- Regrouper des instances du training set en « clusters »
- Parfois hiérarchiquement
- Parfois en décidant du nombre de clusters (semi-)automatiquement

But

- Regrouper des instances du training set en « clusters »
- Parfois hiérarchiquement
- Parfois en décidant du nombre de clusters (semi-)automatiquement

But

- Regrouper des instances du training set en « clusters »
- Parfois hiérarchiquement
- Parfois en décidant du nombre de clusters (semi-)automatiquement

Approches

On distingue deux grandes familles :

- K-means et ses algorithmes dérivés
- Les algorithmes agglomératifs

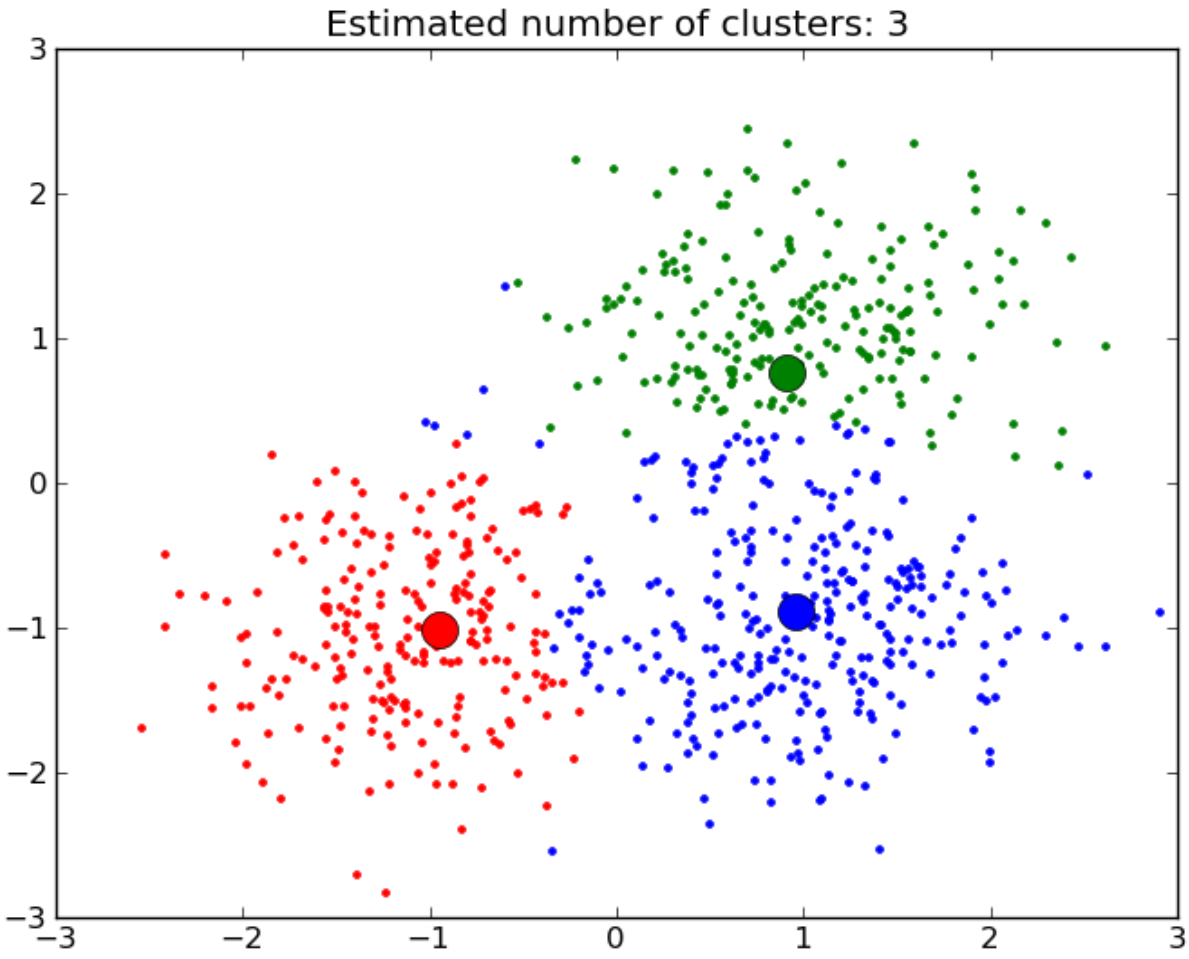
Approches

On distingue deux grandes familles :

- K-means et ses algorithmes dérivés
- Les algorithmes agglomératifs

K-Means — Introduction

Algorithme EM



K-Means — Algorithme

- Choisir k points comme clusters initiaux
- Itérer les phases E (expectation) et M (maximisation) :

  E Attribuer à chaque point le cluster dont le centroïde est le plus proche

  M Réajuster les centroïdes pour qu'ils soient au milieu des points qui leur est attribué

K-Means — Algorithme

- Choisir k points comme clusters initiaux
- Itérer les phases E (expectation) et M (maximisation) :
  - **E** Attribuer à chaque point le cluster dont le centroïde est le plus proche
  - **M** Réajuster les centroïdes pour qu'ils soient au milieu des points qui leur est attribué

K-Means — Algorithme

- Choisir k points comme clusters initiaux
- Itérer les phases E (expectation) et M (maximisation) :

E Attribuer à chaque point le cluster dont le centroïde est le plus proche

M Réajuster les centroïdes pour qu'ils soient au milieu des points qui leur est attribué

K-Means — Algorithme

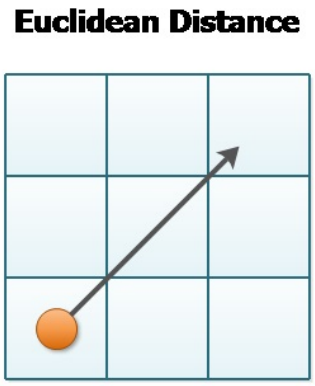
- Choisir k points comme clusters initiaux
- Itérer les phases E (expectation) et M (maximisation) :
  - **E** Attribuer à chaque point le cluster dont le centroïde est le plus proche
  - **M** Réajuster les centroïdes pour qu'ils soient au milieu des points qui leur est attribué

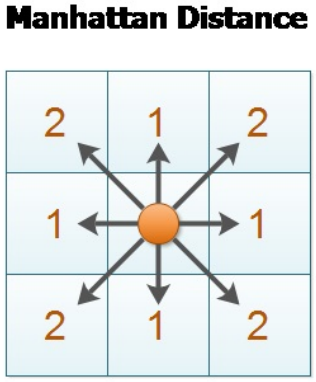
K-Means — Exemple

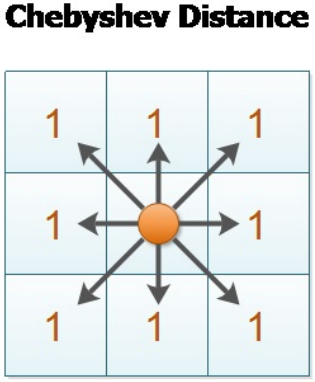
Vidéo de plusieurs apprentissages

K-Means — Distance

distance euclidienne, Manhattan, Chebychev







$\sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2}$ $|x_1 - x_2| + |y_1 - y_2|$ $\max(|x_1 - x_2|, |y_1 - y_2|)$

K-Means++

Améliore l'initialisation :

- Choisir un point comme cluster initial
- Pour les points suivants, itérer :
  - Calculer la distance de chaque point à son cluster le plus proche
  - Choisir un point avec une probabilité proportionnelle à la distance calculée

→ Bien meilleures propriétés de convergence !

K-Means++

Améliore l'initialisation :

- Choisir un point comme cluster initial
- Pour les points suivants, itérer :
  - Calculer la distance de chaque point à son cluster le plus proche
  - Choisir un point avec une probabilité proportionnelle à la distance calculée

→ Bien meilleures propriétés de convergence !

K-Means++

Améliore l'initialisation :

- Choisir un point comme cluster initial
- Pour les points suivants, itérer :
  - Calculer la distance de chaque point à son cluster le plus proche
  - Choisir un point avec une probabilité proportionnelle à la distance calculée

→ Bien meilleures propriétés de convergence !

K-Means++

Améliore l'initialisation :

- Choisir un point comme cluster initial
- Pour les points suivants, itérer :
  - Calculer la distance de chaque point à son cluster le plus proche
  - Choisir un point avec une probabilité proportionnelle à la distance calculée

→ Bien meilleures propriétés de convergence !

K-Means++

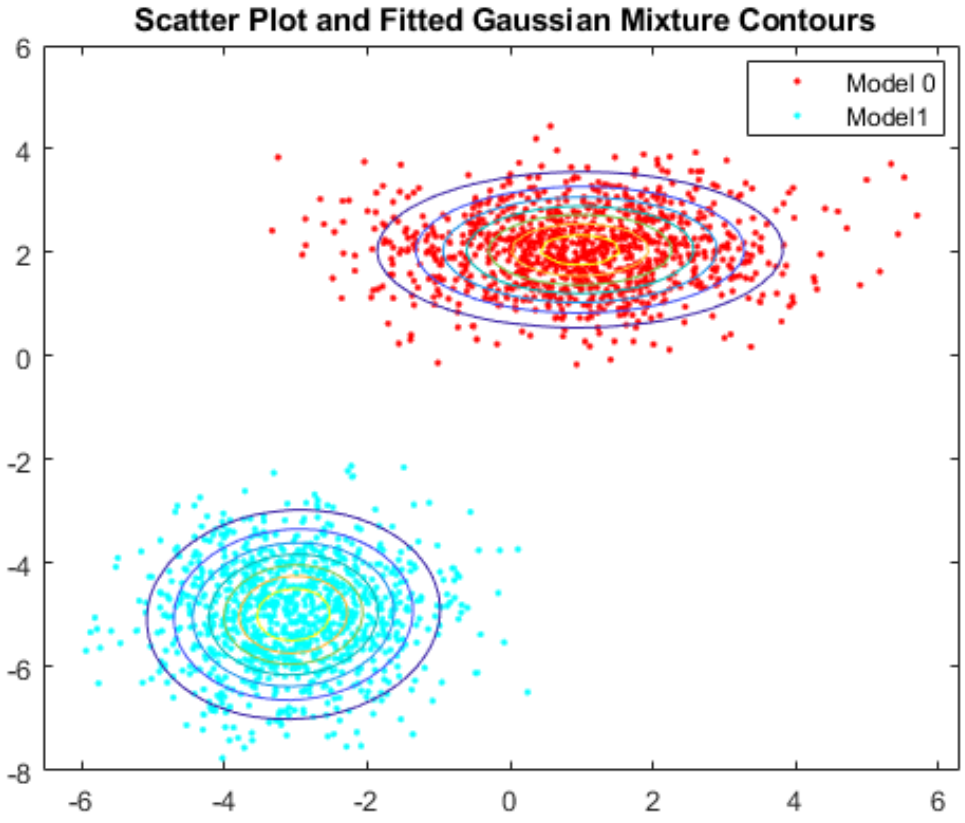
Améliore l'initialisation :

- Choisir un point comme cluster initial
- Pour les points suivants, itérer :
  - Calculer la distance de chaque point à son cluster le plus proche
  - Choisir un point avec une probabilité proportionnelle à la distance calculée

$\rightarrow$ Bien meilleures propriétés de convergence !

Gaussian Mixture Model — Introduction

Les clusters sont représentés par un centre et une matrice de covariance.



Gaussian Mixture Model — Caractéristiques

- « Soft » clustering : chaque cluster est une fonction de densité de probabilité
- Algorithme EM, souvent initialisé avec K-Means

Gaussian Mixture Model — Caractéristiques

- « Soft » clustering : chaque cluster est une fonction de densité de probabilité
- Algorithme EM, souvent initialisé avec K-Means

DBSCAN — Introduction

Density-Based Spatial Clustering of Applications with Noise...

- Gère le bruit (points sans clusters)
- Fonctionne bien dans la pratique
- Ne nécessite pas de nombre de clusters prédéfini

DBSCAN — Introduction

Density-Based Spatial Clustering of Applications with Noise...

- Gère le bruit (points sans clusters)
- Fonctionne bien dans la pratique
- Ne nécessite pas de nombre de clusters prédéfini

DBSCAN — Introduction

Density-Based Spatial Clustering of Applications with Noise...

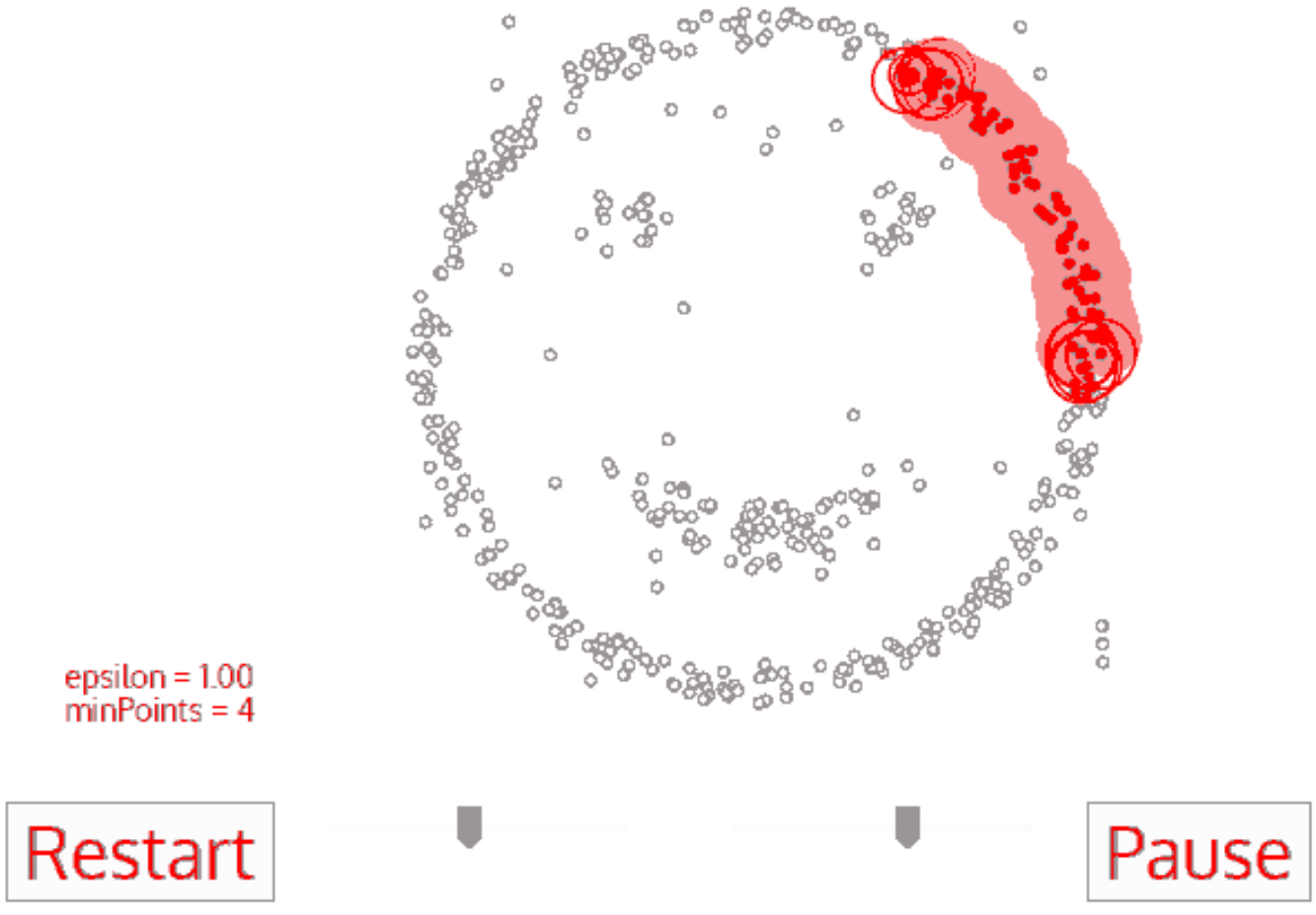
- Gère le bruit (points sans clusters)
- Fonctionne bien dans la pratique
- Ne nécessite pas de nombre de clusters prédéfini

DBSCAN — Procédé

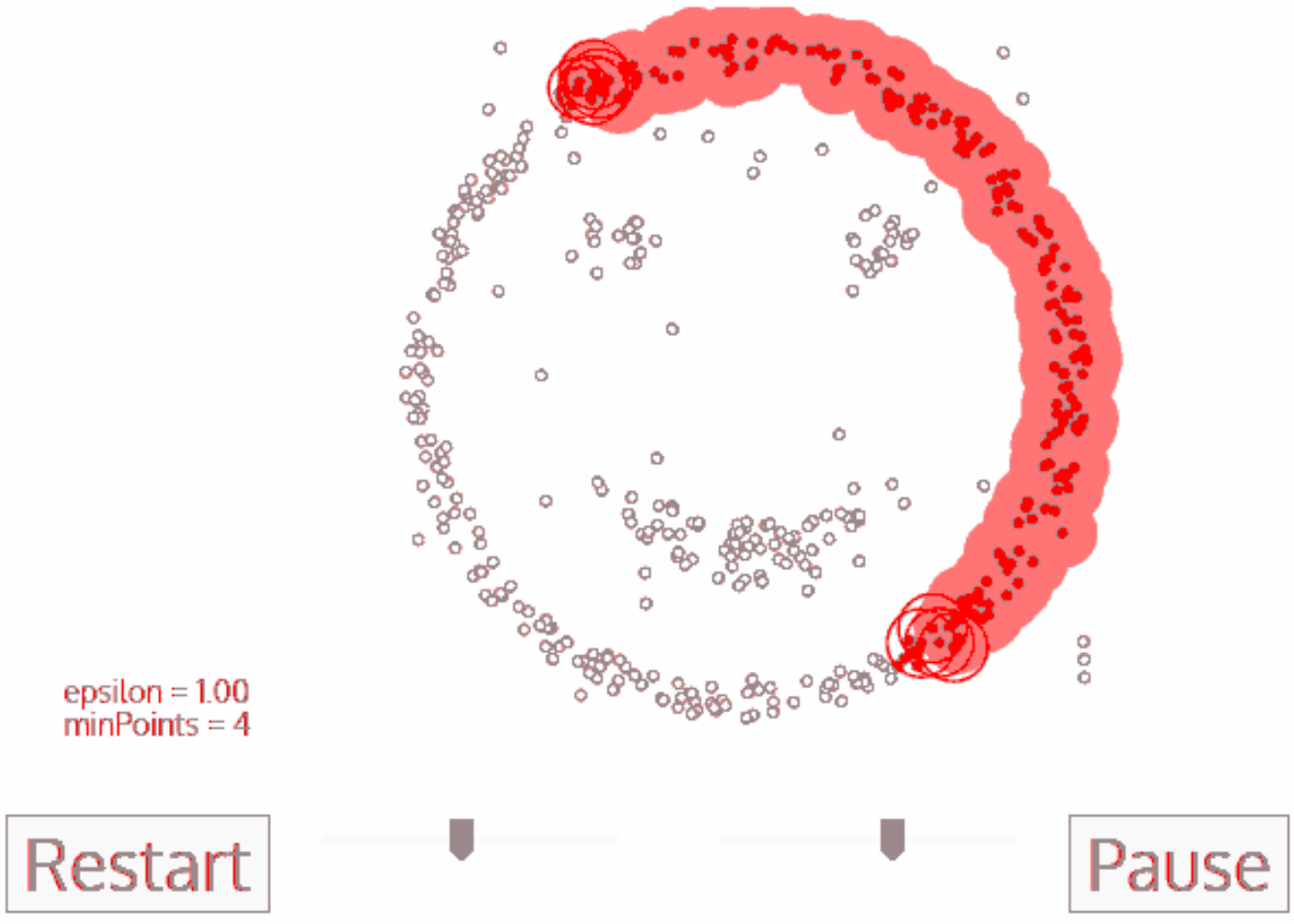
Tant qu'il reste des points non étiquetés :

1. Définir une distance ϵ et un nombre minimum de voisins v
2. Prendre un point non-étiqueté au hasard et regarder son ϵ-voisinage
3. Si il a v voisins ou plus on crée un nouveau cluster
   3.1 Expansion du cluster de proche en proche dans le voisinage
4. Sinon (Bruit)

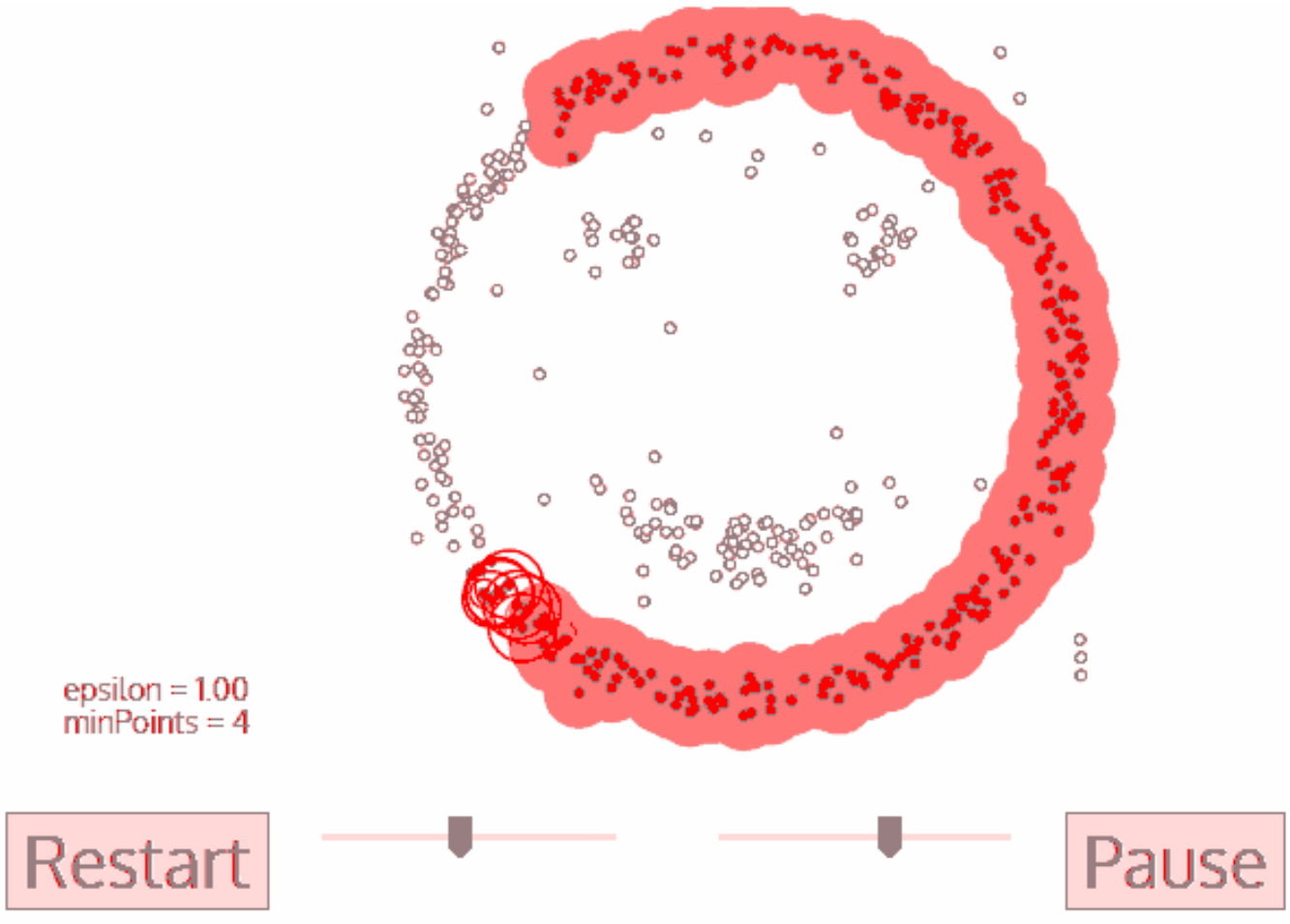
DBSCAN — Démonstration



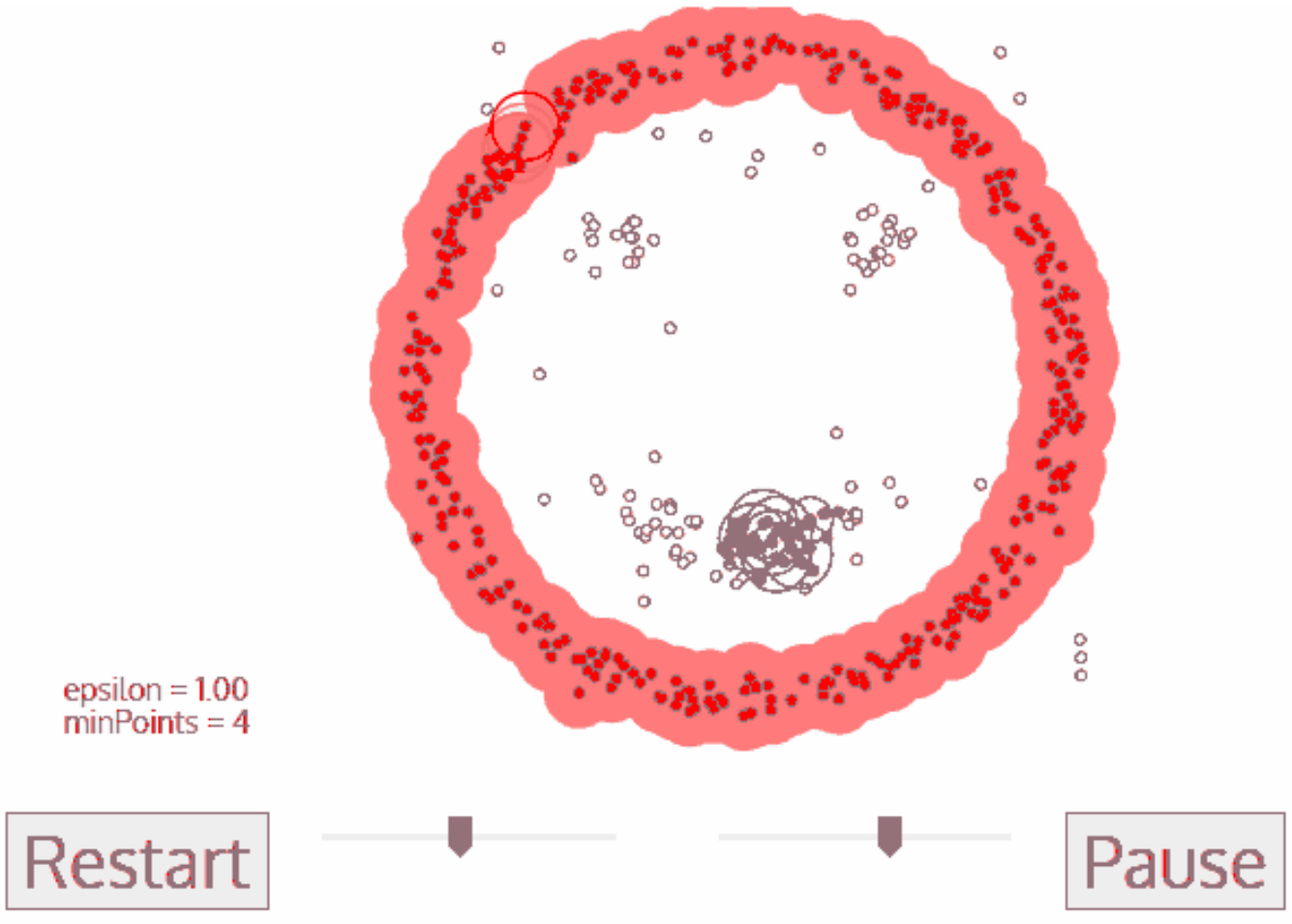
DBSCAN — Démonstration



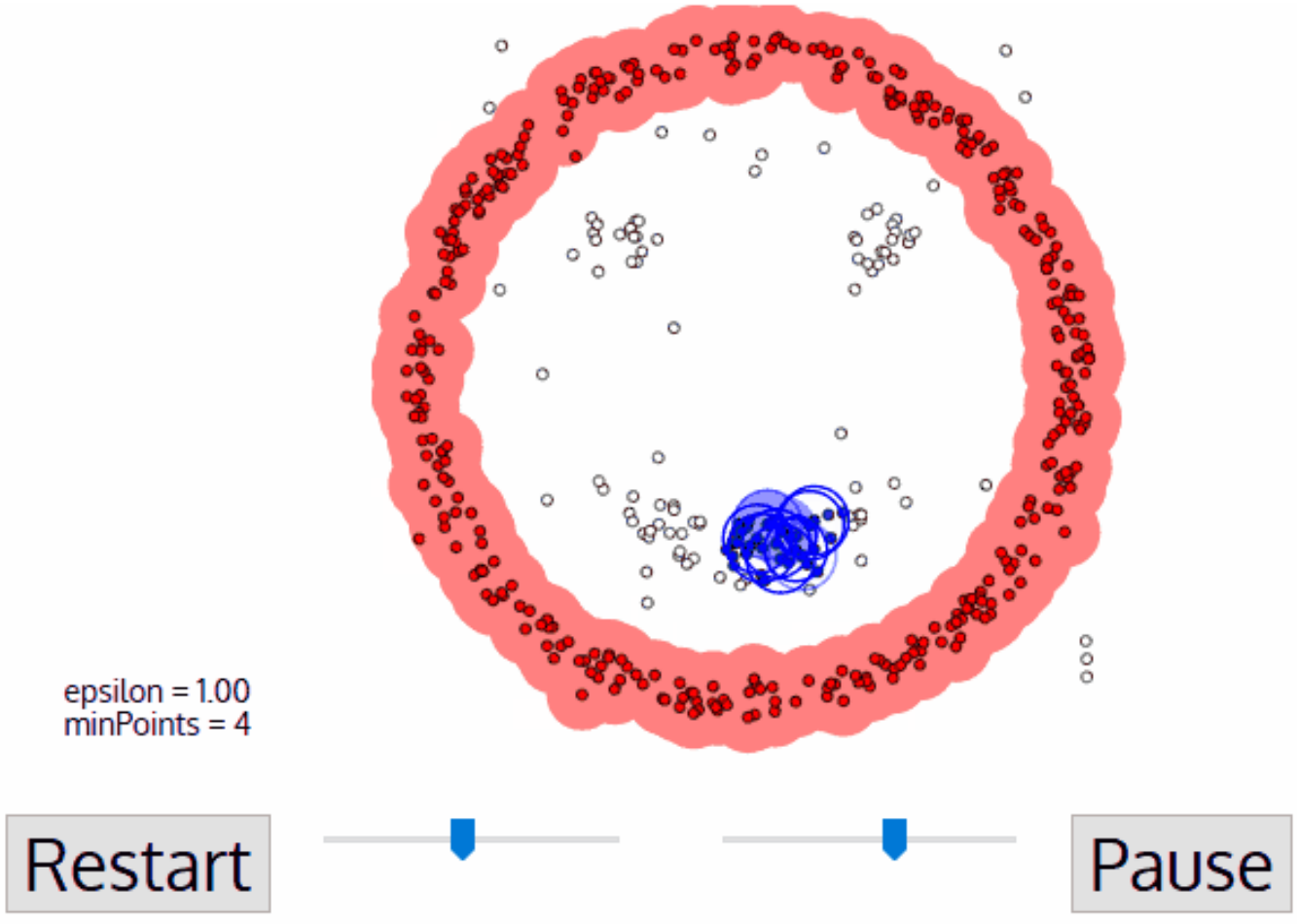
DBSCAN — Démonstration



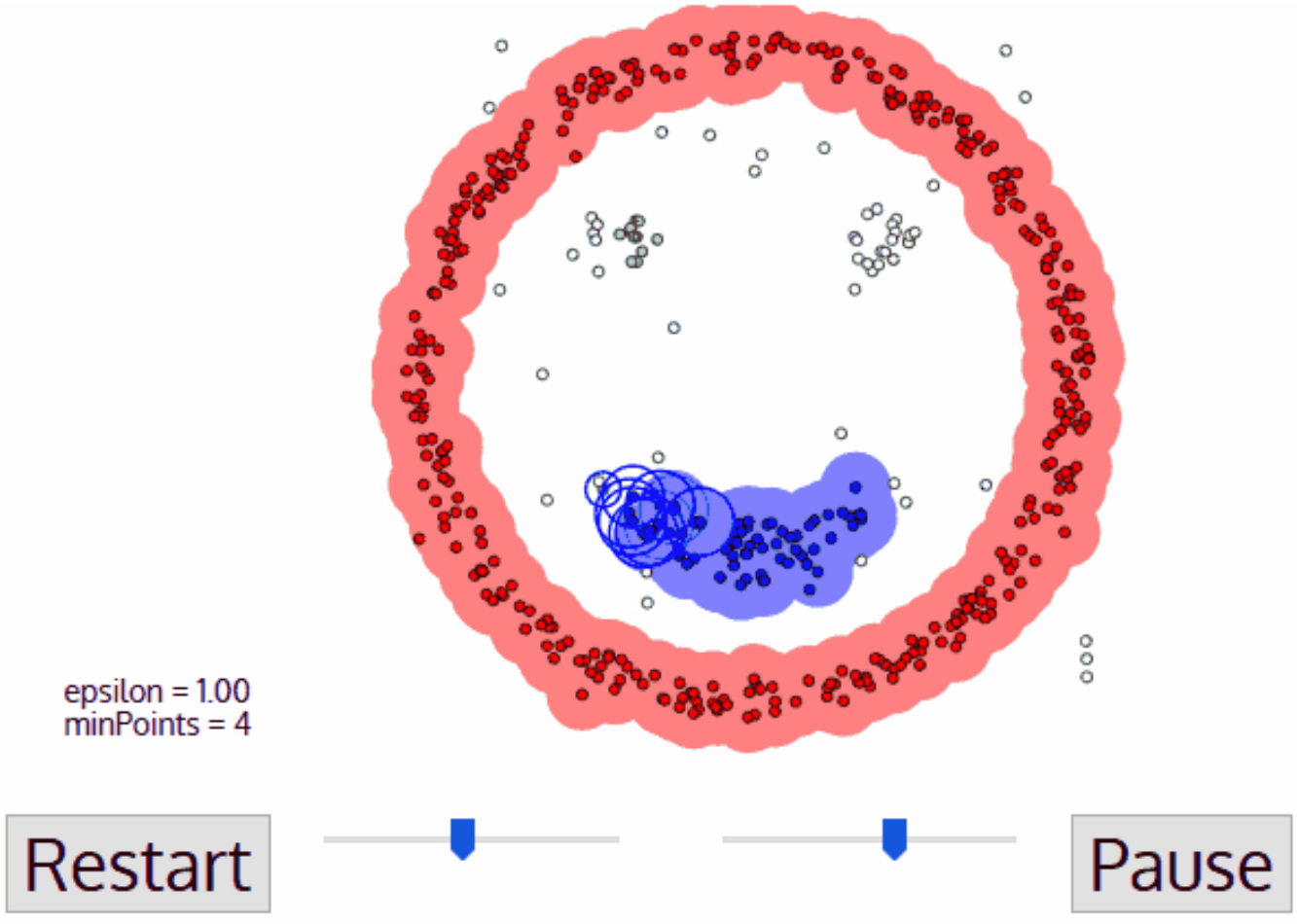
DBSCAN — Démonstration



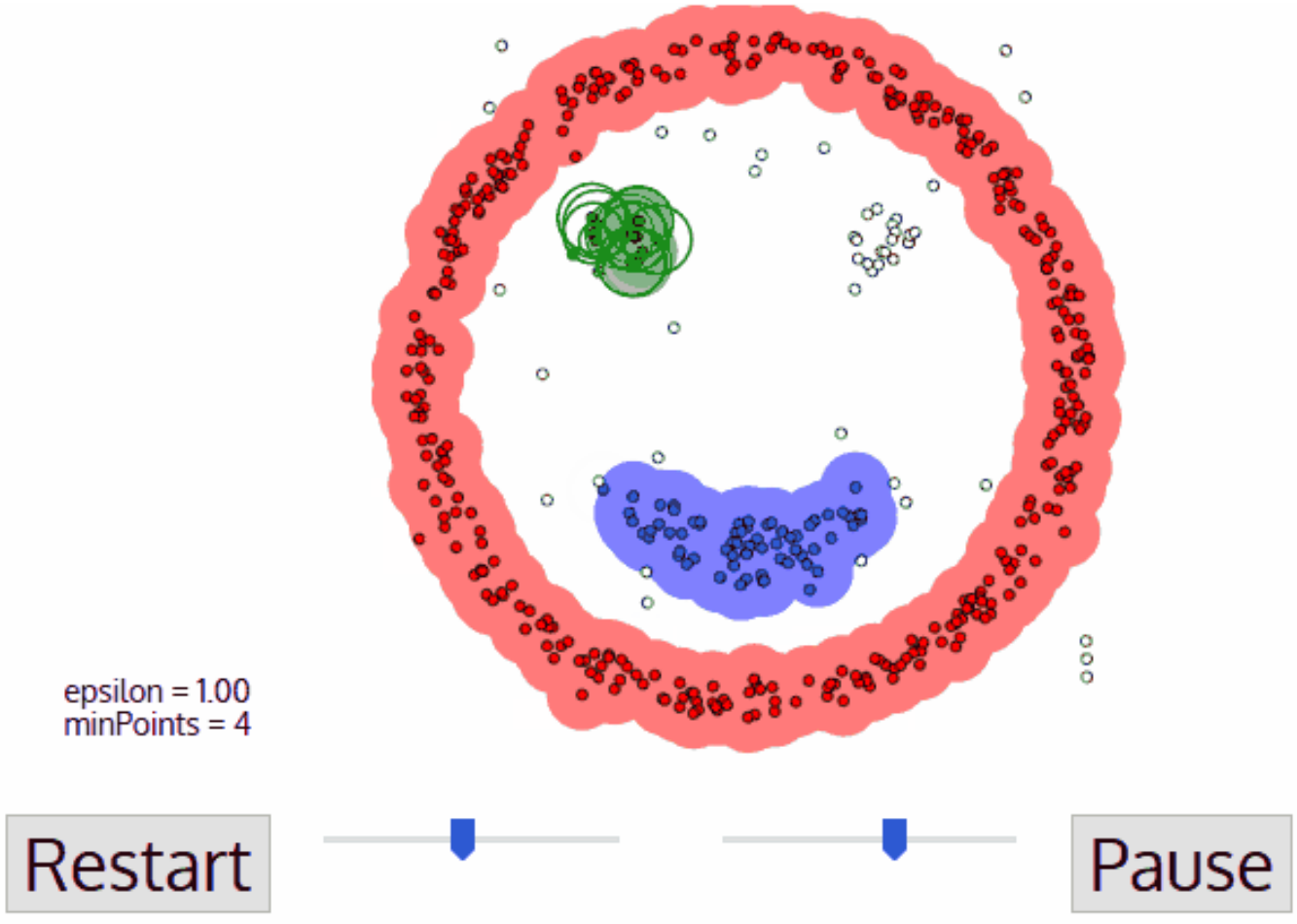
DBSCAN — Démonstration



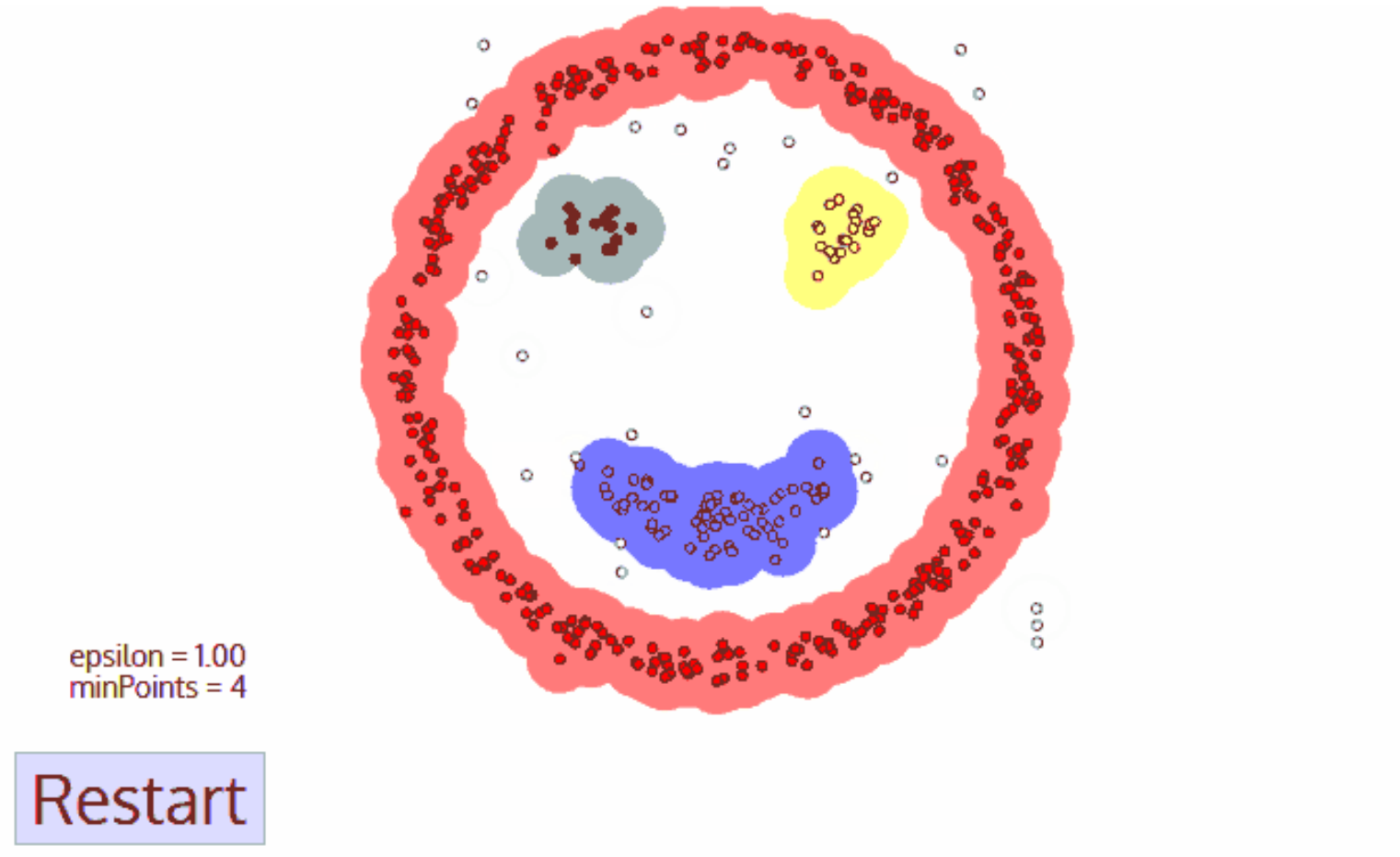
DBSCAN — Démonstration



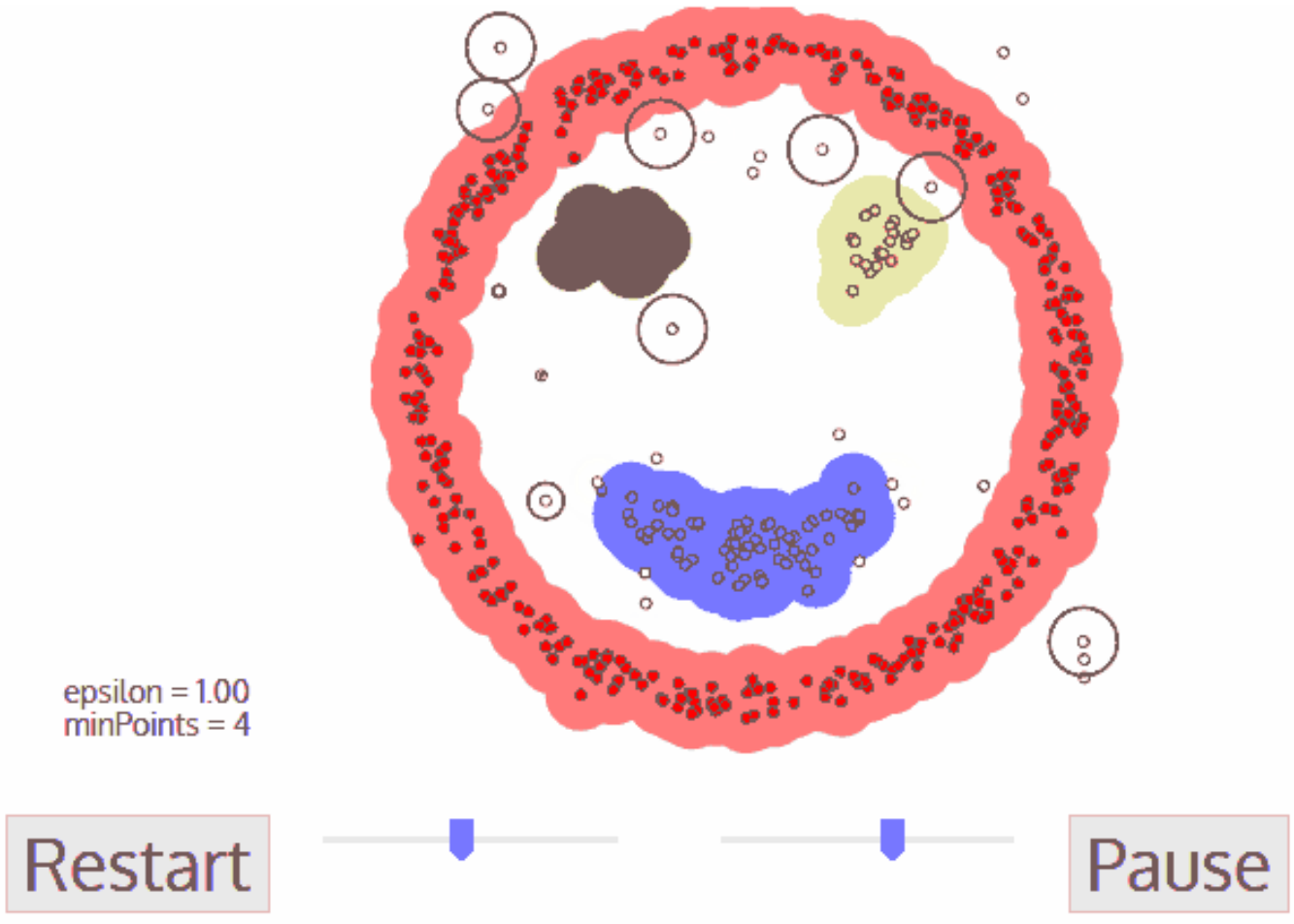
DBSCAN — Démonstration



DBSCAN — Démonstration



DBSCAN — Démonstration



Clustering Hiérarchique

Deux approches :

- Agglomératives (bottom-up)
- Divisantes (top-down), moins utilisées

Clustering Hiérarchique

Deux approches :

- Agglomératives (bottom-up)
- Divisantes (top-down), moins utilisées

Clustering Hiérarchique — Procédé

Méthode Agglomérative

Initialisation Chaque élément est dans une classe distincte

Aggrégation Itérativement, fusion deux par deux les classes les plus similaires

Exploitation On choisit le nombre de clusters à exploiter

Clustering Hiérarchique — Procédé

Méthode Agglomérative

Initialisation Chaque élément est dans une classe distincte

Aggrégation Itérativement, fusion deux par deux les classes les plus similaires

Exploitation On choisit le nombre de clusters à exploiter

Clustering Hiérarchique — Procédé

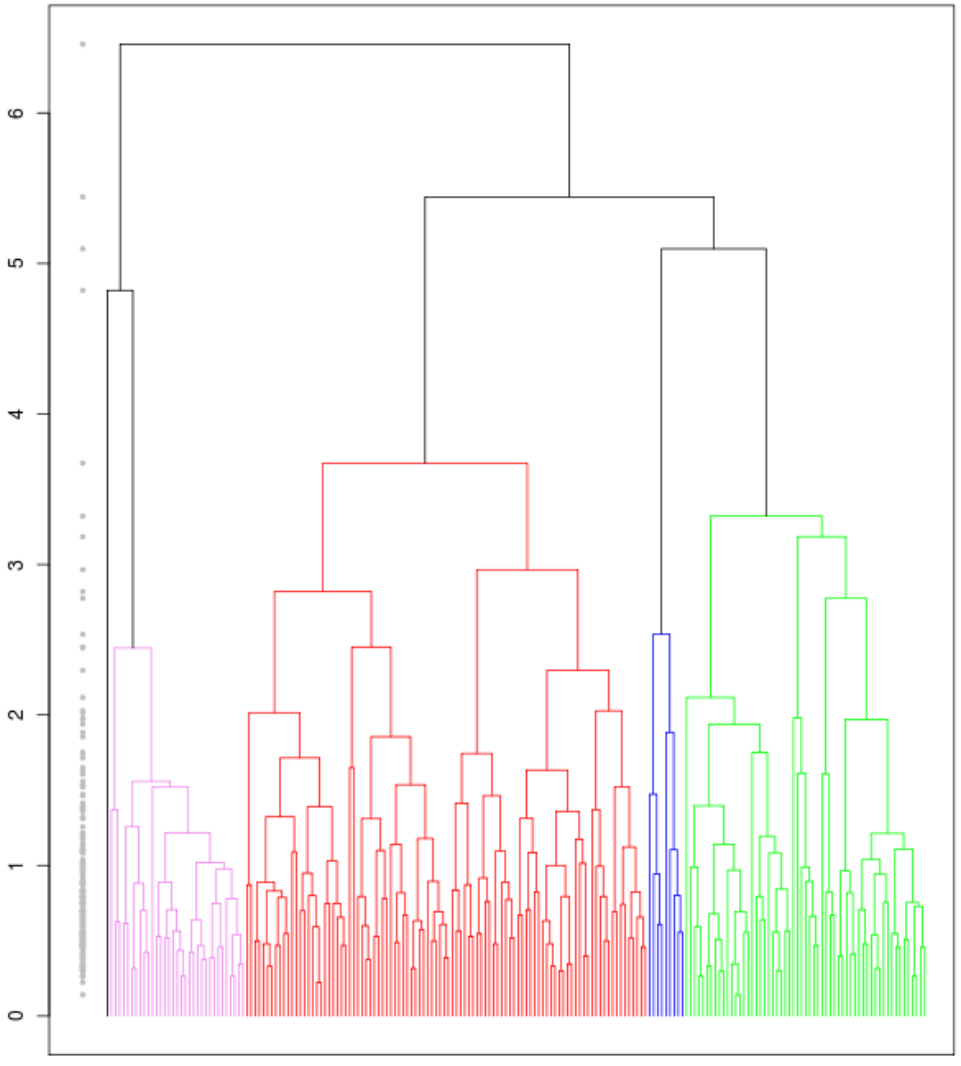
Méthode Agglomérative

Initialisation Chaque élément est dans une classe distincte

Aggrégation Itérativement, fusion deux par deux les classes les plus similaires

Exploitation On choisit le nombre de clusters à exploiter

Clustering Hiérarchique — Résultat



Clustering Hiérarchique — Procédé

Pour décider des clusters les plus similaires (à fusionner donc), plusieurs options :

Saut minimum $\min_{x \in C_1, y \in C_2} D(x, y)$

Saut maximum $\max_{x \in C_1, y \in C_2} D(x, y)$

Saut moyen $\text{mean}_{x \in C_1, y \in C_2} D(x, y)$

Méthode de Ward Maximise l'inertie inter-classe

...

Méthode exacte : $O(n^2) <$ complexité $< O(n^3)$!

Clustering Hiérarchique — Procédé

Pour décider des clusters les plus similaires (à fusionner donc), plusieurs options :

Saut minimum $\min_{x \in C_1, y \in C_2} D(x, y)$

Saut maximum $\max_{x \in C_1, y \in C_2} D(x, y)$

Saut moyen $\text{mean}_{x \in C_1, y \in C_2} D(x, y)$

Méthode de Ward Maximise l'inertie inter-classe

...

Méthode exacte : $O(n^2) <$ complexité $< O(n^3)$!

Clustering Hiérarchique — Procédé

Pour décider des clusters les plus similaires (à fusionner donc), plusieurs options :

Saut minimum $\min_{x \in C_1, y \in C_2} D(x, y)$

Saut maximum $\max_{x \in C_1, y \in C_2} D(x, y)$

Saut moyen $\text{mean}_{x \in C_1, y \in C_2} D(x, y)$

Méthode de Ward Maximise l'inertie inter-classe

...

Méthode exacte : $O(n^2) <$ complexité $< O(n^3)$!

Clustering Hiérarchique — Procédé

Pour décider des clusters les plus similaires (à fusionner donc), plusieurs options :

Saut minimum $\min_{x \in C_1, y \in C_2} D(x, y)$

Saut maximum $\max_{x \in C_1, y \in C_2} D(x, y)$

Saut moyen $\text{mean}_{x \in C_1, y \in C_2} D(x, y)$

Méthode de Ward Maximise l'inertie inter-classe

...

Méthode exacte : $O(n^2) <$ complexité $< O(n^3)$!

Clustering Hiérarchique — Procédé

Pour décider des clusters les plus similaires (à fusionner donc), plusieurs options :

Saut minimum $\min_{x \in C_1, y \in C_2} D(x, y)$

Saut maximum $\max_{x \in C_1, y \in C_2} D(x, y)$

Saut moyen $\text{mean}_{x \in C_1, y \in C_2} D(x, y)$

Méthode de Ward Maximise l'inertie inter-classe

...

Méthode exacte : $O(n^2) <$ complexité $< O(n^3)$!

Clustering Hiérarchique — Procédé

Pour décider des clusters les plus similaires (à fusionner donc), plusieurs options :

Saut minimum $\min_{x \in C_1, y \in C_2} D(x, y)$

Saut maximum $\max_{x \in C_1, y \in C_2} D(x, y)$

Saut moyen $\text{mean}_{x \in C_1, y \in C_2} D(x, y)$

Méthode de Ward Maximise l'inertie inter-classe

...

Méthode exacte : $O(n^2) <$ complexité $< O(n^3)$!

Métriques de clustering — Introduction

Deux cas :

1. Si on dispose de données annotées (même insuffisantes pour apprendre) $\rightarrow$ métriques supervisées de classification
2. Sinon, on évalue la qualité *intrinsèque* des clusters

Métriques de clustering — Introduction

Deux cas :

1. Si on dispose de données annotées (même insuffisantes pour apprendre) $\rightarrow$ métriques supervisées de classification
2. Sinon, on évalue la qualité *intrinsèque* des clusters

Métriques de clustering — Application des métriques supervisées

Un calcul préalable est nécessaire : le mapping optimal entre les clusters et les labels.

Plusieurs possibilités :

- Algorithme hongrois
- Sous-optimal mais bonne première approche : classe la plus représentée dans chaque cluster

Métriques de clustering — Application des métriques supervisées

Un calcul préalable est nécessaire : le mapping optimal entre les clusters et les labels.

Plusieurs possibilités :

- Algorithme hongrois
- Sous-optimal mais bonne première approche : classe la plus représentée dans chaque cluster

Métriques de clustering — Inertie

$$\text{coût} = \sum_i \sum_j \delta_{i,j} \|x_j - \mu_i\|^2$$

où $\delta_{i,j}$ vaut 1 si le cluster μ_i est le plus proche du point x_j, 0 sinon

Métriques de clustering — Silhouette

Points $x = \{x_1, \ldots, x_n\}$, Clusters $\mu = \{\mu_1, \ldots, \mu_k\}$.

$$a(x_i) = \frac{1}{\#\mu_i - 1} \sum_j |x_i - x_j|$$

$$b(x_i) = \min_{i \neq j} \frac{1}{\#\mu_j} \sum_j |x_i - x_j|$$

où :

$\#\mu_i$ est le nombre d'éléments de x dans le cluster μ_i

L'ensembe d'indice j ne représente que ceux des points appartenant au cluster μ_j

$a(x_i)$: distance moyenne aux autres points du cluster contenant x_i

$b(x_i)$: distance moyenne aux points du cluster le plus proche

Métriques de clustering — Silhouette

Points $x = \{x_1, \ldots, x_n\}$, Clusters $\mu = \{\mu_1, \ldots, \mu_k\}$.

$$a(x_i) = \frac{1}{\#\mu_i - 1} \sum_j |x_i - x_j|$$

$$b(x_i) = \min_{i \neq j} \frac{1}{\#\mu_j} \sum_j |x_i - x_j|$$

où :

$\#\mu_i$ est le nombre d'éléments de x dans le cluster μ_i

L'ensembe d'indice j ne représente que ceux des points appartenant au cluster μ_j

$a(x_i)$: distance moyenne aux autres points du cluster contenant x_i

$b(x_i)$: distance moyenne aux points du cluster le plus proche

Métriques de clustering — Silhouette

$$s_i = \frac{b_i - a_i}{\max\{a_i, b_i\}} \quad , \quad s_i = \begin{cases} 1 - a_i/b_i & \text{if } a_i < b_i \\ 0 & \text{if } a_i = b_i \\ b_i/a_i - 1 & \text{if } a_i > b_i \end{cases}$$

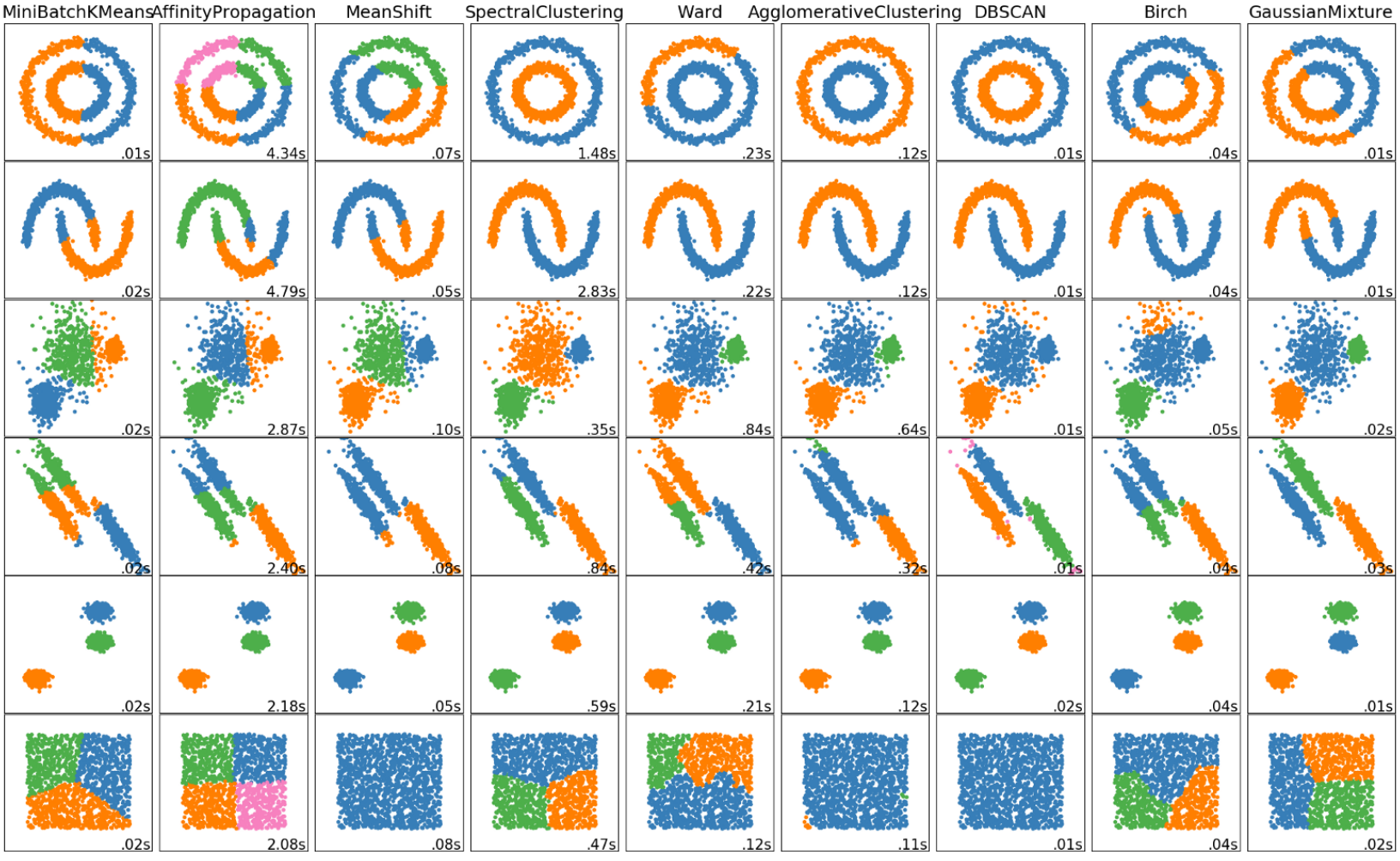
donc $s_i \in [-1, 1]$

$s_i \approx 1 \iff x_i$ bien clusterisé
$s_i \approx 0 \iff x_i$ au bord de 2 clusters
$s_i \approx -1 \iff x_i$ mal clusterisé

Métriques de clustering — Et d'autres options existent

- Calinski-Harabaz index
- Davies-Bouldin Index
- ...

Conclusion


MiniBatchKMeans
AffinityPropagation
MeanShift
SpectralClustering
Ward
AgglomerativeClustering
DBSCAN
Birch
GaussianMixture
.01s 4.34s .07s 1.48s .23s .12s .01s .04s .01s
.02s 4.79s .05s 2.83s .22s .12s .01s .04s .01s
.02s 2.87s .10s .35s .84s .64s .01s .05s .02s
.02s 2.40s .08s .84s .42s .32s .01s .04s .03s
.02s 2.18s .05s .59s .21s .12s .02s .04s .01s
.02s 2.08s .08s .47s .12s .11s .01s .04s .02s

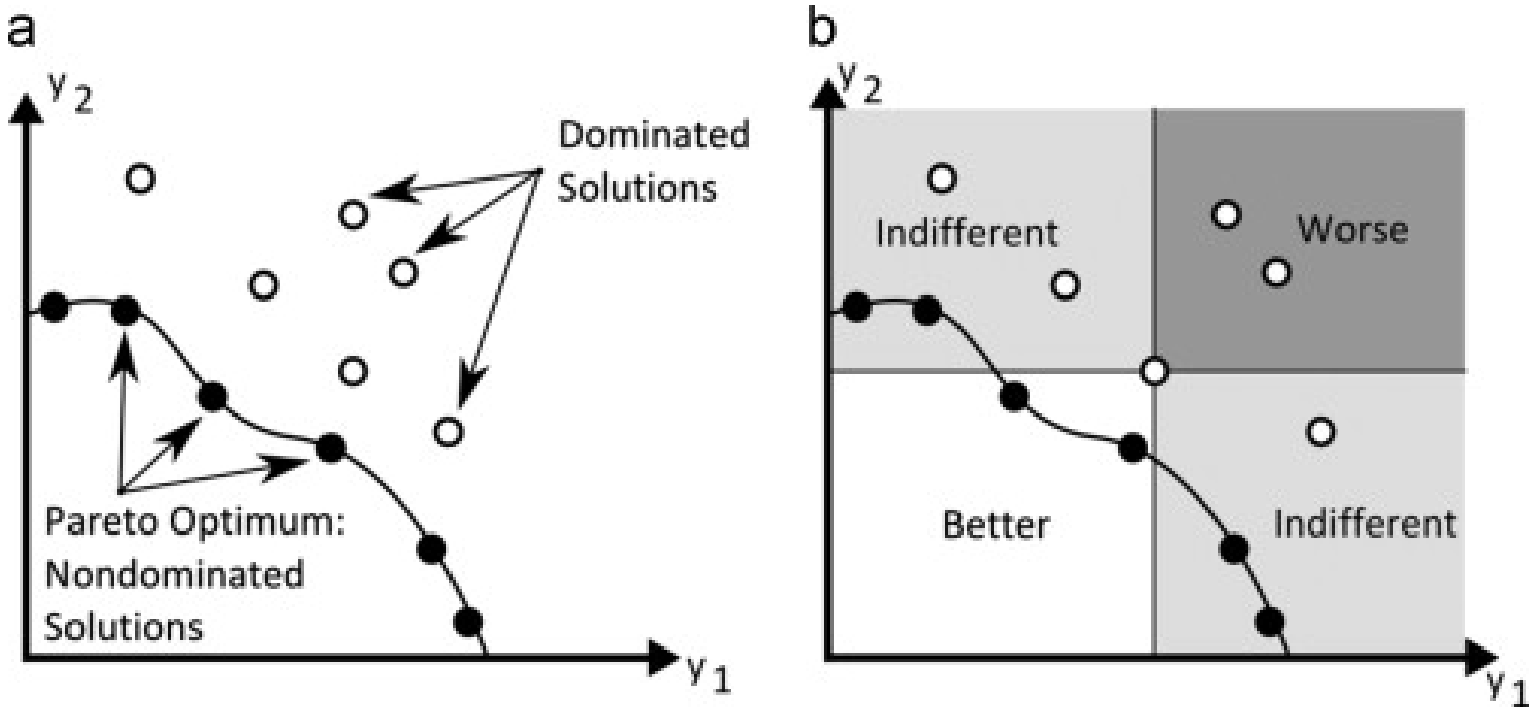
Conclusion

Pas de métrique satisfaisante ! ⇐ Théorie de la Décision :

Problème qui se mesure en plusieurs dimensions

⇒

Pas de solution unique !



Support TP : Clustering

- K-means — Tutoriel
- DBSCAN — Tutoriel

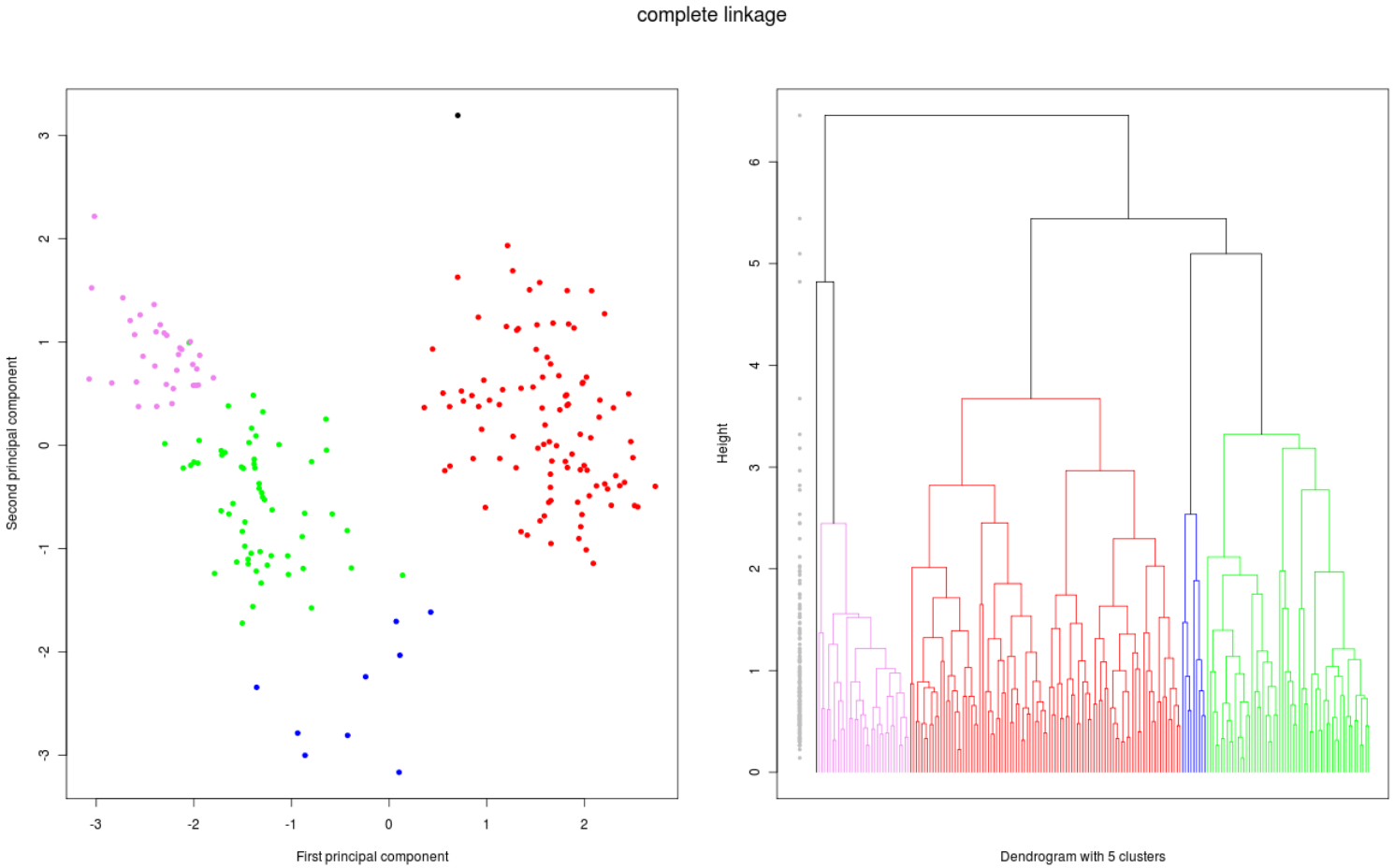
Avez-vous des questions ?

Apprentissage non-supervisé

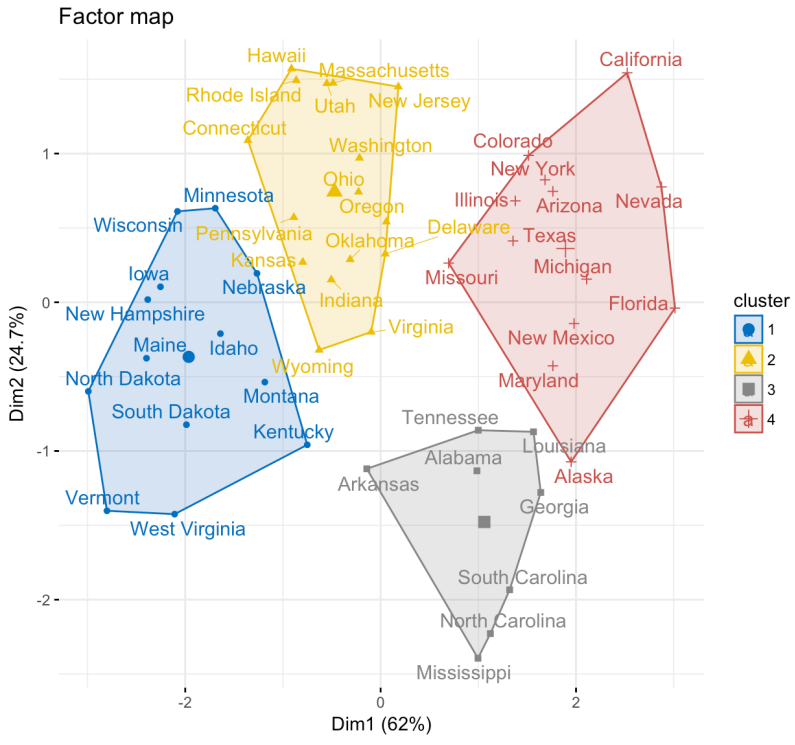
Méthodes hybrides

Classification Hiérarchique sur Composantes Principales

Réduction de dimension par PCA, puis clustering hiérarchique.



Exemple



Obtenus à partir de données relatives aux crimes aux États-Unis.

Colonnes d'origine : Population Totale, Meurtres, Viols, Agression

Clustering par PCA

(Souvenez-vous)

- Calcul de la matrice de covariance (resp corrélation) :

$$\frac{1}{N} \times \overline{X^T} \times \overline{X} \quad (\frac{1}{N} \times \widetilde{X^T} \times \widetilde{X})$$

- Calcul des vecteurs propres de cette matrice : ce sont les combinaisons linéaires des nouvelles features
- Tri des vecteurs par valeur propre décroissante
- Réduction de dimension : on ne conserve que les k premiers vecteurs propres

Clustering par PCA

(Souvenez-vous)

- Calcul de la matrice de covariance (resp corrélation) :

$$\frac{1}{N} \times \overline{X^T} \times \overline{X} \quad (\frac{1}{N} \times \widetilde{X^T} \times \widetilde{X})$$

- Calcul des vecteurs propres de cette matrice : ce sont les combinaisons linéaires des nouvelles features
- Tri des vecteurs par valeur propre décroissante
- Réduction de dimension : on ne conserve que les k premiers vecteurs propres

Clustering par PCA

(Souvenez-vous)

- Calcul de la matrice de covariance (resp corrélation) :

$$\frac{1}{N} \times \overline{X^T} \times \overline{X} \quad (\frac{1}{N} \times \widetilde{X^T} \times \widetilde{X})$$

- Calcul des vecteurs propres de cette matrice : ce sont les combinaisons linéaires des nouvelles features
- Tri des vecteurs par valeur propre décroissante
- Réduction de dimension : on ne conserve que les k premiers vecteurs propres

Clustering par PCA

(Souvenez-vous)

- Calcul de la matrice de covariance (resp corrélation) :

$$\frac{1}{N} \times \overline{X^T} \times \overline{X} \quad (\frac{1}{N} \times \widetilde{X^T} \times \widetilde{X})$$

- Calcul des vecteurs propres de cette matrice : ce sont les combinaisons linéaires des nouvelles features
- Tri des vecteurs par valeur propre décroissante
- Réduction de dimension : on ne conserve que les k premiers vecteurs propres

Clustering par PCA

Considérer les individus comme des features et les features comme des individus.

Les vecteurs propres ayant une grande valeur propre peuvent être considérés comme des centre de cluster d'individus.

Avez-vous des questions ?

Apprentissage non-supervisé

Détection d'anomalies

Détection d'Anomalies

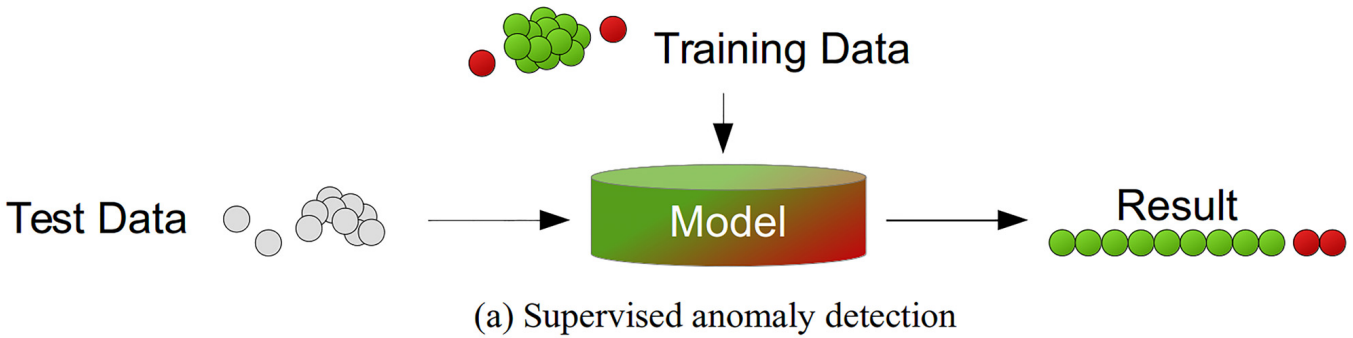
Détection :

- de Fraude
- d'Intrusion/Fuite (physique ou électronique)
- Santé (biologique, geologique, machine, ...)

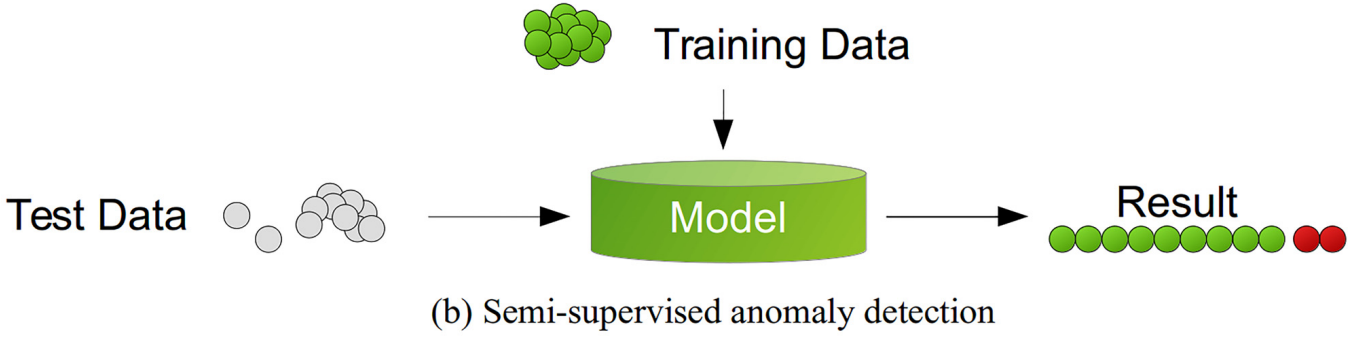
Définition

- une anomalie diffère de la norme par ses features
- les anomalies sont rares comparées aux instances normales

Modes de détection d'anomalie



(a) Supervised anomaly detection



(b) Semi-supervised anomaly detection

(c) Unsupervised anomaly detection

Détection d'Anomalies : Supervisé

Problème de classification normal.

Réseaux de neurones et SVM très performants.

Détection d'Anomalies : Semi-Supervisé

Détection de nouveauté.

Pas traité ici.

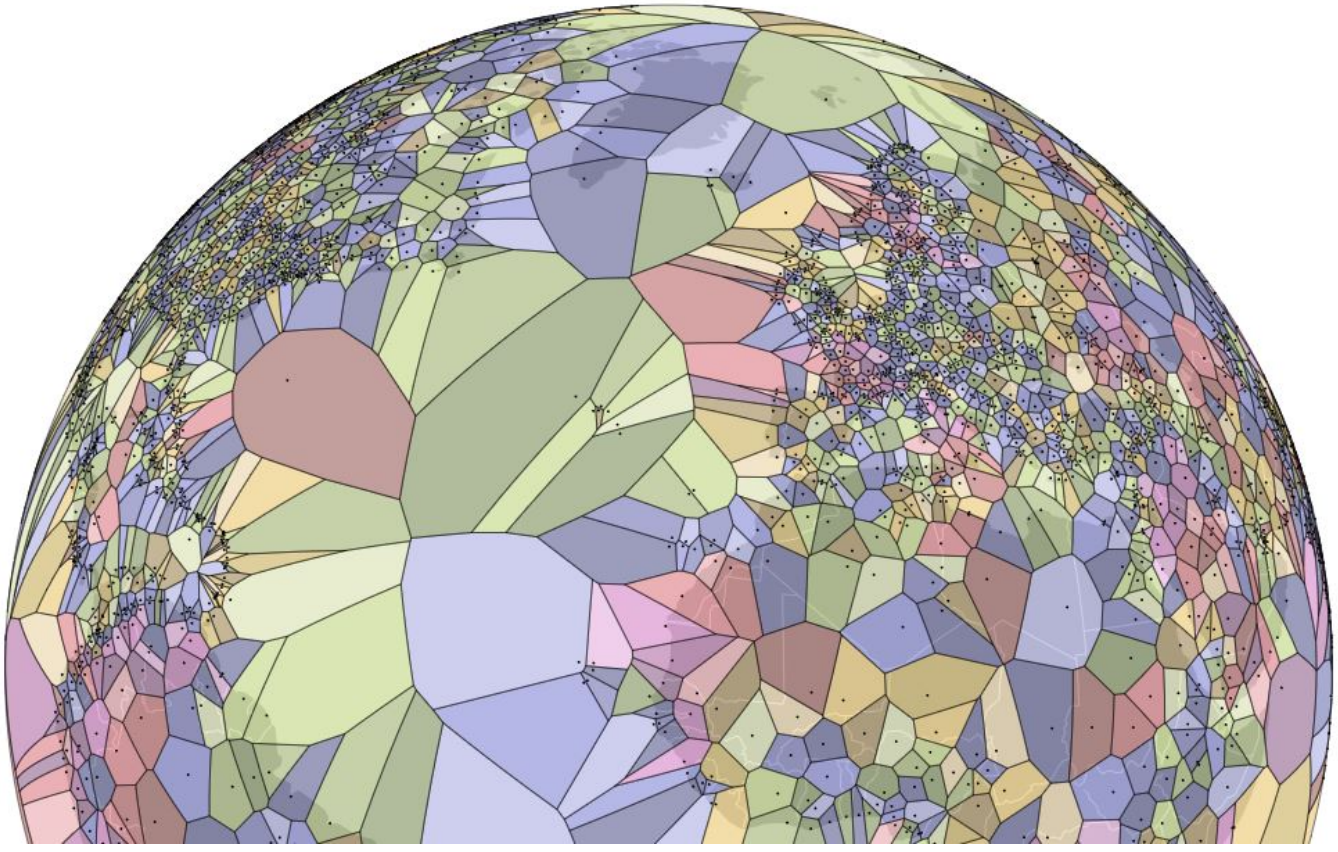
One-class SVM très utilisé.

Détection d'Anomalies : Non-Supervisé

De nombreuses méthodes :

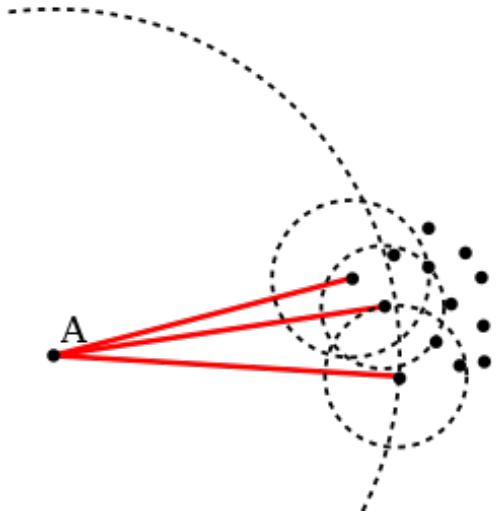
- Local Outlier Factor (LOF)
- Unweighted Cluster-Based Outlier Factor
- Isolation Forest
- Autoencoder
- ...

Détection d'Anomalies



Local Outlier Factor

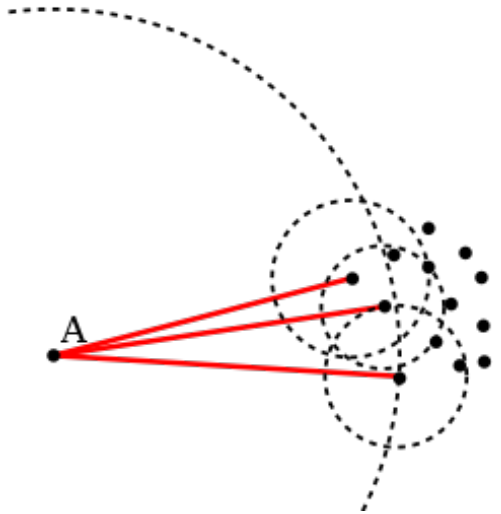
- anomalies locales
- basé sur les k voisins du point
- définit une « atteignabilité » par les distances de ces voisins
- calcule un ratio moyen d'atteignabilité du point et de ses voisins



→ Anomalie si le ratio moyen d'atteignabilité est beaucoup plus faible que celui de ses plus proches voisins

Local Outlier Factor

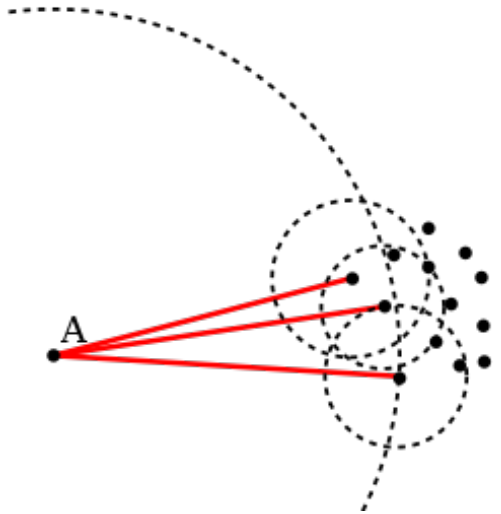
- anomalies locales
- basé sur les k voisins du point
- définit une « atteignabilité » par les distances de ces voisins
- calcule un ratio moyen d'atteignabilité du point et de ses voisins



→ Anomalie si le ratio moyen d'atteignabilité est beaucoup plus faible que celui de ses plus proches voisins

Local Outlier Factor

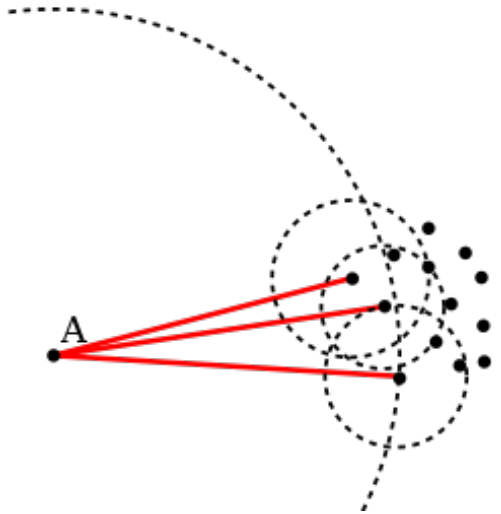
- anomalies locales
- basé sur les k voisins du point
- définit une « atteignabilité » par les distances de ces voisins
- calcule un ratio moyen d'atteignabilité du point et de ses voisins



→ Anomalie si le ratio moyen d'atteignabilité est beaucoup plus faible que celui de ses plus proches voisins

Local Outlier Factor

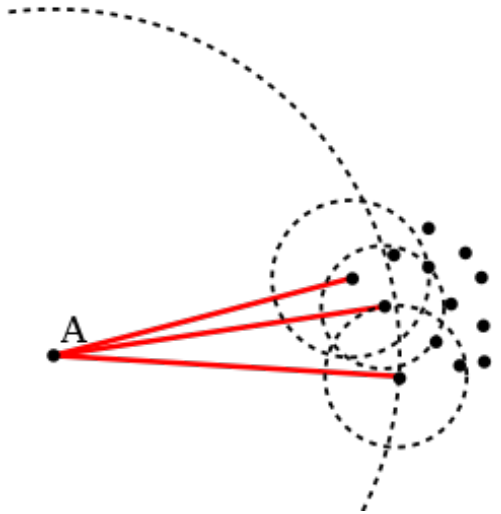
- anomalies locales
- basé sur les k voisins du point
- définit une « atteignabilité » par les distances de ces voisins
- calcule un ratio moyen d'atteignabilité du point et de ses voisins



→ Anomalie si le ratio moyen d'atteignabilité est beaucoup plus faible que celui de ses plus proches voisins

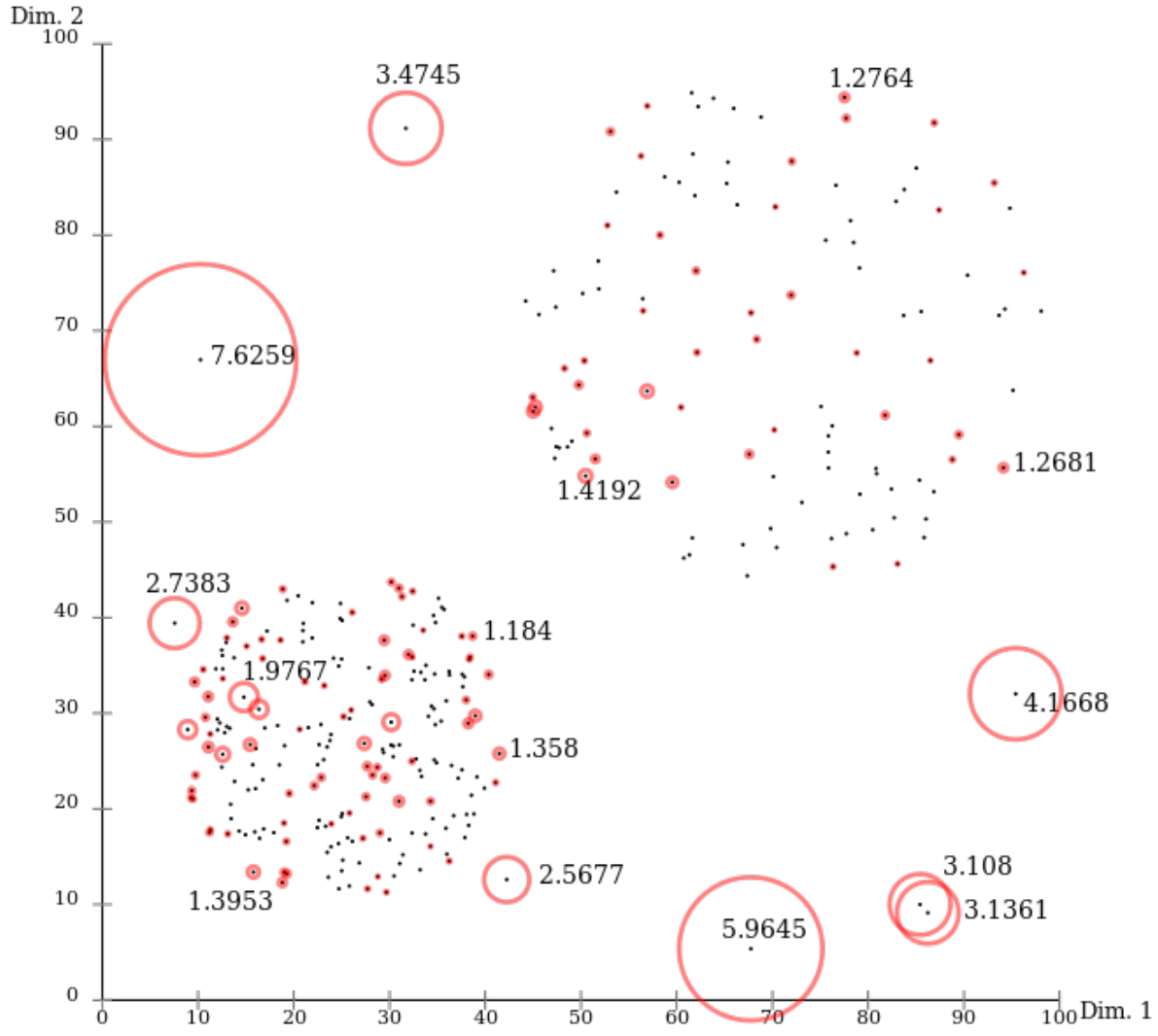
Local Outlier Factor

- anomalies locales
- basé sur les k voisins du point
- définit une « atteignabilité » par les distances de ces voisins
- calcule un ratio moyen d'atteignabilité du point et de ses voisins



$\rightarrow$ Anomalie si le ratio moyen d'atteignabilité est beaucoup plus faible que celui de ses plus proches voisins

Local Outlier Factor



Désavantages

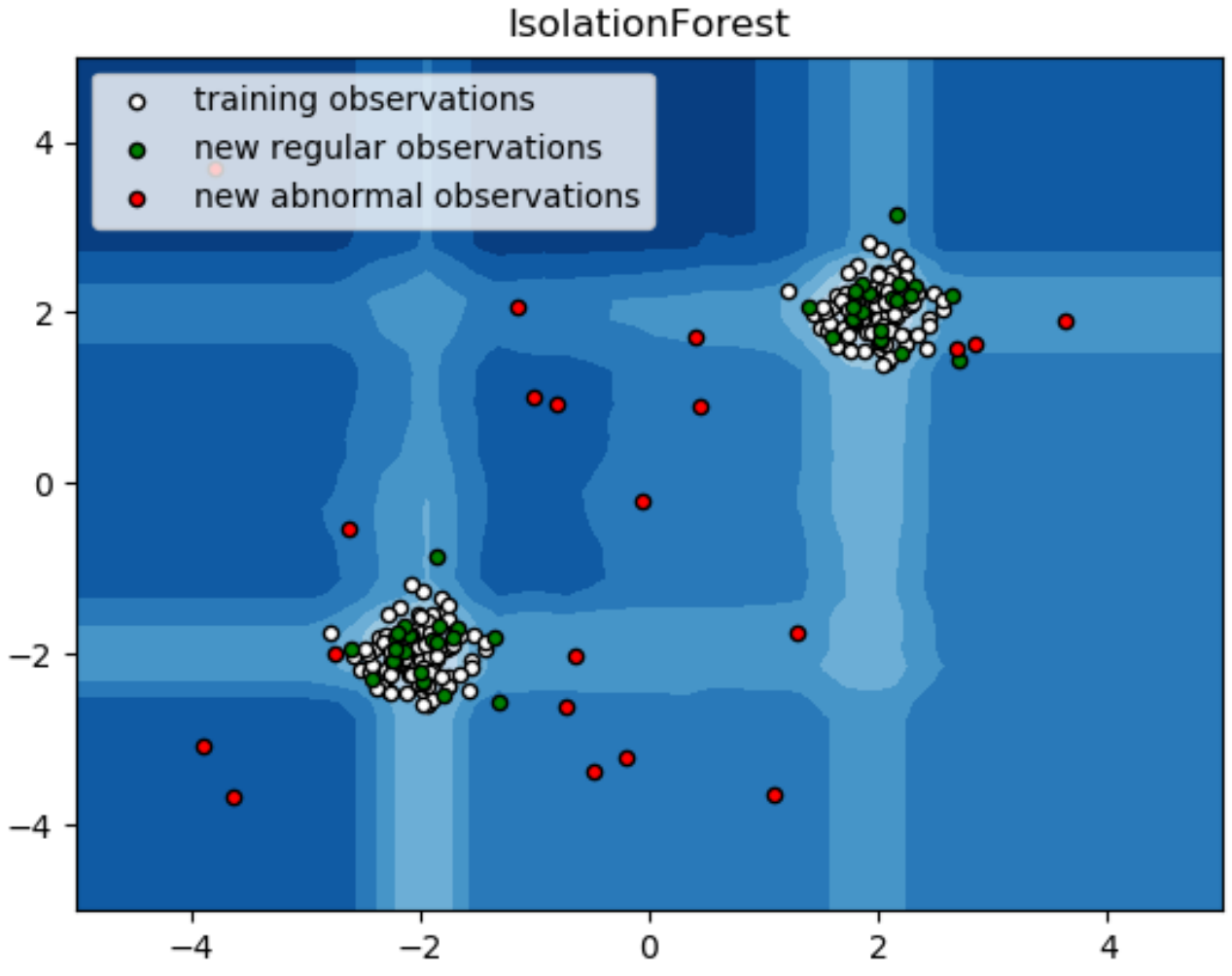
- lent (quadratique)
- a des à priori sur la distribution des données

Isolation tree

- arbre aléatoire (comme random forest mais le split est aléatoire)
- but : isoler une anomalie plus vite qu'un exemple normal
- petit chemin pour arriver à une feuille : anomalie

$\rightarrow$ Se sert du fait que les features des anomalies ne sont pas distribuées comme les autres.

Isolation forest



Support TP : Détection d'anomalie

- Local Outlier Factor — Tutoriel
- Isolation Forest — Tutoriel

Avez-vous des questions ?

Apprentissage non-supervisé

Travaux Pratiques

Instructions — Réduction de dimensionnalité

Utilisation de PCA

Instructions — Réduction de dimensionnalité & Clustering

Clustering & projections linéaires et non-linéaires