Big Data Analytics

Les Algorithmes Non-Supervisés

GIRAUD François-Marie

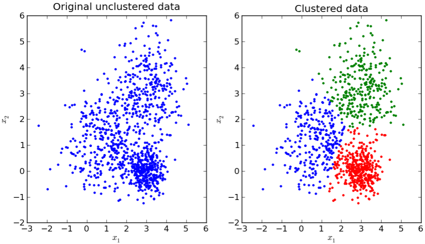
https ://www.orsys.fr/

Big Data Analytics

Données Non-Suppervisées

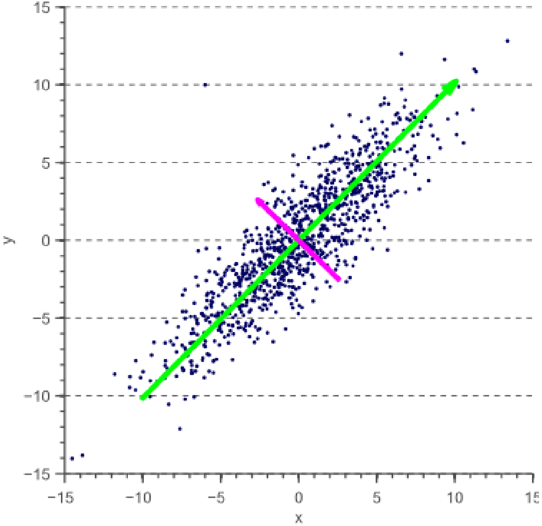
Données Non-Suppervisées

Problème : détection de variables cachées, de “structures” cachées (clusters, variété topologique (manifold), ...)

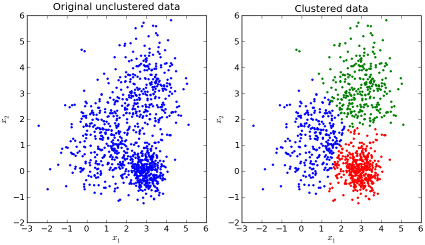


Données Non-Suppervisées

$\approx$ Réduction de Dimensionalité



Données Non-Suppervisées

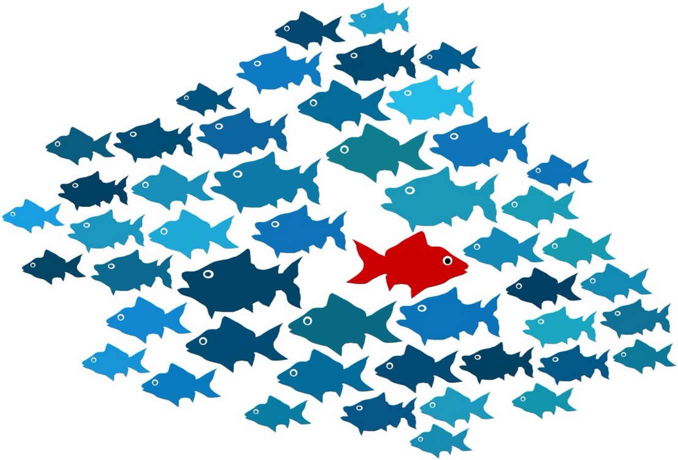


Données Non-Suppervisées

Clustering de Clients

Données Non-Suppervisées

Détection d'anomalie :



Big Data Analytics

Réduction de la dimensionalité

Réduction de la dimensionalité

Comment appréhender des données en grande dimension ?

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

Réduction de la dimensionalité

La malédiction des grandes dimensions !
(Curse of dimensionality)

Réduction de la dimensionalité

- Séléction de dimensions
- Projections linéaires (ACP, LDA, ...)
- Projections non-linéaires (kernels, neural network embeddings, ...)

Réduction de la dimensionalité

Sélection de dimensions :

- Random forest
- SVM
- ...

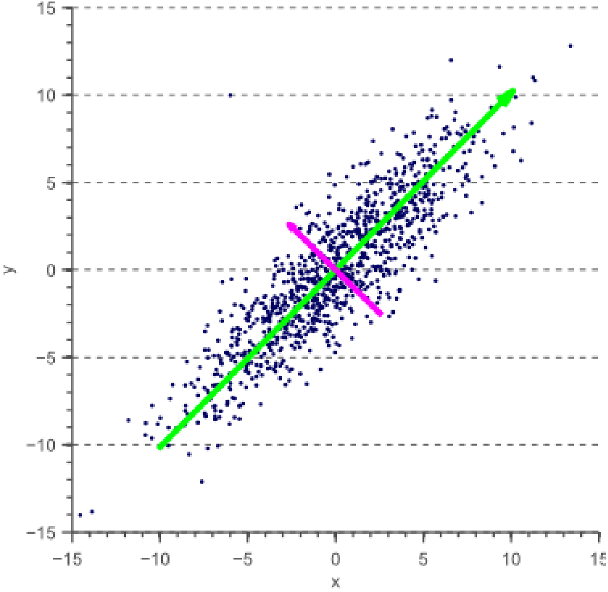
Big Data Analytics

Réduction de la dimensionalité : Projections linéaires

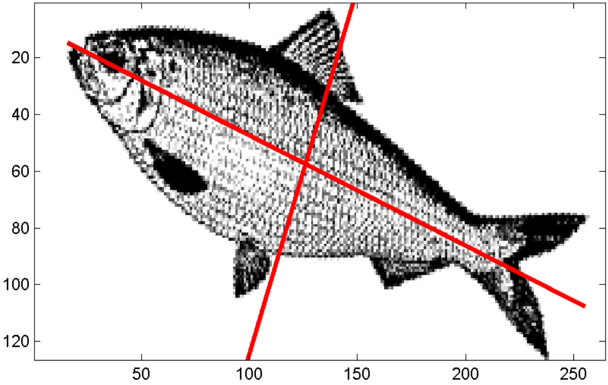
Réduction de la dimensionalité : Projections linéaires

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

Réduction de la dimensionalité : PCA



Réduction de la dimensionalité : PCA



Réduction de la dimensionalité : PCA

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

Réduction de la dimensionalité : PCA

chaque dimension est centrée (et réduite) :

$$\bar{X} = \begin{bmatrix} X_{1,1} - \bar{X}_1 & \dots & X_{1,D} - \bar{X}_D \\ \vdots & \ddots & \vdots \\ X_{N,1} - \bar{X}_1 & \dots & X_{N,D} - \bar{X}_D \end{bmatrix}$$

ou

$$\tilde{X} = \begin{bmatrix} \frac{X_{1,1} - \bar{X}_1}{\sigma(X_1)} & \dots & \frac{X_{1,D} - \bar{X}_D}{\sigma(X_D)} \\ \vdots & \ddots & \vdots \\ \frac{X_{N,1} - \bar{X}_1}{\sigma(X_1)} & \dots & \frac{X_{N,D} - \bar{X}_D}{\sigma(X_D)} \end{bmatrix}$$

Réduction de la dimensionalité : PCA

Matrice de covariance (resp. corrélation) :

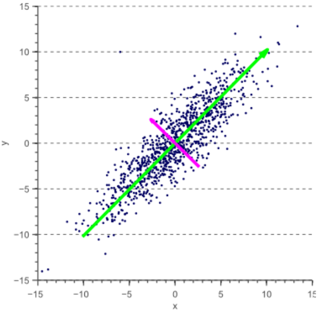
$$\frac{1}{N} * \bar{X}^T * \bar{X} \, , \; (\frac{1}{N} * \tilde{X}^T * \tilde{X})$$

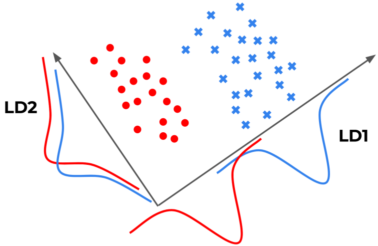
<u>ACP</u> :
Retrouver les valeurs et vecteurs propres de de la matrice de covariance (resp. corrélation), donc diagonaliser la matrice carrée obtenue.

Vecteur propre : vecteur permettant de projeter les données
Valeur propre : “proportion d'information” conservée par la projection suivant le vecteur propre correspondant
Réduction de dimension : On ne projette que suivant le nombre de vecteurs propres voulus

Réduction de la dimensionalité : LDA

Linear Discriminant Analysis





Big Data Analytics

Analyse en composantes - variantes spécifiques

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

Big Data Analytics

Métriques en Non-Supervisé

Métriques en Non-Supervisé

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

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

Métrique : Silouhette

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

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

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

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

Métrique : Silouhette

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

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

Métrique : etc

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

Big Data Analytics

Clustering Hiérarchique

Clustering Hiérarchique

Deux approches :

- Agglomérantes (bottom-up)
- Divisantes (top-down)

Classification Ascendante Hiérarchique (CAH)

Métode Agglomérante

- Chaque élément est dans une classe distincte
- On itère jusqu'à ce qu'on ait le nombre de classes voulues
- On utilise une mesure de dissimilarité inter-classe comme critère d'aggrégation

A chaque itération, on calcule la dissimilarité entre toutes les classes puis on fusionne les plus similaires.

Classification Ascendante Hiérarchique (CAH)

Quelques distances de dissimilarités, après avoir défini une distance D dans l'espace :

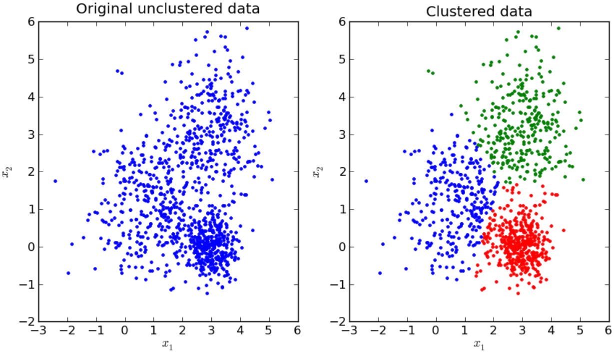
- saut minimum : $dissim(C_1, C_2) = \min_{x \in C_1, y \in C_2} D(x, y)$
- saut maximum : $dissim(C_1, C_2) = \max_{x \in C_1, y \in C_2} D(x, y)$
- saut moyen : $dissim(C_1, C_2) = \underset{x \in C_1, y \in C_2}{moyenne} D(x, y)$
- distance de Ward qui vise à maximiser l'inertie inter-classe
- ...

$O(n^2) <$ complexité $< O(n^3)$!

Big Data Analytics

Expectation-Maximisation

Expectation-Maximisation



Expectation-Maximisation

Input : Données, nombre de clusters, métrique
Initialisation Aléatoire
Jusqu'à clusters "stables" :

1. Calculer les "centres" de chaque cluster
2. Réassigner les clusters à tous les points

Expectation-Maximisation

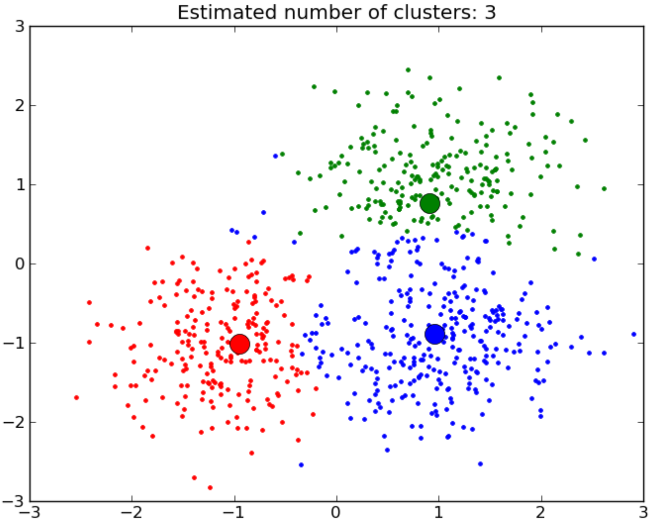
video time

Big Data Analytics

K-Means

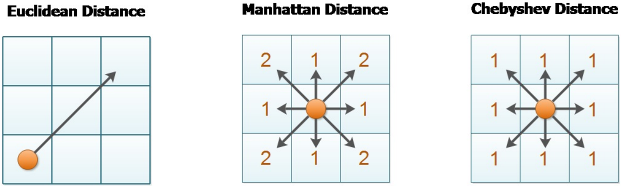
K-Means

Algorithme EM



K-Means

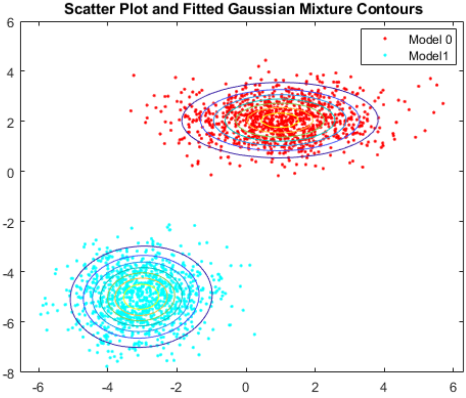
distance euclidienne, Manhattan, Chebychev



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

Gaussian Mixture Model

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



Big Data Analytics

DBSCAN

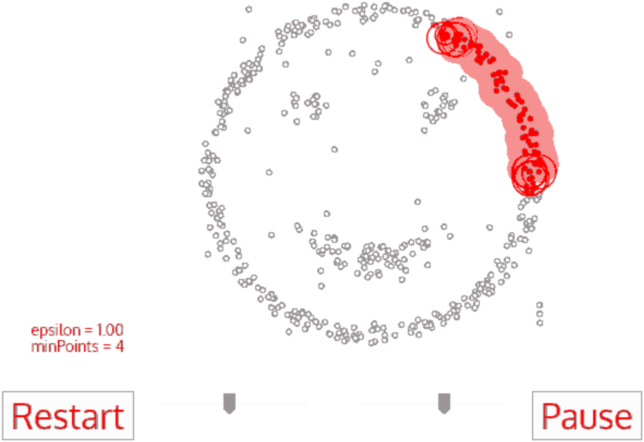
DBSCAN

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

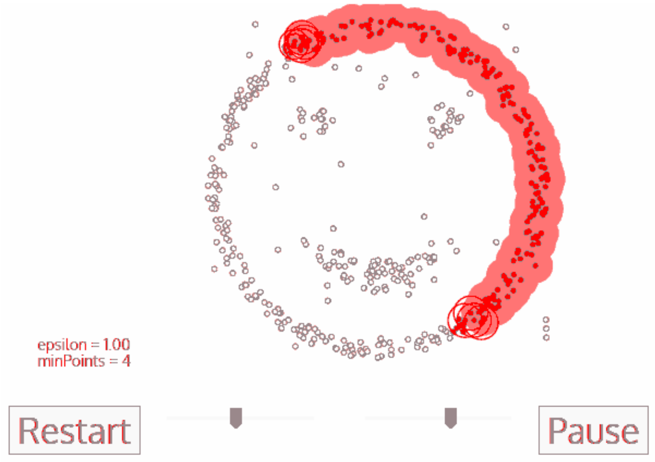
Tant qu'il reste des points non-étiquettés :

1. Prend un point non-étiquettés au hasard et on regarde son voisinage
2. Si (densité $>$ seuil) alors (Nouveau cluster)
    2.1 Expansion du cluster de proche en proche dans le voisinage
3. Sinon (Bruit)

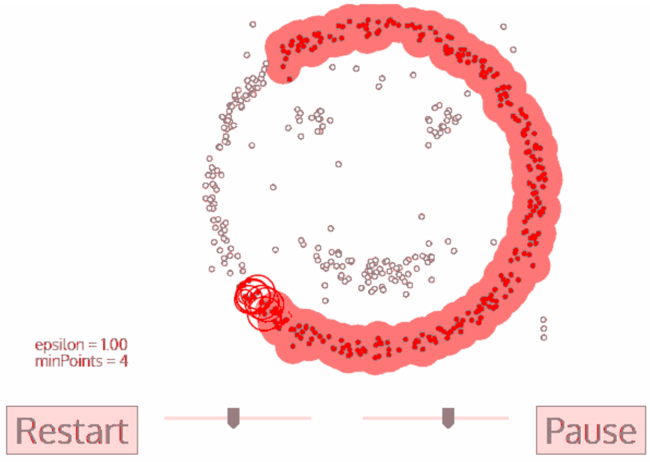
DBSCAN



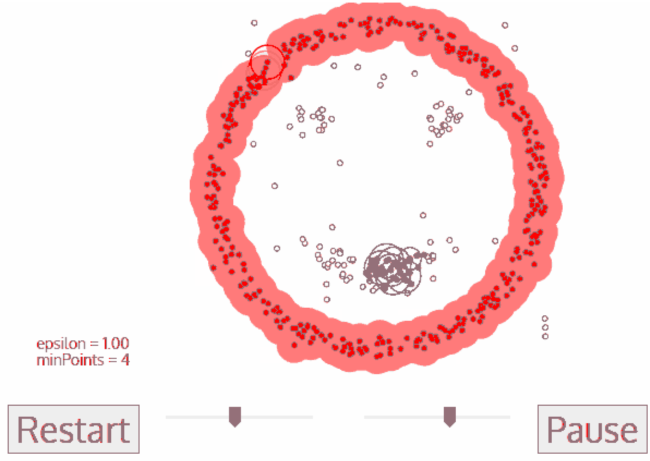
DBSCAN



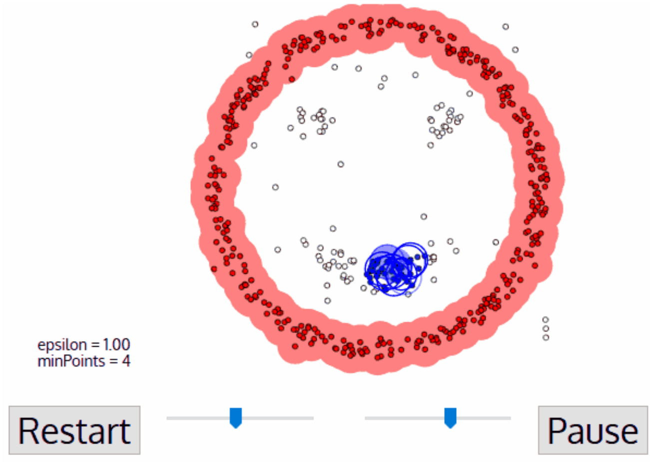
DBSCAN



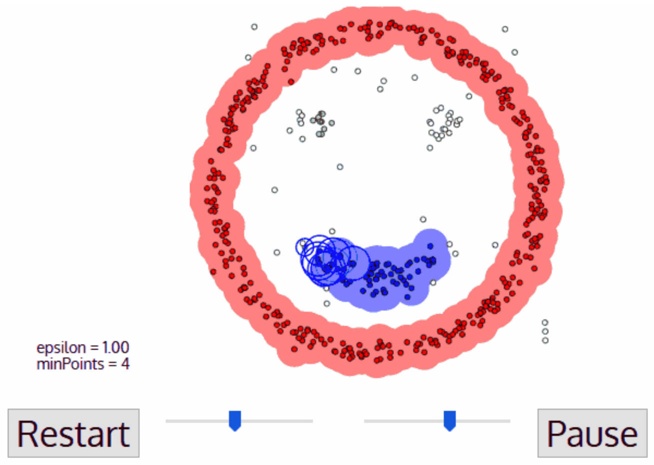
DBSCAN



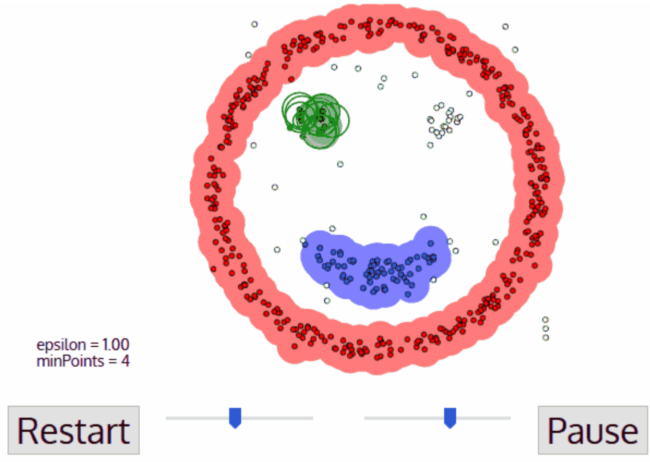
DBSCAN



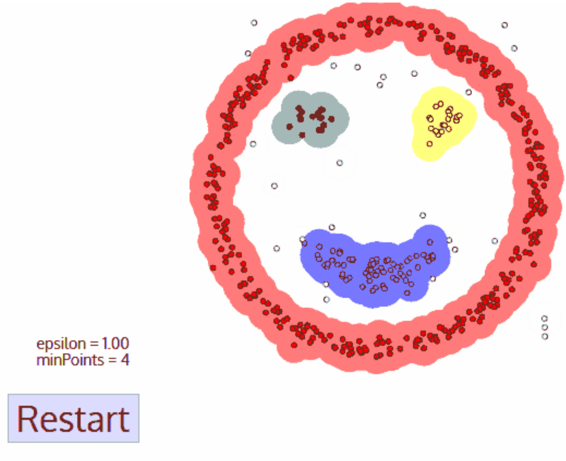
DBSCAN



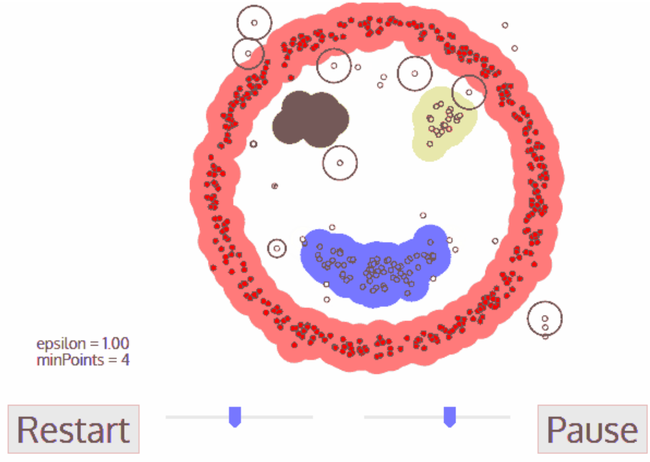
DBSCAN



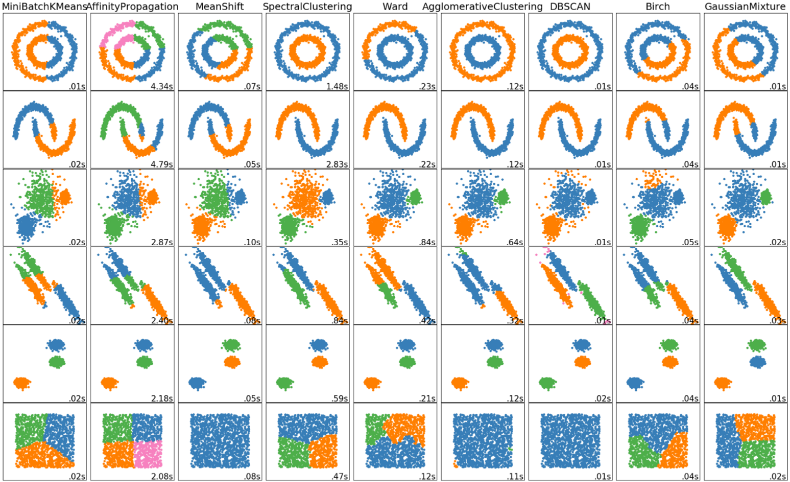
DBSCAN



DBSCAN



Clustering - Conclusions



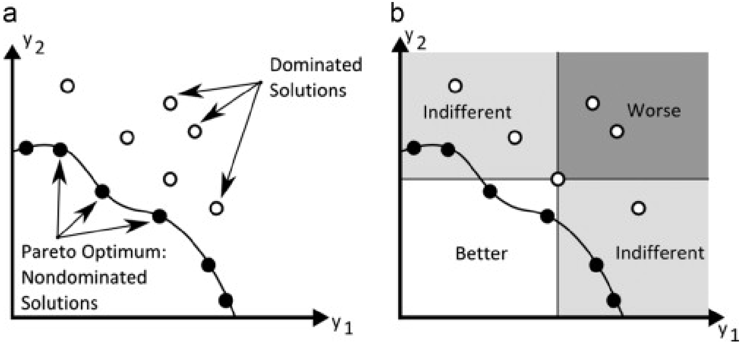
Clustering - Conclusions

Pas de métrique satisfaisante ! $\Leftarrow$ Théorie de la Décision :

Problème qui se mesure en plusieurs dimensions

$\Rightarrow$

Pas de solution unique !

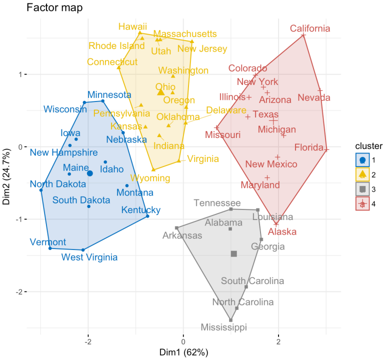


Big Data Analytics

Classification Hiérarchique sur Composantes Principales

Classification Hiérarchique sur Composantes Principales (CHCP)

Après réduction de dimension, on procède à un algorithme de classification hiérarchique.



Obtenus à partir de données relatives aux crimes aux États-Unis.
Colonnes d'origine : Population Totale, Meurtres, Viols, Agression

Big Data Analytics

Clustering par ACP

Réduction de la dimensionalité : PCA

(Souvenez-vous)
Matrice de covariance (resp. corrélation) :

$$\frac{1}{N} * \bar{X}^T * \bar{X}\,,\ (\frac{1}{N} * \tilde{X}^T * \tilde{X})$$

<u>ACP</u> :
Retrouver les valeurs et vecteurs propres de de la matrice de covariance (resp. corrélation), donc diagonaliser la matrice carrée obtenue.

Vecteur propre : vecteur permettant de projeter les données
Valeur propre : “proportion d’information” conservée par la projection suivant le vecteur propre correspondant
Réduction de dimension : On ne projette que suivant le nombre de vecteurs propres voulus

Clustering par ACP

$$\frac{1}{N} * \bar{X} * \bar{X^T} \ , \ (\frac{1}{N} * \tilde{X} * \tilde{X^T})$$

En considérant 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.

Big Data Analytics

Détection d'Anomalies

Détection d'Anomalies

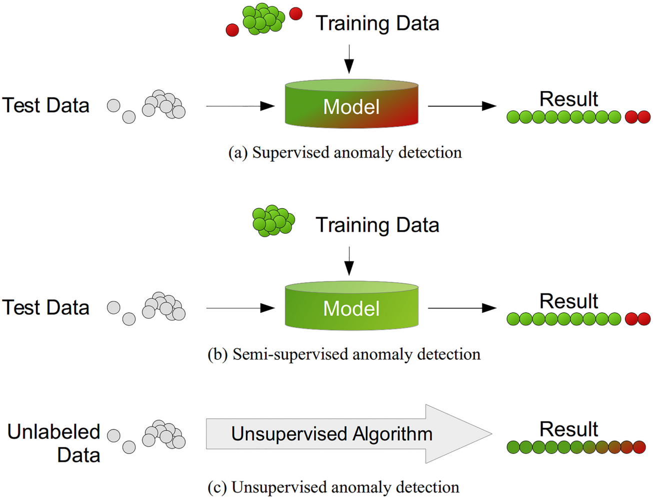
Détection :

- 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



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.

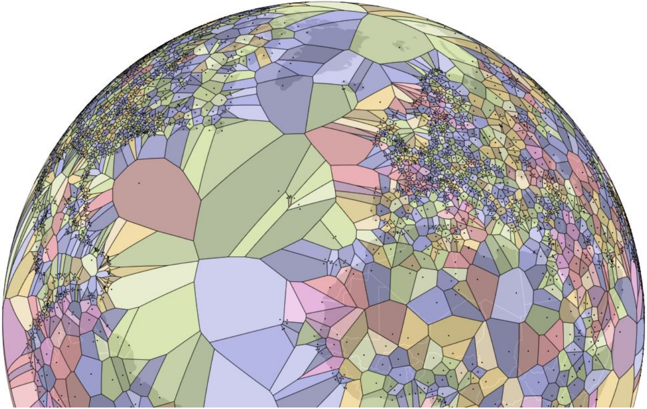
One-class SVM très utilisé.

Détection d'Anomalies : Non-Supervisé

De nombreuses méthodes :

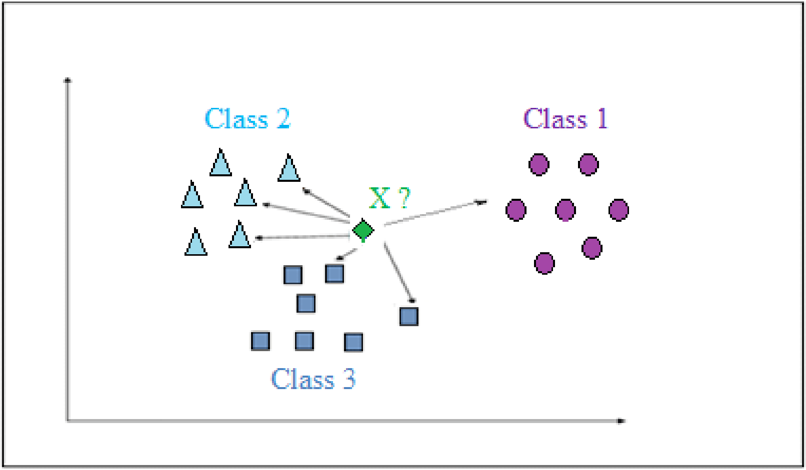
- K-Nearest-Neighbor (KNN)
- Local Outlier Factor (LOF)
- Unweighted Cluster-Based Outlier Factor
- Isolation Forest
- Autoencoder
- ...

Détection d'Anomalies



K-Nearest Neighbor

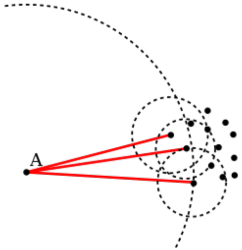
Détection d'Anomalies : KNN



Local Outlier Factor

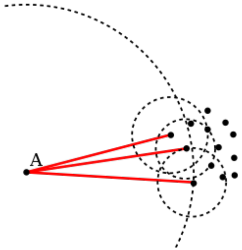
Local Outlier Factor

- anomalies locales



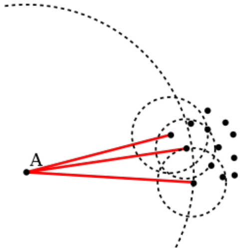
Local Outlier Factor

- anomalies locales
- basé sur les k voisins du point



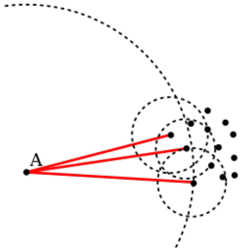
Local Outlier Factor

- anomalies locales
- basé sur les k voisins du point
- définit une « atteignabilité » par les distances de ces voisins



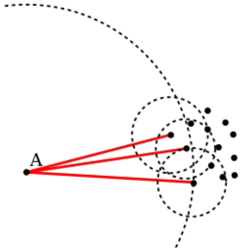
Local Outlier Factor

- 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



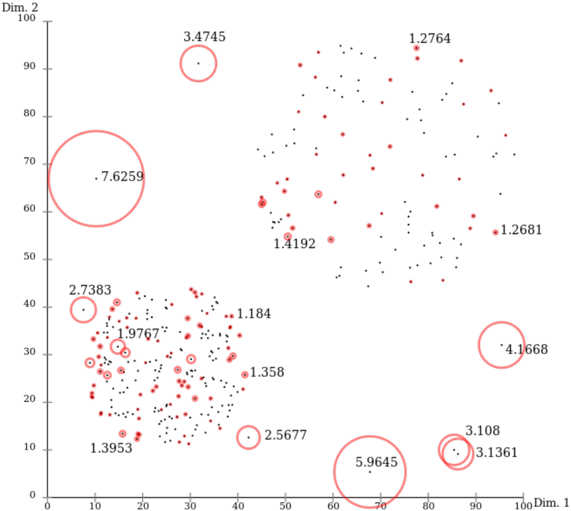
Local Outlier Factor

- 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 forest

Isolation tree

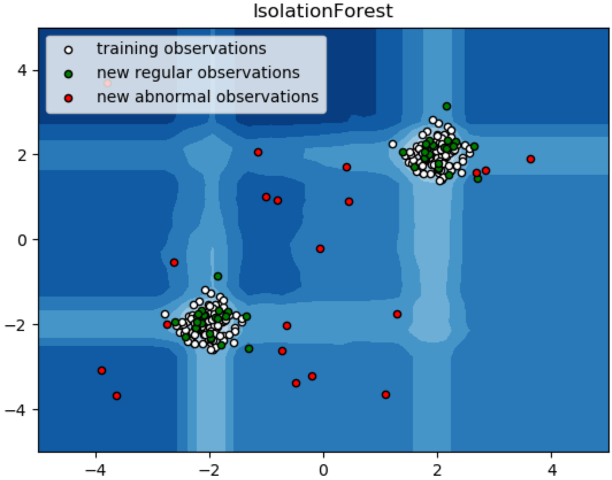
- arbre aléatoire (comme random forest mais le split est aléatoire, ExtraTree)
- but : isoler une anomalie plus vite qu'un exemple normal
- petit chemin pour arriver à une feuille : anomalie

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

Isolation forest

- forêt d'isolation trees
- construits sur des sous-échantillons sans replacement des données
- sous-échantillons plus petits que dans random forest typiquement, pour mieux isoler les anomalies
- converge souvent vite : 100 arbres souvent suffisants

Isolation forest



Travaux Pratiques

Réduction de dimensions et Clustering

Support TP : PCA/LDA

[PCA- iris dataset - Tutoriel]

[PCA-LDA - Tutoriel]

Support TP : Clustering

[kmean - Tutoriel]

[dbscan - Tutoriel]

Support TP : Détection d'anomalie

local-outlier-factor - Tutoriel

isolation-forest - Tutoriel

TP : Réduction dimension

PCA - TP

TP : Clustering

[clustering - TP]