Méthodes d'apprentissage

Introduction aux bases

Aichetou Bouchareb

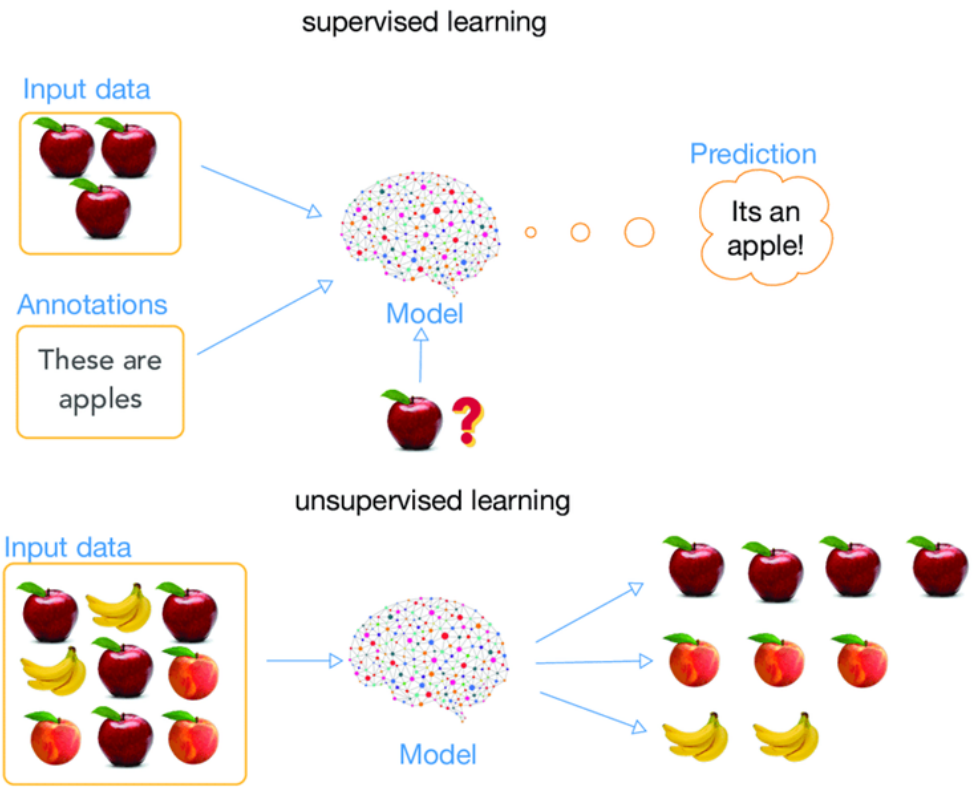
Définitions

L'objectif : données + Modèle $\Rightarrow$ informations utiles et/ou utilisables.

- Un modèle est comme un enfant qui apprend à partir des données (informations) qu'on lui donne

Supervisé ou non-supervisé

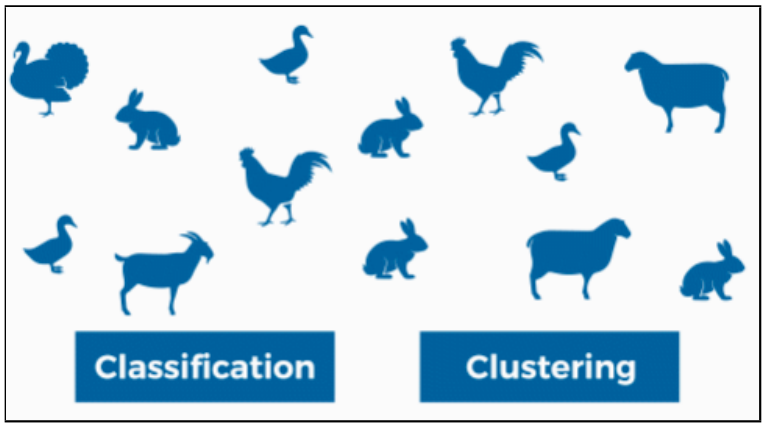
- Deux grandes familles de modèles (Supervisé ou non-Supervisé):

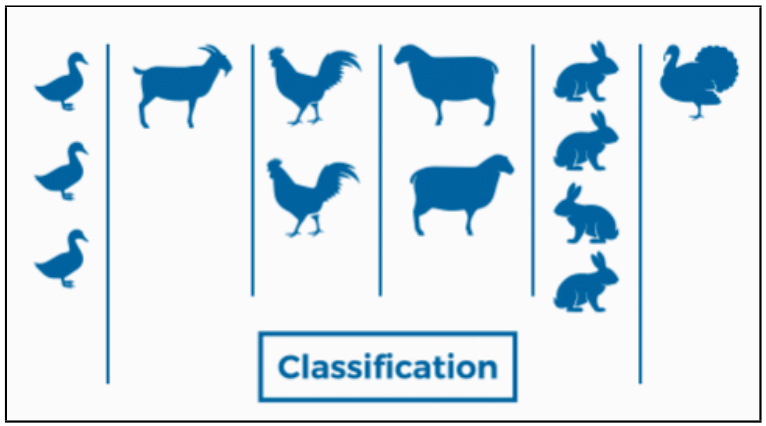


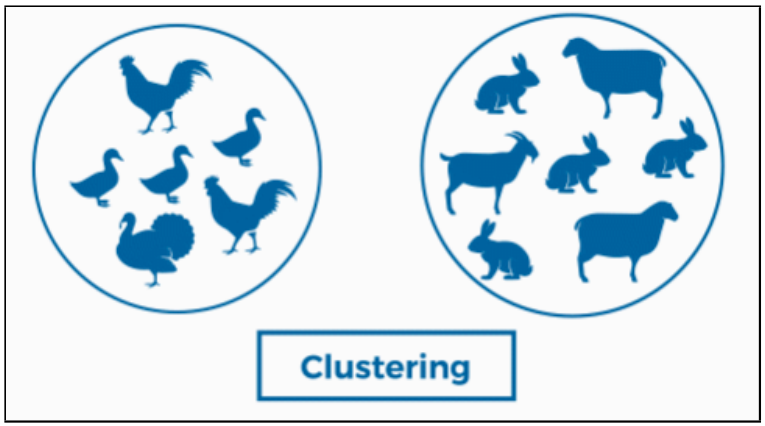
1. Supervisé : on a un ensemble de données $\mathbf{x}, \mathbf{y} = \{(x_i, y_i)\}$ où $\mathbf{y}$ est appelée variable cible (target ou label en anglais). Objectif : meilleur approximation de y et prédire la cible y_j pour une nouvelle donnée x_j jamais vue.
2. Non-Supervisé : on a un ensemble de données $\mathbf{x} = \{(x_i)\}$. Objectif : classer les données, réduire l'information, interpreter, . . ., etc.

Supervisé ou non-supervisé

- Deux grandes familles de modèles (Supervisé ou non-Supervisé):

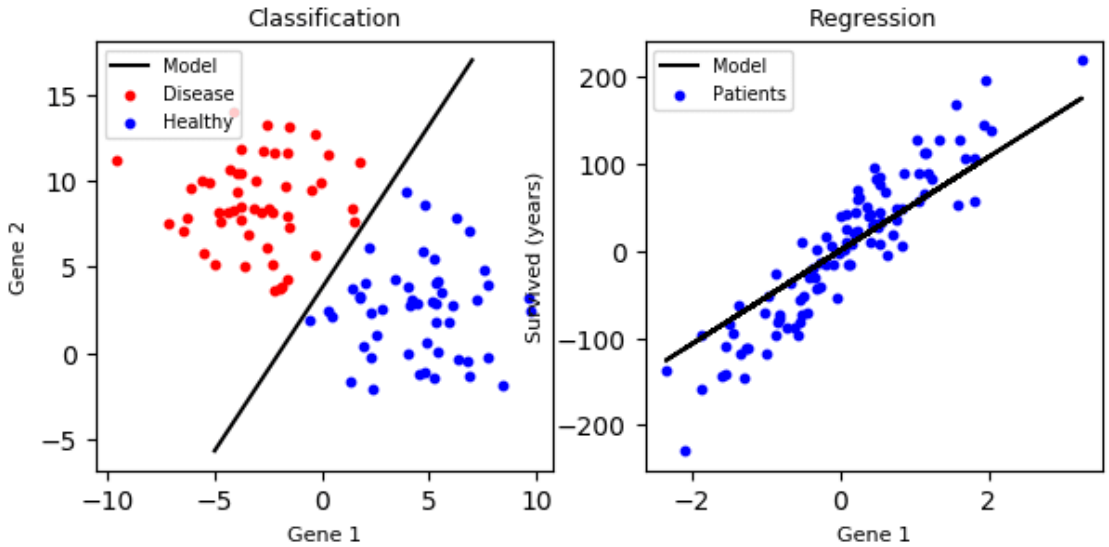


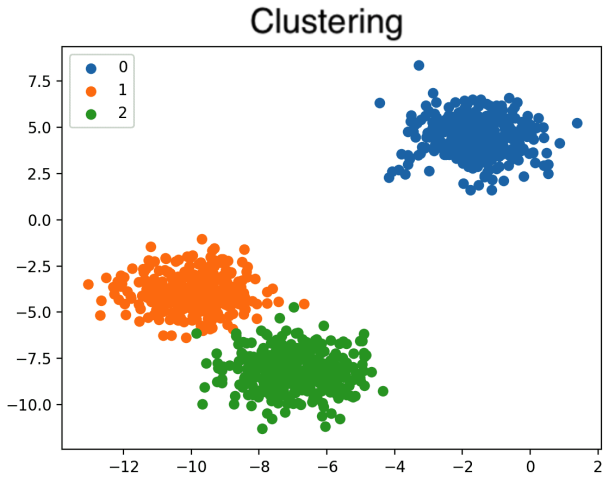




Supervisé ou non-supervisé

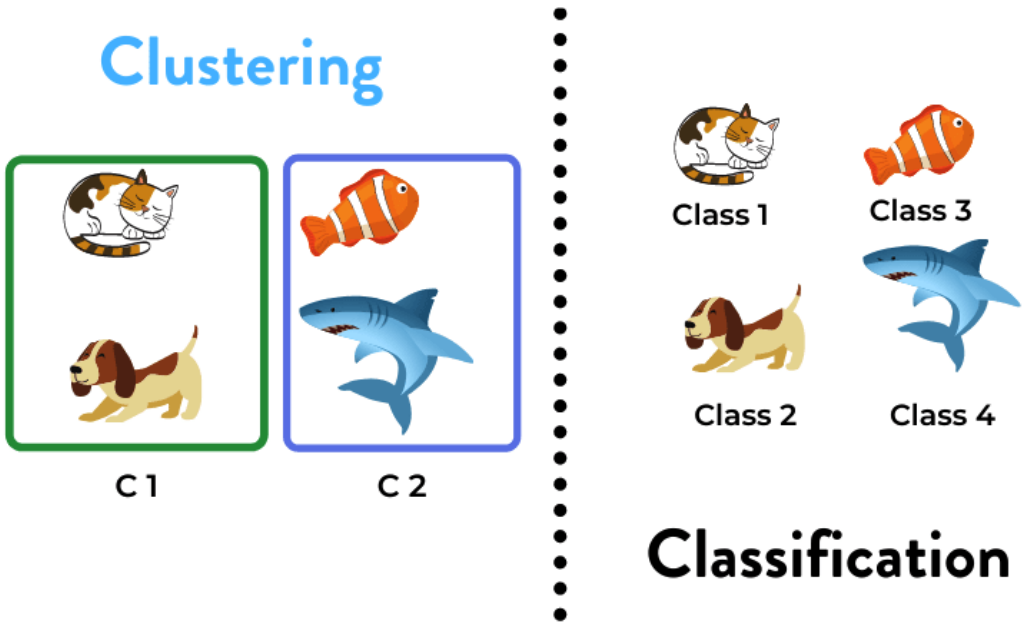
- Classification (Classification supervisée), Regression, Clustering (Classification non supervisée)

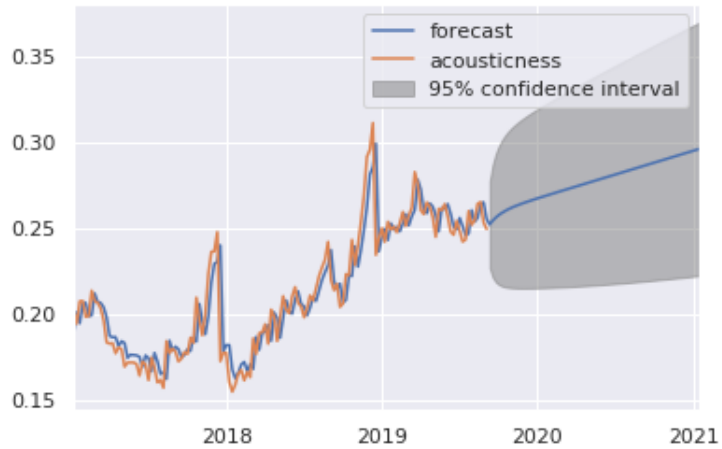




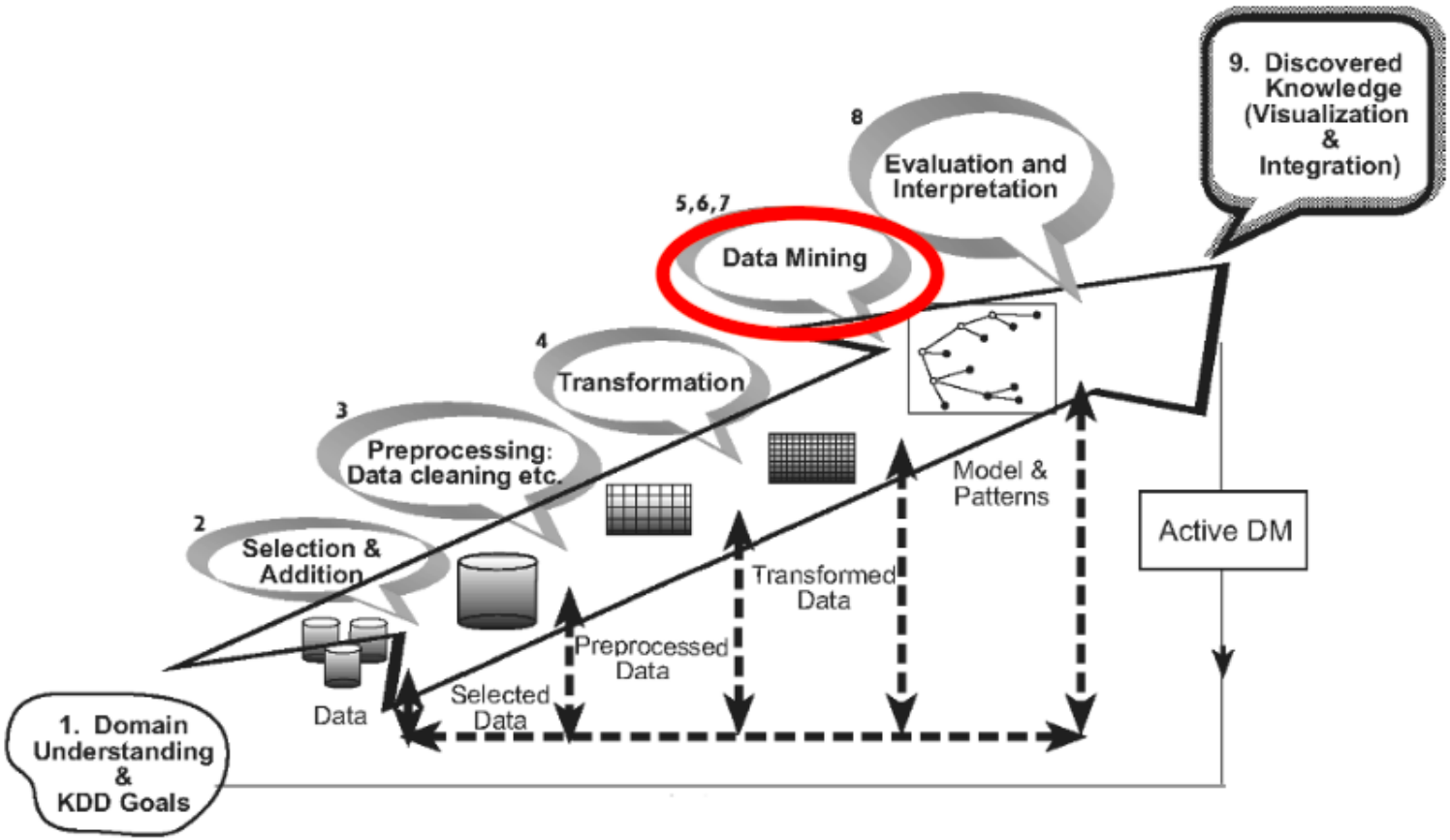
Supervisé ou non-supervisé

- Classification (Classification supervisée), Regression, Clustering (Classification non supervisée)
  1. Classification (Classification supervisée) : $\mathbf{y}$ donné et catégoriel.
  2. Regression (Classification supervisée) : $\mathbf{y}$ donné et numérique.
  3. Clustering (Classification non supervisée) : $\mathbf{y}$ n'est pas donné et on doit en trouver un (catégoriel).





Le processus d'analyse des données (Data Mining)



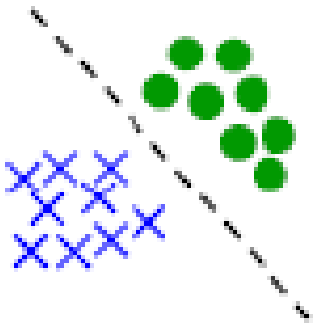
1. Recuperer les données et comprendre le domaine : installation des capteurs si nécessaire (camera, thermostat, base des données)
2. Nettoyage et pré-traitement: Nettoyage et sélection des données, découpage pour l'apprentissage, et le test, ... etc.
3. Transformation : extraction des features (sélection des variables caractéristique de description, discrétisation des variables numériques, projection et réduction des dimensions)
4. Modélisation (Data mining) : choix de l'approche la plus adaptée (classification, régression, clustering, ... etc.)
5. Validation et évaluation (cross-validation, mesures de performance)
6. Interprétation des résultats.
7. Répéter si besoin.

Exemples des données

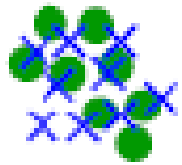
SepalLength	SepalWidth	PetalLength	PetalWidth	Class
5.4	3.0	4.5	1.5	Iris-versicolor
5.5	4.2	1.4	0.2	Iris-setosa
5.5	3.5	1.3	0.2	Iris-setosa
5.5	2.3	4.0	1.3	Iris-versicolor
5.5	2.4	3.8	1.1	Iris-versicolor
5.5	2.4	3.7	1.0	Iris-versicolor
5.7	2.5	5.0	2.0	Iris-virginica
5.6	2.8	4.9	2.0	Iris-virginica
5.5	2.5	4.0	1.3	Iris-versicolor
5.5	2.6	4.4	1.2	Iris-versicolor
5.6	2.9	3.6	1.3	Iris-versicolor
5.6	3.0	4.5	1.5	Iris-versicolor

Exemples de sélection des variables

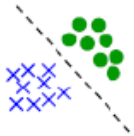
On apprend un modèle à partir des données. Il faut donc bien choisir les données.



"Good" features

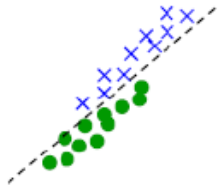


"Bad" features

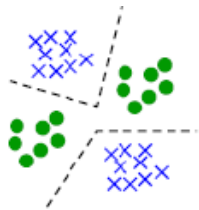


Linear separability

Non-linear separability



Highly correlated features



Multi-modal

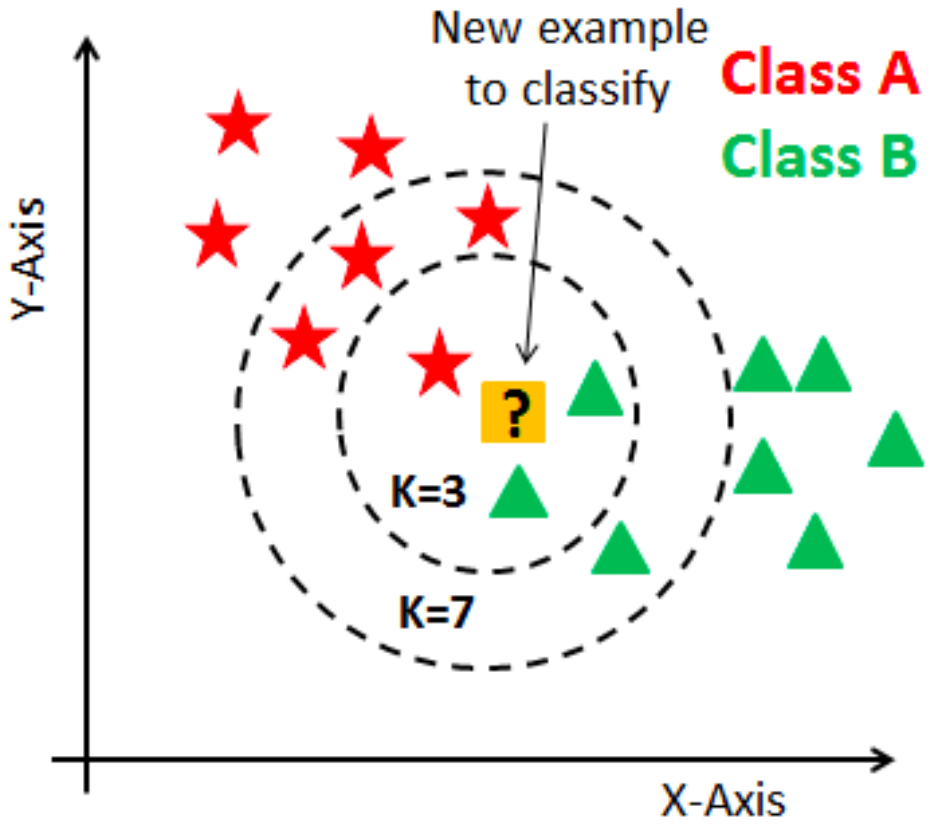
Exemples des transformations des données

1. Discrétisation de certaines variables.
2. ACP, ACM pour réduire la dimensionnalité et pour transformer les données (de catégorielles en numeriques par exemple).

Exemples des méthodes d'analyse des données (Data Mining)

1. Classification (Classification supervisée) : K-PPV
2. Régression (Classification supervisée) : Régression linéaire, K-PPV
3. Clustering (Classification non supervisée) : K-means et CAH

La méthode du K-plus proche voisin

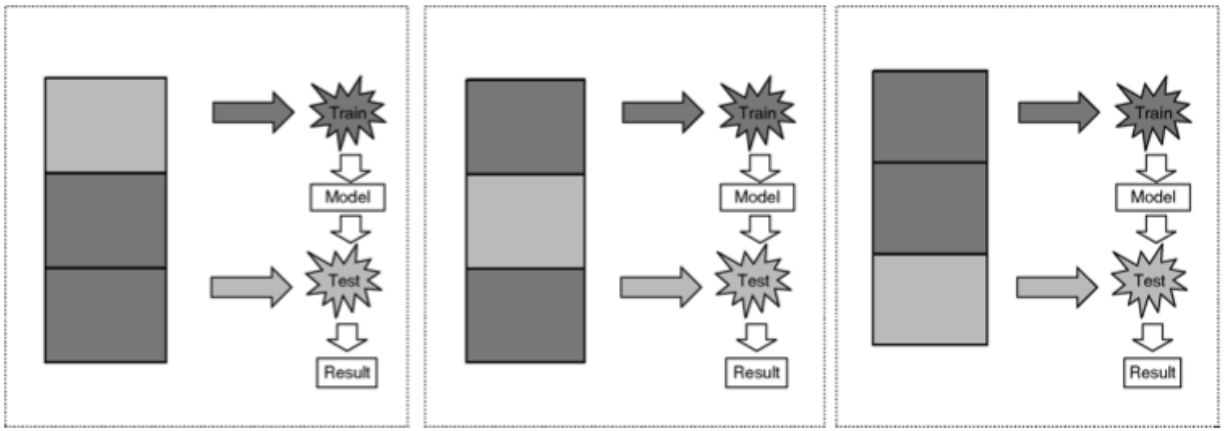


Évaluation et interprétation

- Non-Supervisée : interprétation des résultats et validation par des experts.
- Supervisé : il est souvent coutume de séparer les données en deux (apprentissage et test). Évaluation du modèle en apprentissage et en test (un ensemble n'ayant pas servi a la création du modèle) car un modèle peut être performant sur l'ensemble utilise pour sa création et moins performant sur les nouveaux jeux des données (sur-apprentissage).

Attention

Les performances en test dépendent de l'ensemble choisi pour le test. Pour éviter cet effet, on a souvent recours a la cross-validation.



Cross validation

Plusieurs approches pour faire la cross validation :

- Hold-Out Validation (2-Fold) : un ensemble de train (apprentissage) et un ensemble de test.
- K-Fold Cross-Validation : données séparées en K ensembles, a chaque fois, le modèle est appris sur $K-1$ ensembles et testé sur la K^{eme} partie.
- Leave-One-Out Cross-Validation : à chaque fois, le modèle est appris sur $N-1$ données et testé sur le N^{eme} individu (N modèles sont donc appris).
- K-Fold Cross-Validation répétée : répéter la K-Fold Cross-Validation un certain nombre de fois et utiliser une agrégation pour définir une mesure agrégée de performance (exemple : moyenne).

Double objectif

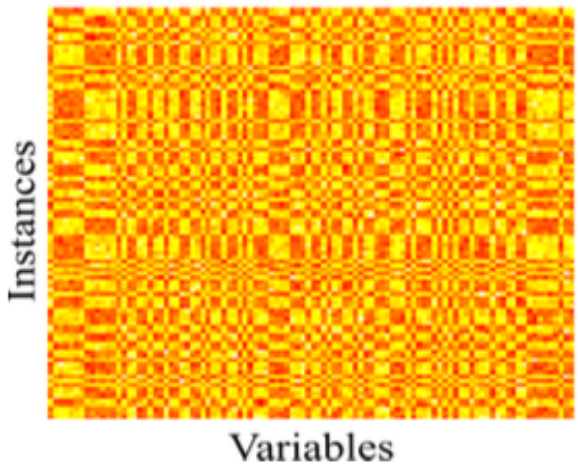
La cross-validation sert a juger la pertinence du modèle et sa capacité de generalisation (bonne performance sur des nouvelles données jamais vues en apprentissage) mais aussi à comparer deux ou plusieurs modèles.

Plan

Objectif et données

Objectif

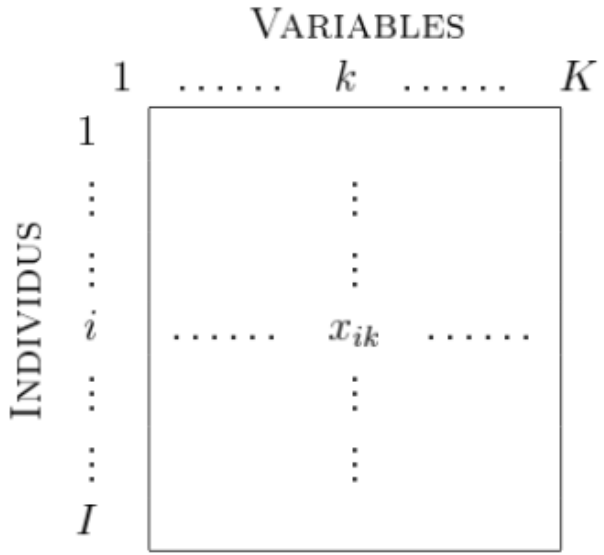
Rassembler les éléments (individus ou variables) qui se ressemblent et/ou de séparer ceux qui diffèrent. C'est-à-dire qu'il s'agit de créer des classes homogènes les plus éloignées les unes des autres $\Rightarrow$ mieux interpréter une grand quantité de données.



Données

Les données

Les données sont souvent organisées comme une matrice X où x_{ik} est la valeur de la variable k pour l'individu i, I représente le nombre d'individus et K représente le nombre de variables.



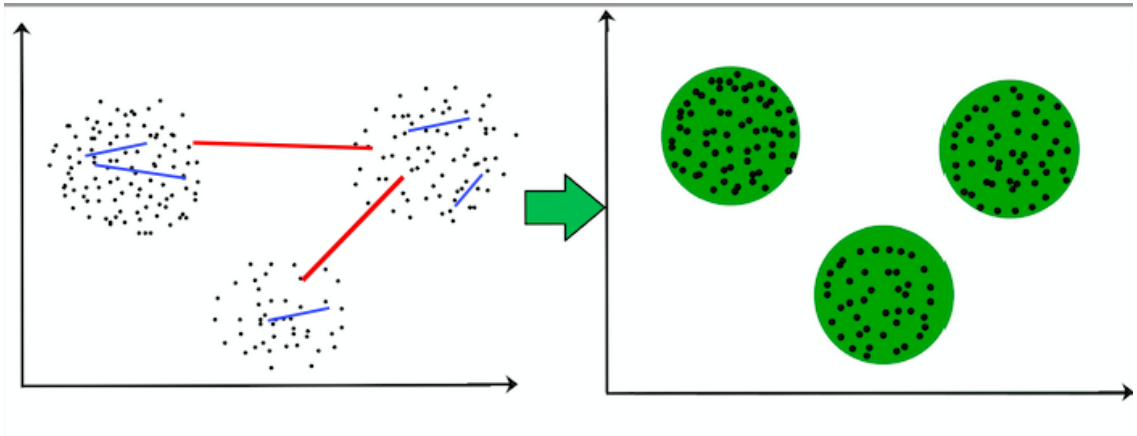
- Les variables peuvent être quantitatives continues ou issues de tableaux de contingences, ou binaires issues de tableaux logiques, ou encore qualitatives.
- Grouper des individus ou variables qui se ressemblent $\Rightarrow$ Il faut choisir une mesure de similarité ou dissimilarité et qui doit être adaptée aux types de données.

Mesures de similarité

Une mesure de dissimilarité

Une mesure de dissimilarité est une fonction d similaire à une distance à l'exception que l'inégalité triangulaire n'est pas exigée. Ces mesures peuvent être des distances dans le cas de variables quantitatives. Sur un ensemble E, la dissimilarité est une fonction de $E \times E$ dans R qui vérifie

- Positive $d(x, y) \geq 0, \ \forall \ x \in E, \forall \ y \in E$
- Identique : $d(x, y) = 0$ ssi $x = y$.
- Symmetric : $d(x, y) = d(y, x)$.



- Une similarité mesure la ressemblance entre les individus alors que la

Exemples

1. Variables quantitatives :
   - Distance euclidienne (la distance usuelle dans R^K) :

$$d(x,y) = \sqrt{\sum_{i=1}^{K} (x_i - y_i)^2}.$$

   - Distance de Minkowsky (generalisation de la distance euclidienne):

$$d(x,y) = \Big(\sum_{i=1}^{K} | x_i - y_i |^n \Big)^{\frac{1}{n}}.$$

   Pour $n = 2$, la distance de Minkowsky devient la distance euclidienne.
   Pour $n = 1$, elle correspond a la distance de Manhattan :

$$d(x,y) = \Big(\sum_{i=1}^{K} | x_i - y_i | \Big)$$

   - Distance de Mahalanobis (diffère de la distance euclidienne par le fait qu'elle prend en compte la variance et la corrélation):

$$d(x) = \sqrt{(x-\mu)^T \Sigma^{-1} (x-\mu)}$$

Pearson correlation	$D_{ij} = (1 - r_{ij})/2$, where $r_{ij} = \dfrac{\sum_{l=1}(x_{il} - \overline{x_i})(x_{jl} - \overline{x_j})}{\sqrt{\sum_{l=1}(x_{il} - \overline{x_i})^2 \sum_{l=1}(x_{jl} - \overline{x_j})^2}}$
Point symmetry distance	$D_{ir} = \min_{\substack{j=1,\ldots,N \\ \text{and } j \neq i}} \dfrac{\|(\mathbf{x}_i - \mathbf{x}_r) + (\mathbf{x}_j - \mathbf{x}_r)\|}{\|(\mathbf{x}_i - \mathbf{x}_r)\| + \|(\mathbf{x}_j - \mathbf{x}_r)\|}$
Cosine similarity	$S_{ij} = \cos\alpha = \dfrac{\mathbf{x}_i^T \mathbf{x}_j}{\|\mathbf{x}_i\| \|\mathbf{x}_j\|}$

Les trois dernières distances sont plus adaptées pour mesurer la similarité entre les variables.

2. Tableau de contingence : distance du χ^2

$$\chi^2(l,k) = \sqrt{\sum_j \frac{1}{x_{+j}} \left(\frac{x_{lj}}{x_{l+}} - \frac{x_{kj}}{x_{k+}} \right)^2}$$

ou $x_{+j} = \sum_i x_{ij}$ et $x_{i+} = \sum_j x_{ij}$ représentant les sommes sur les lignes et colonnes, respectivement.