Introduction au Machine Learning (ML)

ASSOHOUN Egomli Stanislas

M2 Informatique

UFR Environnement -Département PCMI - Université Jean Lorougnon Guédé de Daloa (UJLoG)

PRESENTATION

- Cours :
  - ***Responsable*** : ASSOHOUN Egomli Stanislas
  - ***Contact*** :
    - **49016196**
    - **Stanislas.assohoun@ujlg.edu.ci**
  - ***Cours*** :
    - **CM : 4 séances**
    - **TD-TP : 6 séances**

Présentation

- Resources et évaluations :
  - Ressources
    - Support de cours
    - Support de TD - TP
    - articles
    - Python, anaconda, miniconda, conda
  - Contrôles :
    - CC : 1 projet + 1 contrôle de TP
    - Examen (1ère session) :
    - Examen (2eme session)
    - TP : 1 projet + 1 contrôle de TP

Introduction

Ce cours a pour but de vous aider à acquérir les bases du Machine Learning (Aprentissage Machine). Vous y trouverez les outils pour construire des modèles pour résoudre des problèmes business basés sur des scénarios du monde réel. Vous apprendrez comment tirer profit du Machine Learning pour créer de la valeur ajoutée dans ces problèmes business.

Dans le cadre de cette introduction allez apprendre:

- ✓ Avoir la bonne intuition du Machine Learning
- ✓ Implémenter des modèles de Machine Learning sur Python
- ✓ Créer de la valeur ajoutée dans des problèmes business grâce au Machine Learning
- ✓ Faire des prédictions précises
- ✓ Faire du clustering
- ✓ Gérer et tirer des idées, des notions et des concepts à partir des données

Machine Learning, pourquoi ?

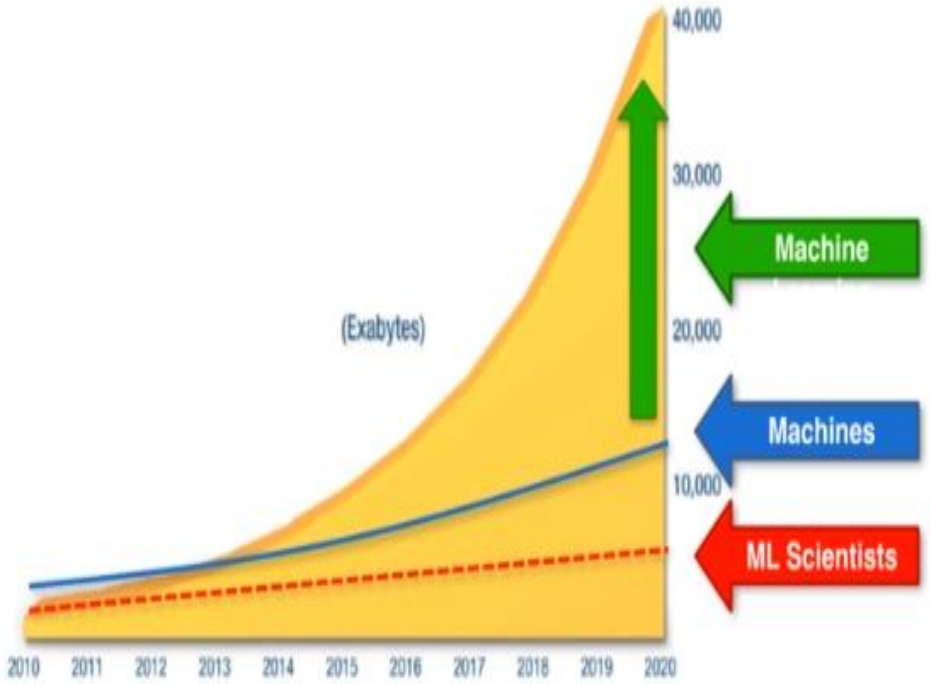
2005 – 130 EXABYTES

2010 – 1,200 EXABYTES

2015 – 7,900 EXABYTES

2020 – 40,900 EXABYTES

Les données dans le monde augmentent de manière exponentiel.



- **Croissance des données de 2010 – 2020 (x 40)**
- **Capacité de traitement des machine actuelle (un peut plus de 10 000 EB**
- **Capacité de traitement des scientist ML (≈ 5 0000 EB)**
- **Nécessité d'aligner machine et scientist ML sur les quantités des données**

Plan

- CH 1 : Généralités
- CH 2 : Régression
- CH 3 :Classification
- CH 4 : Clustering

CH I- GENERALITES

CH – I : GENERALITES

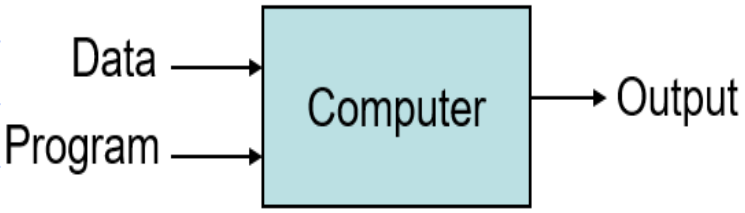
I-1 Qu'est-ce que l'apprentissage machine?

✓ un des champs d'étude de l'intelligence artificielle ,

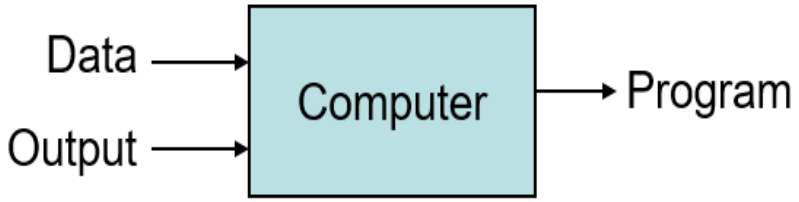
✓la discipline scientifique concernée par le développement, l'analyse et l'implémentation de méthodes automatisables qui permettent à une machine (au sens large) d'évoluer grâce à un processus d'apprentissage

✓permet de remplir des tâches qu'il est difficile ou impossible de remplir par des moyens algorithmiques plus classiques

Programmation traditionnelle



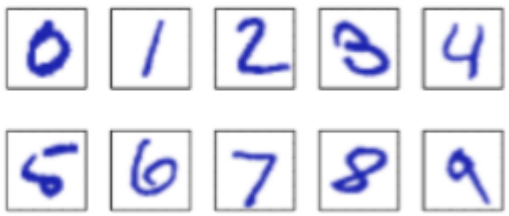
Machine Learning



CH – I : GENERALITES

I-1 Qu'est-ce que l'apprentissage machine?

Exemple : Comment reconnaître des caractères manuscrits?



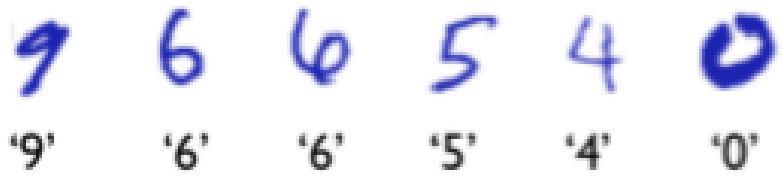
- par énumération de règles
  - si intensité pixel à la position ... alors c'est un « 4 » ,
  - long et fastidieux, difficile de couvrir tous les cas
- en demandant à la machine d'apprendre
  - lui laisser faire des essais et apprendre de ses erreurs ,
  - apprentissage machine (machine-learning)

CH – I : GENERALITES

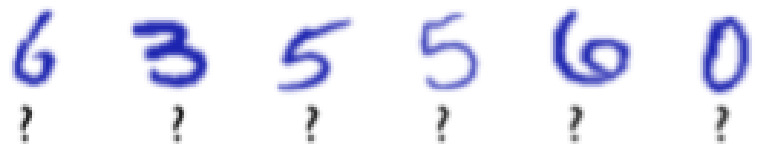
I-1 Qu'est-ce que l'apprentissage machine?

Comment ca marche :

- On donne à l'algorithme des données d'entrainement



- l'algorithme d'apprentissage machine apprend un modèle capable de généraliser à de nouvelles données.



CH – I : GENERALITES

I-1 Qu'est-ce que l'apprentissage machine?

Notations:

- On appel ensemble d'entraînement (trainning set) :
  - D_{train} $\{(x_1,t_1), \ldots ,(x_N,t_N)\}$
    - x_n une observation (entrée du système)
    - t_n la cible correspondante (sortie du système)
- L'apprentissage machine fournit un modèle y(x) qui prédit t en fonction de x : $y(x_n) = \hat{t}_n$
- L'objectif est de trouver un modèle tel que : $y(x_n) = \hat{t}_n \approx t_n$
- On mesure la qualité de l'apprentissage (la qualité du modèle) sur un ensemble de test (test set) :
  - D_{test} $\{(x_{N+1},t_{N+1}), \ldots ,(x_{N+M},t_{N+M})\}$

CH – I : GENERALITES

I-2 Deux types d'apprentissage

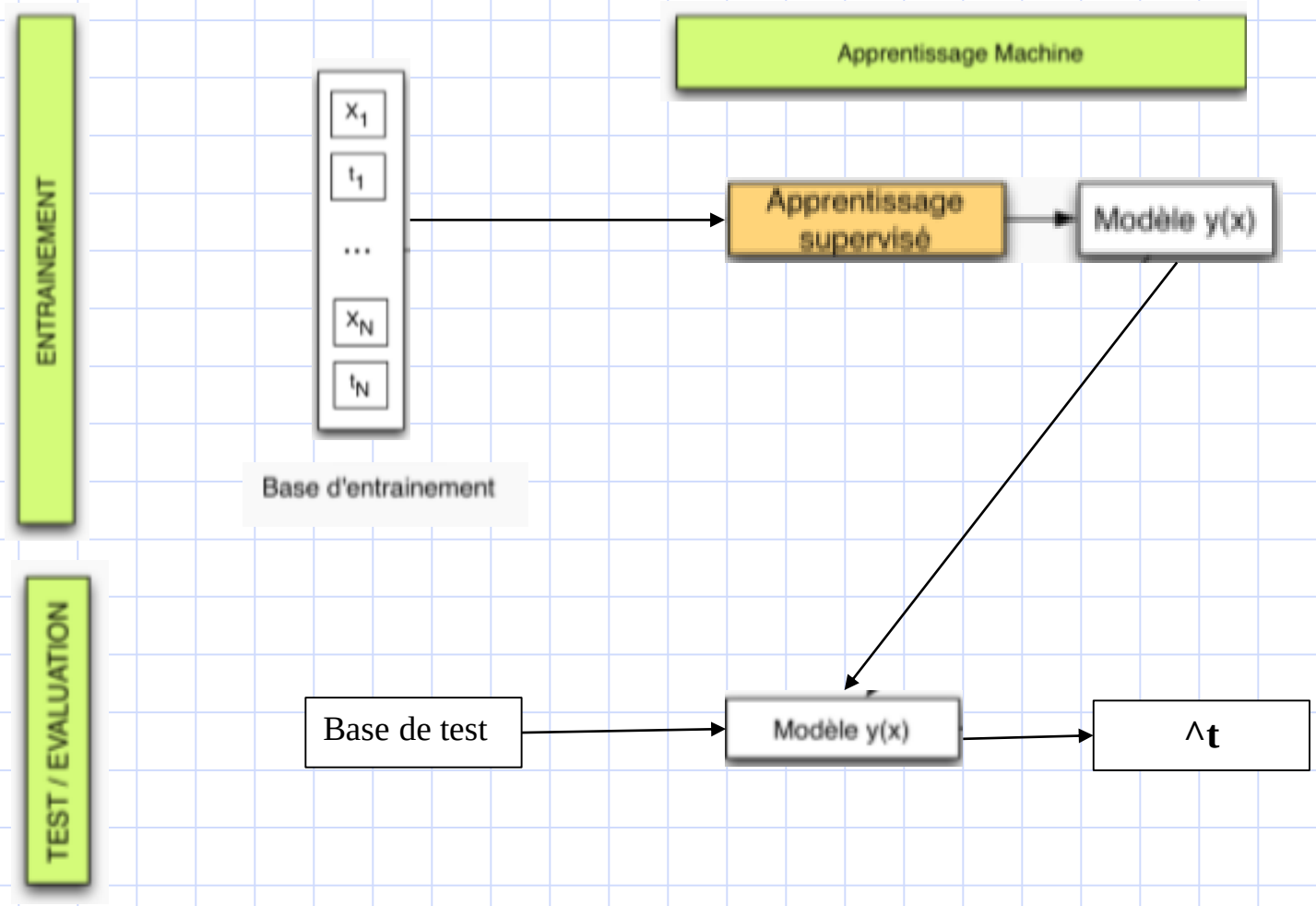
Apprentissage supervisé

- Nous considérons un ensemble d'observations (entrées du système) $\{(x_1, \ldots, x_N)\}$
  - Nous donnons également à la machine les cibles (sorties du système) souhaitées $(t_1, \ldots, t_N)$
  - Dtrain {(x1,t1), ... ,(xN,tN)}
- L'objectif de la machine est d'apprendre les cibles (sorties) correctes pour de nouvelles observations (entrées)

CH – I : GENERALITES

I-2 Deux types d'apprentissage

Exemple d'apprentissage supervisé



CH – I : GENERALITES

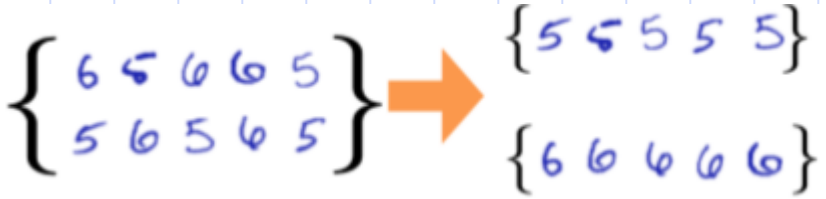
I-2 Deux types d'apprentissage

Apprentissage non-supervisé

Nous considérons un ensemble d'observations (entrées du système)

$\{(x_1, \ldots, x_N)\}$

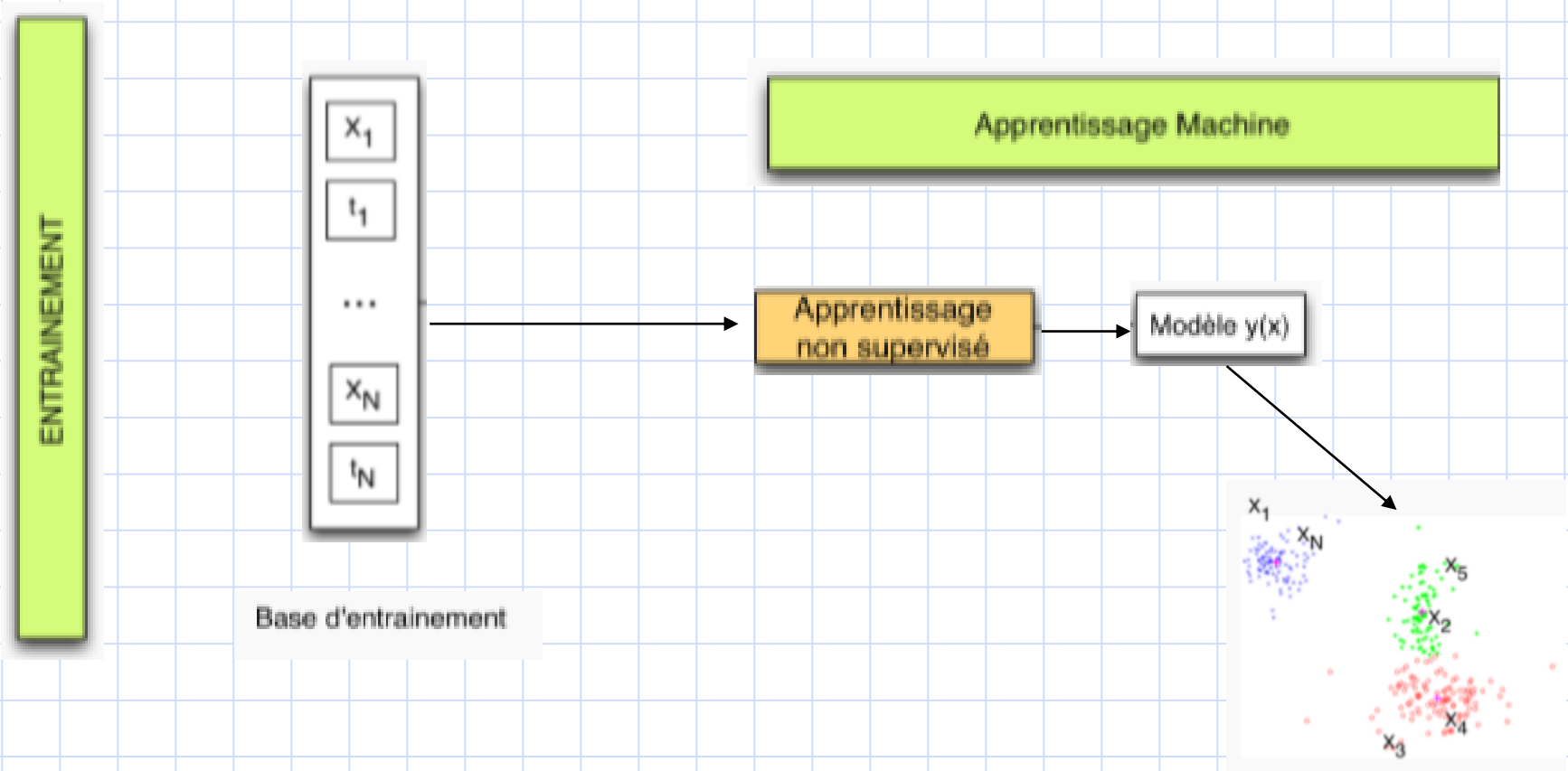
- Nous ne donnons pas à la machine les cibles
- Dtrain {x1, ... ,xN}
- L'objectif de la machine est de créer un modèle de x, un partitionnement (clustering) des données
  - Utilisation? analyse de données, prise de décisions



CH – I : GENERALITES

I-2 Deux types d'apprentissage

Exemple d'apprentissage non supervisé



CH – I : GENERALITES

I-3 Deux grandes cibles pour l'apprentissage supervisé

La régression

La cible est un nombre réel : t_n appartient à IR

Exemples :

- Economie (prédiction de valeur en bourse) :
  - x= activité économique de la journée, t= la valeur d'une action demain
- Audio (reconnaissance de tempo) :
  - x= le contenu spectral du signal, t= le tempo du morceau

La classification

La cible est un indice de classe : t_n appartient à $\{1, \ldots, C_n\}$

Exemples :

- Image (reconnaissance de caractères)
  - x= vecteur d'intensité des pixels, t= l'identité du caractère
- Audio (reconnaissance de parole) :
  - x : le contenu spectral du signal audio, t= le phonème prononcé

CH – I : GENERALITES

I-4 trois grandes methodes pour l'apprentissage supervisé

Système de communication

Imaginons un système de communication dont l'entrée est Y et la sortie X.

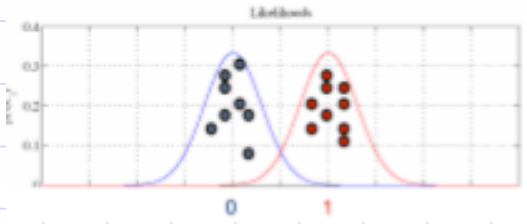
- on observe uniquement la sortie X
- on souhaite retrouver Y (non-observable) à partir de X (observable)
- → on infère Y à partir de X

CH – I : GENERALITES

I-4 trois grandes méthodes pour l'apprentissage supervisé

Solution1: Approche générative

- On apprend la fonction qui génère
  - les valeurs de X quand Y = 0 : P(X|Y = 0)
  - les valeurs de X quand Y = 1 : P(X|Y =1)
  - → on infère Y à partir de X



- On en déduit les probabilités P(Y = 0|X) et P(Y =1|X)
- On décide que Y = 0 si P(Y =0|X) > P(Y =1|X)
- Ceci conduit à une fonction de décision g(x)
  - g(x) est une conséquence des modèles génératifs

En résumé :

nous partons de l'hypothèse qu'il existe une famille de modèles paramétriques permettant de générer X connaissant Y

Exemple : apprentissage Bayesien, modèle de Markov caché, réseaux de neurones artificiels.

CH – I : GENERALITES

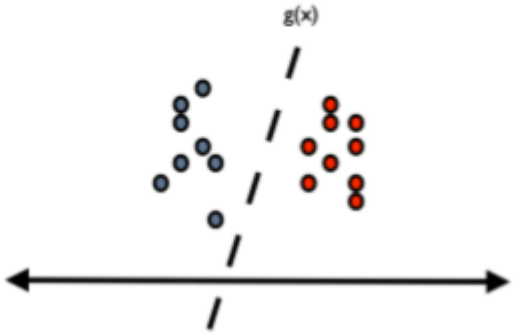
I-4 trois grandes méthodes pour l'apprentissage supervisé

Solution2: Approche discriminante

❑On apprend directement la fonction de décision g(x) qui sépare le mieux

- les valeurs de X correspondant à Y = 0 et
- les valeurs de X correspondant à Y = 1

→ On ne considère pas la manière dont X est généré à partir de Y



En résumé :

❑ nous n'avons pas d'hypothèse sur le modèle sous-jacent à X mais nous étudions comment séparer ses valeurs

Exemple : analyse linéaire discriminante, machine à vecteur support (SVM)

CH – I : GENERALITES

I-4 trois grandes méthodes pour l'apprentissage supervisé

Solution3: Approche par exemplification

- On possède une série d'exemples de couples assignant une observation X à une cible Y :

D_{train} {(x1,t1), ... ,(x_N,t_N)}

- pour une nouvelle observation x*, on cherche les observations X de la base d'entrainement les plus proches de x*,
- on assigne à x* le y correspondant aux X les plus proches
- Exemple : K-plus-proche-voisin

CH – I : GENERALITES

I-5 Résumé

Deux grands types d'apprentissage

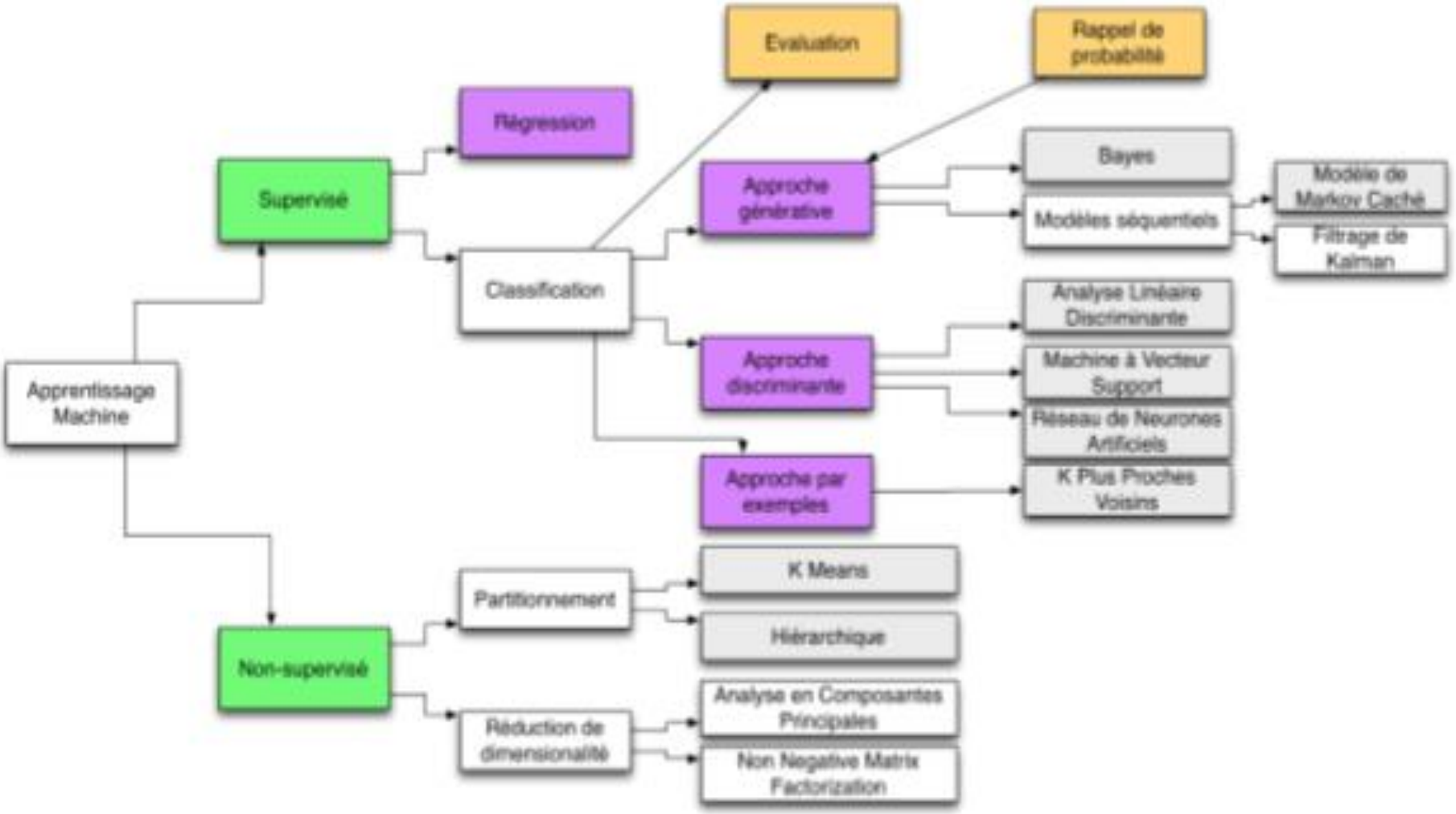
Supervisé → deux grandes cibles

- Régression
- Classification → trois grandes approches
  - Approche générative :
    - nous partons de l'hypothèse qu'il existe une famille de modèles paramétriques permettant de générer X connaissant Y
    - Exemple : apprentissage Bayesien, modèle de Markov caché
  - Approche discriminante :
    - nous n'avons pas d'hypothèse sur le modèle sous-jacent à X mais nous étudions comment les séparer
    - Exemple : analyse linéaire discriminante, machine à vecteur support (SVM), réseaux de neurones artificiels (ANN, Deep Learning)
  - Approche par exemplification :
    - K-plus-proche-voisin

- Non-supervisé

CH – I : GENERALITES

I-5 Résumé



CH – I : GENERALITES

I-6 Généralisation

❑Apprentissage

▪Apprendre un modèle (génératif ou discriminant) à partir des observations X et des valeurs à prédire Y

▪Le modèle doit permettre une bonne prédiction de Y en fonction des observations X

❑ Généralisation

▪Capacité du modèle à prédire correctement des valeurs Y* en fonction de X* en dehors de l'ensemble d'apprentissage

▪Sur-apprentissage (over-fitting)

▪ En pratique on évalue les performances d'un modèle appris en séparant :

▪Ensemble d'entrainement (training-set) : {X,Y}

▪ Ensemble d'entrainement (training-set) : {X*,Y*}

NB :

▪ Sous-apprentissage : grande perte sur l'ensemble d'entraînement

▪Sur-apprentissage : , apprentissage "par coeur" de l'ensemble d'entraînement

CH – I : GENERALITES

III- Bien démarrer un projet de machine learning

Un projet de machine learning commence généralement avec un jeu de données et un problème à résoudre. Celui-ce se décrit par trois éléments, des données (X), une cible (Y) et une fonction d'erreur qui permette d'évaluer la distance entre la prédiction et la cible. Une fois qu'on a cela, les premières étapes débutent avec presque toujours les mêmes questions :

- Etape 1 : quel est le type de problème ?
  - Supervisé : régression, classification, ranking, ...
  - non supervisé : clustering, réduction du nombre de dimension, système de recommandations, ...

Il n'est pas rare qu'un projet requiert un assemblage de modèles de types différents. La première étape consiste à imaginer un chemin entre les données initiales et la valeur à prédire. Le problème est rarement non supervisé car on cherche le plus souvent à reproduire un processus humain. L'aspect non supervisé intervient sous la forme d'une étape intermédiaire.

- Etape 2 : quelles sont les données ? (1/3)
  - Est-ce une table classique ou un graphe ?
  - Y a-t-il une dimension temporelle ?

CH – I : GENERALITES

III- Bien démarrer un projet de machine learning

❑ Etape 2 : quelles sont les données ? (2/3)

- Est-ce une table classique ou un graphe ?
- Y a-t-il une dimension temporelle ?
- Nombre d'observations ?
- Nombre de variables (ou features) ?
- Quelles sont les variables connues, les variables à prédire ?
- Valeurs manquantes ?
- Variables catégorielles, discrètes ou continues, Encoder les catégories ?
- Corrélations avec la cible à prédire ?

La plupart des algorithmes d'apprentissages utilisent des données numériques, il faut convertir les variables catégorielles au format numérique.

❑ Etape 3 : séparation train/test

Il faut faire attention à deux ou trois détails. Par exemple, si le problème est un de problème de classification, il faut faire attention que toutes les classes à prédire sont bien représentées dans les deux bases.

CH – I : GENERALITES

III- Bien démarrer un projet de machine learning

❑ Etape 4 : apprentissage d'un modèle

Apprendre un modèle tout de suite pour avoir une idée de la difficulté du problème. On privilégiera les modèles linéaires et les arbres et décisions si on souhaite obtenir un modèle interprétable. On optera pour les forêts aléaoires dans les autres cas. Ces modèles présentent l'avantage de s'apprendre rapidement et de marcher sur tout type de données, discrètes continues…

❑ Etape 5 : mesure de la performance

On mesure la performance du modèle sur la base de test. Il existe certaines façons standard de le faire en fonction des types de problèmes :

- Classification : matrice de confusion, courbe ROC, précision / rappel, ...
- Régression : erreur de prédiction, graphe XY valeur à prédire / valeur prédite, ...
- Clustering

Un modèle peut être considéré comme bon par un indicateur (R^2 par exemple) et pourtant ne pas être assez bon pour l'usage qu'on doit en faire (prédictions de séries temporelles). Si la performance globale convient, on s'arrête souvent ici. Dans le cas contraire, il faut retourner à l'étape 4 :

CH – I : GENERALITES

III- Bien démarrer un projet de machine learning

- La base d'apprentissage contient peut-être des points aberrants.
- Le modèle a besoin de plus de variables, combinaison non linéaires des variables existantes (polynômes, fonctions en escalier, …), recoupement de la base de données avec une autre base.
- Les valeurs manquantes empêchent le modèle d'apprendre.

❑ Etape 6 : ajouter des variables

- Passer au logarithme lorsque les variables ont des valeurs extrêmes, cela réduit leur importance.
- Si les données peuvent être groupées : ajouter des moyennes, somme, nombre par groupes.
- Chercher l'information qui pourrait aider un modèle à corriger une erreur en particulier.

❑ Etape 7 : validation du modèle

On regarde sur quelques exemples bien choisis que le modèle proposent une réponse acceptable. On applique des méthodes du type validation croisée.

CH II- REGRESSION ET ARBRES DE DECISION

CH – II : A-REGRESSIONS

La régression utilise une ou plusieurs variables explicatives (x) pour prédire une variable de réponse (y). La partie "simple" est que nous n'utiliserons qu'une seule variable explicative. S'il existe deux variables explicatives ou plus, une régression linéaire multiple est nécessaire. En régression, la variable explicative est toujours x et la variable de réponse est toujours y. Tous les deux et doivent être des variables quantitatives.

I- La régression linéaire

I-1- Présentation

L'algorithme de régression linéaire est un algorithme d'apprentissage supervisé c'est-à-dire qu'à partir de la variable cible ou de la variable à expliquer (Y), le modèle a pour but de faire une prédiction grâce à des variables dites explicatives (X) ou prédictives.

La variable cible (Y) est quantitative tandis que la variable X peut être quantitative ou qualitative.

L'objectif est de trouver une fonction dite de prédiction ou une fonction coût qui décrit la relation entre X et Y c'est-à-dire qu'à partir de valeurs connues de X, on arrive à donner une prédiction des valeurs de Y.

La fonction recherchée est de la forme : Y= f(X) avec f(X) une fonction linéaire

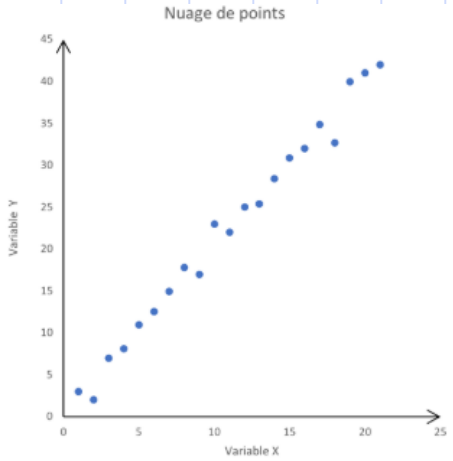
CH – II : A- REGRESSIONS

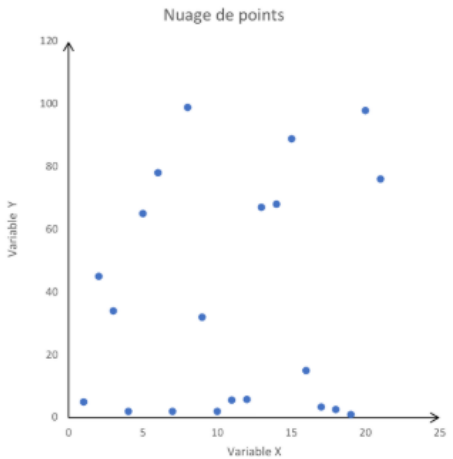
À partir d'un échantillon qui représente notre data set (jeu de données), on répartit les données en deux groupes, les données d'entraînement et les données de test.

La première catégorie de données servira pendant la phase d'apprentissage du modèle alors que le second sera utilisé pour évaluer la qualité de prédiction du modèle. Le but n'est donc pas de construire une fonction qui prédira avec une précision optimale les valeurs des variables cibles mais une fonction qui se généralisera au mieux pour prédire des valeurs de données qui n'ont pas encore été observées.

I-2- Représentation graphique

Le but est de savoir si le modèle linéaire est oui ou non pertinent pour l'étude de notre phénomène. Le graphique est au départ un nuage de points et on relève la tendance qu'a la forme de ce nuage de points.





Au vu de ces deux graphiques, il semble approprié d'utiliser le modèle linéaire pour la première image et pas pour la deuxième qui ne laisse transparaitre aucune tendance connue.

CH – II : A- REGRESSIONS

II- Modelisation

II-1- Modélisation de la régression linéaire

Modélisation	Nature de la régression
Une seule variable explicative X	Régression simple
Plusieurs variables explicatives X_j $(j=1,\ldots,q)$	Régression multiples

Le modèle de régression linéaire analyse les relations entre la variable dépendante ou variable cible Y et l'ensemble des variables indépendantes ou explicatives X. Cette relation est exprimée comme une équation qui prédit les valeurs de la variable cible comme une combinaison linéaire de paramètres.

❑ Modèle linéaire simple

$Y=aX+b+\varepsilon$ où $f(X)= aX+b$

avec :

- Y, la variable cible, aléatoire dépendante
- a et b, les coefficients (pente et ordonnée à l'origine) à estimer
- X, la variable explicative, indépendante
- ε, une variable aléatoire qui représente l'erreur

CH – II : A- REGRESSIONS

II- Modelisation

II-1- Modélisation de la régression linéaire

Modélisation	Nature de la régression
Une seule variable explicative X	Régression simple
Plusieurs variables explicatives X_j $(j=1,\ldots,q)$	Régression multiples

Le modèle de régression linéaire analyse les relations entre la variable dépendante ou variable cible Y et l'ensemble des variables indépendantes ou explicatives X. Cette relation est exprimée comme une équation qui prédit les valeurs de la variable cible comme une combinaison linéaire de paramètres.

❑ Modèle de regression linéaire simple

$Y=aX+b+\varepsilon$ où $f(X)= aX+b$

avec :

- Y, la variable cible, aléatoire dépendante
- a et b, les coefficients (pente et ordonnée à l'origine) à estimer
- X, la variable explicative, indépendante
- ε, une variable aléatoire qui représente l'erreur

CH – II : A-REGRESSIONS

- Modèle de régression linéaire multiple

$Y=ax1+bx2+cx3+\cdots+K+\varepsilon$ où $f(X)= aX+b$

avec :

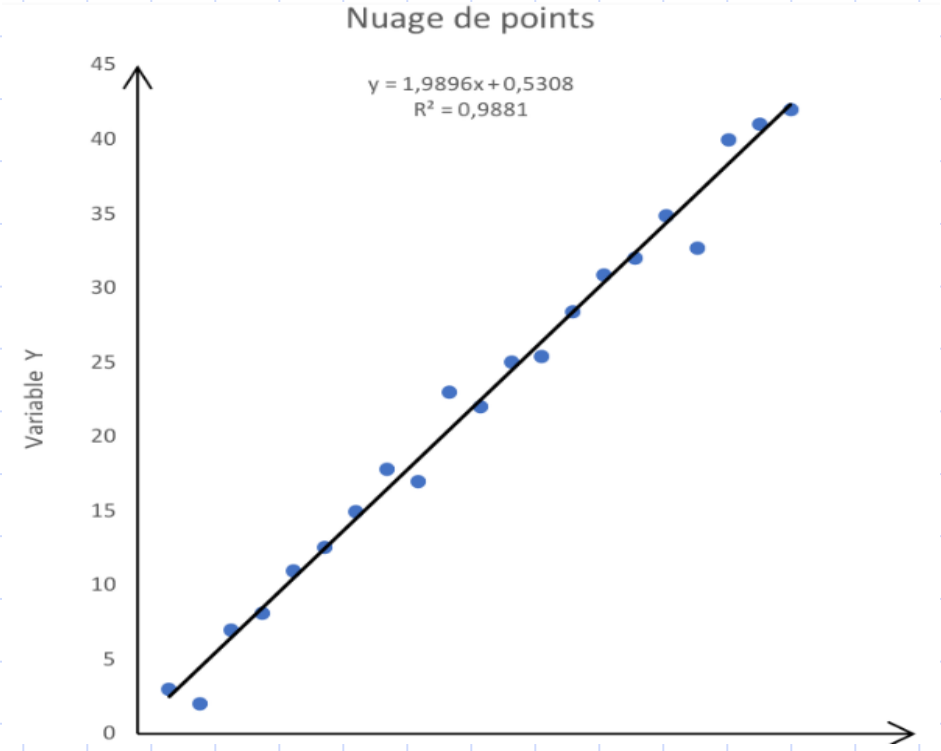
- Avec :Y, la variable cible, aléatoire dépendante
- a,…,K les coefficients (pente et ordonnée à l'origine) à estimer
- X=(x1,…,xq), la variable explicative, indépendante
- ε, une variable aléatoire qui représente l'erreur

ε est appelé résidus, c'est l' erreur commise, c'est-à-dire l'écart entre la valeur Yi observée et la valeur aiXi+b donnée par la relation linéaire.

En effet, même si une relation linéaire est effectivement présente, les données mesurées ne vérifient pas en général cette relation exactement. Pour ce faire, on tient compte dans le modèle mathématique des erreurs observées.

CH – II : A- REGRESSIONS

- Droite des moindres carré et coefficient de corrélation de Bravais-Pearson au carré



Sur ce graphique, la droite de régression linéaire ou la droite des moindres carrés de Y en X représente la droite d'ajustement linéaire, celle qui résume le mieux la structure du nuage de points pendant la phase d'apprentissage.

Elle rend minimale la somme des carrés des erreurs d'ajustement

CH – II : A- REGRESSIONS

- Droite des moindres carré et coefficient de corrélation de Bravais-Pearson au carré

Le terme R^2 de l'image représente le coefficient de corrélation de Bravais-Pearson au carré.

Ce coefficient mesure l'intensité de la relation linéaire entre Y et X .

Le coefficient de corrélation est un nombre toujours compris entre -1 et 1.

- Si R est proche de 1 : il y a une forte liaison linéaire entre les variables et les valeurs prises par Y ont tendance à croître quand les valeurs de X augmentent.
- Si R est proche de 0 : il n'y a pas de liaison linéaire
- Si R est proche de -1 : il y a une forte liaison linéaire et les valeurs prises par Y ont tendance à décroître quand les valeurs de X augmentent.

Le coefficient de corrélation mesure la qualité de la droite d'ajustement linéaire mais ne représente en aucun cas une cause de la relation logique entre X et Y. Seul le data scientist pourra estimer de la relation logique entre les deux variables.

Bien que le coefficient de corrélation soit supérieur à 0, il n'y a aucun lien logique entre les deux phénomènes. Il faut donc faire une distinction entre corrélation et causalité.

CH – II : A- REGRESSIONS

- Droite des moindres carré et coefficient de corrélation de Bravais-Pearson au carré

Le terme R^2 de l'image représente le coefficient de corrélation de Bravais-Pearson au carré.

Ce coefficient mesure l'intensité de la relation linéaire entre Y et X .

Le coefficient de corrélation est un nombre toujours compris entre -1 et 1.

- Si R est proche de 1 : il y a une forte liaison linéaire entre les variables et les valeurs prises par Y ont tendance à croître quand les valeurs de X augmentent.
- Si R est proche de 0 : il n'y a pas de liaison linéaire
- Si R est proche de -1 : il y a une forte liaison linéaire et les valeurs prises par Y ont tendance à décroître quand les valeurs de X augmentent.

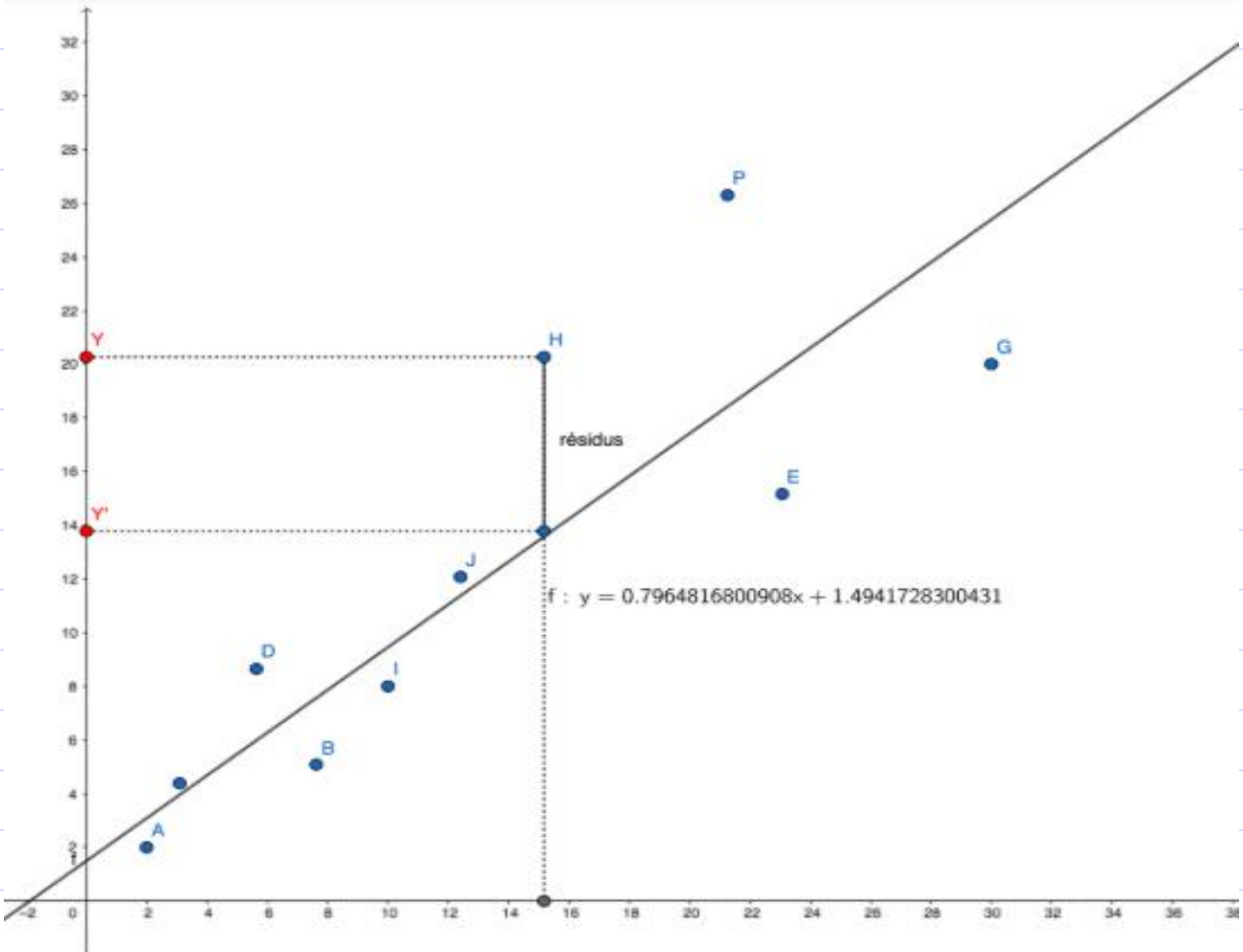
Le coefficient de corrélation mesure la qualité de la droite d'ajustement linéaire mais ne représente en aucun cas une cause de la relation logique entre X et Y.

Bien que le coefficient de corrélation soit supérieur à 0, il n'y a aucun lien logique entre les deux phénomènes. Il faut donc faire une distinction entre corrélation et causalité.

CH – II : A- REGRESSIONS

II-2- estimation des coefficients par la méthode des moindre carré

Nous expliquons comment ces paramètres sont ajustés afin d'estimer la variable de sortie Y.



CH – II : A- REGRESSIONS

II-2- estimation des coefficients par la méthode des moindre carré

Le principe des moindres carrés ordinaires consiste à choisir les valeurs de a et b qui minimisent les erreurs de prédiction ou les résidus sur un jeu de données d'apprentissage:

$$\varepsilon = \sum_{i=0}^{P} \left(Y_i - (aX_i + b)\right)^2$$

Minimiser cette expression revient à résoudre un problème d'optimisation, voici la forme des estimateurs notés â et $\hat{b}$ qui sont égaux à

$$\hat{a} = \frac{\sum_{i=1}^{p}(X_i - \bar{X})(Y_i - \bar{Y})}{(X_i - \bar{X})^2} = \frac{c_{xy}}{s_x^2}$$

$$\hat{b} = \bar{Y} - \hat{a}\bar{X}$$

Où c_{xy} est la covariance empirique entre les X_i et les Y_i et S_x^2 est la variance empirique des Xi.

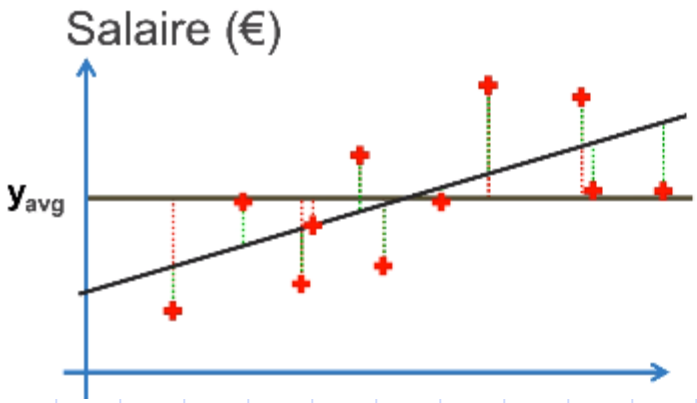
L'expression de $\hat{b}$ indique que la droite de régression linéaire passe par le centre de gravité du nuage de points $(\bar{X},\bar{Y})$.

CH – II : A- REGRESSIONS

III- Evaluer un modèle de régression

III-1 coefficient de détermination

Régression Linéaire Simple:



$$SS_{res} = SUM (y_i - \hat{y}_i)^2$$

$$SS_{tot} = SUM (y_i - y_{avg})^2$$

$$R^2 = 1 - \frac{SS_{res}}{SS_{tot}}$$

SS_{res} = somme des carrés des résidus (qui est a minimiser)

SS_{tot} = somme des carrés total (sommes de carrés des résidus par rapport a la moyenne) qui est toujours >0

R^2 nous dit que plus on parvient a minimiser SS_{res} plus R^2 sera proche de 1 et notre droite de régression sera proche de l'ensemble de observation.

R^2 – Qualité de la prediction

CH – II : A- REGRESSIONS

III- Evaluer un modèle de régression

III-2 Adjusted R^2

Remarques:

- Plus on ajoute des variables R^2 ne va jamais diminuer
- Quelque soit la variable ajouté (meme si elle ne contribue pas a ameliorer le modele de prediction), il aura tendance a augmenter le R^2 · Pour régler ce problème on utilisera le Adjusted R^2

$$\text{Adj } R^2 = 1 - (1 - R^2)\frac{n-1}{n-p-1}$$

p - nombre de régresseurs
n – taille de l'échantillon

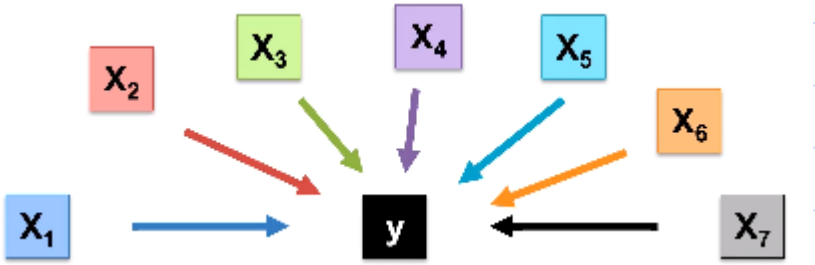
- Le Adjusted R^2 contient un facteur de pénalisation pour réguler l'augmentation du R^2 lorsqu'on ajoute une nouvelle variable qui n'ameliore pas notre modèle.

- Si on ajoute un redresseur qui améliore significativement la qualité du modèle alors le R^2 va considérablement augmenter bien plus que le ration (n-1)/(n-p-1) ne va le faire diminuer. Donc Le Adjusted R^2 n'augmente qui si la variable indépendante ajoutée est significative

CH – II : A- REGRESSIONS

IV- Sélection des variables explicatives

En présence de p variables explicatives dont on ignore celles qui sont réellement influentes, on doit rechercher un modèle d'explication de Y à la fois performant (résidus les plus petits possibles) et économique (le moins possible de variables explicatives).



- Quelles variables choisir.
- Comment choisir de manière optimale les variables à intégrer au modèle.

IV-1 Cinq (05) méthodes pour construction de modèle

1-All-in

2-Backward Elimination

3-Forward selection

4-Bidirectional elimination

5-Score comparison

NB : 1-2-3 la stepwise régression ou méthodes pas à pas

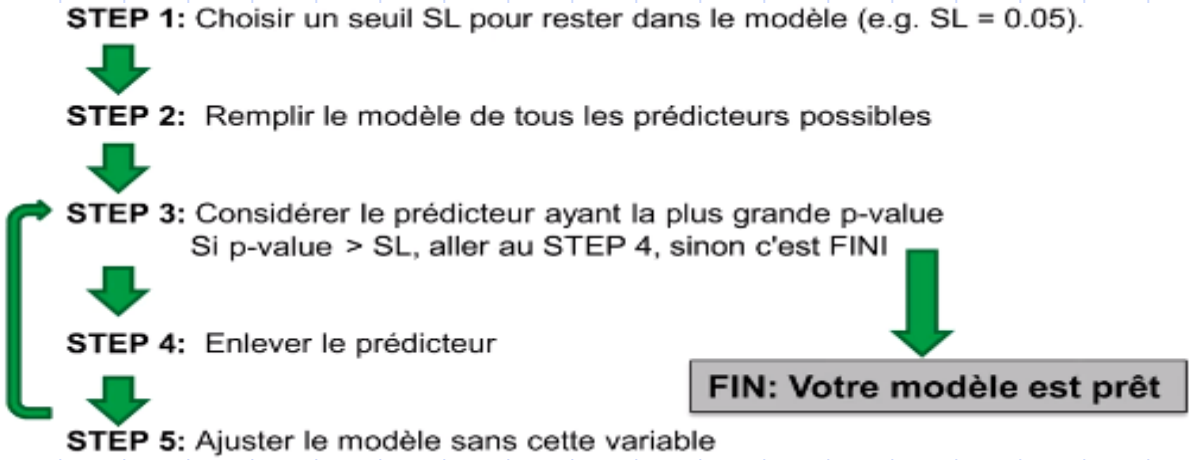
CH – II : A- REGRESSIONS

IV- Sélection des variables explicatives

IV-1-1 All-in : utiliser toutes les variables a votre disposition

- Les variables indépendantes sont données d'avance
- Vous connaissez vos variables indépendantes
- Vous n'avez pas le choix
- A utiliser pour préparer la backward elimination

IV-1-2 Backward Elimination

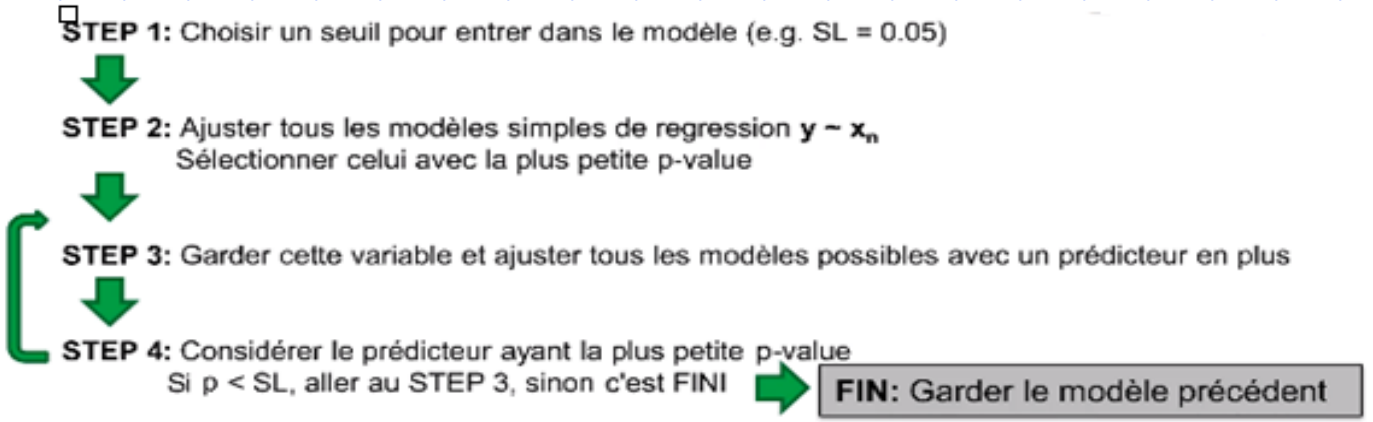


NB : lorsqu'on retourne au step3 il faut recalculer les p-value avec les variables indépendantes restantes.

CH – II : A- REGRESSIONS

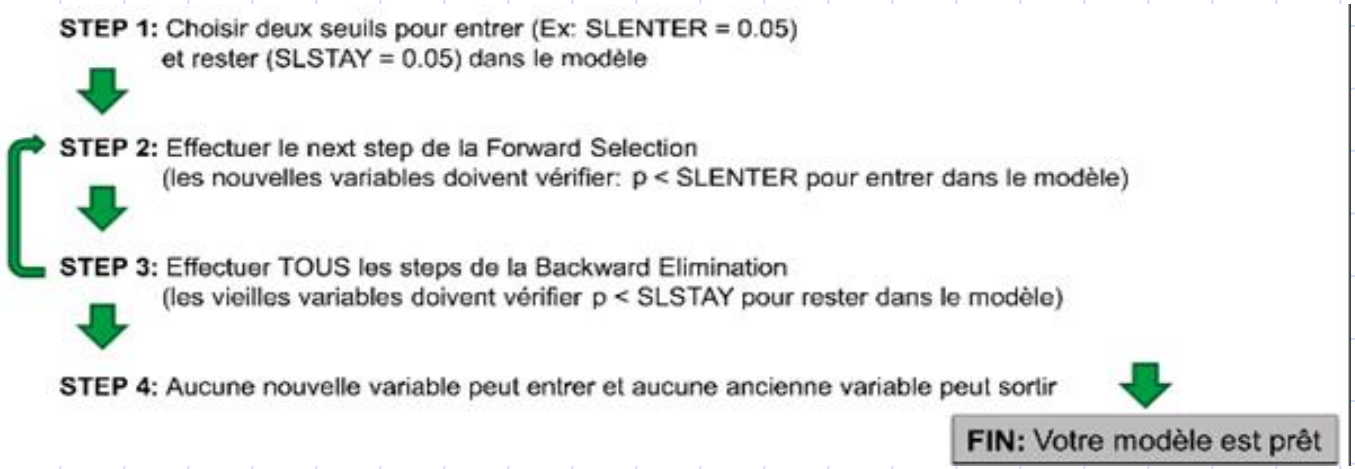
IV- Sélection des variables explicatives

IV-1-3 Forward sélection



NB : on ne retient que le modèle de l'avant dernière étape

IV-1-4 Bidirectional élimination

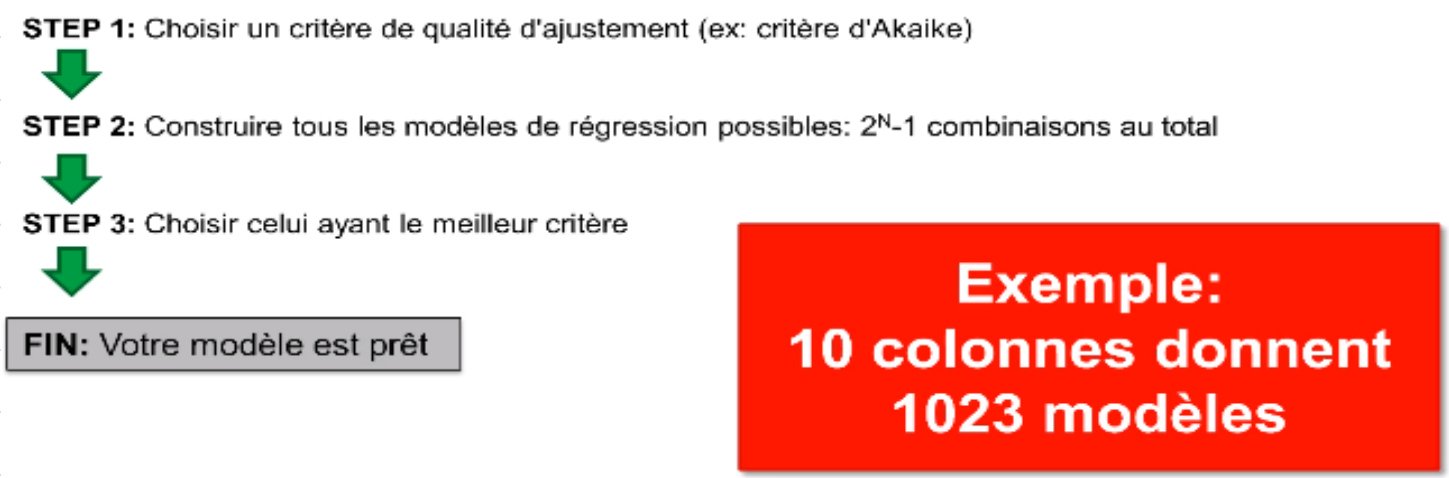


CH – II : A- ARBRES DE DÉCISION

IV- Sélection des variables explicatives

IV-1-5 Score comparaison

C'est la meilleur méthode mais la plus consommatrice en terme de ressource. Difficile a mettre en œuvre lorsqu'on a un nombre élevé de variables



V- Hypothèses d'un modèle de régression linéaire

Qu'elle soit simple ou multiple, la régression linéaire suppose qu'un certains nombre d'hypothèses soient vérifiées.

1- Exogénéité 2- Homoscédasticité 3- Erreurs indépendantes

4- Normalité des erreurs 5- Non colinéarité des prédicteurs

CH – II : B- ARBRES DE DECISION

INTRODUCTION

Les arbres de décision sont utilisés pour la prédiction ou l'explication d'une variable cible (Y)(variable cible, variable dépendante) à partir d'un ensemble de variable explicatives (X) (input variables, variables indépendantes)

Le résultat est un ensemble de règles simples qui permettent de réaliser des prévisions, de segmenter la population ou d'identifier qu'elles sont les variables qui discriminent le plus la variable cible.

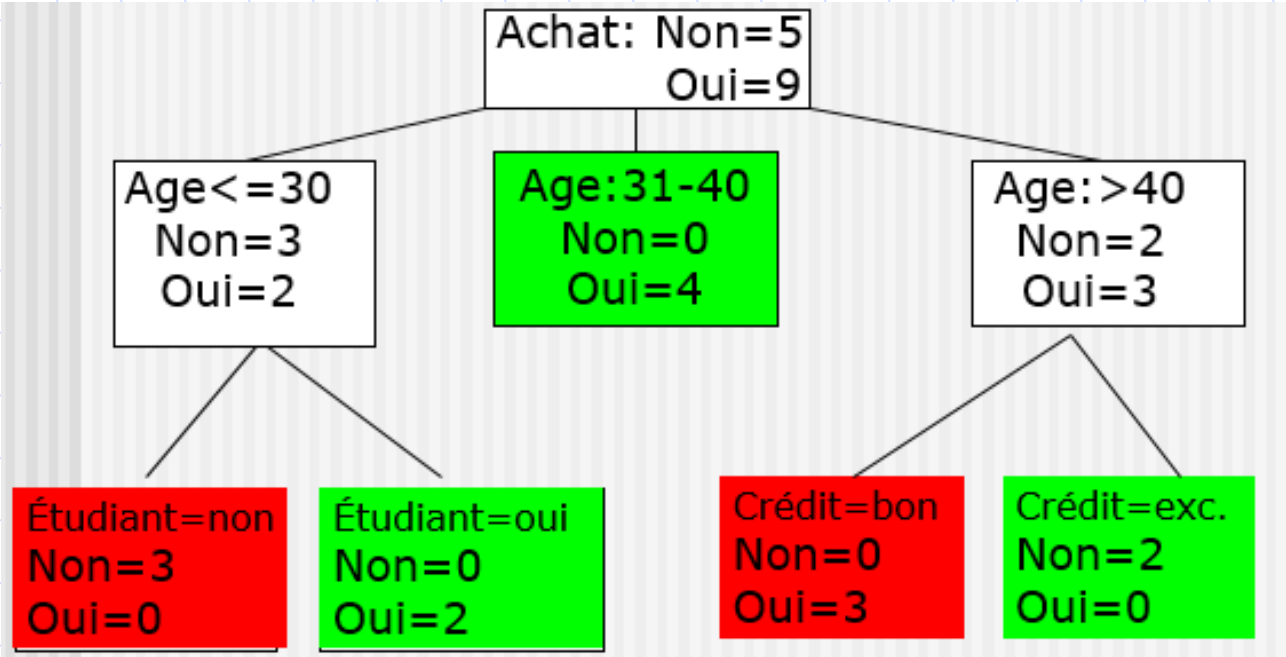
Exemple

age	revenu	etudiant	credit	achat
<=30	eleve	non	bon	non
<=30	eleve	non	excellent	non
31-40	eleve	non	bon	oui
>40	moyen	non	bon	oui
>40	faible	oui	bon	oui
>40	faible	oui	excellent	non
31-40	faible	oui	excellent	oui
<=30	moyen	non	bon	non
<=30	faible	oui	bon	oui
>40	moyen	oui	bon	oui
<=30	moyen	oui	excellent	oui
31-40	moyen	non	excellent	oui
31-40	eleve	oui	bon	oui
>40	moyen	non	excellent	non

CH – II : B- ARBRES DE DECISION

INTRODUCTION

Exemple d'arbre de décision



Algorithmes et logiciels les plus répandus pour construire les arbres de décision:

- CHAID Chi-Square Automatic Interaction Detection (1975)
- CART Classification And Regression Trees (Breiman et al., 1984)
- Knowledge seeker

CH – II : B- ARBRES DE DECISION

I- Régression vs. Classification (CART)

I-1 Contexte de CART

La méthode CART été formalisées dans un cadre générique de sélection de modèle par Breiman et col. (1984) sous l'acronyme de CART: Classification and Regression Tree.

L'acronyme CART correspond à deux situations bien distinctes selon que la variable à expliquer, modéliser ou prévoir est :

1. qualitative (classification)
2. quantitative (régression)

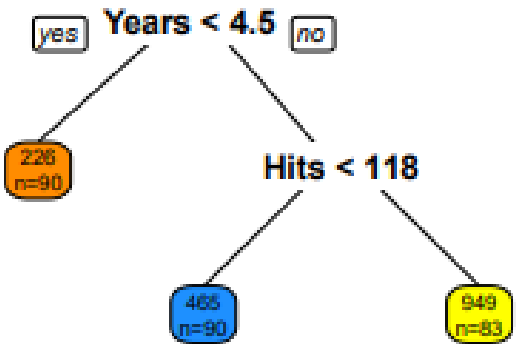


Fig.: Régression

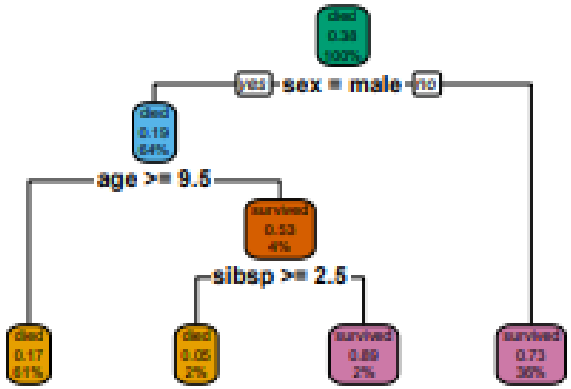


Fig.: Classification

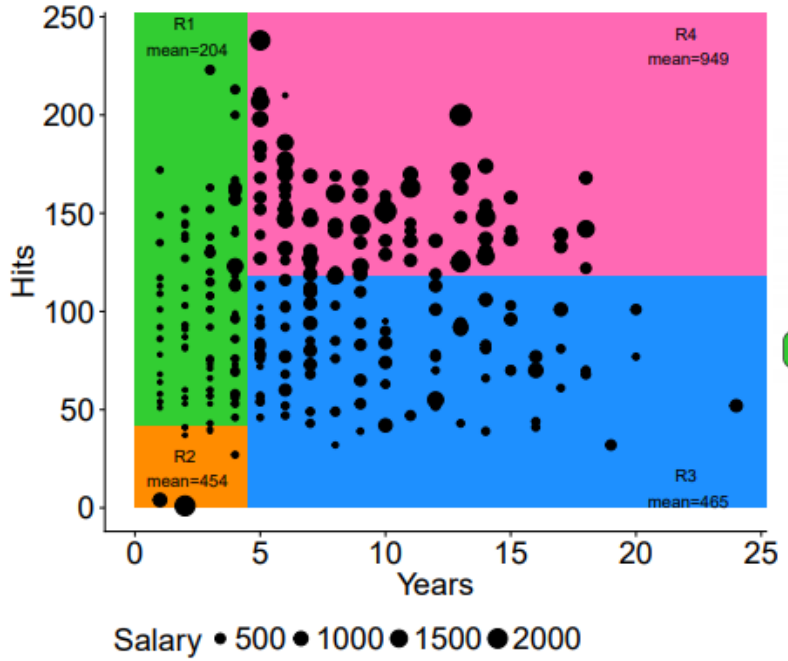
CH – II : B- ARBRES DE DECISION

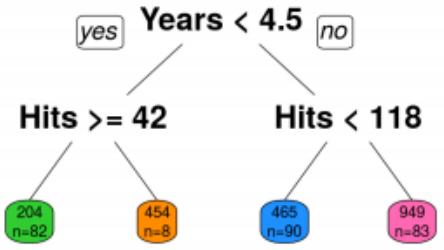
I- Régression vs. Classification (CART)

I-2 Comment fonctionne CART

Il y a deux étapes:

1. Nous divisons l'espace prédicteur, c'est-à-dire l'ensemble des valeurs possibles X1, X2, . . . , Xp - en J régions exhaustives et non chevauchantes, R1, R2, . . . , RJ.
2. Pour chaque observation qui tombe dans la région Rj, nous faisons la même prévision, qui est simplement la moyenne des valeurs de réponse à Rj





CH – II : B- ARBRES DE DECISION

I- Régression vs. Classification (CART)

I-2 Comment fonctionne CART

Les détails

L'algorithme considéré nécessite:

1. La définition d'un critère permettant de sélectionner la meilleure division parmi toutes celles admissibles pour les différentes variables.

2. Une règle permettant de décider qu'un noeud est terminal: il devient ainsi une feuille.

3. Élagage (pruning) de l'arbre optimal pour éviter le sur-ajustement.

I-3 Sélectionner la meilleure division

L'objectif est de trouver les divisions R1, . . . , RJ qui minimisent la fonction de perte :

$$\sum_{j=1}^{J} \sum_{i \in R_j} (y_i - \hat{y}_{R_j})^2$$

^yRj : la moyenne de la variable réponse dans la région Rj

CH – II : B- ARBRES DE DECISION

I- Régression vs. Classification (CART)

I-3 Sélectionner la meilleure division de façon «greedy»

- Au début, toutes les observations appartiennent à la même région.
- La division du nœud crée deux enfants, gauche et droit.
- On cherche pour chaque nœud la division, ou plus précisément la variable (Xj) et la règle de division (s), qui contribuera à la plus forte décroissance de l'hétérogénéité des nœuds enfants à gauche (R1) et à droite (R2)

$$R_1(j,s) = \{X|X_j < s\} \qquad \text{et} \qquad R_2(j,s) = \{X|X_j \geq s\}$$

L'objectif est de trouver les valeurs de j et s qui minimisent la fonction de perte :

$$\sum_{i:x_i \in R_1(j,s)} (y_i - \hat{y}_{R_1})^2 + \sum_{i:x_i \in R_2(j,s)} (y_i - \hat{y}_{R_2})^2$$

- Règle d'arrêt
  - Pour éviter un découpage inutilement fin, le nombre d'observations minimal qui doivent exister dans un nœud pour qu'une tentative de division soit effectuée (ex: minsplit = 20 par défaut dans le rpart).
  - La croissance de l'arbre s'arrête à un noeud donné, qui devient donc terminal ou feuille, lorsqu'il contient le nombre d'observations minimum. (minbucket = minsplit/3 par défaut dans rpart)

CH – II : B- ARBRES DE DECISION

II- Sélectionner la meilleure division de façon «greedy»

- Au début, toutes les observations appartiennent à la même région.
- La division du nœud crée deux enfants, gauche et droit.
- On cherche pour chaque nœud la division, ou plus précisément la variable (Xj) et la règle de division (s), qui contribuera à la plus forte décroissance de l'hétérogénéité des nœuds enfants à gauche (R1) et à droite (R2)

$$R_1(j,s) = \{X|X_j < s\} \qquad \text{et} \qquad R_2(j,s) = \{X|X_j \geq s\}$$

L'objectif est de trouver les valeurs de j et s qui minimisent la fonction de perte :

$$\sum_{i:x_i \in R_1(j,s)} (y_i - \hat{y}_{R_1})^2 + \sum_{i:x_i \in R_2(j,s)} (y_i - \hat{y}_{R_2})^2$$

- Règle d'arrêt
  - Pour éviter un découpage inutilement fin, le nombre d'observations minimal qui doivent exister dans un nœud pour qu'une tentative de division soit effectuée (ex: minsplit = 20 par défaut dans le rpart).
  - La croissance de l'arbre s'arrête à un noeud donné, qui devient donc terminal ou feuille, lorsqu'il contient le nombre d'observations minimum. (minbucket = minsplit/3 par défaut dans rpart)

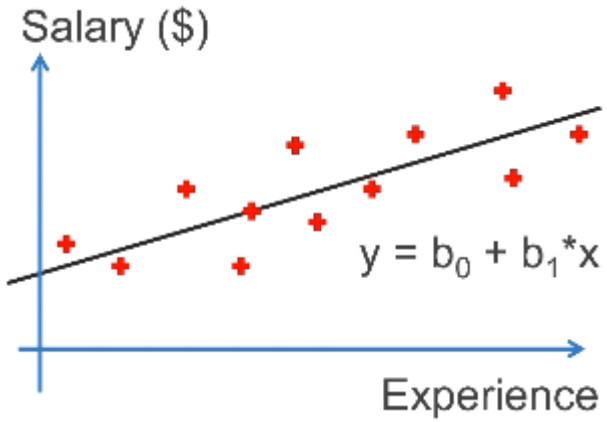
CH III- CLASSIFICATION

CH – III : CLASSIFICATION

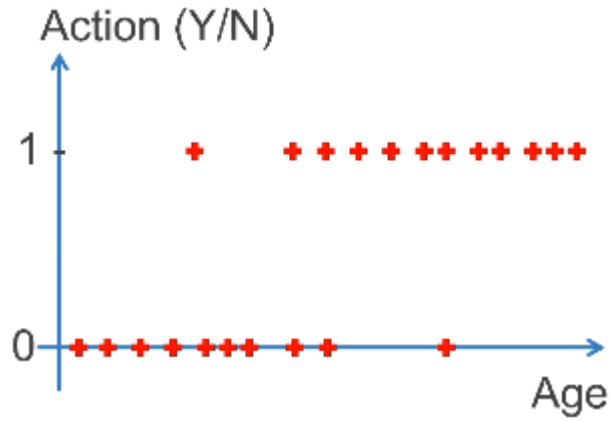
I- Régression logistique

I-1 Intuition de la régression logistique

On connait ceci (cas 1)



Ceci est nouveau (cas 2)



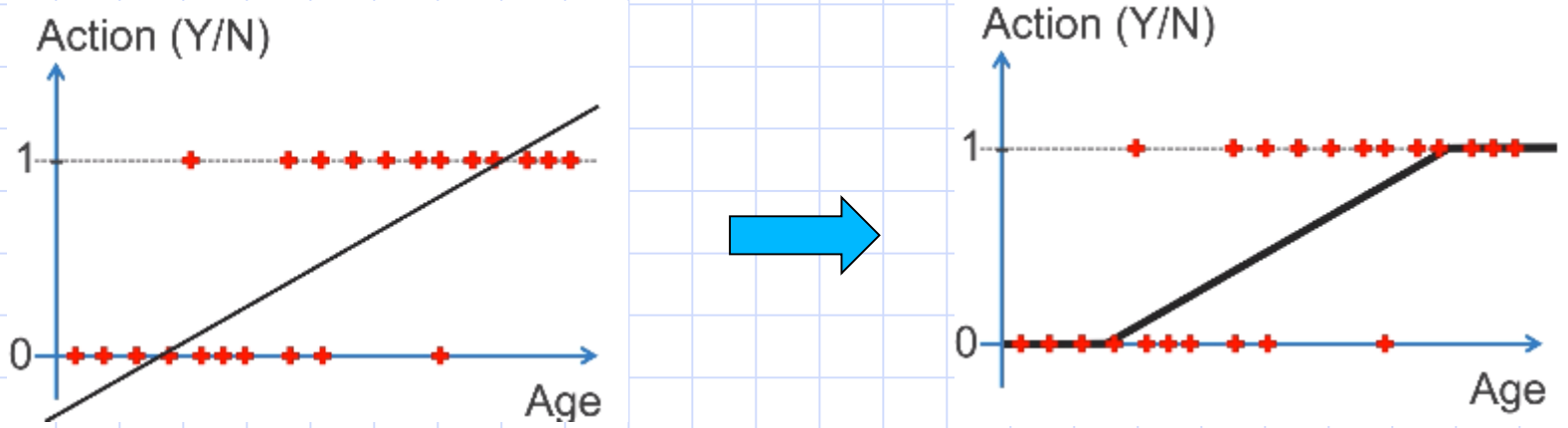
- Cas 1: représentation de l'évolution du salaire en fonction du nombre d'années d'expérience il est modélisable par un modèle de régression linéaire simple.
- Cas 2: représente l'action de cliquer sur une offre (0 = le client a cliquer sur le lien de l'offre; 1 = le client n'a pas cliquer sur le clien de l'offre) en fonction de l'âge. On remarque que les personnes plus âgées on tendance a cliquer sur l'offre ce qui n'est pas le cas pour les personne moins âgées.
- Comment trouver un modèle ou chaque valeur prédite est proche de la valeur réelle observé ?

CH – III : CLASSIFICATION

I- Régression logistique

I-1 Intuition de la régression logistique

Considérons un modèle de régression linéaire simple pour traduire cette situation



- La régression linéaire n'est pas adapté a cette situation.
- Considérons maintenant la probabilité qu'un client donnés click sur l'offre
  - Cette probabilité est faible pour les âges de gauche et plus élevée a droite
  - Pour donner un sens a cette probabilité nous allons procéder a des ajustement pour les parties de la droite en dessous de 0 et au dessous de 1. Basé sur cette approche probabiliste procédons a un aplatissement (voir la figure de gauche)
- Nous venons d'améliorer ce modèle et prédit mieux l'action de cliquer oui ou non avec l'âge

NB: le premier pilier de la régression logistique est cette notion de probabilité

CH – III : CLASSIFICATION

I- Régression logistique

I-2 Régression logistique

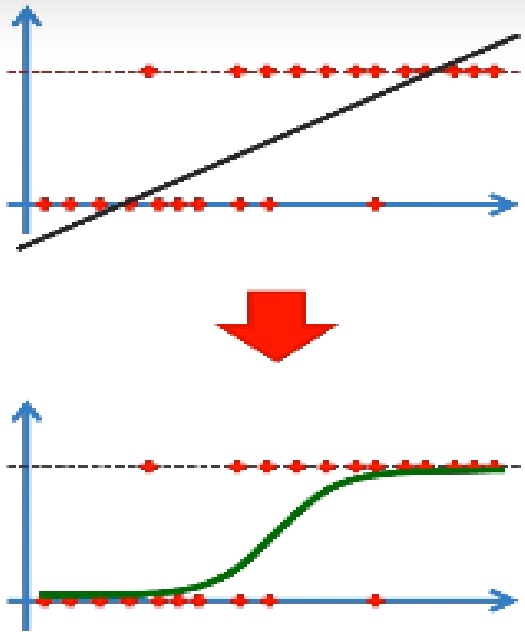
Procedons a un changement de variable en utilisant la fonction sigmoïde

$$y = b_0 + b_1 * x$$

Sigmoid Function

$$p = \frac{1}{1 + e^{-y}}$$

$$\ln\left(\frac{p}{1-p}\right) = b_0 + b_1 * x$$



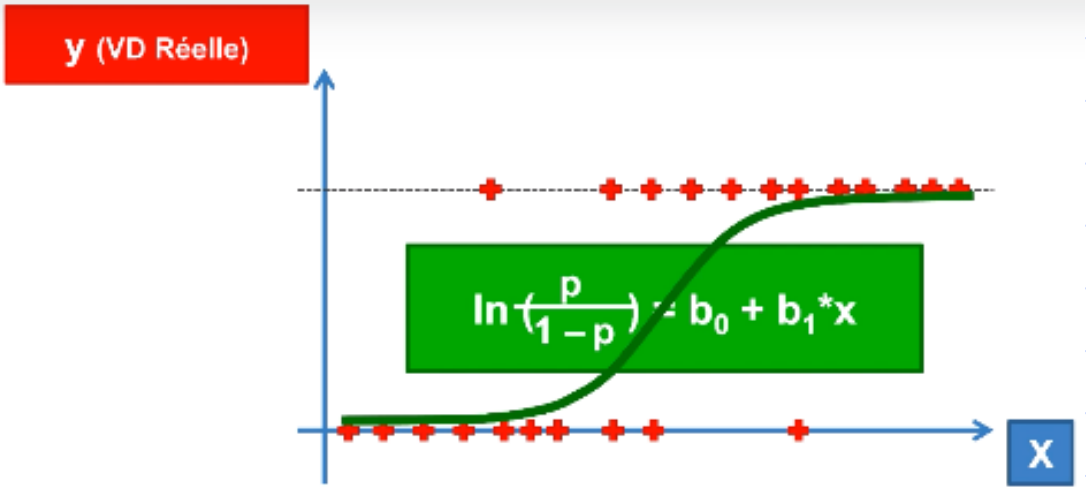
- L'equation en bas n'est rien d'autre que celui de la regression logistique .
- Le graphe de haut devient ce nouveau graphe qui est celui de la fonction de regression logistique

CH – III : CLASSIFICATION

I- Régression logistique

I-2 Régression logistique

En effet en les observations de notre data set et l'equation de la regression igistique on obtient la courbe de regression logistique comme figuré ci-dessous.



- C'est notre nouveau modèle de régression logistique qui remplace notre droite de régression linéaire. Elles sera utilisée pour prédire des variables binaire de type oui ou non.
- Quelle est son utilité ?

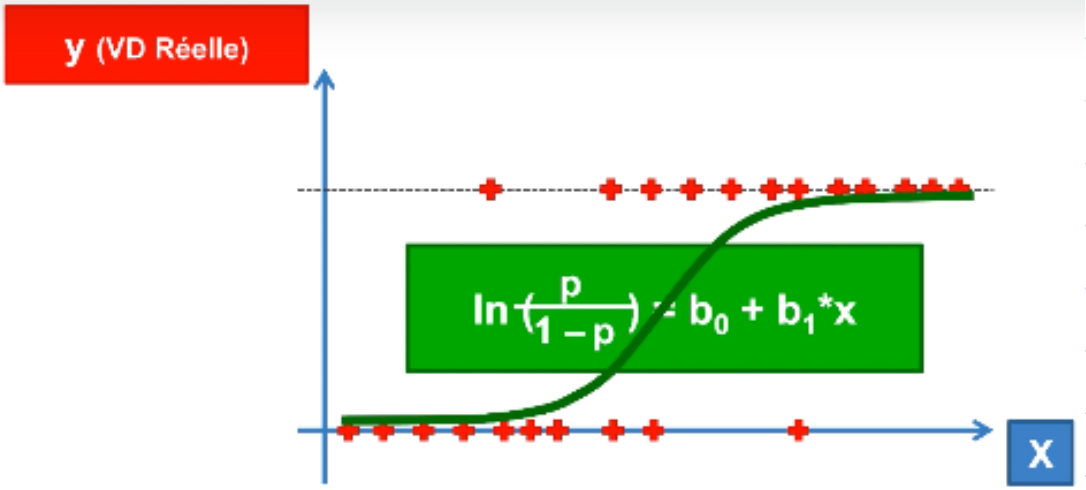
Nous allons l'utiliser pour prédire la probabilité qu'un évènement se produise.

CH – III : CLASSIFICATION

I- Régression logistique

I-2 Régression logistique

En effet en les observations de notre data set et l'equation de la regression igistique on obtient la courbe de regression logistique comme figuré ci-dessous.



- C'est notre nouveau modèle de régression logistique qui remplace notre droite de régression linéaire. Elles sera utilisée pour prédire des variables binaire de type oui ou non.
- Quelle est son utilité ?

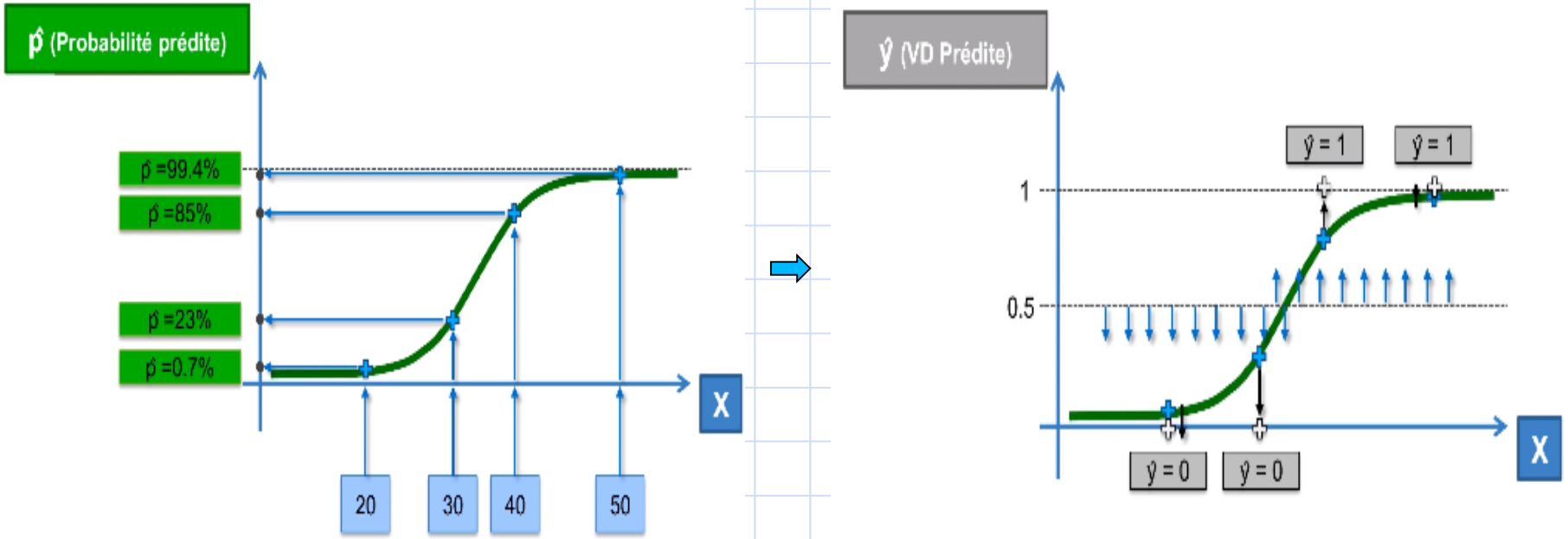
Nous allons l'utiliser pour prédire la probabilité qu'un évènement se produise.

CH – III : CLASSIFICATION

I- Régression logistique

I-3 Prediction avec la fonction de régression logistique

- Prenons quatre âges au hasard



- On prédit ainsi des score de probabilité qui sont tres utile pour casser nos observations.
- Pour obtenir les valeurs prédites alors qu'on des probabilités prédites, fixons un seuil (alpha = 50% qui est communément utilisé)

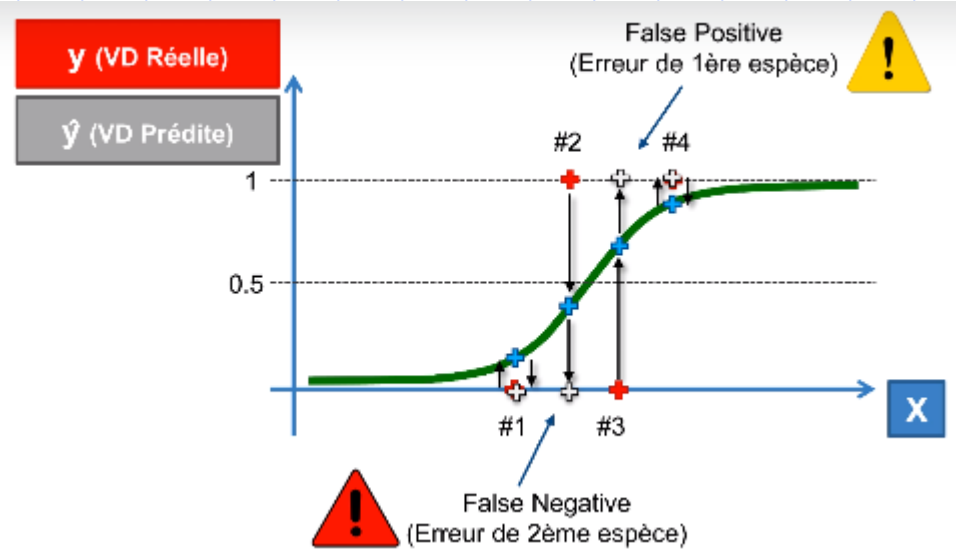
Si la probabilité prédite < 50% la probabilité prédite est 0 sinon la probabilité prédite est 1.

NB: avec la regression logistique on peut jouer avec ce seuil en fonction du problème a résoudre.

CH – III : CLASSIFICATION

II- False Positive et False Negative

- Prenons quatre âges de nos observations et regardons les valeurs prédites selon le modèle



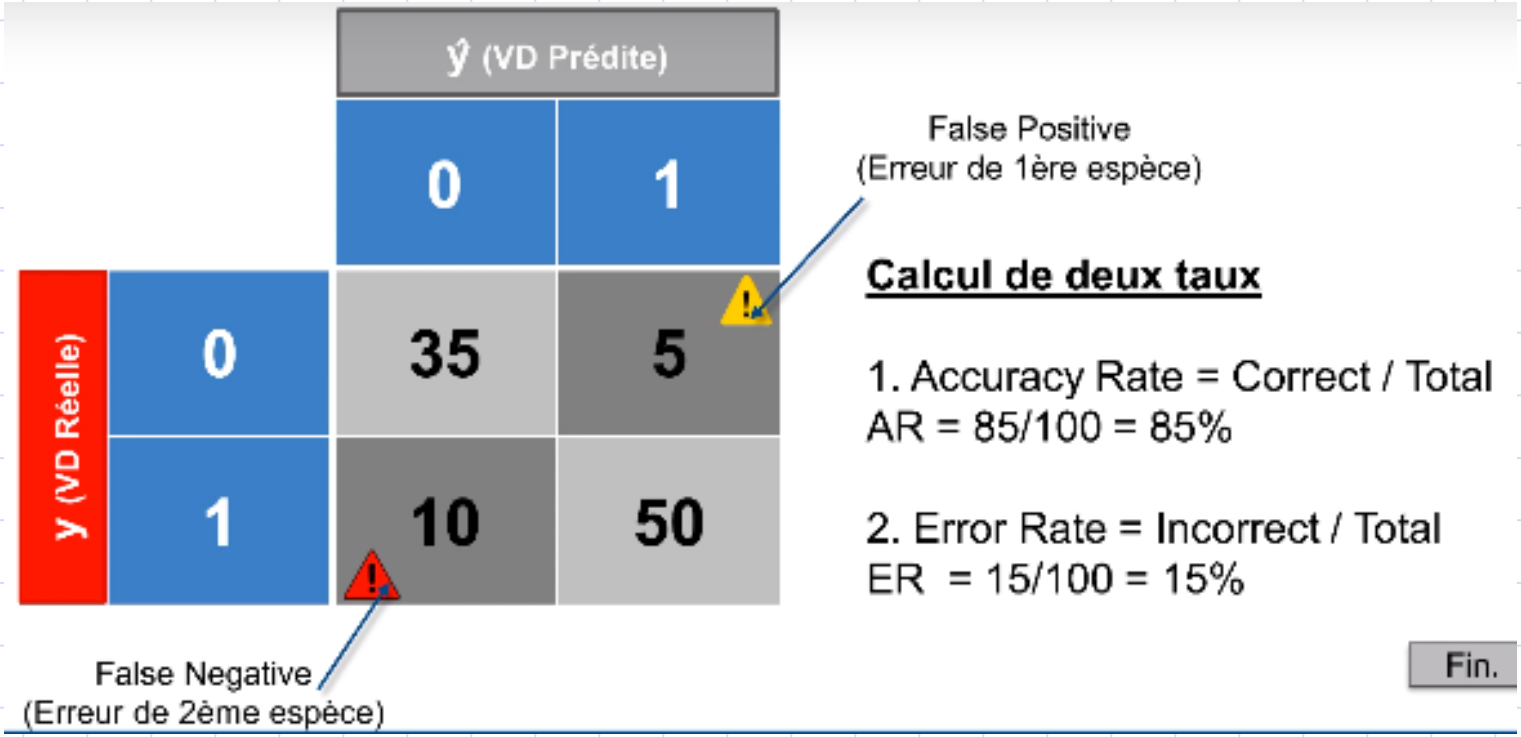
- Pour #1 et #4 le modèle a bien prédit les valeurs réelles.
- Pour #2 et #3 le modèle s'est trompé
- Les erreurs de type #2 sont appelées False négative ou (erreur de 2eme espèce) c'est-à-dire que la valeur prédite est 0 mais elle est fausse
- Les erreurs de type #3 sont appelées False positive ou (erreur de 1ere espèce) c'est-à-dire que la valeur prédite est 1 mais elle est fausse

NB : le false négative et plus dangereux que le false positive

CH – III : CLASSIFICATION

III- Matrice de confusion

- Soit la matrice ci-dessous contenant les prédictions de notre modèle comparé au valeurs réelles



- La matrice de confusion donnes une idée de la manière dont le modèle a fonctionné
- La premières diagonale est celle des bonnes choses (le modèle ne s'est pas trompé)
- La seconde diagonale est celle des mauvaise choses (le modèle s'est trompé)
- Acurracy rate et Errol rate nous donne une idée de la précision du modèle.

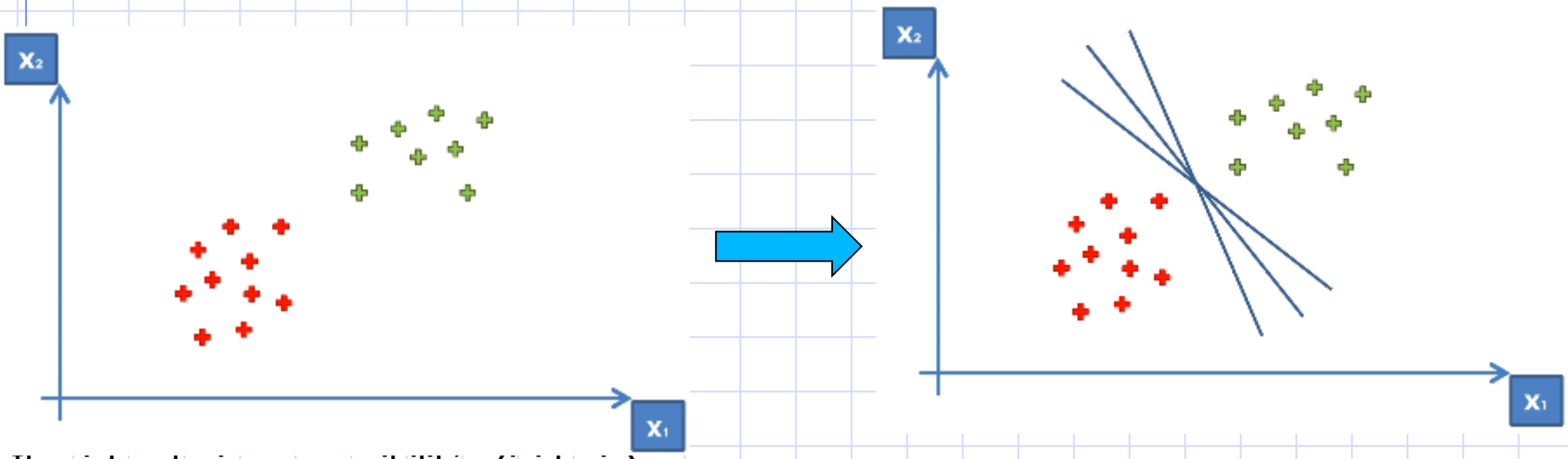
CH – III : CLASSIFICATION

IV- SVM (Support Vector Machine)

- Le SVM (Support Vector Machine) ou Séparateur a Vaste Marge a été introduit et 1960 et repris dans les années 1990 pour devenir aujourd'hui très populaire à cause de son efficacité.

IV-1 – Intuition (1)

Soit séparer les deux groupe de point ci-dessous par le SVM qui est une modèle a séparateur linéaire. Comment peut-on séparer ces deux catégories (la classe rouge et la classe verte) de points par un séparateur linéaire ?



Il existe plusieurs possibilités (ici trois).

Mais quelles la meilleur droite c-a-d quelle est la droite qui va séparer la plus pertinemment ces deux classes de points?

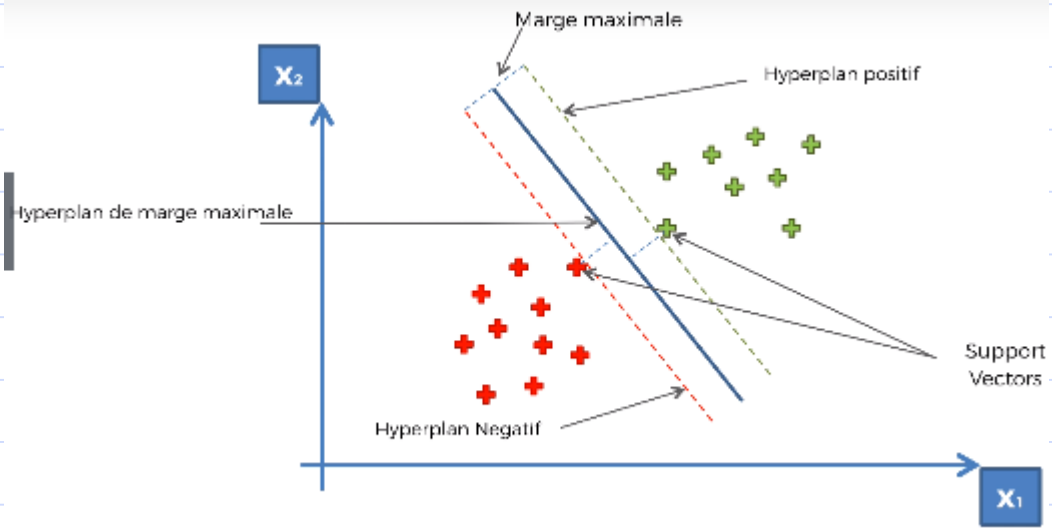
CH – III : CLASSIFICATION

IV- SVM (Support Vecteur Machine)

IV-1 – Intuition (2)

C'est ce que va essayer de faire le modèle SVM, il va trouver la droite optimale pour séparer ces deux classes. Le modèle SMV pour trouver cette droite optimale repose sur les concepts ci-dessous:

- Les marge maximale construite a partir des deux points les plus proches. Avec pour droite frontière de prédiction est la droite équidistante des droite rouge et verte
- Les deux points les plus proches de la frontière de prédiction appelés les supports vectors. Qui caractérisent modèle.
- l'hyperplan de marge maximale (NB: des un espace de dimension N un hyperplan est de dimension N-1)



CH – III : CLASSIFICATION

IV- SVM (Support Vecteur Machine)

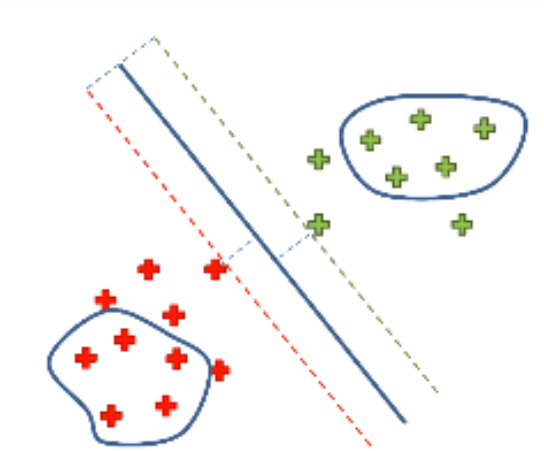
IV-2 –Pour le SVM est-il si speciale ?

En apprentissage machine on donne généralement un ensemble de points pour construire le modèle d'apprentissage (trainnig set) pour la plupart des modèle de ML . La machine va construire son apprentissage en construisant les corrélations entre les points du training. Quant il aura compris ces correlations il vales utiliser pour faire de nouvelles prédictions sur le test set.

En revanche pour notre modèle SVM

- Au lieu de construire son apprentissage sur un ensemble de points
- Il regarder uniquement les points les plus proches dans chacune des deux classes
- Il va construire son analyse en utilisant uniquement ces deux points

C'est un très bon modèle de classification si les points sont linéairement séparables

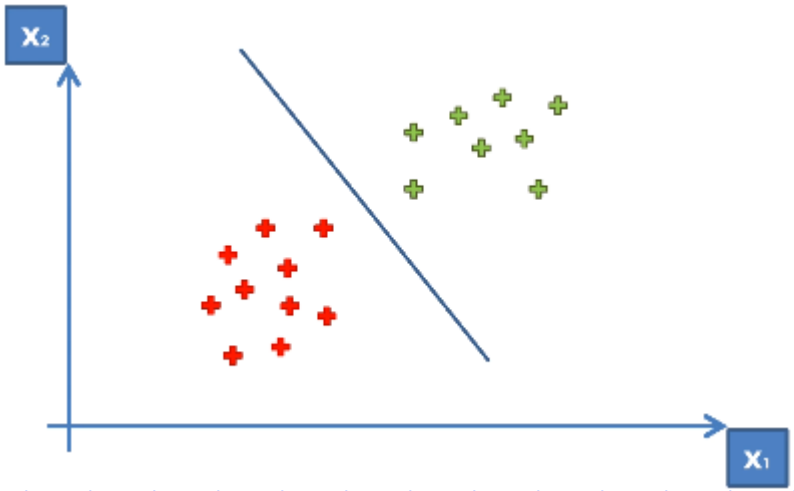


CH – III : CLASSIFICATION

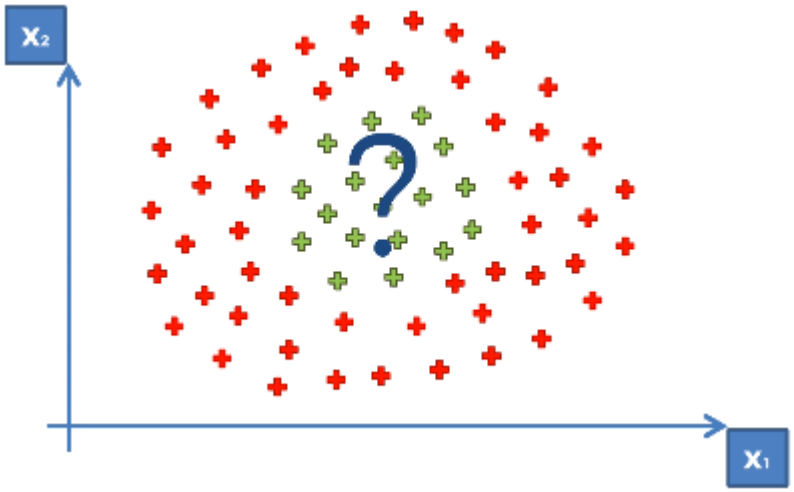
V- Kernel SVM

V-1 –Intuition

Points linéairement séparables



Points non linéairement séparables



Que faire lorsque les points ne sont plus linéairement séparable?

Le Kernel SVM va améliorer l'algorithme du SVM pour séparer des points qui ne sont pas linéairement séparables.

V-2 –Principe

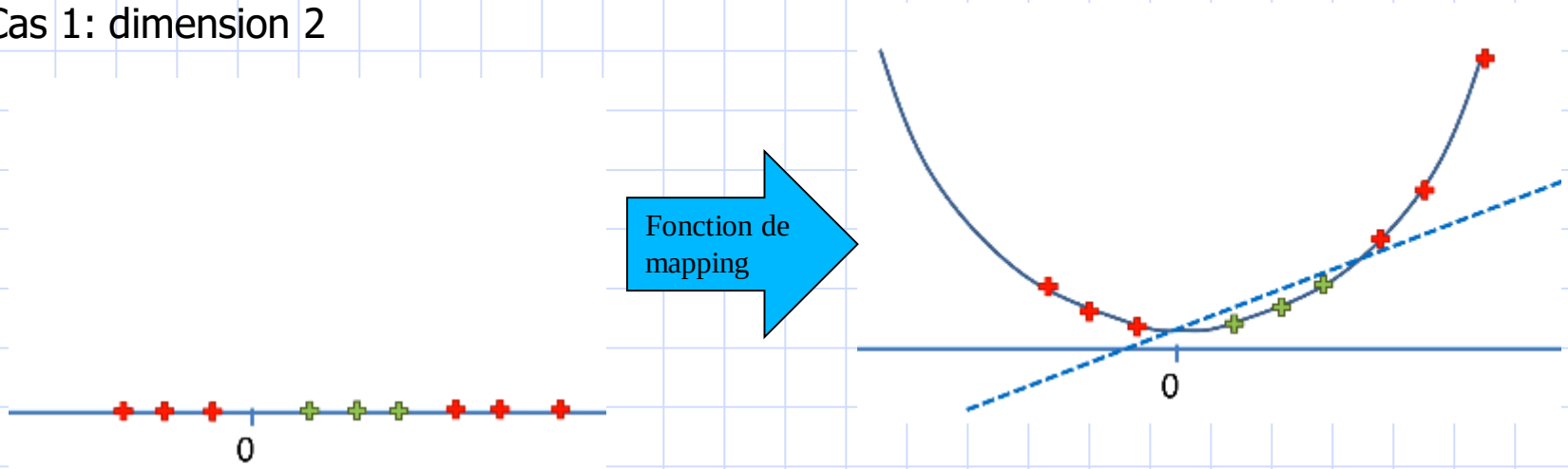
- Mapper nos points dans un espace de plus grande dimension
- Les séparer par un hyperplan

CH – III : CLASSIFICATION

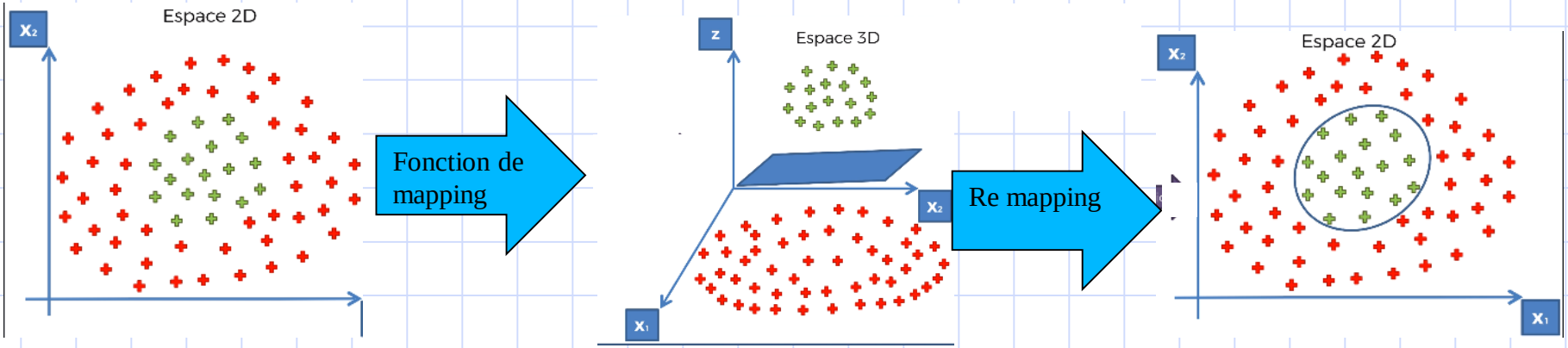
V- Kernel SVM

V-2 –Principe (2)

- Cas 1: dimension 2



- Cas 2 : dimension 3



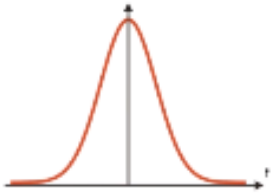
CH – III : CLASSIFICATION

V- Kernel SVM

V-2 –Principe (3)

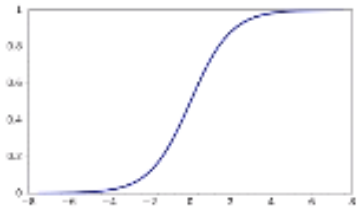
- Le mapping vers un espace a plus grande dimension peut demander énormément de calcul. On dit qu'il est compute intensive.
- Une solution a ce problème est utiliser le Gaussian RBF Kernel

V-3 – Les solutions Kernel



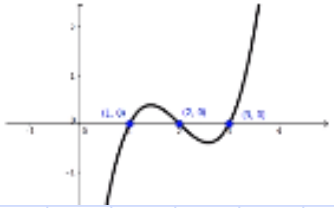
Gaussian RBF Kernel

$$K(\vec{x}, \vec{l}^i) = e^{-\frac{\|\vec{x}-\vec{l}^i\|^2}{2\sigma^2}}$$



Sigmoid Kernel

$$K(X, Y) = \tanh(\gamma \cdot X^T Y + r)$$



Polynomial Kernel

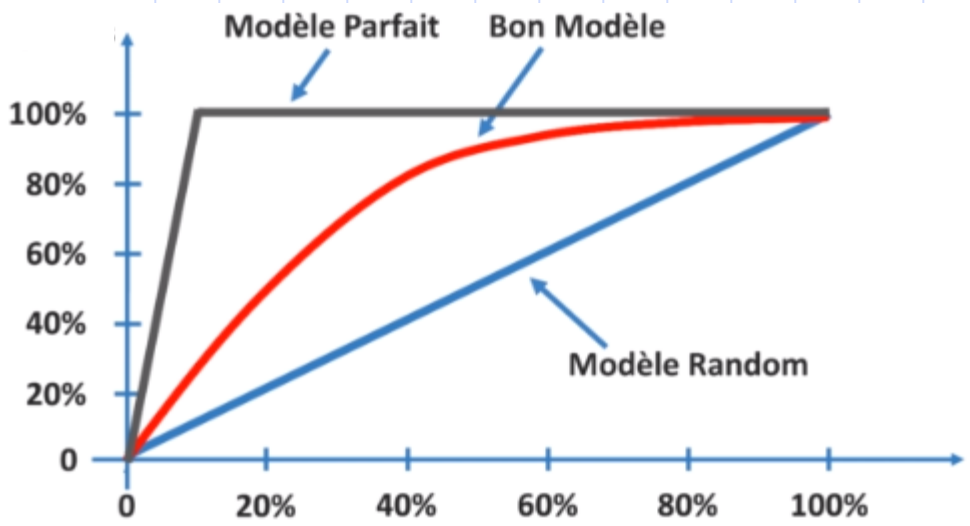
$$K(X, Y) = (\gamma \cdot X^T Y + r)^d, \gamma > 0$$

CH – III : CLASSIFICATION

VI- Validation des modeles de classification

VI-1 – CAP (Cummulative Acuracy Profil)

- Le CAP d'un modèle est une courbe qui est utiliseé pour la validation des performances d'un modele de classification.
- La validation se fait en comparant la courbe CAP du modèle à l'étude avec la courbe CAP d'un modèle parfait et la courbe CAP d'un modèle aléatoire (imparfait).



- Il parait assez intuitif de dire que plus la courbe rouge est proche de cette en gris et plus votre modèle est performant. Inversement plus elle est proche de la courbe bleue et moins notre modèle est performant. Cette approche est purement qualitative mais comment peut-on quantifier cela?

CH – III : CLASSIFICATION

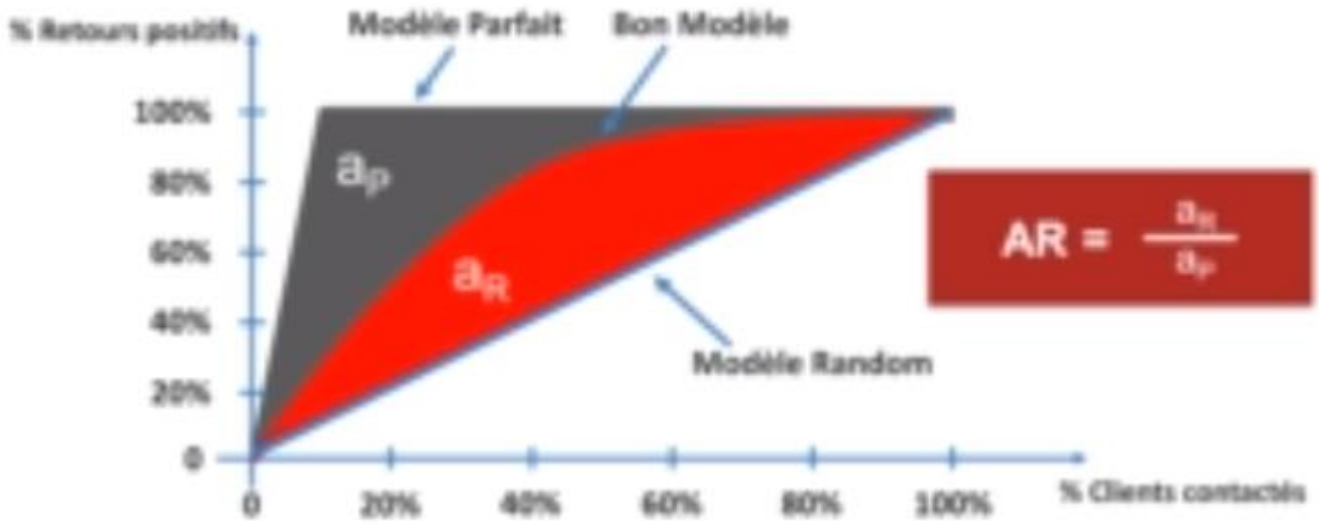
VI- Validation des modeles de classification

VI-1 – CAP (Cummulative Acuracy Profil)

- Méthode 1: utilisation du ration de precision

On utiliser le ratio entre deux métriques :

- a_p = aire entre le modèle parfait et le modèle random
- a_r = aire entre le bon modèle et le modèle random



- 0 <= AR <= 1
- Plus AR est proche de 0 et moins le modèle est bon
- Plus AR est proche de 1 et plus le modèle est bon
- Cependant il n'est pas facile de quantifier cette metrique

CH – III : CLASSIFICATION

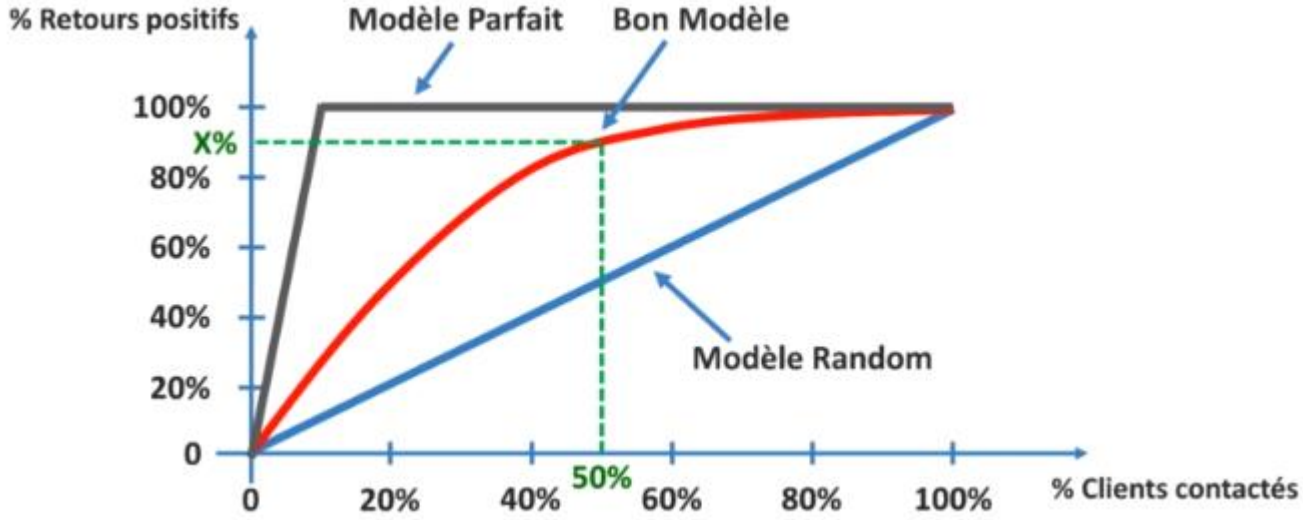
VI- Validation des modeles de classification

VI-1 – CAP (Cummulative Acuracy Profil)

- Méthode 2: utilisation de la verticale a 50%

On lieu de regarder ces aires considerons la droite verticale considerontla verticale a 50% (axe des abscisse)

- Projetons le point a 50% de l'axe des abscisse sur la courbe
- Projetons le point ontenu sur l'axe des ordonnées



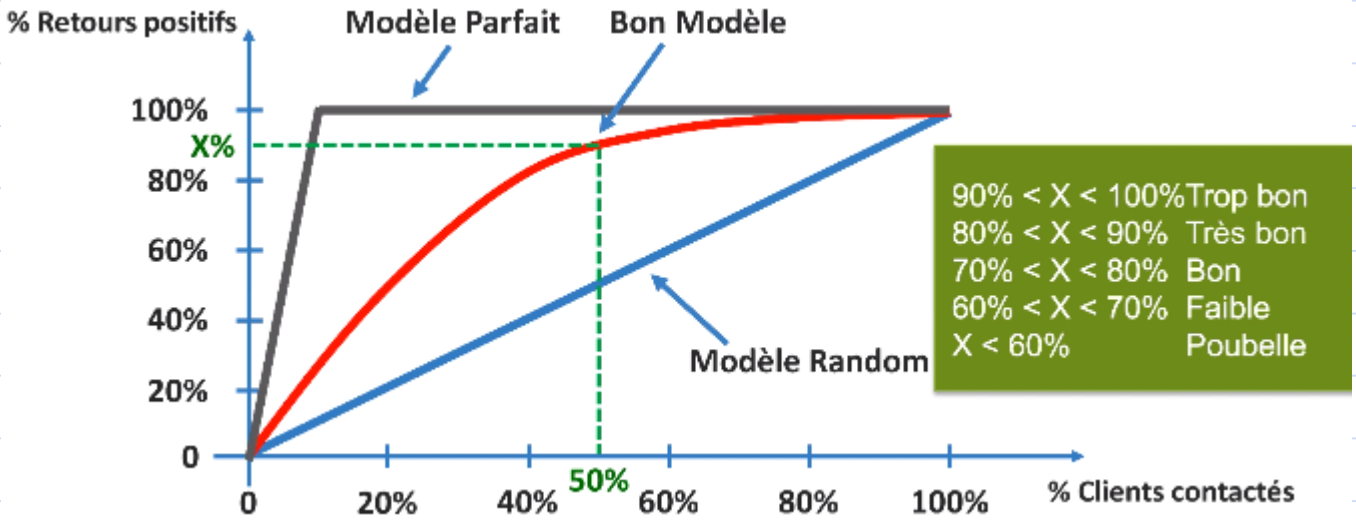
- x% représente 50% des observations qui on la plus grande probabilité d'indexer la variable dépendante dans le modèle
- De ce x% on peut mettre en place une regle pour evaluer notre modele.

CH – III : CLASSIFICATION

VI- Validation des modeles de classification

VI-1 – CAP (Cummulative Acuracy Profil)

- Méthode 2: utilisation de la verticale a 50%



- NB :
- Si 70 < X < 80 le modèle est bon et va nous apporter une bonne valeur ajouté
- Si 80 < X < 90 le modèle est très bon c-a-d qu'il est géniale
- Si 90 < X < 1000 le modèle est trop bon (to good to be true en anglais) attention a l'over fitting c-a-d le sur-apprentissage.
- NB: cette approche est très utile pour comparer deux modèles.

CH IV- CLUSTERING

CH – IV : CLUSTERING

Bienvenue à la Partie - Clustering !

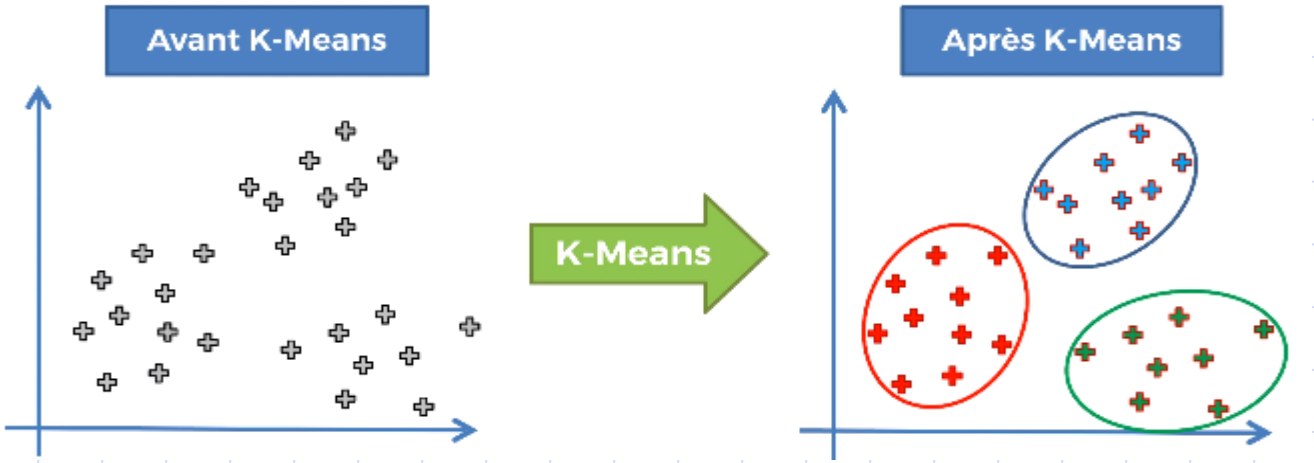
Le Clustering est du Machine Learning non supervisé, c'est à dire qu'il n'y a pas de variable dépendante à prédire. Autrement dit, on ne sait pas ce qu'on cherche, et on essaie d'identifier des groupes d'observations similaires, appelés clusters.

Nous allons former notre intuition et faire l'implémentation des modèles K-Means, le modèle le plus populaire et le plus efficace de clustering.

I- Le modèles K-Means

I-1 Intuition

Le modele K-means va chercher a identifier des groupes de variables indépendantes (VI) appelées clusters en se basant uniquement sur les VI . Que fait le K-means?



Ces clusters, traités comme des variable categorielles seront nos variables dependentes

CH – IV : CLUSTERING

I-2 – L'algorithme K-means

STEP 1: Choisir le nombre K de clusters

STEP 2: Sélectionner au hasard K points, les centroids

STEP 3: Assigner chaque point au centroid le plus proche ➡ Cela forme K clusters

STEP 4: Calculer et placer le nouveau centroid de chaque cluster

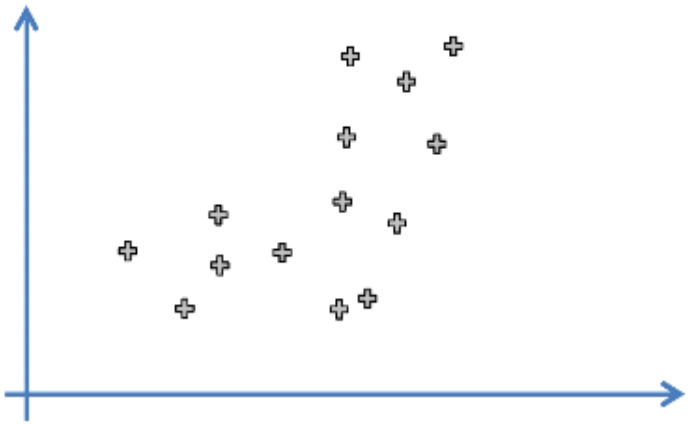
STEP 5: Réassigner chaque point au nouveau centroid le plus proche.
Si au moins un point a été réassigné, retourner au STEP 4, sinon: Votre modèle est prêt

La distance utilisée ici est quelconque (ex: distance euclidienne)

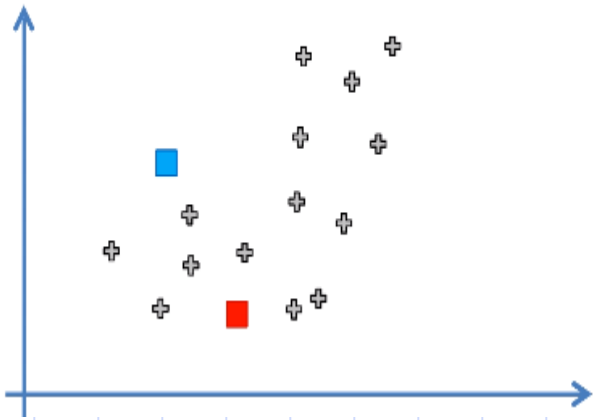
CH – IV : CLUSTERING

I-3 – Mise en œuvre graphique

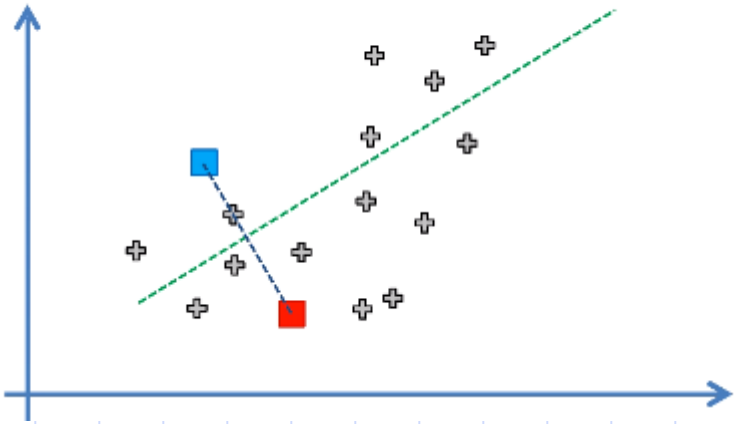
STEP 1: Choisir le nombre K de clusters: K = 2



STEP 2: Sélectionner au hasard K points, les centroids (pas nécessairement du dataset)



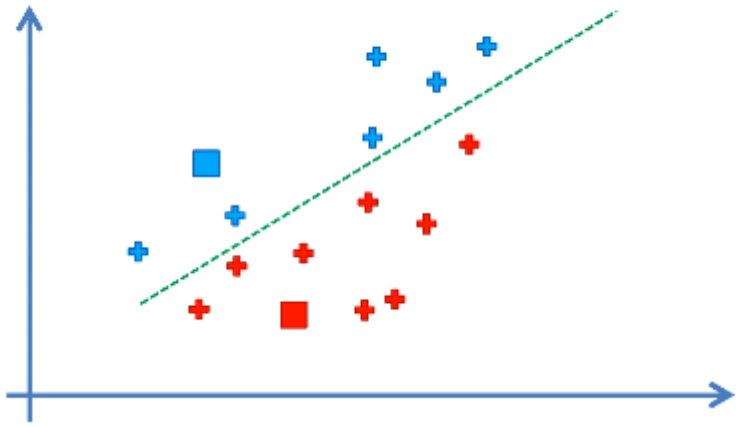
STEP 3: Assigner chaque point au centroid le plus proche ➞ Cela forme K clusters



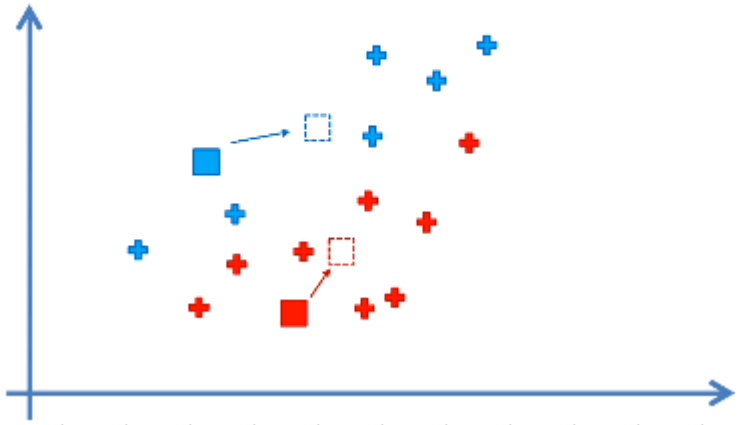
CH – IV : CLUSTERING

I-3 – Mise en œuvre graphique

STEP 3: Assigner chaque point au centroid le plus proche ➡ Cela forme K clusters



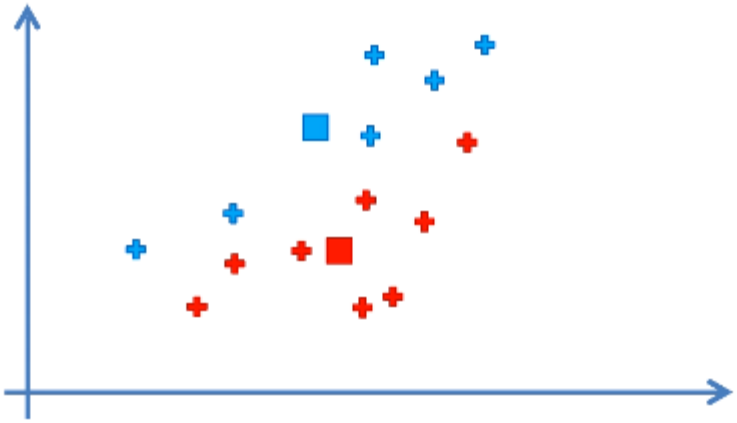
STEP 4: Calculer et placer le nouveau centroid de chaque cluster



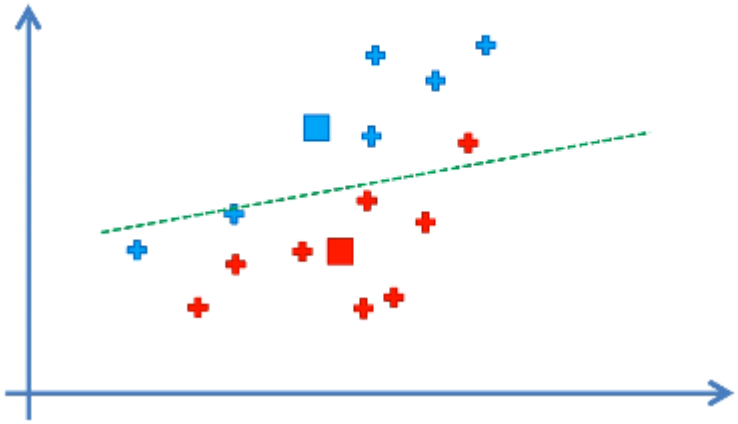
CH – IV : CLUSTERING

I-3 – Mise en œuvre graphique

STEP 5: Réassigner chaque point au nouveau centroid le plus proche.
Si au moins un point a été réassigné, retourner au STEP 4, sinon FIN.



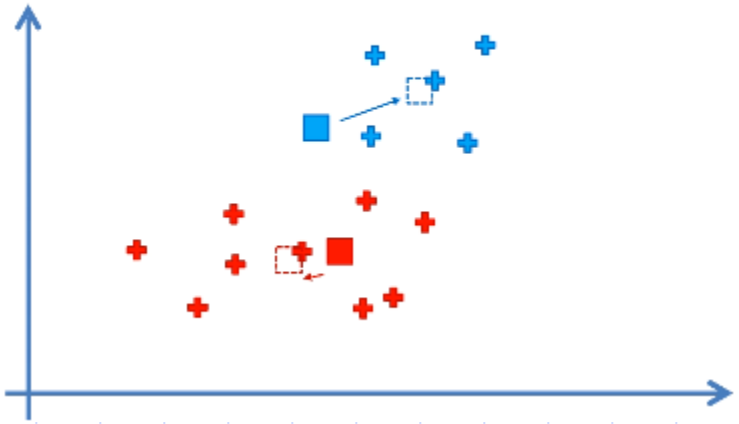
STEP 5: Réassigner chaque point au nouveau centroid le plus proche.
Si au moins un point a été réassigné, retourner au STEP 4, sinon FIN.



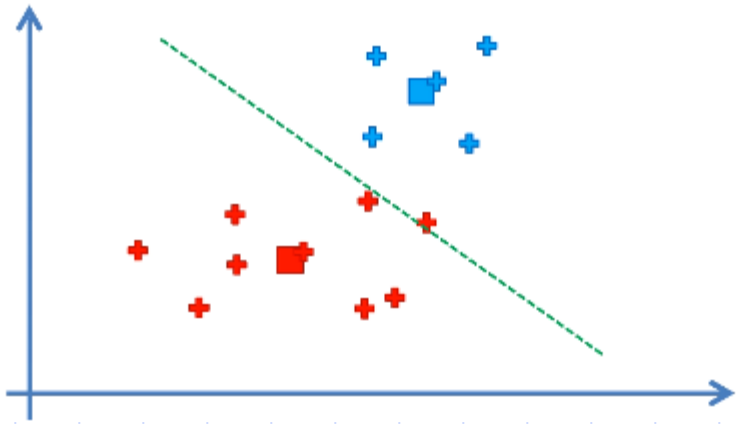
CH – IV : CLUSTERING

I-3 – Mise en œuvre graphique

STEP 4: Calculer et placer le nouveau centroid de chaque cluster



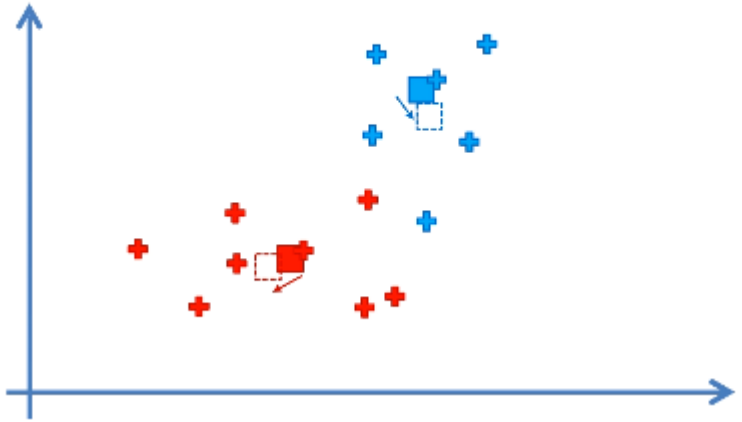
STEP 5: Réassigner chaque point au nouveau centroid le plus proche.
Si au moins un point a été réassigné, retourner au STEP 4, sinon FIN.



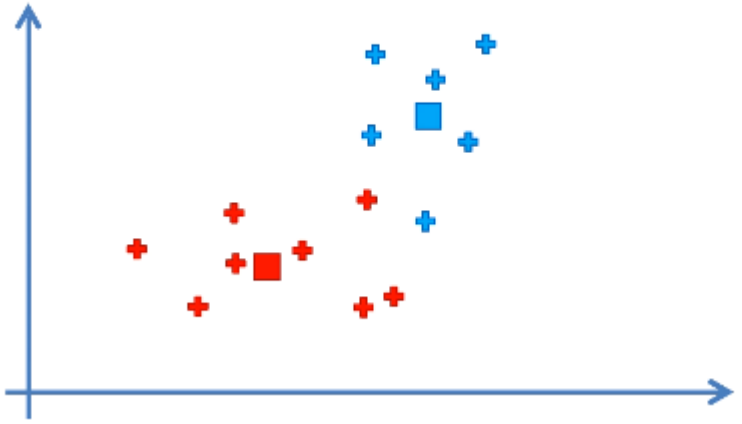
CH – IV : CLUSTERING

I-3 – Mise en œuvre graphique

STEP 4: Calculer et placer le nouveau centroid de chaque cluster



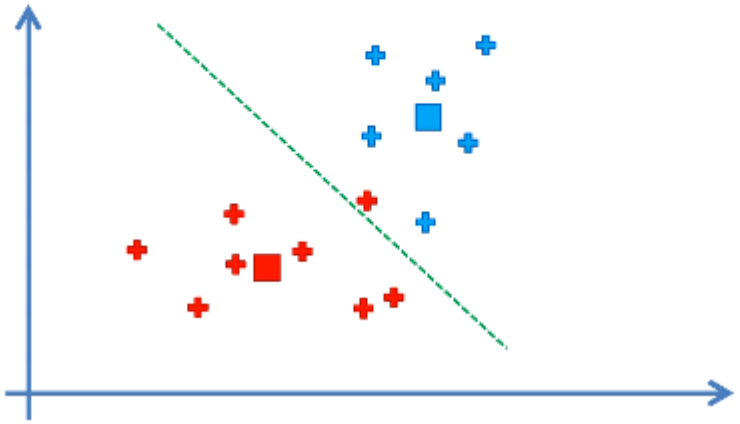
STEP 5: Réassigner chaque point au nouveau centroid le plus proche.
Si au moins un point a été réassigné, retourner au STEP 4, sinon FIN.



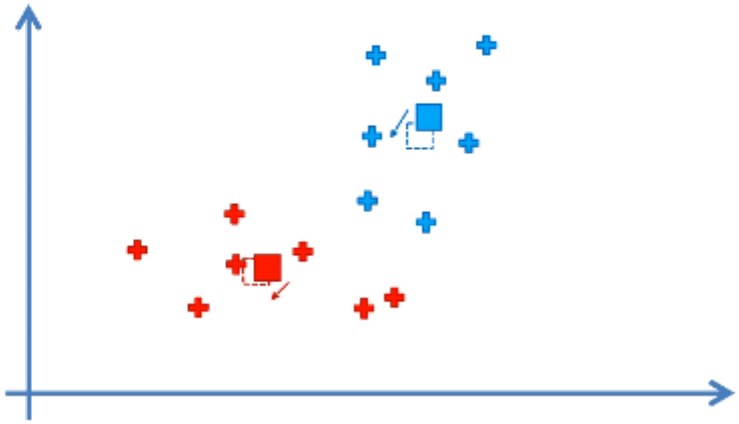
CH – IV : CLUSTERING

I-3 – Mise en œuvre graphique

STEP 5: Réassigner chaque point au nouveau centroid le plus proche.
Si au moins un point a été réassigné, retourner au STEP 4, sinon FIN.



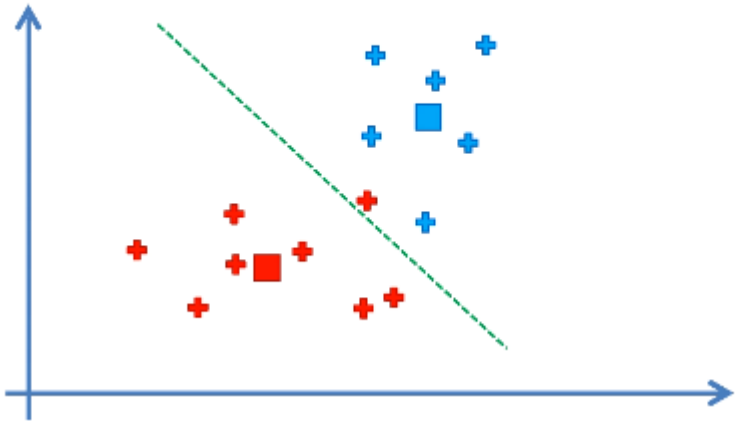
STEP 4: Calculer et placer le nouveau centroid de chaque cluster



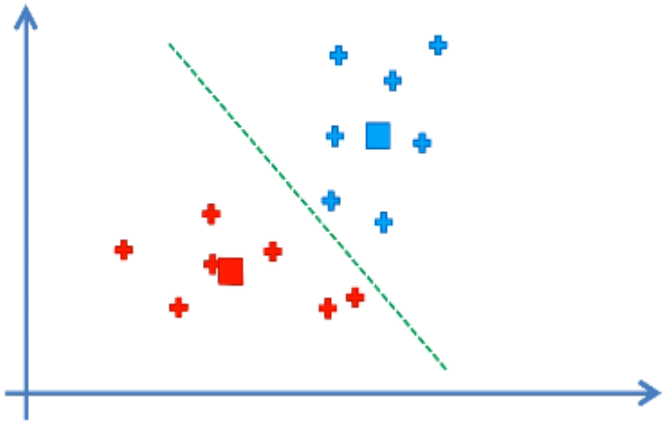
CH – IV : CLUSTERING

I-3 – Mise en œuvre graphique

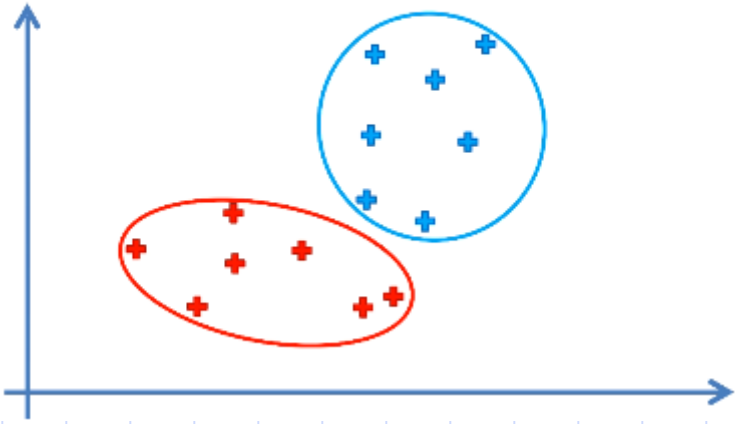
STEP 5: Réassigner chaque point au nouveau centroid le plus proche.
Si au moins un point a été réassigné, retourner au STEP 4, sinon FIN.



STEP 5: Réassigner chaque point au nouveau centroid le plus proche.
Si au moins un point a été réassigné, retourner au STEP 4, sinon FIN.



FIN: Votre modèle est prêt



CH – IV : CLUSTERING

I-3 – Piège de l'initialisation aléatoire des centroides

La sélection aléatoire des centroides peut biaisé la construction de notre modèle de classification.

Comment résoudre ce problème ?

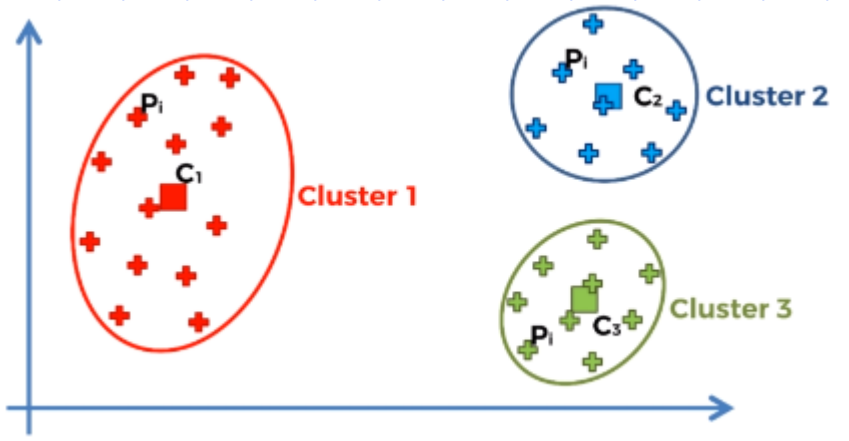
La solution k-means++ est implementé par la plus par des logiciels Ml parcequelle parmet de choisir le nombre cluster optinale.

II – Choisir le bon nombre de cluster

Soit le graphique ci-contre.

Pour un nombre de cluster données on obtient Differentes valeurs de clustering pour mesurer la pertinance du modele.

Quelles mesure pouvons nous choisir?



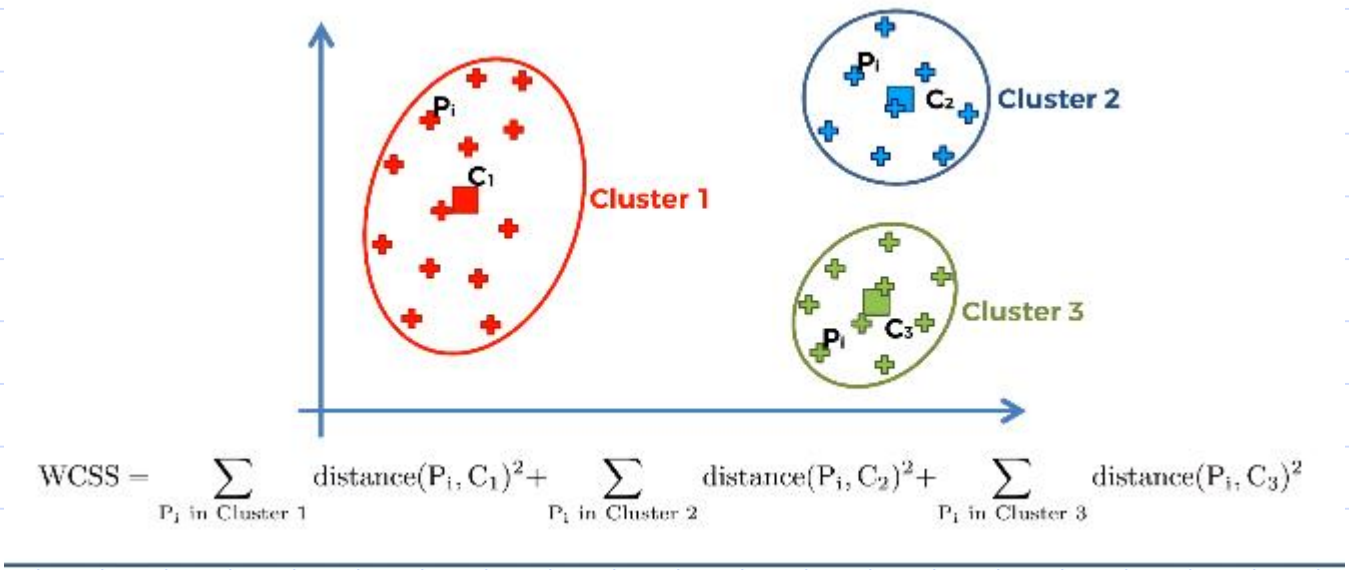
CH – IV : CLUSTERING

II – Choisir le bon nombre de cluster

Il existe une metrique pour evaluer la pertinance du clustering, c'est le WCSS

$$\text{WCSS} = \sum_{P_i \text{ in Cluster 1}} \text{distance}(P_i, C_1)^2 + \sum_{P_i \text{ in Cluster 2}} \text{distance}(P_i, C_2)^2 + \sum_{P_i \text{ in Cluster 3}} \text{distance}(P_i, C_3)^2$$

Ainsi



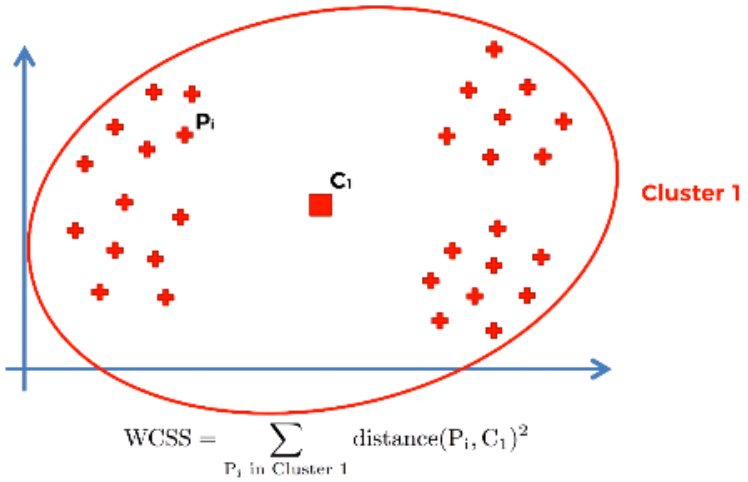
$$\text{WCSS} = \sum_{P_i \text{ in Cluster 1}} \text{distance}(P_i, C_1)^2 + \sum_{P_i \text{ in Cluster 2}} \text{distance}(P_i, C_2)^2 + \sum_{P_i \text{ in Cluster 3}} \text{distance}(P_i, C_3)^2$$

En quoi est-il utile pour évaluer la pertinence de notre modèle ?

CH – IV : CLUSTERING

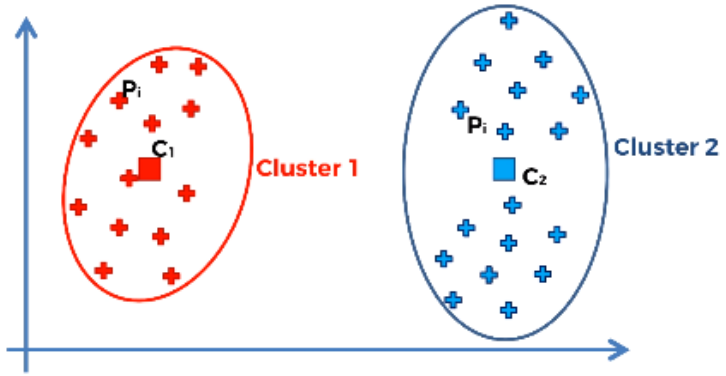
II – Choisir le bon nombre de cluster

Pour 1 cluster:



$$WCSS = \sum_{P_i \text{ in Cluster 1}} distance(P_i, C_1)^2$$

pour 2 Clusters



$$WCSS = \sum_{P_i \text{ in Cluster 1}} distance(P_i, C_1)^2 + \sum_{P_i \text{ in Cluster 2}} distance(P_i, C_2)^2$$

- Pour 1 cluster: Il y a des points très éloigné du centroide donc le WCSS sera très élevé
- Pour 2 Cluster : les points d'observation étant plus proche de leurs centroides on aura un WCSS moins élevé que précédemment
- Avec 3 clusters c'est la même observation
- .....

Mais quelle est la limite de cette observation ?

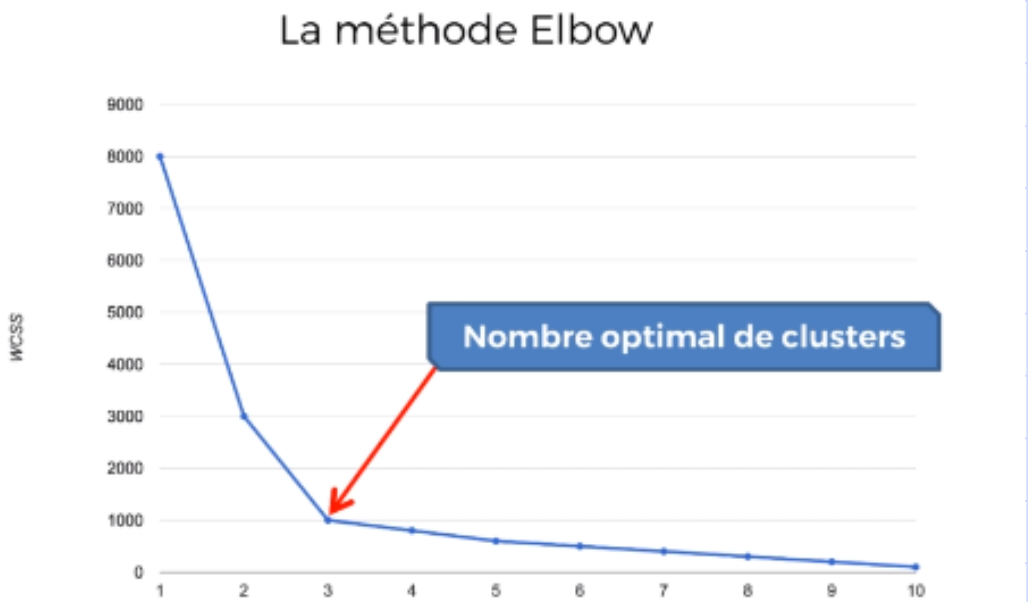
- Lorsque le nombre de clusters augment WCSS tend vers 0.

Comment choisir le nombre de clusters optimale ?

CH – IV : CLUSTERING

II – Choisir le bon nombre de cluster (La methode Elbow ou du coude)

Représentons l'évolution du WCSS en fonction du nombre de cluster



- WCSS part d'une valeur très élevées pour 1 cluster et chute pour tendre vers 0 (10 clusters)
- Cette chute est importante pour les trois premiers clusters puis est moins prononcées après
- C'est cette particularité qui nous permet de choisir le nombre de clusters optimale.
- ici le nombre optimale de cluster est 3.

NB: il existe d'autres méthodes pour déterminer le nombre optimale de clusters