Chapitre 2:
Apprentissage Automatique

1

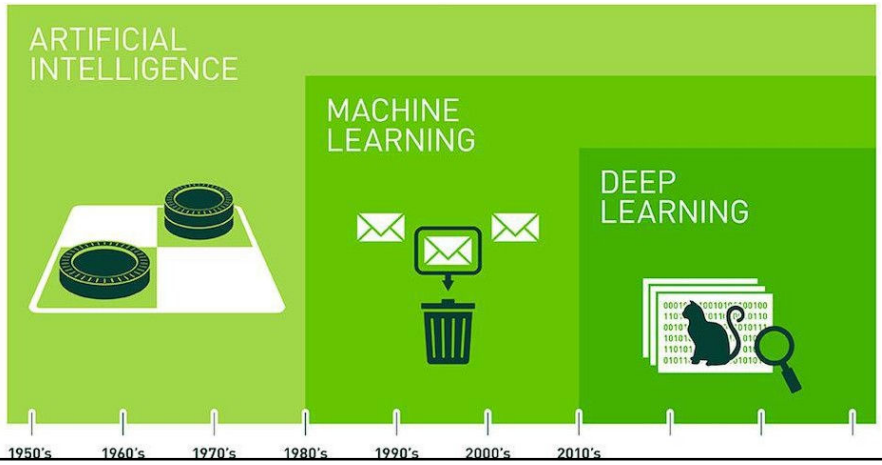
Apprentissage automatique

Apprentissage automatique = Apprentissage artificiel
= Machine learning

Wikipedia: « L'apprentissage automatique fait référence au **développement**, à l'**analyse** et à **l'implementation de méthodes** qui permettent à une **machine** (au sens large) d'évoluer grâce à un **processus d'apprentissage**, et ainsi de remplir des tâches qu'il est difficile ou impossible de remplir par des moyens algorithmiques plus classiques».

Définition

Le Machine Learning (ML) est une branche de l'IA qui permet aux machines d'apprendre à partir des données, sans être explicitement programmées pour chaque tâche.



Définitions

C'est un champ d'étude de l'intelligence artificielle qui se fonde sur des approches mathématiques et statistiques pour donner aux ordinateurs la capacité d' « apprendre » à partir de données, c'est à-dire d'améliorer leurs performances à résoudre des tâches sans être explicitement programmés pour chacune. Plus largement, il concerne la conception, l'analyse, l'optimisation, le développement et l'implémentation de telles méthodes.

3

Définitions

Ensemble de techniques permettant l'extraction de connaissances sous la forme de **modèles** à partir de grandes masses de données

Ces modèles peuvent être de nature :

- **Descriptive :** permettant d'expliquer le comportement actuel des données
- **Prédictive :** comportement futur des données

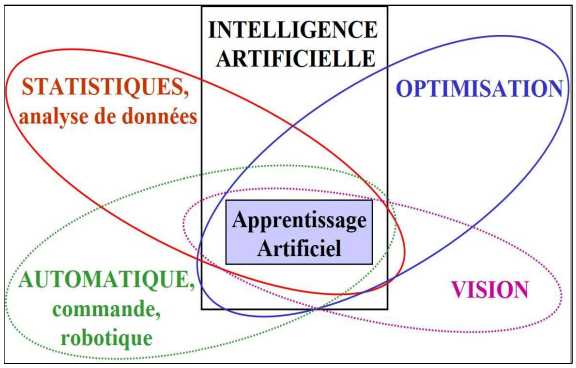
3

Apprentissage automatique: domaine pluri- disciplinaire

L'apprentissage automatique (AA) (**ML**) est à la croisée de plusieurs disciplines :

- Les **statistiques** : pour l'inférence de modèles à partir de données.
- Les **probabilités** : pour modéliser l'aspect aléatoire inhérent aux données et au problème d'apprentissage.
- L'**intelligence artificielle** : pour étudier les tâches simples de reconnaissance de formes que font les humains (comme la reconnaissance de chiffres par exemple), et parce qu'elle fonde une branche de l'AA dite symbolique qui repose sur la logique et la représentation des connaissances.
- L'**optimisation** : pour optimiser un critère de performance afin, soit d'estimer des paramètres d'un modèle, soit de déterminer la meilleure décision à prendre étant donné une instance d'un problème.
- L'**informatique** : puisqu'il s'agit de programmer des algorithmes et qu'en AA ceux-ci peuvent être de grande complexité et gourmands en termes de ressources de calcul et de mémoire

Apprentissage automatique: domaine pluri-disciplinaire



Principe de l'apprentissage

Le ML consiste à déduire des connaissances en utilisant des données dites d'entraînement. Par exemple:

« Étant donnée une collection de paires (entrées, sorties) appelées exemples d'apprentissage, comment apprendre (élaborer) un modèle qui puisse prédire correctement une sortie étant donnée une nouvelle entrée. »

Apprentissage automatique

Le Machine Learning est un concept qui se base sur le principe que :

- Il existe des algorithmes génériques
- ces algorithmes sont alimentés par les données
- sans avoir besoin de construire ou de développer un code spécifique. Ils peuvent à partir des données construire leurs propres logiques et générer automatiquement des codes

Apprentissage automatique

- Le terme a été défini dès 1959 par Arthur Lee Samuel
- Un champ d'études qui donne aux ordinateurs la capacité **d'apprendre** des tâches pour lesquelles ils ne sont pas spécifiquement programmés.

  ➢ Doter la machine d'un mécanisme d'apprentissage

  ➢ Concevoir des programmes pouvant s'améliorer automatiquement avec l'expérience

Pourquoi etudier l'apprentissage automatique?

Le ML permet:

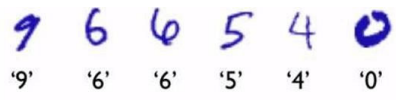
- Développement de systèmes capables de s'adapter aux changements,
- Conception d'une intelligence artificielle autonome et évolutive
- Anticipation et prédiction basées sur l'analyse des données

Apprentissage automatique

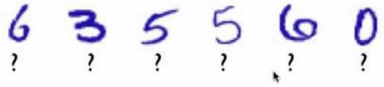
Principe d'apprentissage: « données d'entrées vs. généralisation «

Les algorithmes d'apprentissage précèdent comme suit:

- On fournit à l'algorithme des données d'entrainement:



- Et l'algorithme retourne un «programme» capable de se généraliser à de nouvelles données

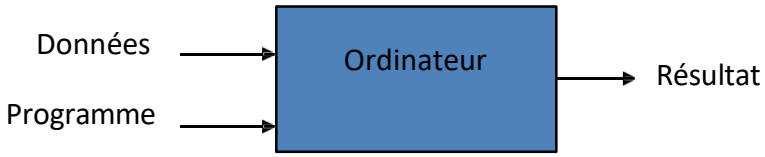


Programmation Vs apprentissage automatique

- Programmation traditionnelle

- Apprentissage automatique (ML)

Programmation traditionnel



Apprentissage automatique

Problème: Estimation du prix d'une maison

Pour illustrer l'importance de ML. Supposons qu'on veut ecrire un programme qui permet d'estimer le prix d'une maison.

Prix ??

Prix ??

Programme estimation prix maison !!

Fonction estimation_prix_vente_maison(***nombre_de_chambres, superficie, quartier***)

```
prix = 0
# A côté de chez moi, le prix moyen au m² est de 1800 dinars
prix_par_metre_carre = 1800 if quartier == 'Ezzahra'
# mais quelques quartiers sont plus chers...
prix_par_metre_carre = 3000 elseif quartier== 'Lac2'
# ...et d'autres moins
prix_par_metre_carre= 500
# commençons avec un prix estimé à partir de la superficie du lieu
prix = prix_par_metre_carre * superficie
# ajustons maintenant notre prix en fonction du nombre de chambres
if nombre_de_chambres == 0:
# les studios sont moins chers (moins de murs)
prix = prix—20000
else:
# les lieux avec plus de chambres ont en général plus de valeurs
prix = prix + (nombre_de_chambres * 1000)
…….
……..
return prix
```

Programme estimation prix maison !!

Fonction estimation_prix_vente_maison(***nombre_de_chambres, superficie, quartier***)

```
prix = 0
# A côté de chez moi, le prix moyen au m² est de 1800 dinars
prix_par_metre_carre = 1800 if quartier == 'Ezzahra'
# mais quelques quartiers sont plus chers...
prix_par_metre_carre = 3000 elseif quartier== 'Lac2'
# ...et d'autres moins
prix_par_metre_carre= 500
# commençons avec un prix estimé à partir de la superficie du lieu
prix = prix_par_metre_carre * superficie
# ajustons maintenant notre prix en fonction du nombre de chambres
if nombre_de_chambres == 0:
# les studios sont moins chers (moins de murs)
prix = prix—20000
else:
# les lieux avec plus de chambres ont en général plus de valeurs
prix = prix + (nombre_de_chambres * 1000)
…….
……..
return prix
```

Problème!! ?

Programme estimation prix maison !!

- ☹ Le programmeur doit être connaisseur du domaine de vente des maisons
- ☹ La qualité de prédiction dépondra de l'expertise du programmeur
- ☹ Le nombre de paramètre à inclure dans le programme est réduit (exemple: vous ne pouvez pas spécifier tout les quartiers !!!)
- ☹ Si les prix changes, on reprogramme tout !!! L'année prochaine même estimation ??

Programme estimation prix maison !!

Pour palier à ces probelèmes :

→Il faut apprendre à la machine la manière dont un agent immobilier détermine le prix d'une maison

Comment un agent immobilier peut-il déterminer le prix exacte d'une maison ?

- Un agent immobilier qui vend des maisons depuis longtemps, peut déterminer le prix d'une maison par similitude :

- Exemple: il sait qu'une maison dans le même quartier ayant les mêmes caractéristiques a été vendu à un tel prix. Donc la maison vaut le même prix

On doit reprendre ce principe pour la machine

Format des données en apprentissage

Variables, caractères, caractéristique, attributs, descripteurs, champs, etc.
décrit quantité ou une qualité observée d'une instance

$$X(n,p) = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \vdots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{bmatrix}$$

instances , Individus, observations, objets, enregistrements, etc
L' ensemble de valeurs pris par *x*.

L'espace de toutes les observations possibles est noté ***X***

p le nombre de variables
n le nombre d'instances

Format des données en apprentissage

Nom	Grade	Années	Titulaire
David	Assistant	3	non
Marie	Assistant	7	oui
Jean	Professeur	2	oui
Jim	Prof. Associé	7	oui
Pierre	Assistant	6	non
Anne	Prof associé	3	non

Instances

X
- ***n= 6***
- ***P=4***

Ensemble d'enseignants

Format des données en apprentissage

On veut savoir qui sont les enseignants titulaires

Nom	Grade	Années	Titulaire
David	Assistant	3	non
Marie	Assistant	7	oui
Jean	Professeur	2	oui
Jim	Prof. Associé	7	oui
Pierre	Assistant	6	non
Anne	Prof associé	3	non

Variable à prédire
Attribut classe
Variable endogène

Variables prédictives
Descripteurs
Variables exogènes

Format des données en apprentissage

- Les colonnes sont appelées **variables d'entrée, attributs ou caractéristiques.**
- Titulaire (que nous essayons de prédire) est appelée variable résultat ou **la cible**.
- Une ligne du tableau est appelée **un exemple d'entrainement ou instance**.
- Le tableau en entier est appelé **l'ensemble d'entrainement**.

Format des données en apprentissage: exemple

Training set of housing prices (Portland, OR)

Size in feet² (x)	Price ($) in 1000's (y)
2104	460
1416	232
1534	315
852	178
…	…

Notation:

m = Number of training examples

x's = "input" variable / features

y's = "output" variable / "target" variable

(x,y) one training example (one raw)

$(x^{(i)},y^{(i)})$ i th training example

l'ensemble d'entrainement???

Format des données en apprentissage: exemple

Training set of housing prices (Portland, OR)

Size in feet² (x)	Price ($) in 1000's (y)
2104	460
1416	232
1534	315
852	178
...	...

m

Notation:

- **m** = Number of training examples
- **x**'s = "input" variable / features
- **y**'s = "output" variable / "target" variable

(x,y) one training example (one raw)

$(x^{(i)},y^{(i)})$ i th training example

Format des données en apprentissage: exemple

Training set of housing prices (Portland, OR)

Size in feet² (x)	Price ($) in 1000's (y)
2104	460
1416	232
1534	315
852	178
...	...

m

Notation:

- **m** = Number of training examples
- **x**'s = "input" variable / features
- **y**'s = "output" variable / "target" variable

(x,y) one training example (one raw)

$(x^{(i)},y^{(i)})$ i th training example

Example

$x^{(1)}$ =?

$y^{(2)}$ =?

$x^{(4)}$ =?

Format des données en apprentissage: exemple

Training set of housing prices (Portland, OR)

Size in feet² (x)	Price ($) in 1000's (y)
2104	460
1416	232
1534	315
852	178
...	...

m

Notation:

- **m** = Number of training examples
- **x**'s = "input" variable / features
- **y**'s = "output" variable / "target" variable

(x,y) one training example (one raw)

$(x^{(i)},y^{(i)})$ i th training example

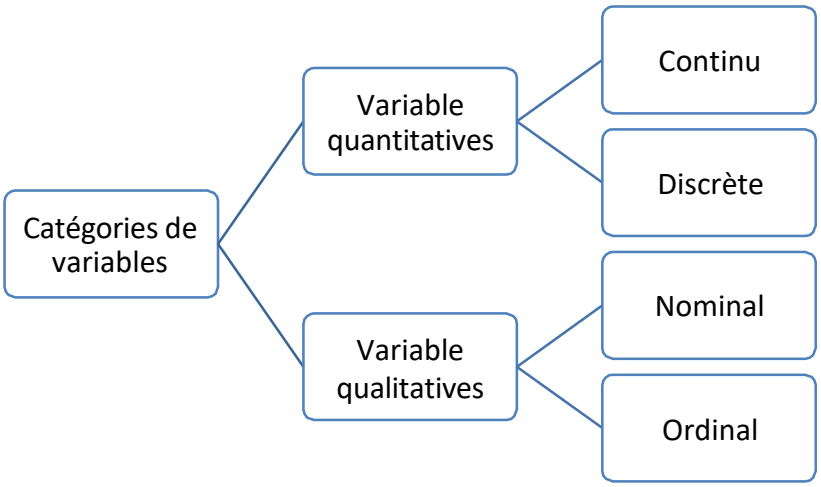
Example

$x^{(1)}$ 2104

$y^{(2)}$ 232

$x^{(4)}$ 852

Format des données en apprentissage



Format des données en apprentissage

- **Les variables qualitatives:**
  - sont des variables représentées par des qualités (sexe, état civil, travail)
  - s'expriment en modalités (sexe = femme\homme)
- **Les variables quantitatives:**
  - sont des variables représentées par des quantités (l'âge, le poids, la taille.) souvent mesurées ou comptées
  - Elles s'expriment en valeurs

Format des données en apprentissage

- **Les variables qualitatives nominales:**
  - il n'y a aucun ordre précis.
  - Exemple:

    Sexe **{féminin, masculin}**.

    Saveur de crème glacée **{chocolat ,vanille, pistache}**.
- **Les variables qualitatives ordinales:**
  - sont des variables qui contiennent un ordre.
  - Exemple

    Degré de satisfaction : **{insatisfait, satisfait, très satisfait}**.

    Niveau d'étude : **{primaire, secondaire, universitaire}**.

Format des données en apprentissage

- **Les variables quantitatives continues:**
  - Peuvent prendre n'importe quelle valeur numérique dans un intervalle.
  - Exemple: **Taille, Température**
- **Les variables quantitatives discrètes:**
  - Peuvent seulement prendre des valeurs numériques spécifiques mais ces valeurs numériques ont une interprétation quantitative claire
  - Exemple: **le nombre de plaintes de clients, Nombre d'heures supplémentaires**

Format des données en apprentissage

Exercice: les variables suivantes sont-elles quantitatives ou qualitatives?

- Les marques des voitures garées sur un parking de supermarché

.

Format des données en apprentissage

Exercice: les variables suivantes sont-elles quantitatives ou qualitatives

- Les marques des voitures garées sur un parking de supermarché

Qualitative nominale

- Les nationalités des touristes se rendant au festival de Carthage.

Format des données en apprentissage

Exercice: les variables suivantes sont-elles quantitatives ou qualitatives

- Les marques des voitures garées sur un parking de supermarché

Qualitative nominale

- Les nationalités des touristes se rendant au festival de Carthage.

Qualitative nominale

- L'âge des auditeurs de Radio Mosaïque.

Format des données en apprentissage

Exercice: les variables suivantes sont-elles quantitatives ou qualitatives

- Les marques des voitures garées sur un parking de supermarché

Qualitative nominale

- Les nationalités des touristes se rendant au festival de Carthage.

Qualitative nominale

- L'âge des auditeurs de Radio Mosaïque.

Quantitative continu

- Les températures matinales relevées chaque jour sous abri à Nabeul

Format des données en apprentissage

Exercice: les variables suivantes sont-elles quantitatives ou qualitatives

- Les marques des voitures garées sur un parking de supermarché

Qualitative nominale

- Les nationalités des touristes se rendant au festival de Carthage.

Qualitative nominale

- L'âge des auditeurs de Radio Mosaïque.

Quantitative continu

- Les températures matinales relevées chaque jour sous abri à Nabeul

Quantitative continu

Format des données en apprentissage

Exercice: les variables suivantes sont-elles quantitatives ou qualitatives

- Les marques des voitures garées sur un parking de supermarché

Qualitative nominale

- Les nationalités des touristes se rendant au festival de Carthage.

Qualitative nominale

- L'âge des auditeurs de Radio Mosaïque.

Quantitative continu

- Les températures matinales relevées chaque jour sous abri à Nabeul

Quantitative continu

Qu'est ce qu'on fait dans le ML?

Dans ML , on cherche à résoudre des problèmes. Par exemple:

- **Exemple 1 :** Supposons que l'on dispose d'un certain nombre d'images représentant des chiens, et d'autres représentant des chats. Comment classer automatiquement une nouvelle image dans une des catégories « chien » ou « chat » ?

4

Qu'est ce qu'on fait dans le ML?

Dans ML , on cherche à résoudre des problèmes. Par exemple

- **Exemple 1 :** Supposons que l'on dispose d'un certain nombre d'images représentant des chiens, et d'autres représentant des chats. Comment classer automatiquement une nouvelle image dans une des catégories « chien » ou « chat » ?

- **Exemple 2 :** Supposons que l'on dispose d'une collection d'articles de journaux. Comment identifier des groupes d'articles portant sur un même sujet ?

4

Qu'est ce qu'on fait dans le ML?

Dans ML , on cherche à résoudre des problèmes. Par exemple

- **Exemple 1 :** Supposons que l'on dispose d'un certain nombre d'images représentant des chiens, et d'autres représentant des chats. Comment classer automatiquement une nouvelle image dans une des catégories « chien » ou « chat » ?

- **Exemple 2 :** Supposons que l'on dispose d'une collection d'articles de journaux. Comment identifier des groupes d'articles portant sur un même sujet ?

- **Exemple 3 :** Supposons que l'on dispose d'une base de données regroupant les caractéristiques de logements dans une ville : superficie, quartier, étage, prix, année de construction, nombre d'occupants, montant des frais de chauffage. Comment prédire la facture de chauffage à partir des autres caractéristiques pour un logement qui n'appartiendrait pas à cette base?

4

Sources de données et applications

Le machine Learning est exploité grâce à la grande quantité de données disponible (Big Data):

- **Données textuelles** (réseaux sociaux, articles de presses, échanges des mails, livres sur le web …)
- **Données vidéos** (médias partagées, télévision sur le web, vidéo surveillance …)
- **Données images** (images partagées, images sur le web …)
- **Données numériques** (transactions bancaires, historique d'achats des clients,données d'une entreprise, données de bâtiments …)

…

5

Termes et concepts de base

- **Ensemble de données (Dataset)** : fait référence à un ensemble de données utilisées dans les tâches d'apprentissage automatique. Chaque donnée est appelée un échantillon. L'événement ou l'attribut qui reflète la performance ou la nature d'un échantillon dans un certain aspect est appelé une caractéristique (**feature**).

- **Ensemble d'apprentissage (training set)** : fait référence à un ensemble de données utilisé dans le processus d'apprentissage, où chaque échantillon est appelé échantillon d'apprentissage. Le processus d'apprentissage d'un modèle à partir de données est appelé apprentissage (**training**).

Termes et concepts de base

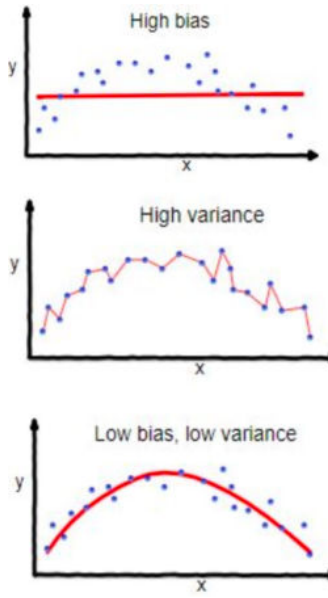
- **Ensemble de test (Test set)** : le test fait référence au processus d'utilisation du modèle appris pour la prédiction. L'ensemble de données utilisé est appelé ensemble de test et chaque échantillon est appelé échantillon de test.

- **Capacité de généralisation (Generalization capability):** l'objectif de l'apprentissage automatique est que le modèle appris fonctionne bien sur les nouveaux échantillons, et pas seulement sur ceux sur lesquels le modèle a été formé. La capacité de bien performer sur de nouveaux échantillons est appelée capacité de généralisation.

Termes et concepts de base

- **Erreur** : fait référence à la différence entre le résultat de l'échantillon prédit par le modèle appris et le résultat de l'échantillon réel.
  - ***Erreur d'apprentissage*** : erreur du modèle sur l'ensemble d'apprentissage
  - ***Erreur de généralisation*** : erreur sur le nouvel échantillon. Évidemment, nous préférons un modèle avec une erreur de généralisation plus faible.
- **Sous-apprentissage (Underfitting)** : se produit lorsque l'erreur d'entraînement est trop importante.
- **Surapprentissage (Overfitting)** : se produit lorsque l'erreur d'apprentissage du modèle appris est faible mais que l'erreur de généralisation est importante (faible capacité de généralisation).

Termes et concepts de base



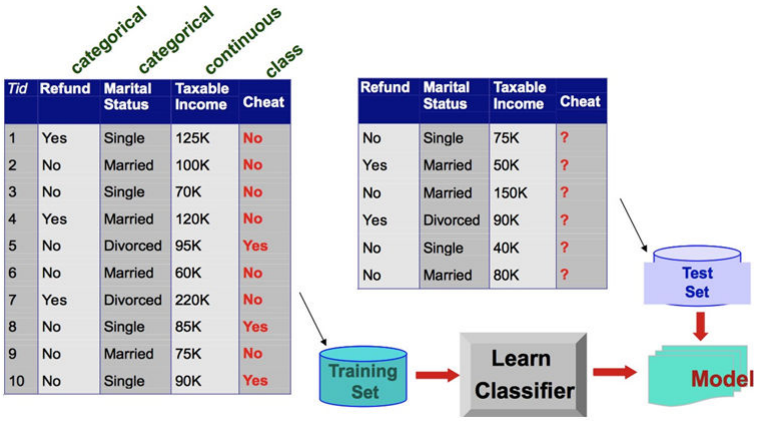
Termes et concepts de base

- **Capacité d'un modèle** : fait référence à la capacité de s'adapter à une grande variété de fonctions. Les algorithmes d'apprentissage automatique fonctionnent généralement mieux lorsque leur capacité est appropriée à la véritable complexité de la tâche qu'ils doivent effectuer et à la quantité de données d'entraînement qui leur sont fournies. Les modèles avec une capacité insuffisante sont incapables de résoudre des tâches complexes. Les modèles avec une capacité élevée peuvent résoudre des tâches complexes, mais lorsque leur capacité est supérieure à celle nécessaire pour résoudre la tâche actuelle, ils peuvent être surdimensionnés.

Jeux de données

Deux jeux de données sont utilisés :

- **Base d'apprentissage (Training set)** : 80% de la base initiale utilisée pour apprendre le modèle
- **Base de test (Test set)** : 20 % de la base initiale pour tester le modèle



Exemple

Nom	Grade	Années	Titulaire
David	Assistant	3	non
Marie	Assistant	7	oui
Jean	Professeur	2	oui
Jim	Prof. associé	7	oui
Pierre	Assistant	6	non
Anne	Prof. associé	3	non
Paul	Assistant	8	oui
Rose	Assistant	1	non
Adam	Prof. associé	2	non
Jack	Prof. associé	5	oui

Échantillon d'apprentissage (70%) Utilisé pour la construction du modèle

Échantillon test 30% Utilisé pour l'évaluation du modèle

Présentation de l'environnement Python pour le ML

Pourquoi Python ?

- Langage polyvalent, simple à apprendre et largement utilisé dans la communauté scientifique.
- Grande richesse de bibliothèques dédiées à l'analyse de données et au Machine Learning.

Environnements de travail

- Google Colab : Idéal pour les tests rapides et pour partager du code sans installation locale.
- Jupyter Notebook : Pour une approche interactive et pédagogique.

Bibliothèques essentielles

- NumPy : Pour les opérations sur les tableaux et la manipulation de données numériques.
- Pandas : Pour la gestion et l'analyse des jeux de données.
- Matplotlib/Seaborn : Pour la visualisation des données.
- Scikit-learn : Pour la mise en œuvre de modèles ML (régression, classification,

Bibliothèques essentielles

➢ **NumPy (Python numérique)**

NumPy est la base du calcul numérique. Il fournit de puissants tableaux à N dimensions et des outils pour travailler avec eux, rendant efficaces la manipulation des données et les opérations numériques.

Caractéristiques:
- Fournit des fonctions rapides et précompilées pour les routines numériques.
- Permet une informatique orientée tableau pour une meilleure efficacité.
- Prend en charge une approche orientée objet pour la manipulation des données.
- Calculs compacts et plus rapides avec vectorisation.

Applications:
- Largement utilisé dans l'analyse de données pour ses capacités numériques.
- Crée de puissants tableaux à N dimensions pour le stockage de données structurées.
- Constitue la base d'autres bibliothèques, telles que SciPy et scikit-learn.
- Peut servir d'alternative à MATLAB lorsqu'il est utilisé avec SciPy et Matplotlib

Bibliothèques essentielles

- **Pandas**

Pandas est votre bibliothèque incontournable pour la manipulation et l'analyse des données. Il propose des structures de données telles que DataFrames et Series, simplifiant le traitement des données.

Caractéristiques:
- Outils de manipulation de données pour nettoyer et transformer les données.
- Fonctionnalité de série chronologique pour gérer les données liées au temps.
- Fonctionnalités d'alignement des données pour fusionner des ensembles de données.
- Gère les données manquantes avec élégance.

Applications:
- Exploration et nettoyage des données, notamment avec des données tabulaires.
- Analyse et visualisation de données tabulaires.
- Analyse de données de séries chronologiques pour les prévisions et l'analyse des tendances.
- o Prétraitement des données pour les tâches d'apprentissage automatique.

Bibliothèques essentielles

- **Scikit-Learn**

Scikit-Learn est une bibliothèque polyvalente proposant une large gamme d'algorithmes d'apprentissage
automatique pour la classification, la régression, le clustering, etc.

Caractéristiques:
- Des outils simples et efficaces pour l'analyse et la modélisation des données.
- API cohérente pour un développement de modèles facile.
- Divers algorithmes d'apprentissage automatique pour diverses tâches.
- Sélection et évaluation de modèles pour l'optimisation des performances.

Applications:
- Tâches de classification et de régression en apprentissage supervisé.
- Clustering et réduction de dimensionnalité pour l'apprentissage non supervisé.
- Sélection du modèle et réglage des paramètres pour optimiser les performances du modèle.
- Évaluation et comparaison de modèles pour choisir le meilleur modèle pour votre tâche

Bibliothèques essentielles

- **Matplotlib**

  Matplotlib est la bibliothèque de confiance pour créer des tracés statiques, animés et interactifs en Python. C'est parfait pour visualiser des données.

  Caractéristiques:

  - Bibliothèque complète pour différents types de parcelles.
  - Styles et thèmes d'intrigue personnalisables.
  - Chiffres de qualité de publication pour les rapports et les publications.
  - Intégration avec les notebooks Jupyter pour le traçage interactif.

  Applications:

  - Visualisation des données pour l'analyse exploratoire des données.
  - Création de tableaux et de graphiques pour des présentations et des rapports.
  - Créer des visualisations interactives pour des applications Web.
  - Traçage de données pour la recherche scientifique et la communication de données

Bibliothèques essentielles

- **Né en mer**

  Seaborn est une interface de niveau supérieur construite sur Matplotlib, offrant des graphiques statistiques attrayants.

  Caractéristiques:

  - Interface de haut niveau pour créer des graphiques statistiques élégants.
  - Thèmes et palettes de couleurs intégrés pour une personnalisation facile.
  - Fonctions de visualisation de modèles de régression linéaire.
  - Intégration transparente avec les structures de données Pandas.

  Applications:

  - Visualisation de données élégante pour explorer et présenter les données.
  - Visualiser les relations et les modèles dans les données.
  - Présenter les résultats statistiques de manière engageante.
  - Créer des tableaux et des graphiques informatifs et visuellement attrayants

Bibliothèques essentielles

- **TensorFlow**

TensorFlow est une bibliothèque d'apprentissage profond open source, développée par Google. Il est
largement utilisé pour les tâches d'apprentissage automatique basées sur les réseaux neuronaux.

Caractéristiques:
- Cadre d'apprentissage profond avec des applications polyvalentes.
- API de haut niveau comme Keras pour un développement rapide de modèles.
- TensorBoard pour visualiser les réseaux de neurones.
- Prend en charge l'informatique distribuée pour les tâches à grande échelle.

Applications:
- Réseaux de neurones profonds pour la reconnaissance d'images et la détection d'objets.
- Modèles de traitement du langage naturel pour l'analyse de texte.
- Prévision de séries chronologiques à l'aide de réseaux de neurones récurrents.
- Créer des modèles d'apprentissage profond personnalisés pour des tâches spécifiques

Bibliothèques essentielles

- **PyTorch**

PyTorch est une autre bibliothèque d'apprentissage en profondeur connue pour ses graphiques de calcul dynamiques et son interface conviviale.

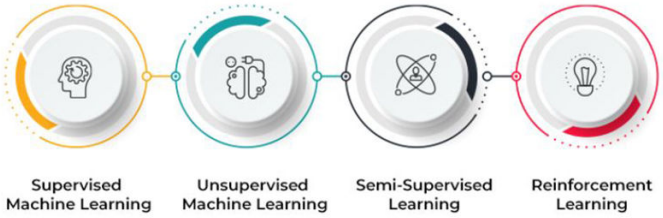
Caractéristiques:
- Graphiques de calcul dynamique pour une conception de modèle flexible.
- Tenseurs pour les calculs numériques et les calculs de gradient.
- Module de réseau neuronal pour créer des modèles d'apprentissage en profondeur.
- Prise en charge forte de l'accélération GPU pour un entraînement plus rapide.

Applications:
- Largement utilisé dans les milieux universitaires et de recherche pour les projets d'apprentissage profond.
- Implémentation de tâches de traitement du langage naturel et de vision par ordinateur.
- Formation de modèles d'apprentissage profond personnalisés avec des architectures flexibles.
- Résoudre des problèmes complexes d'apprentissage automatique avec des calculs efficaces

Les types d'apprentissage en ML

Pour donner à un ordinateur la capacité d'apprendre, on utilise des **méthodes d'apprentissage** qui sont fortement inspirées de la façon dont nous, les êtres humains, apprenons à faire des choses. Parmi ces méthodes, on compte :

- L'apprentissage **supervisé** (Supervised Learning)
- L'apprentissage **non supervisé** (Unsupervised Learning)
- L'apprentissage par **renforcement** (Reinforcement Learning)



Approches d'apprentissage automatique

1. **Supervised learning (Apprentissage supervisé):**
   - les classess sont connues a priori
   - Learning by examples: C'est-à-dire , On a des:

     Données d'entraînement + résultats souhaités (labels))
   - Modélisation prédictive

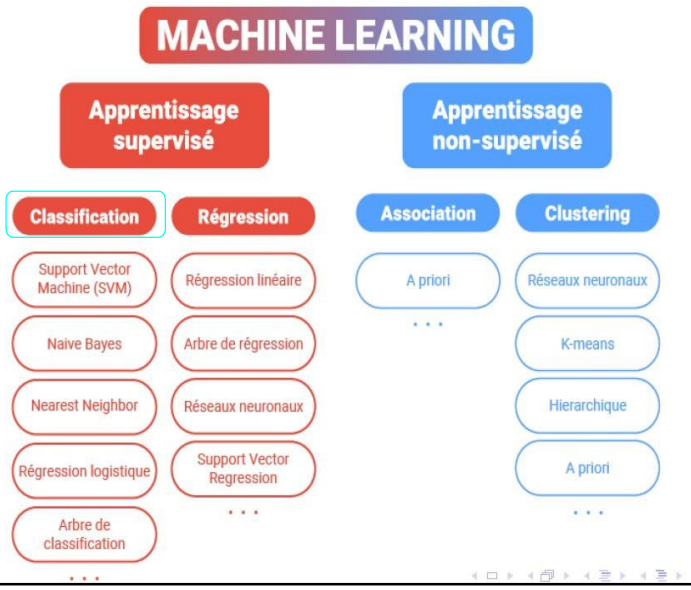
   ➡ **Objectif: Modéliser et prévoir**
2. **Unsupervised learning (Apprentissage non supervisé):**
   - On doit chercher les groupes et leur signification (Label).
   - Learning by observations: C'est-à-dire, On a des:

     Données d'entraînement sans résultats souhaités
   - Modélisation descriptive, exploratoire

   ➡ **Objectif: Organiser et synthétiser**

- **Autres:** Semi supervised, Reinforcement learning, deep learning

Approches d'apprentissage automatique



Apprentissage supervisé

Apprentissage supervisé

Imaginez que vous commenciez à apprendre le chinois.

Pour ce faire, il vous faudra soit acheter un livre de traduction chinois-français, ou bien trouver un professeur de chinois.

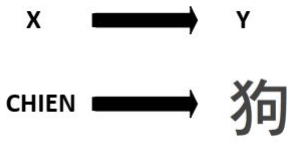
Apprentissage supervisé

Le rôle du professeur ou du livre de traduction sera de **superviser** votre apprentissage en vous fournissant des **exemples** de traductions français-chinois que vous devrez mémoriser.

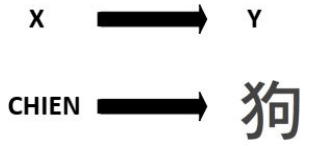
On parle ainsi **d'apprentissage supervisé** lorsque l'on fournit à une machine beaucoup **d'exemples** qu'elle doit étudier.

Apprentissage supervisé

Comme pour apprendre la langue chinoise, on parle d'apprentissage supervisé lorsque l'on fournit à une machine beaucoup **d'exemples** (x, y) dans le but de lui faire apprendre la **relation** qui **relie x à y**.



Apprentissage supervisé



En Machine Learning, on compile ces **exemples** (x, y) dans un tableau que l'on appelle **Dataset** :

- La variable y porte le nom de ***target*** (la cible). C'est la valeur que l'on cherche à prédire.
- La variable x porte le nom de ***feature*** (facteur). Un facteur influence la valeur de y, et on a en général beaucoup de ***features*** ($x1$, $x2$, …) dans notre Dataset que l'on regroupe dans une matrice **X.**

Apprentissage supervisé

Ci-dessous, un **Dataset** qui regroupe des exemples d'appartements avec leur prix y ainsi que certaines de leurs **caractéristiques** (*features*).

Target y	Features x_1	x_2	x_3
Prix	Surface m2	N chambres	Qualité
€ 313,000.00	124	3	1.5
€ 2,384,000.00	339	5	2.5
€ 342,000.00	179	3	2
€ 420,000.00	186	3	2.25
€ 550,000.00	180	4	2.5
€ 490,000.00	82	2	1
€ 335,000.00	125	2	2

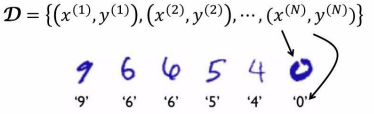
Apprentissage supervisé

Principe de l'apprentissage supervisé

Les algorithmes d'apprentissage supervisé procèdent comme suit:

- On fournit à l'algorithme des données d'entrainement:

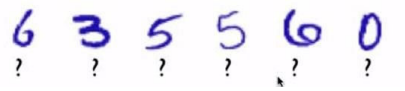
$$\mathcal{D} = \{(x^{(1)}, y^{(1)}), (x^{(2)}, y^{(2)}), \cdots, (x^{(N)}, y^{(N)})\}$$



- On appelle $x(i)$ l'entrée et $y(i)$ la cible du i-iéme exemple.
- Un élément de D est appelé exemple d'apprentissage ou une instance de données.

Apprentissage supervisé

- L'algorithme retourne un «programme» capable de se généraliser à de nouvelles données:

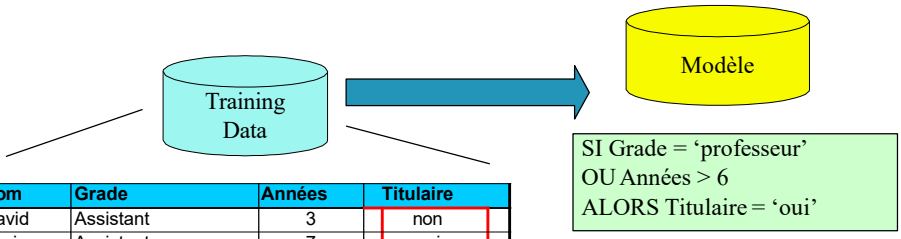


On note le «programme» généré par l'algorithme d'apprentissage *f(x)*.

- On appelle *f(x)* un modèle ou une hypothèse.
- On utilise souvent un ensemble de test *Dtest* pour mesurer la performance du modèle *f(x)*.

Apprentissage supervisé

- On cherche à produire automatiquement des règles à partir d'une base de données d'**apprentissage** contenant des « exemples » (en général des cas déjà traités et validés).



Nom	Grade	Années	Titulaire
David	Assistant	3	non
Marie	Assistant	7	oui
Jean	Professeur	2	oui
Jim	Prof. Associé	7	oui
Pierre	Assistant	6	non
Anne	Prof associé	3	non

Apprentissage supervisé

Population Ω:
(objet de l'étude)
ensemble de paires ***(xi , yi),***

- ***X***: X***=(x1|…|xp)*** variables exogènes
- ***Y*** : variable à prédire (endogène)

On veut construire une fonction de prédiction telle que
$y_i = f(x_i)$

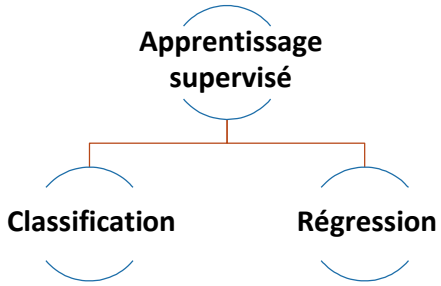
L'objectif de l'apprentissage:
Trouver une bonne approximation ***h*** de ***f*** en utilisant un échantillon de Ω telle que l 'on minimise l 'erreur théorique.

Types d'apprentissage supervisé

- Si ***f*** est une *fonction continue* **(variable réponse continue)**
  - Prédiction
- Si ***f*** est une *fonction discrète* ***(variable réponse qualitative)***
  - Classement (catégorisation ou en anglais « classification »)

Apprentissage supervisé: Classification & Régression

L'apprentissage supervisé est généralement effectué dans le contexte de la classification et de la régression.



Classification vs Régression

Classification

Un problème de classification survient lorsque la variable de sortie est une catégorie, telle que «rouge», «bleu» ou «maladie» et «pas de maladie».

Régression

Un problème de régression se pose lorsque la variable de sortie est une valeur réelle, telle que «dollars» ou «poids».

Classification vs Régression

Classification

- Les étiquettes de données (sorties) sont des valeurs discrètes: appartenant à un ensemble fini de valeurs
- Étiquette= classe

Régression

- Les étiquettes de données (tsorties) sont des valeurs continues: appartenant à un ensemble infini de valeurs

Exemples de problèmes de Classification vs Régression

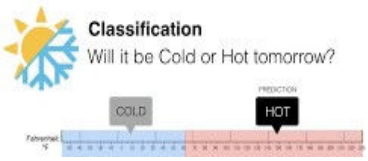
classification :

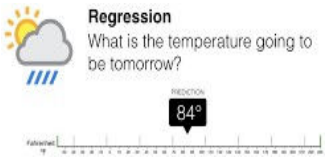
- En finance et dans le secteur bancaire pour la détection de la fraude par carte de crédit (fraude, pas fraude).
- Détection de courrier électronique indésirable (spam, pas spam).
- Dans le domaine du marketing utilisé pour l'analyse du sentiment de texte (heureux, pas heureux).
- En médecine, pour prédire si un patient a une maladie particulière ou non.

Régression :

- Prédire le prix de l'immobilier
- Prédiction du montant des ventes d'une entreprise compte tenu du contexte économique.
- Prédiction du prix de vente d'une maison en fonction de plusieurs critères
- Prédiction de la taille d'un enfant à un certain âge
- Prédiction de la consommation électrique dans une ville étant donné des conditions météorologiques

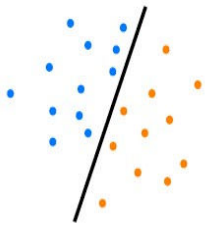
Exemples de problèmes de Classification vs Régression



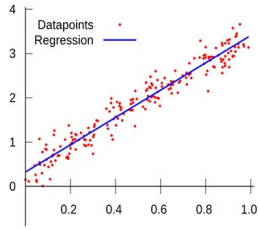


Apprentissage supervisé: Classification & Régression

Classification:
chercher les frontières séparant les données de différentes classes

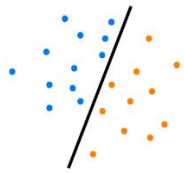


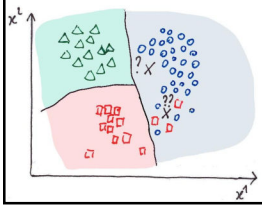
Régression:
chercher la courbe qui colle le plus possible aux données



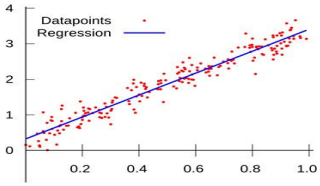
Apprentissage supervisé: Classification & Régression

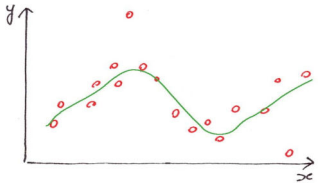
Classification:
chercher les frontières séparant les données de différentes classes





Régression:
chercher la courbe qui colle le plus possible aux données





Prédiction: Régression

L'analyse de la régression

- L'analyse de la régression est une méthode statistique qui permet d'étudier le type de relation pouvant exister entre une certaine variable (dépendante) dont on veut expliquer les valeurs et une ou plusieurs autres variables qui servent à cette explication (variables indépendantes)
  - Régression linéaire simple: une variable indépendante
- En d'autres termes, l'analyse de la régression permet d'étudier les variations de la variable dépendante en fonction des variations connues des variables indépendantes

En bref

Définition :

Méthode statistique pour modéliser la relation entre variables

Objectif :

Prédire une variable dépendante (Y) à partir de variable(s) indépendante(s) (X)

Régression simple vs multiple

Une analyse de régression est :

- dite simple si elle permet de prédire les valeurs d'une variable dite dépendante (expliquée (Y)) à partir des valeurs prises par une autre variable dite indépendante (explicative (X)).

- dite multiple si elle permet de prédire les valeurs d'une variable dite dépendante (expliquée (Y)) à partir des valeurs prises par plusieurs autres variables dites indépendantes (explicatives (X_j)).

Nuage de points ou diagramme de dispersion

Définition :

C'est la représentation graphique dans le plan cartésien de l'ensemble des paires de données (x_i,y_i). Ces données proviennent d'une série statistique de deux variables obtenues à partir d'une étude menée sur un échantillon ou sur une population.

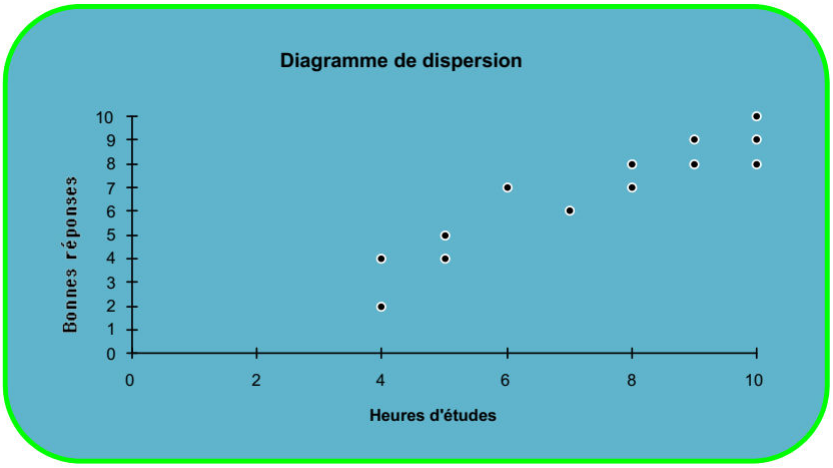
Exemple : Nuage de points ou diagramme de dispersion

Supposons que le nombre d'heures d'études nécessaires pour preparer l'examen final en statistiques et le nombre de bonnes réponses obtenues par chaque étudiant sont donnés dans le tableau suivant :

Heures d'étude (X)	5	8	6	9	10	8	5	4	10	4	10	7	9
Bonne réponde (Y)	5	8	7	9	10	7	4	4	8	2	9	6	8

Dessiner le nuage de points ou le diagramme de dispersion

Exemple : Nuage de points ou diagramme de dispersion …



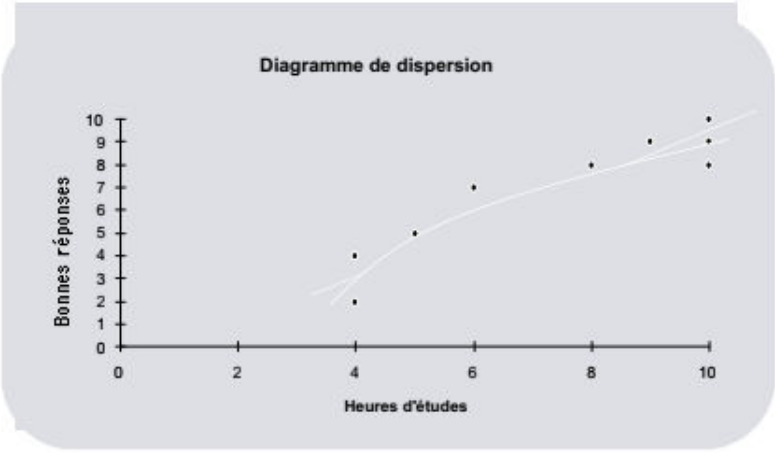
Objectif d'une analyse de régression simple

Une fois la représentation graphique effectuée, il est facile de soupçonner l'existence d'une certaine relation entre les deux variables (caractères étudiés). Il faut maintenant chercher à exprimer cette relation à l'aide d'une équation mathématique. C'est a dire:

On essaie de trouver la forme mathématique de la fonction f

$$Y = f(X)$$

Objectif d'une analyse de régression simple



Remarque: La forme de la courbe est tres proche de la forme d'une droite

La regression linéaire

En termes algébriques, le modèle serait défini comme suit :

***y=wx+b*,**

- **Y:** est la valeur que nous voulons prédire.
- **W:** est la pente de la droite.
- ***x:*** correspond à notre valeur d'entrée.
- **B:** est l'ordonnée à l'origine.

Régression linéaire

En général, L'hypothèse d'une régression linéaire s'écrit sous la forme suivante :

$$y = f(x) = w_0 + w_1 x_1 + \cdots$$

où $x = (x_1, x_2, \ldots, x_D)$.

Régression linéaire

Définition:

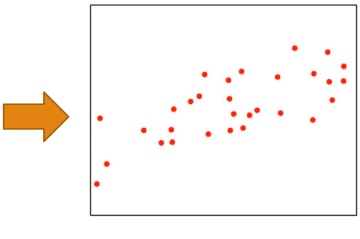
- Nous appelons régression linéaire l'ajustement d'une droite au nuage statistique d'une série de couples de données.

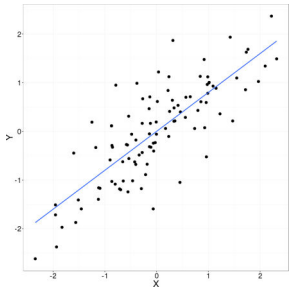
- Ainsi, une régression linéaire simple va permettre de résumer, d'interpréter et de prévoir les variations d'un caractère dit dépendant (Y) en fonction d'un autre dit indépendant (X) et ce en utilisant une droite.

Exemple régression

Exemple: (régression linéaire)

x	y
0.86	2.49
0.09	0.83
-0.85	-0.25
0.87	3.10
-0.44	0.87
-0.43	0.02
-1.10	-0.12
0.40	1.81
-0.96	-0.83
0.17	0.43





Exemple

Si vous cherchez à prédire le cours de la bourse, le prix d'un appartement, ou bien l'évolution de la température sur Terre, alors vous cherchez en fait à résoudre un problème de régression.

Si vous disposez d'un Dataset (x,y) alors vous pouvez utiliser l'apprentissage supervisé pour développer un modèle de régression.

Dans ce qui suit on va voir comment développer votre premier modèle de Machine Learning

Propriétés d'une droite de régression

- L'équation peut être utile pour prédire une valeur de Y pour n'importe quelle valeur de X ;
- La droite passe toujours par le point Xmoy. et Ymoy.;
- L'ordonnée à l'origine «**b**» donne la valeur de Y quand X égale zéro;
- La pente «**w** » mesure les variations de Y par rapport aux variations de X : la pente peut être nulle, positive ou négative;

Régression linéaire simple estimation des paramètres

Il y a plusieurs méthodes d'estimation des paramètres:

- Méthode des moindres carrés «Least square»
- Méthode du maximum de vraisemblance
- Méthode du meilleur estimateur linéaire non biaisé

La méthode des moindres carrés est la plus utilisée.