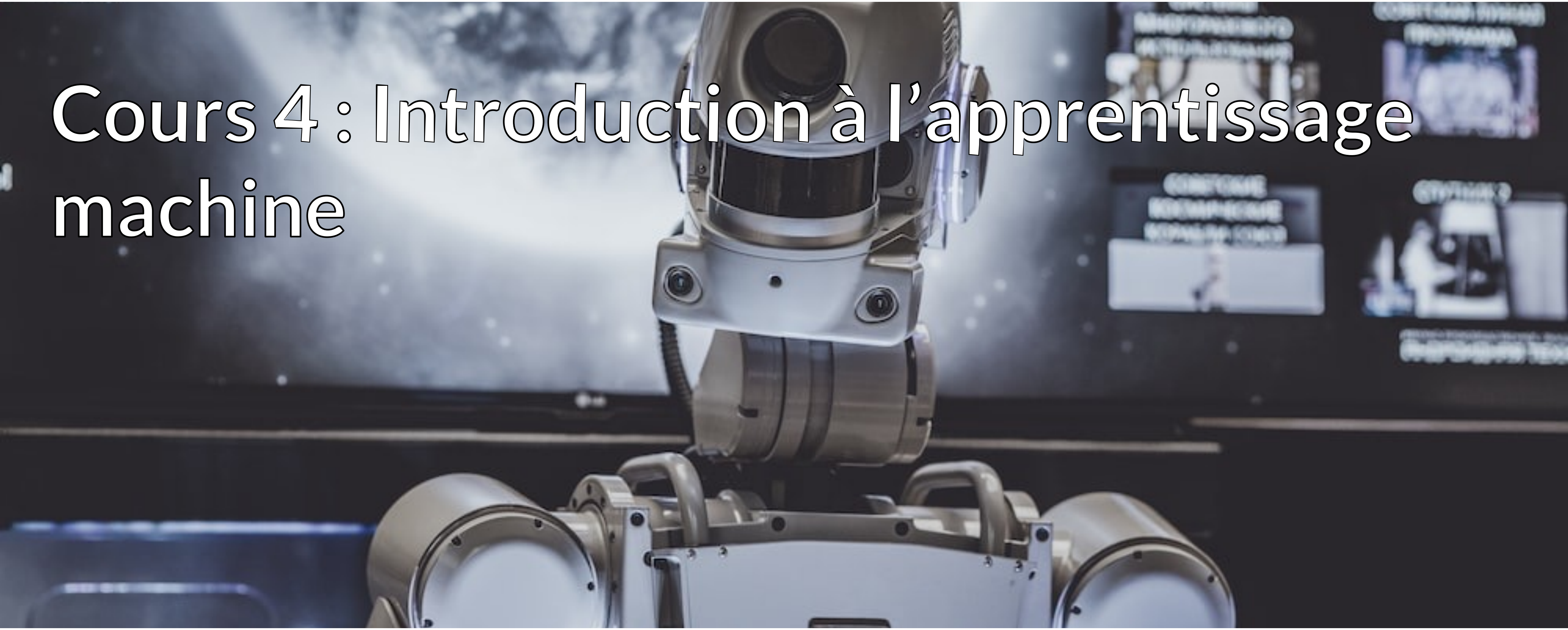
ÉCOLE D'INGÉNIEURS
ENGINEERING SCHOOL

ING3 – **Calcul embarqué**

Cours 4 : Introduction à l'apprentissage machine

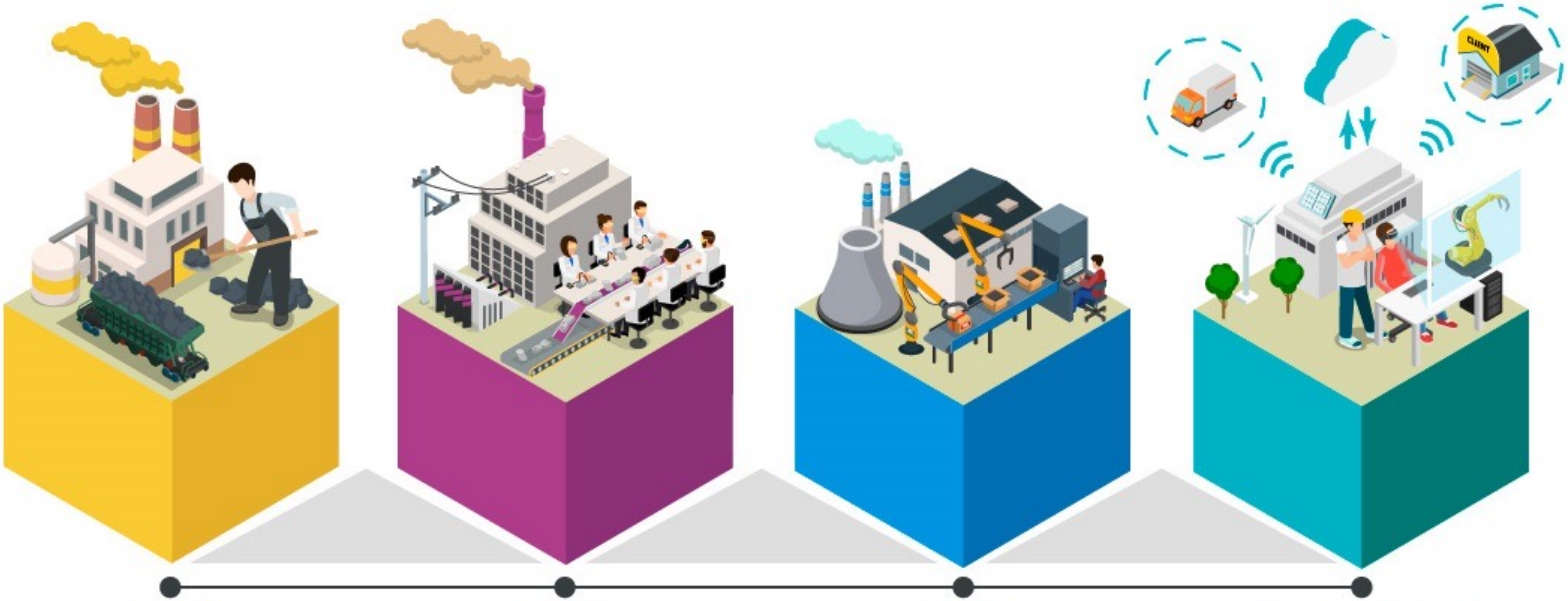


PÔLE ÉLECTRONIQUE

Dr Foued DERRAZ
fderraz@ece.fr

Dr Maxime SCHNEIDER
maxime.schneider@ece.fr

La 4ème révolution industrielle



1765
1ère révolution
LA PRODUCTION MÉCANIQUE
Portée par la machine à vapeur

1870
2ème révolution
LA PRODUCTION DE MASSE
Poussée par l'énergie électrique et pétrolière

1969
3ème révolution
LA PRODUCTION AUTOMATISÉE
Soutenue par l'électronique et les technologies informatiques

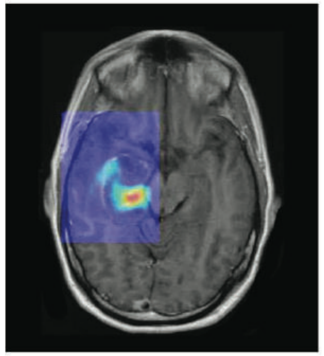
Aujourd'hui
4ème révolution
INTRODUCTION DE NOUVELLES TECHNOLOGIES
Internet des objets, Intelligence artificielle, Cloud, Big Data, etc... et des systèmes cyber-physiques

Intelligence artificielle : exemples d’application

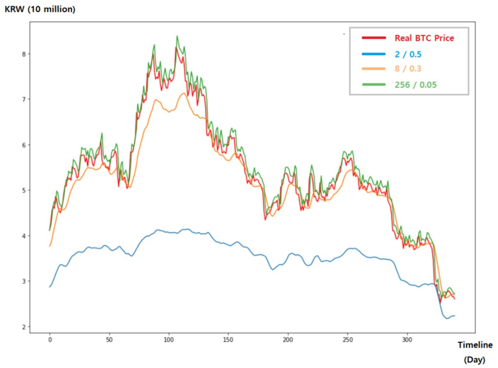
Détection de fraude

E-commerce

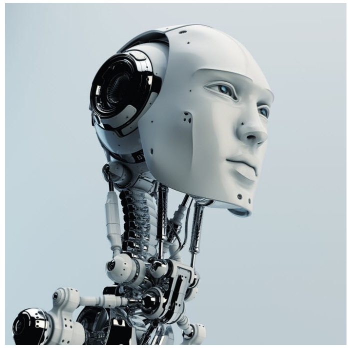
Véhicule autonome



Santé



Finance



Robotique

Cours 4 : Introduction à l’apprentissage machine

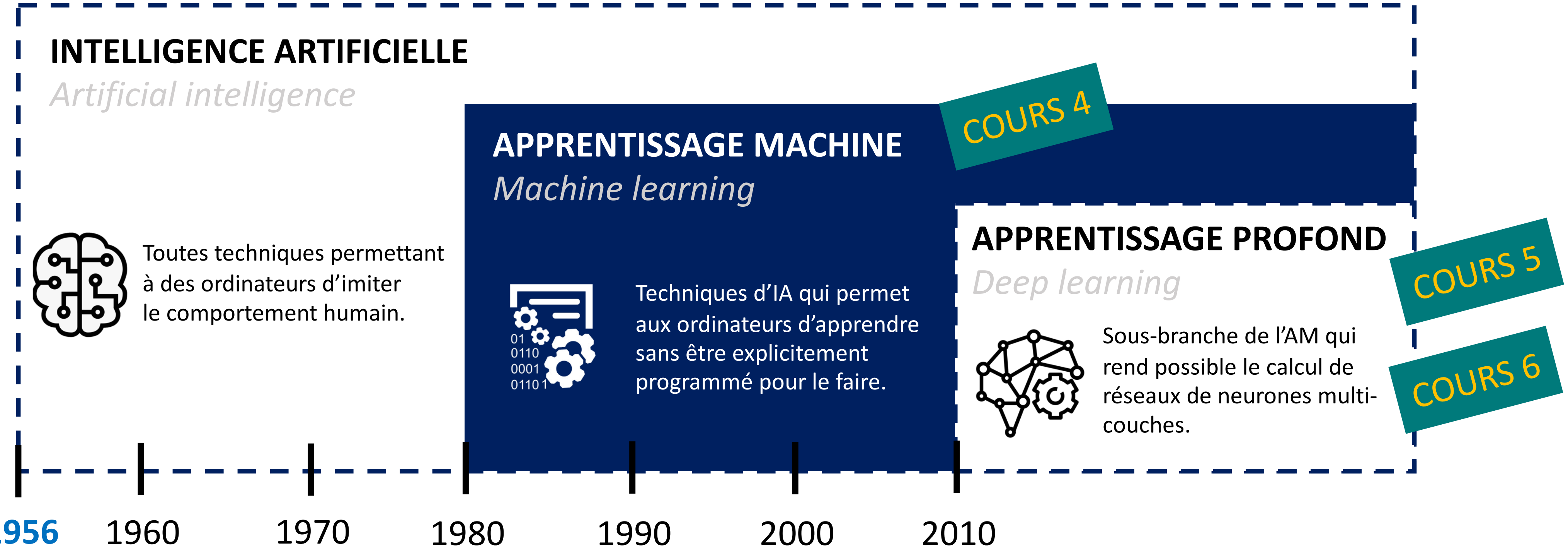
1 L’apprentissage machine

Qu’est-ce que l’intelligence artificielle ?

■ L’**intelligence artificielle** est un programme ou un modèle non humain qui peut résoudre des tâches sophistiquées, basé sur l’expérimentation.

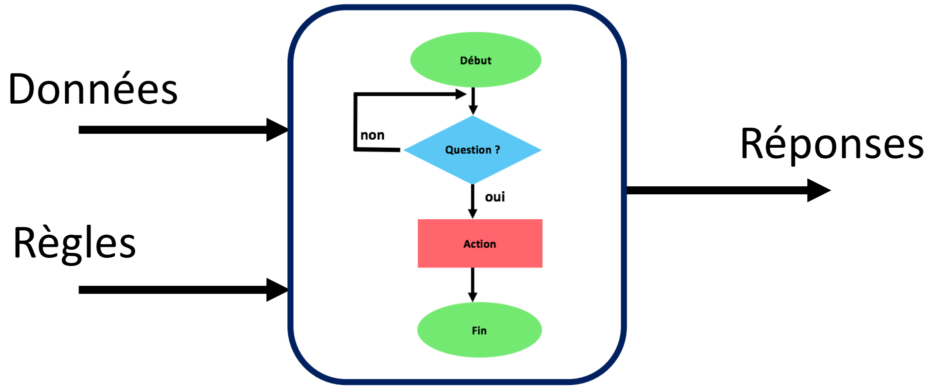
■ Officiellement, l’**apprentissage machine** (AM, *machine learning*) est un sous-domaine de l’intelligence artificielle. Ces dernières années, ces deux termes sont parfois utilisés de manière interchangeable par abus de langage.

■ Arthur Samuel (pionnier l’IA et des jeux vidéos) : « **L’apprentissage machine est un domaine d’étude qui donne aux ordinateurs la capacité d’apprendre sans être explicitement programmés.** »



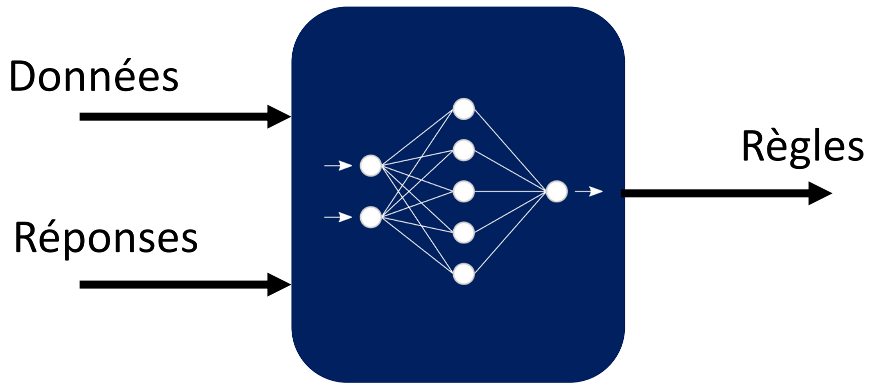
Programmation classique vs AM (1 / 2)

Programmation traditionnelle



■ Dans la **programmation classique**, on alimente l'entrée, on programme la logique et on exécute le programme pour obtenir la sortie.

Apprentissage machine



■ Dans l'**apprentissage automatique**, on alimente l'entrée et la sortie et l'exécutons sur la machine pendant la l'apprentissage.
■ la machine crée alors sa propre logique, qui est évaluée pendant le test.

Programmation classique vs AM (2 / 2)

Avantages

- Temps de développement beaucoup plus rapide (des jours contre des mois)
- Création de nouvelles possibilités
- Possibilité de trouver des modèles cachés
- Apprentissage local
- Diminution des faux positifs ou faux négatifs

Défis

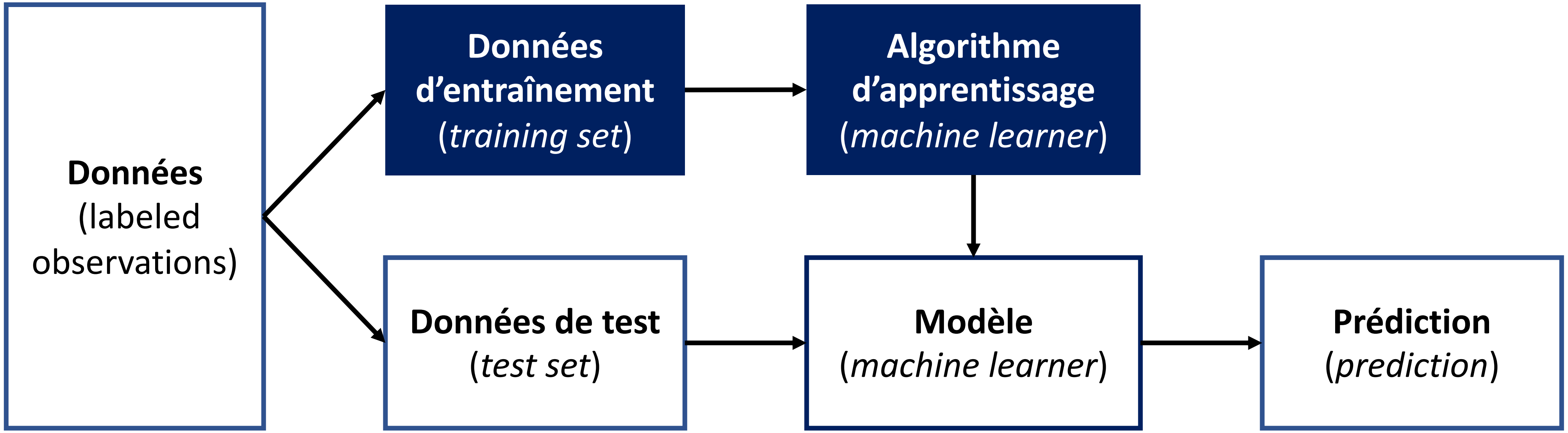
- Il faut des données pour entraîner le modèle (étiquetées / sélectionnées)
- Couvre tous les scénarios possibles
- Le développement du bon modèle ou de la bonne inférence peut être complexe.
- Peut nécessiter l'expertise d'un scientifique et d'un spécialiste en ML
- Empreinte mémoire plus importante

Terminologie (1 / 2)

■ Un **modèle** (ou **hypothèse**) est une **représentation spécifique** apprise à partir de données en appliquant un algorithme d'apprentissage automatique.

■ Une **caractéristique** est une propriété individuelle mesurable de nos données.

■ Un ensemble de caractéristiques numériques peut être décrit de manière pratique par un **vecteur de caractéristiques**. Les vecteurs de caractéristiques sont introduits dans le modèle (AM supervisé).
Par exemple, pour prédire un fruit, il peut y avoir des caractéristiques comme la couleur, l'odeur, le gout, etc.

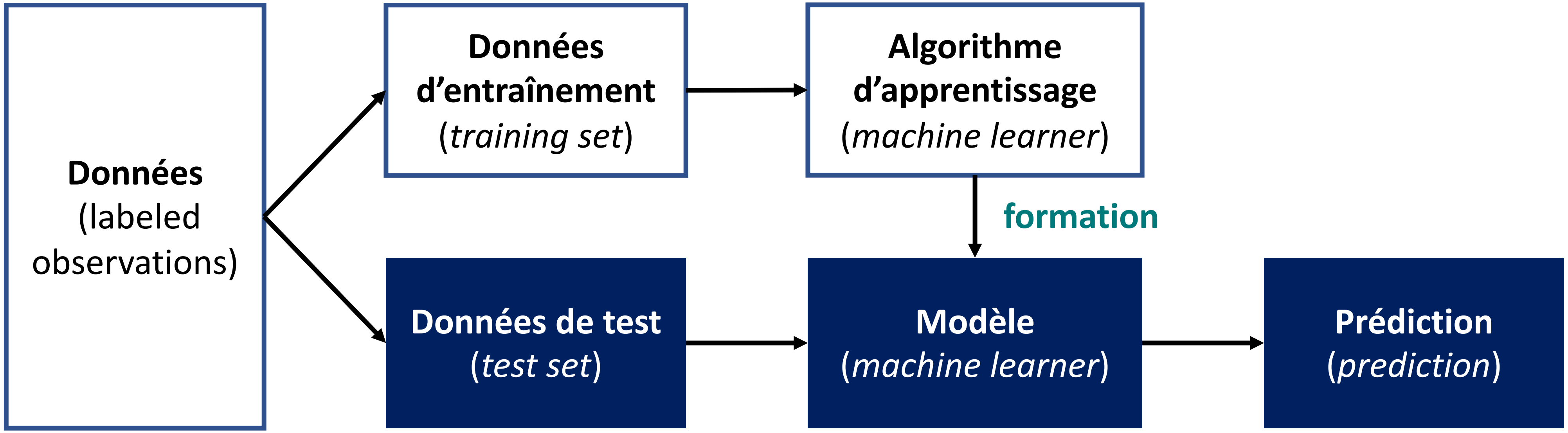


Terminologie (2 / 2)

- Une **cible** (= un **label**) est variable ou une étiquette cible est la valeur à prédire par notre modèle. Pour l'exemple des fruits abordé dans la section sur les caractéristiques, l'étiquette de chaque ensemble d'entrées serait le nom du fruit, comme pomme, orange, banane, etc.

- Une **formation** : L'idée est de donner un ensemble d'entrées (caractéristiques) et les sorties attendues (étiquettes), ainsi après la formation, nous aurons un modèle (hypothèse) qui fera correspondre les nouvelles données à l'une des catégories formées.

- Une **prédiction** : Une fois que notre modèle est prêt, il peut être alimenté par un ensemble d'entrées auxquelles il fournira une sortie prédite (étiquette).



Classification

■ Selon si l'algorithme est entraîné sur des données d'entraînement bien définies (vecteur de caractéristiques) ou non, on parle d'**apprentissage machine supervisé** ou **non supervisé**.

Apprentissage Machine

COURS 4 — Apprentissage supervisé

- Régression
  - **Linéaire**
  - Polynomiale
- Arbre de décision
  - Dichotomiseur itératif
- Classification
  - KNN
  - Arbres
  - **Régression logistique**
  - **SVM**
  - Random forest

COURS 4 — Apprentissage non supervisé

- Partitionnement
  - SVD
  - PCA
  - **K-moyennes**
- Analyse d'association
  - Apriori
  - FP growth
- Modèles graphiques
  - HMM
  - GHMM

Continue (Partitionnement, Analyse d'association)

Catégoriel (Modèles graphiques)

COURS 5 — Apprentissage par renforcement

Classification : Apprentissage supervisé (1 / 3)

Démarche

❶ Subdivision des observations en :

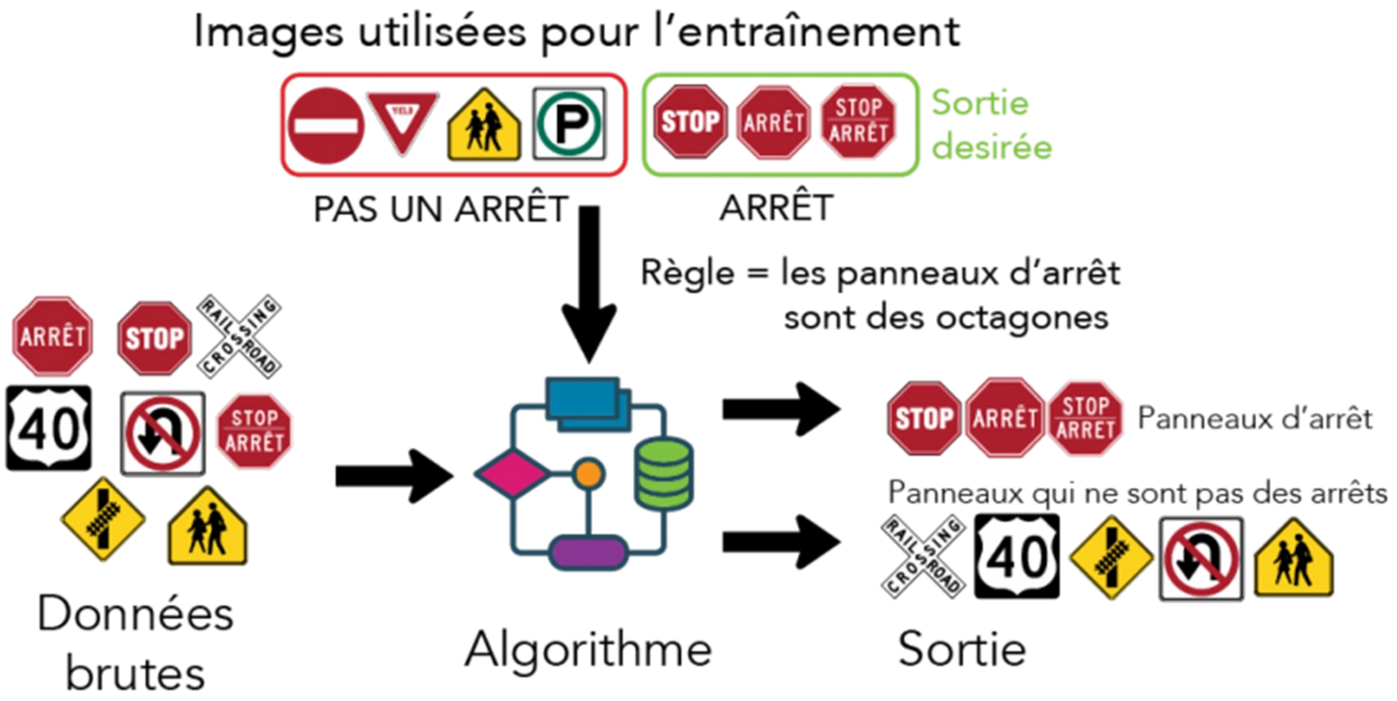
- Observations pour l'apprentissage
- Observations pour la validation

❷ Apprentissage du modèle

❸ Validation du modèle

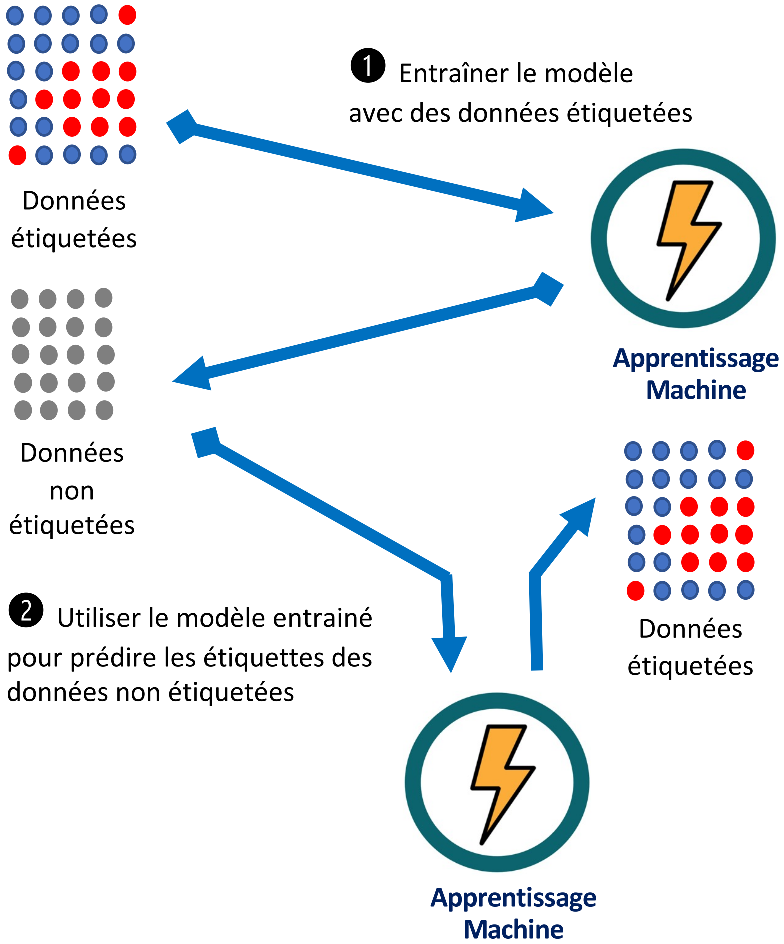
❹ Prédiction sur de nouvelles observation

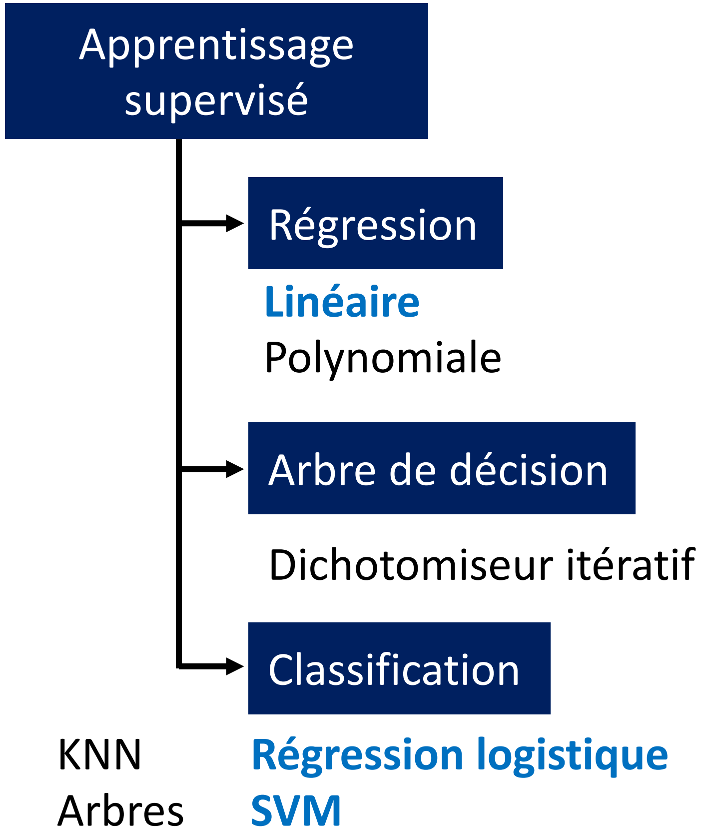
EXEMPLE



Classification : Apprentissage supervisé (2 / 3)

- Défini pour un **modèle** entraîné sur un ensemble de données **étiquetées**.
- L'ensemble **étiqueté** est défini par les caractéristiques et les prédictions.





Classification : Apprentissage supervisé (3 / 3)

■ **Régression:** Il s'agit d'une tâche d'apprentissage supervisé où la sortie a une **valeur continue**.

■ **Classification** : Il s'agit d'un type d'apprentissage supervisé où la sortie a des **étiquettes définies** (valeur discrète). Il peut s'agir d'une classification binaire ou multi-classes. Dans la classification **binaire**, le modèle prédit soit 0 ou 1 : oui ou non, mais dans le cas de la classification **multi-classes**, le modèle prédit plus d'une classe.

■ L'inconvénient le plus fondamental de tout algorithme d'**apprentissage supervisé** est que l'ensemble de données doit être étiqueté à la main, soit par un ingénieur en apprentissage automatique, soit par un scientifique des données. Il s'agit d'un *processus* très *coûteux*, surtout lorsqu'il s'agit de traiter de grands volumes de données.

EXEMPLES

Quelle va être la vitesse du vent demain ?
Quelle sera la température de demain ?

Temperature	Pressure	Relative Humidity	Wind Direction	Wind Speed
10.69261758	986.882019	54.19337313	195.715087	3.278597116
13.59184184	987.8729248	48.0648859	189.295120	2.909167767
17.70494885	988.1119385	39.11965597	192.927383	2.973036289
20.95430404	987.8500366	30.66273218	202.075286	2.965289593
22.9278274	987.2833862	26.06723423	210.658920	2.798230886
24.04233986	986.2907104	23.46918024	221.118850	2.627005816
24.41475295	985.2338867	22.25082295	233.791198	2.448749781
23.93361956	984.8914795	22.35178837	244.350433	2.454271793
22.68800023	984.8461304	23.7538641	253.086471	2.418341875
20.56425726	984.8380737	27.07867944	264.507110	2.318677425
17.76400389	985.4262085	33.54900114	280.782745	2.343950987

Régression

Est-ce que cette personne a le profil d'un acheteur potentiel ?
Va t-il faire chaud demain ?

User ID	Gender	Age	Salary	Purchased
15624510	Male	19	1900	0
15810944	Male	35	2000	1
15668575	Female	26	4300	0
15603246	Female	27	5700	0
15804002	Male	19	7600	1
15728773	Male	27	5800	1
15598044	Female	27	8400	0
15694829	Female	32	15000	1
15600575	Male	25	3300	1
15727311	Female	35	6500	0
15570769	Female	26	8000	1
15606274	Female	26	5200	0

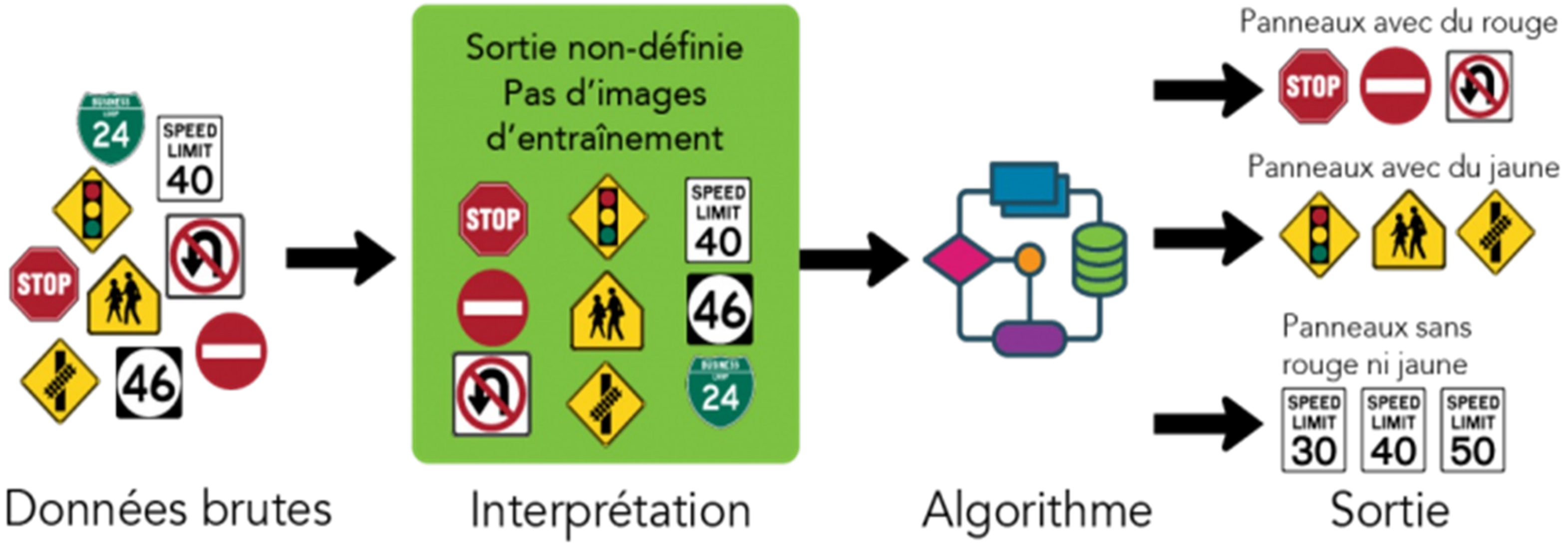
Classification

Classification : Apprentissage non supervisé (1 / 3)

■ L'**apprentissage non-supervisé** est la formation d'une machine à l'aide d'informations qui ne sont ni classées ni étiquetées et permet à l'algorithme d'agir sur ces informations sans guide.

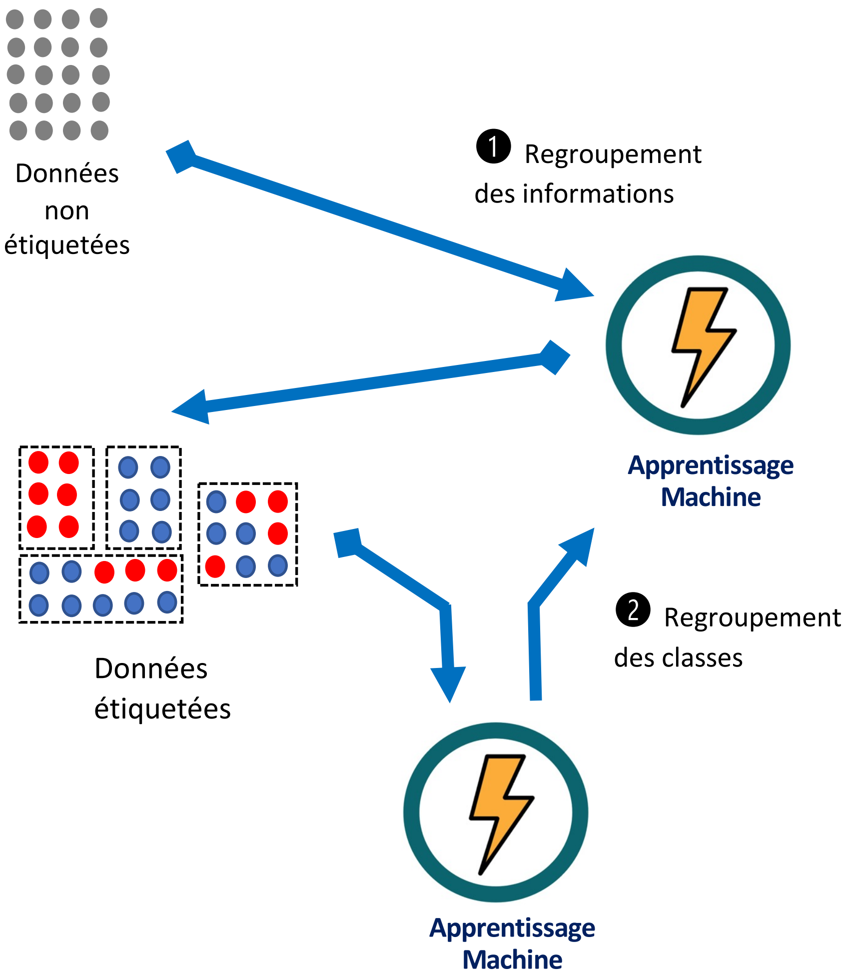
■ Dans ce cas, la tâche de la machine consiste à regrouper des informations non triées en fonction de similitudes, de modèles et de différences, sans formation préalable des données.

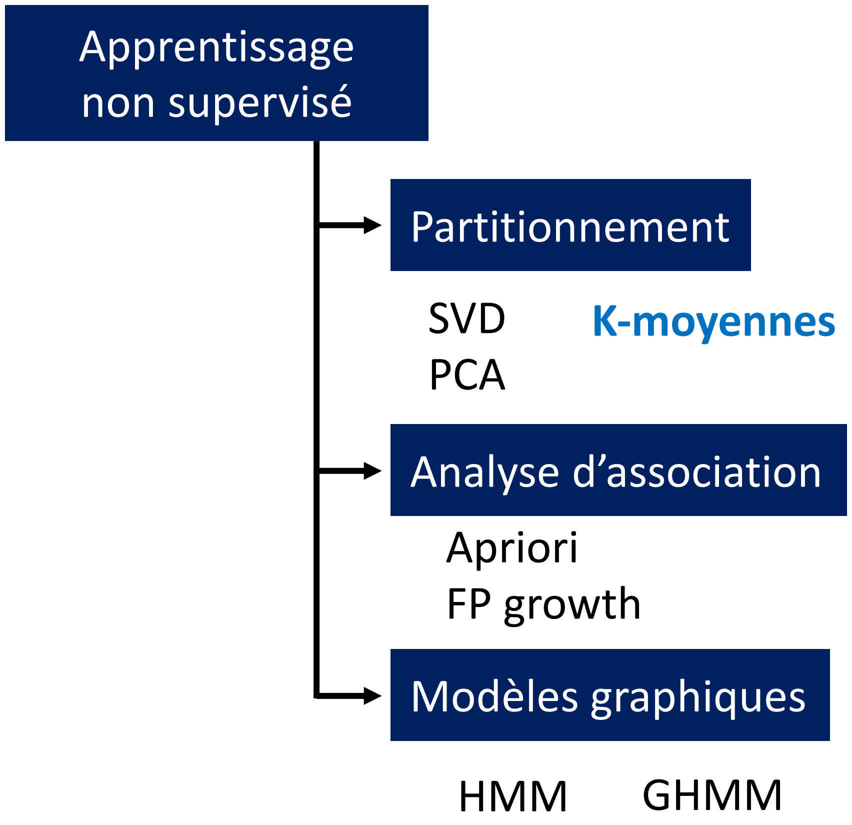
■ L'apprentissage automatique non supervisé est plus difficile que l'apprentissage supervisé en raison de l'absence d'étiquettes.



Classification : Apprentissage non supervisé (2 / 3)

- Défini pour un **modèle** entraîné sur un ensemble de données **non étiquetées**.





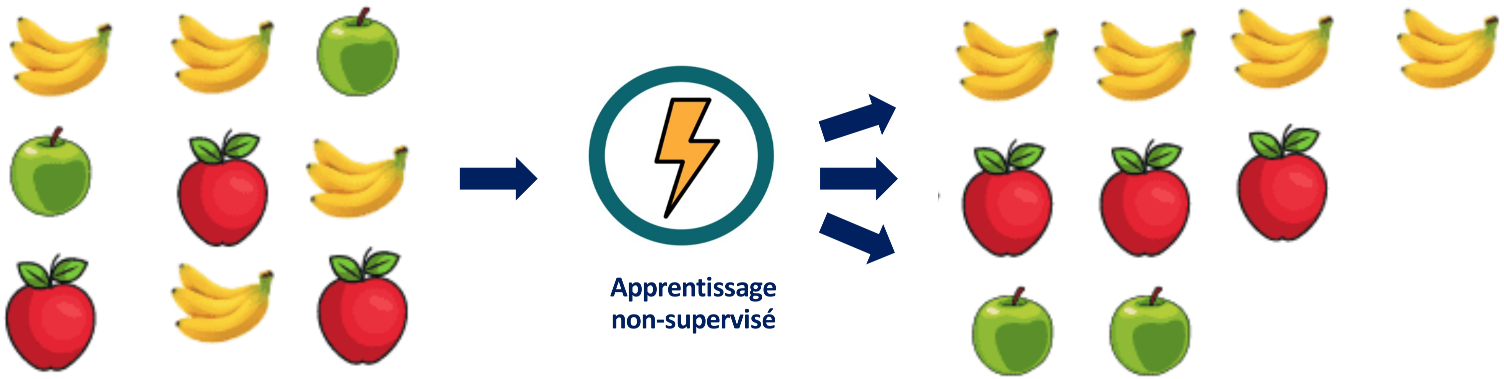
Classification : Apprentissage non supervisé (3 / 3)

■ **Regroupement** : Un problème de clustering consiste à découvrir les regroupements inhérents aux données, par exemple en regroupant les clients par comportement d'achat.

■ **Association** : Un problème d'apprentissage de règles d'association consiste à découvrir des règles qui décrivent de grandes parties de vos données, telles que les personnes qui achètent X ont aussi tendance à acheter Y.

■ L'inconvénient le plus fondamental de tout **apprentissage non supervisé** est que son **spectre d'application est limité**.

EXEMPLE **Déterminer le type de fruit sans avoir au préalable des données étiquetées.**



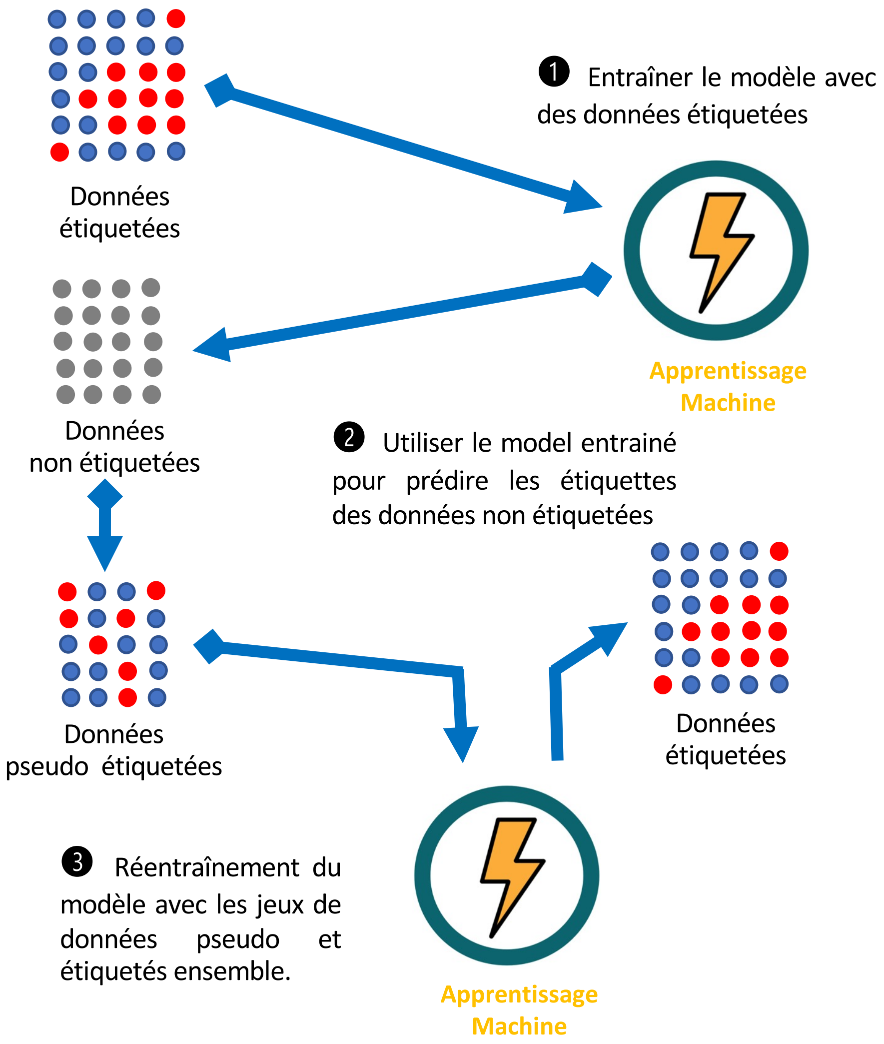
Classification : Apprentissage semi-supervisé

■ Pour contrer ces inconvénients, le concept d'**apprentissage semi-supervisé** a été introduit.

■ Dans ce type d'apprentissage, l'algorithme est entraîné sur une combinaison de données étiquetées et non étiquetées.

■ En général, cette combinaison contient une très petite quantité de données étiquetées et une très grande quantité de données non étiquetées.

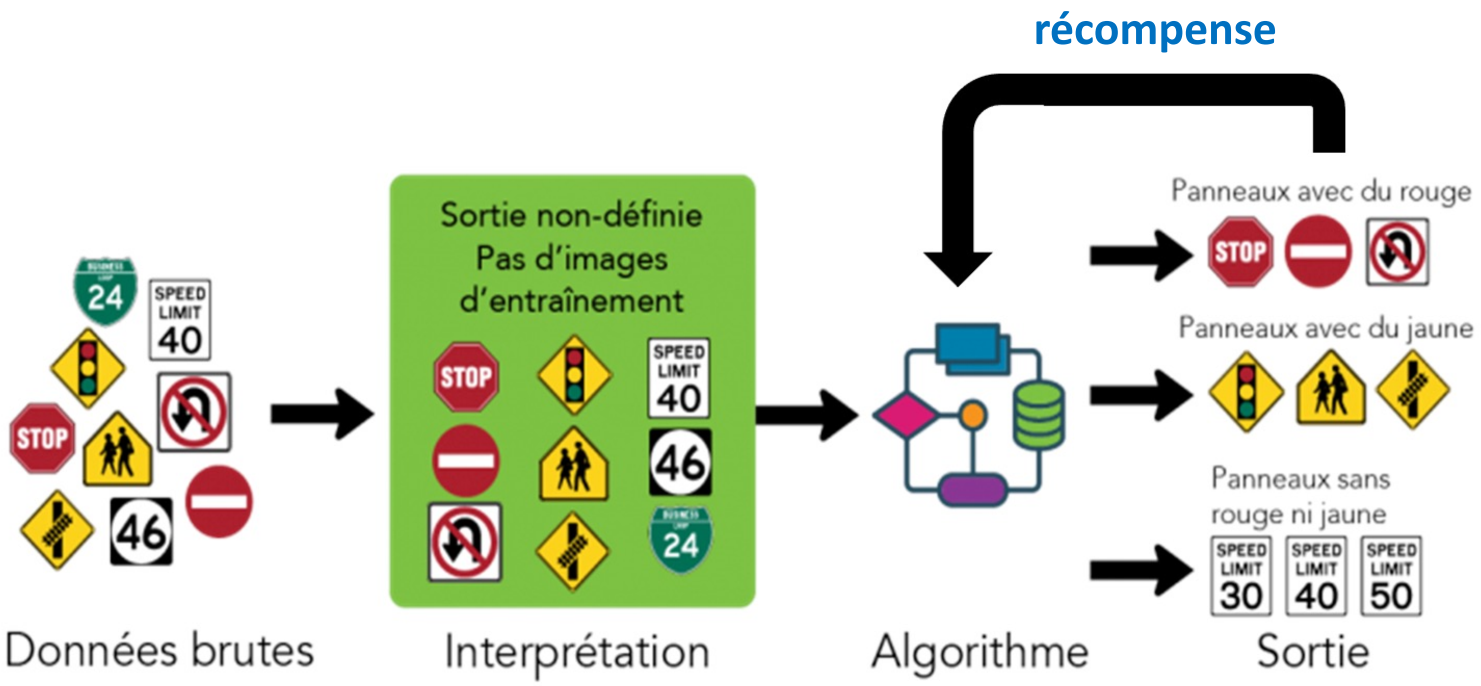
■ Dans l'apprentissage semi-supervisé, les données étiquetées sont utilisées pour apprendre un modèle et, à l'aide de ce modèle, les données non étiquetées sont étiquetées, ce que l'on appelle le **pseudo-étiquetage**, et en utilisant l'ensemble des données, le modèle est formé pour une utilisation ultérieure.



Classification : Apprentissage par renforcement

■ L'**apprentissage par renforcement** (ou auto supervisé) est la formation continue d'une machine à partir d'experiences de façon à optimizer une recompense quantitative au cours du temps.

■ Le système prend des décisions en fonction de son état courant et reçoit une **récompense** (positive ou négative) en fonction de son action. L'objectif est de bâtir un comportement décisionnel qui permet d'optimiser les gains au cours du temps.



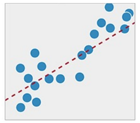
Comparatif des méthodes



	Supervisé	Non-supervisé	Par renforcement
Caractéristiques clés	Les données fournies sont étiquetées par des humains.	Les machines (ordinateurs) cherchent des tendances dans les données qui sont difficiles à étiqueter pour les humains.	La machine crée ses propres données et supervise elle-même son apprentissage.
Champs d'utilisation	La tâche à exécuter est connue et exige un temps considérable.	Nous voulons découvrir des tendances inconnues dans les données.	Nous avons un objectif, mais ignorons peut-être le meilleur moyen de l'atteindre.

2 L'apprentissage supervisé

Régression linéaire (1 / 6)



régression

■ La **régression linéaire** est utilisée pour montrer ou prédire la relation entre deux variables ou facteurs afin de **prédire une quantité continue**.

■ L'hypothèse de départ est que **le phénomène a un comportement linéaire**.

Fonction de prédiction linéaire

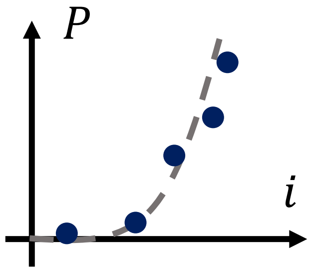
$$h_\theta(x) = \theta_0 + \theta_1 x$$

EXEMPLES

→ Ventes en fonction du budget invsetit dans la communication pour un Youtuber
→ Évolution du salaire en fonction des années d'expériences
→ Nombre d'accidents de la route en fonction de la population
→ Courant en fonction de la tension aux bornes d'un composant
→ Quantité de farine pour faire 13 crêpes connaissant la recette pour 4 crêpes

FAUX EXEMPLES

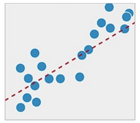
Chien ou chat ?
→ classification



Quelle est la puissance dissipée si $i = 20\ mA$?
$P = ri^2$ non linéaire
→ pas de régression linéaire

Dégâts d'un séisme en fonction de la magnitude de Richter
→ Échelle de Richter = discret donc non applicable.

Régression linéaire (2 / 6)



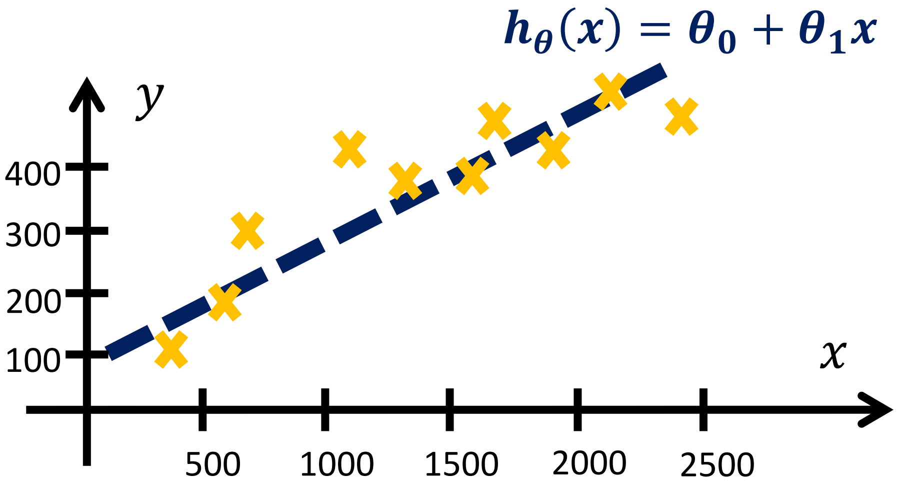
- Le facteur qui est prédit est appelé **variable dépendante**
- Les facteurs qui sont utilisés pour prédire la variable dépendante sont appelés **variables indépendantes**.
- On cherche un couple de paramètres (θ_0, θ_1) pour que la fonction $h_\theta(x)$ soit proche de y pour notre exemple d'entraînement (x, y).

Fonction coût pour une régression simple

$$J(\theta_0, \theta_1) = \frac{1}{2m} \sum_{i=1}^{m} \left(h_\theta\left(x^{(i)}\right) - y^{(i)}\right)^2$$

- Hypothèse : $h_\theta(x) = \theta_0 + \theta_1 x$ (relation linéaire)
- Paramètres : θ_0, θ_1

- Objectif : minimiser la fonction de coût $J(\theta_0, \theta_1)$ = erreur quadratique moyenne



Remarque :

- $h_\theta\left(x^{(i)}\right) - y^{(i)}$ est la différence entre le point de mesure et la fonction de prédiction h_θ
- On l'élève au carré pour avoir la norme euclidienne plutôt que de prendre la valeur absolue

Régression linéaire (3 / 6)

Algorithme

Répéter jusqu'à convergence :

{

$$temp0 := \theta_0 - \alpha \frac{\partial}{\partial \theta_0} J(\theta_0, \theta_1)$$

$$temp1 := \theta_1 - \alpha \frac{\partial}{\partial \theta_1} J(\theta_0, \theta_1)$$

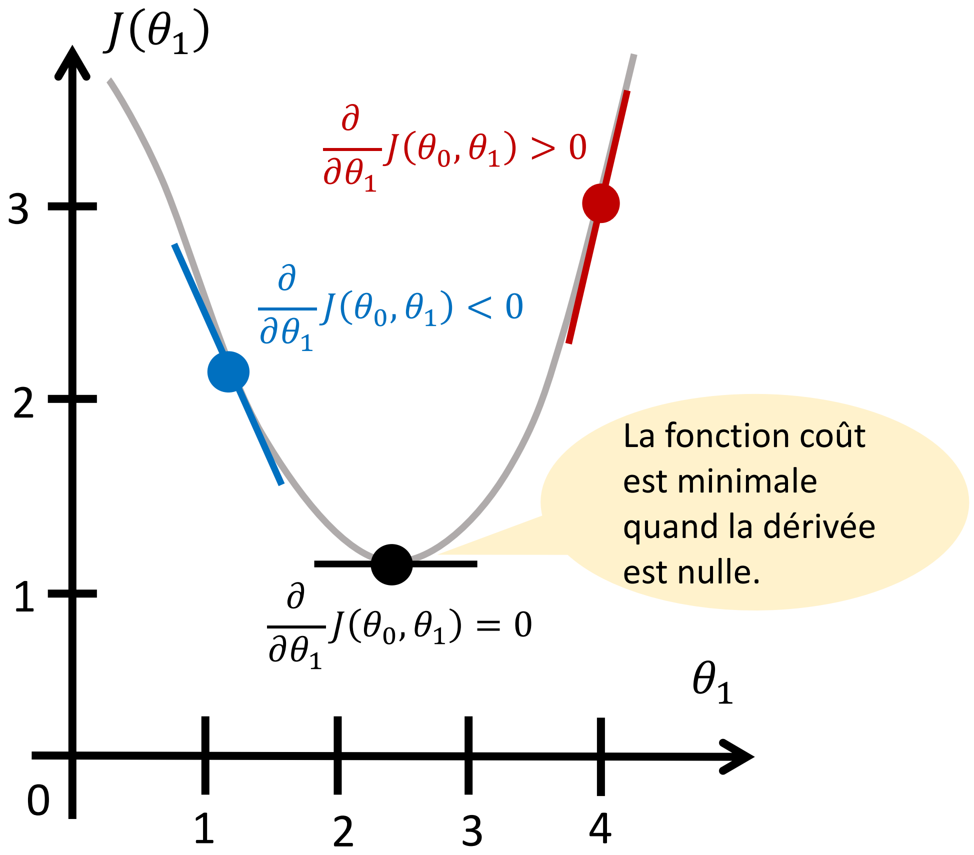
$$\theta_0 := temp0$$

$$\theta_1 := temp1$$

}

α : taux d'apprentissage

$\frac{\partial}{\partial \theta_i} J(\theta_0, \theta_1)$: dérivée partielle de la fonction coût

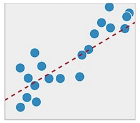


Remarque :

$$\frac{\partial}{\partial \theta_0} J(\theta_0, \theta_1) = \frac{1}{m} \sum_{i=1}^{m} \left(h_\theta\left(x^{(i)}\right) - y^{(i)}\right)$$

$$\frac{\partial}{\partial \theta_1} J(\theta_0, \theta_1) = \frac{1}{m} \sum_{i=1}^{m} \left(h_\theta\left(x^{(i)}\right) - y^{(i)}\right) x^{(i)}$$

Régression linéaire (4 / 6)

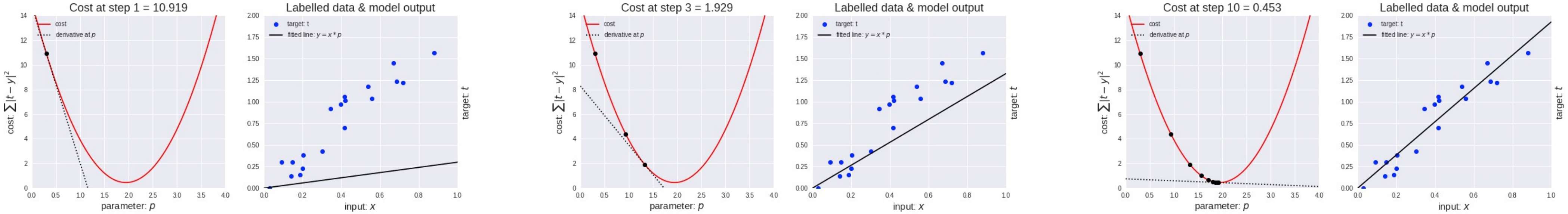


régression

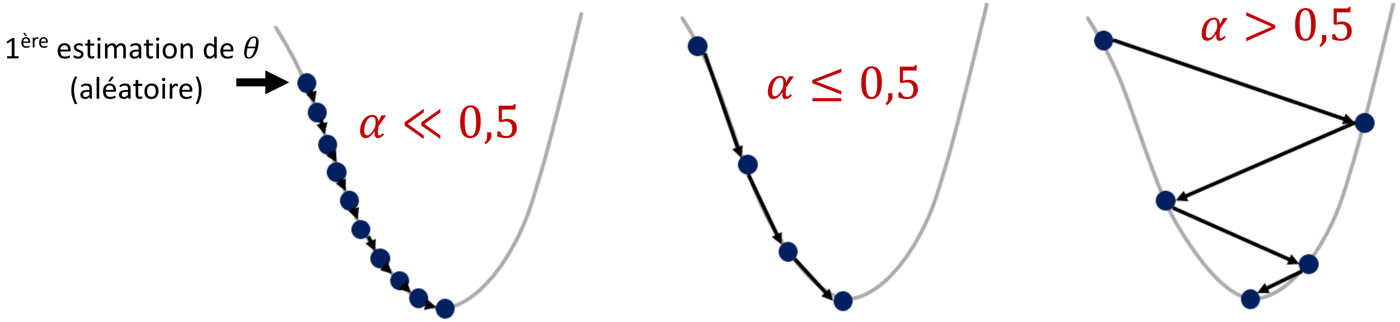
- On parle de **descente de gradient** lorsque l'on cherche à optimiser la fonction de coût $J(\theta_0, \theta_1)$:

$$\theta_i = \theta_i - \alpha \frac{\partial}{\partial \theta_i} J(\theta_0, \theta_1)$$

- Cette technique a pour avantage de moins surcharger la mémoire qu'une résolution directe.



- Le choix du **taux d'apprentissage** (*learning rate*) est important et conditionne le temps de calcul.

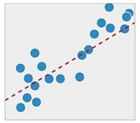


α trop petit
→ Long à converger

α optimal
→ Temps de calcul optimal

α trop grand
→ Peut diverger

Régression linéaire (5 / 6)



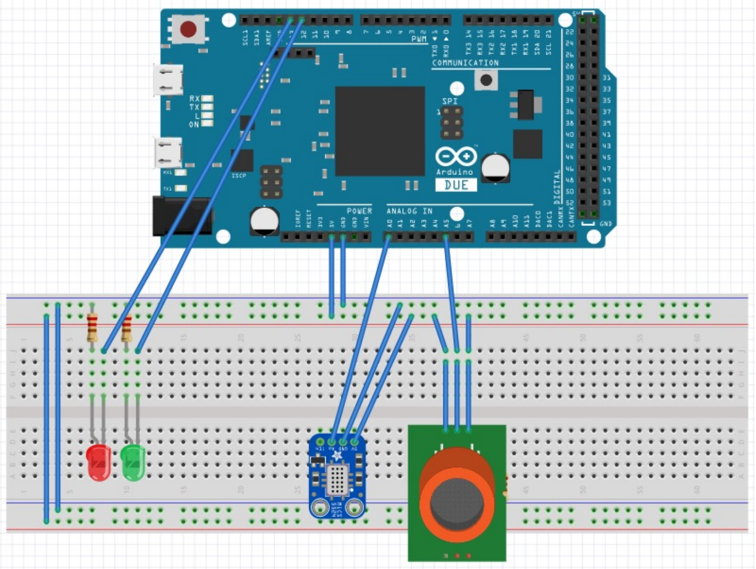
régression

- La recherche du couple de paramètres (θ_0, θ_1) correspond à la **phase d'apprentissage**.
- Par la suite, on peut prédire des valeurs en appliquant le modèle (équation de la droite de prédiction) :

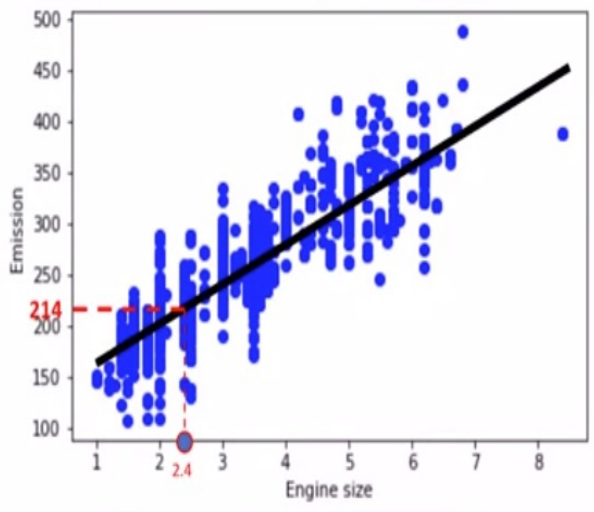
$$h_\theta(x) = \theta_0 + \theta_1 x$$

EXEMPLE

Prévision des émissions de CO_2 en fonction de la taille du moteur à l'aide d'une régression linéaire simple



	ENGINESIZE	CYLINDERS	FUELCONSUMPTION_COMB	CO2EMISSIONS
0	2.0	4	8.5	196
1	2.4	4	9.6	221
2	1.5	4	5.9	136
3	3.5	6	11.1	255
4	3.5	6	10.6	244
5	3.5	6	10.0	230
6	3.5	6	10.1	232
7	3.7	6	11.1	255
8	3.7	6	11.6	267
9	2.4	4	9.2	?



Régression linéaire (6 / 6)

régression

Algorithme

❶ Initialisation

$n \leftarrow 0;$
$\theta_0^0 \leftarrow 0;$
$\theta_1^0 \leftarrow 0;$
$m \coloneqq taille\ de\ vecteur\ data;$
erreur:=0;
lire vecteur x(1:m);
lire vecteur y (1:m);

❷ *Entraînement*

do
{

$$\theta_0^{n+1} := \theta_0^n - \frac{\alpha}{m}\sum_{i=1}^{m}\left(h_\theta\left(x^{(i)}\right) - y^{(i)}\right);$$

$$\theta_1^{n+1} := \theta_1^n - \frac{\alpha}{m}\sum_{i=1}^{m}\left(h_\theta\left(x^{(i)}\right) - y^{(i)}\right) x^{(i)};$$

$$n \coloneqq n + 1;$$

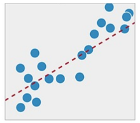
$$erreur = erreur + \left(h_\theta\left(x^{(i)}\right) - y^{(i)}\right)^2;$$

}
$while\ (erreur < 1e - 6)$
$\theta_0 := \theta_0^{n+1};$
$\theta_1 := \theta_1^{n+1};$

❸ *Prédiction*

lire x(m+1);
y(m+1):= $\theta_1 x(m+1) + \theta_0$

Régression linéaire : Exemple



EXEMPLE Température en fonction de l'altitude

Un relevé de gradients de température dans des stations d'altitudes différentes donne :

i	$x^{(i)}$	$y^{(i)}$
1	2	0
2	3	1
3	4	2
4	7	?

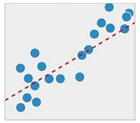
On émet l'hypothèse que la relation entre l'altitude et la température est linéaire, selon l'équation de prédiction suivante :

$$h_\theta = \theta_0 + \theta_1 x$$

Calculer le couple de coefficients (θ_0, θ_1) permettant de prédire le gradient de température à une altitude donnée, et prédire celle pour une altitude de 7 *u.a.*.

On prendra $\alpha = 0{,}25$

Régression linéaire : Exemple



régression

On a $\alpha = 0{,}25$ et $m = 3$

La descente de gradient donne

$$\frac{\partial \theta_j}{\partial t} = -\frac{\partial}{\partial \theta_j} J(\theta) \quad \text{avec} \quad \begin{cases} \frac{\partial}{\partial \theta_0} J(\theta) = -\frac{1}{\mathrm{m}} \sum_{i=1}^{4} \left(h_\theta\left(x^{(i)}\right) - y^{(i)}\right). \\ \frac{\partial}{\partial \theta_1} J(\theta) = -\frac{1}{\mathrm{m}} \sum_{i=1}^{4} \left(h_\theta\left(x^{(i)}\right) - y^{(i)}\right) x^{(i)} \end{cases}$$

i	$x^{(i)}$	$y^{(i)}$
1	2	0
2	3	1
3	4	2
4	7	?

Une approximation par différence finie donne :

$$\frac{\partial \theta_j}{\partial t} \approx \frac{\theta_j^{n+1} - \theta_j^n}{\alpha} \quad \Leftrightarrow \quad \begin{cases} \theta_0^{n+1} = \theta_0^n - \frac{\alpha}{m} \sum_{i=1}^{4} \left(h_\theta\left(x^{(i)}\right) - y^{(i)}\right). \\ \theta_1^{n+1} = \theta_1^n - \frac{1}{4} \sum_{i=1}^{4} \left(h_\theta\left(x^{(i)}\right) - y^{(i)}\right) x^{(i)} \end{cases}$$

$$\alpha = \Delta t = 0{,}25$$

On fixe les conditions initiales sont : $\theta_0^0 = 0$ et $\theta_1^0 = 0$

$$\boldsymbol{h_\theta\left(x^{(i)}\right) = 0}$$

Itération $n = 1$

$$\theta_0^1 = -\frac{\alpha}{\mathrm{m}} \sum_{i=1}^{\mathrm{m}} \left(h_\theta\left(x^{(i)}\right) - y^{(i)}\right) = \frac{0+1+2}{12} = \mathbf{0,25}$$

$$\theta_1^1 = -\frac{\alpha}{\mathrm{m}} \sum_{i=1}^{m} \left(h_\theta\left(x^{(i)}\right) - y^{(i)}\right) x^{(i)} = \frac{0+3+8}{12} = \mathbf{0,915}$$

$$\boldsymbol{h_\theta\left(x^{(i)}\right) = 0{,}25 + 0,915 * x^{(i)}}$$

Itération $n = 2$

$$\theta_0^2 = \theta_0^1 - \frac{\alpha}{\mathrm{m}} \sum_{i=1}^{\mathrm{m}} \left(h_\theta\left(x^{(i)}\right) - y^{(i)}\right) = 0{,}25 - \frac{2{,}08 + 1{,}995 + 1{,}91}{12} = \mathbf{-0,25}$$

$$\theta_1^2 = \theta_1^1 - \frac{\alpha}{\mathrm{m}} \sum_{i=1}^{\mathrm{m}} \left(h_\theta\left(x^{(i)}\right) - y^{(i)}\right) x^{(i)} = 0{,}915 - \frac{9{,}32 + 16{,}47 + 25{,}28}{12} = \mathbf{1,48}$$

$$\boldsymbol{h_\theta\left(x^{(i)}\right) = -0{,}25 + 1,48 * x^{(i)}}$$

Régression linéaire : Exemple

régression

Itération $n = 3$

$$\theta_0^2 = \theta_0^1 - \frac{\alpha}{m}\sum_{i=1}^{m}\left(h_\theta\left(x^{(i)}\right) - y^{(i)}\right) = -0{,}25 - \frac{2{,}71 + 3{,}69 + 4{,}17}{12} = \mathbf{-1,45}$$

$$\theta_1^2 = \theta_1^1 - \frac{\alpha}{m}\sum_{i=1}^{m}\left(h_\theta\left(x^{(i)}\right) - y^{(i)}\right)x^{(i)} = 1{,}485 - \frac{5{,}42 + 9{,}57 + 1468}{12} = \mathbf{2,47}$$

$$\boldsymbol{h_\theta\left(x^{(i)}\right) = -1{,}45 + 2{,}47 * x^{(i)}}$$

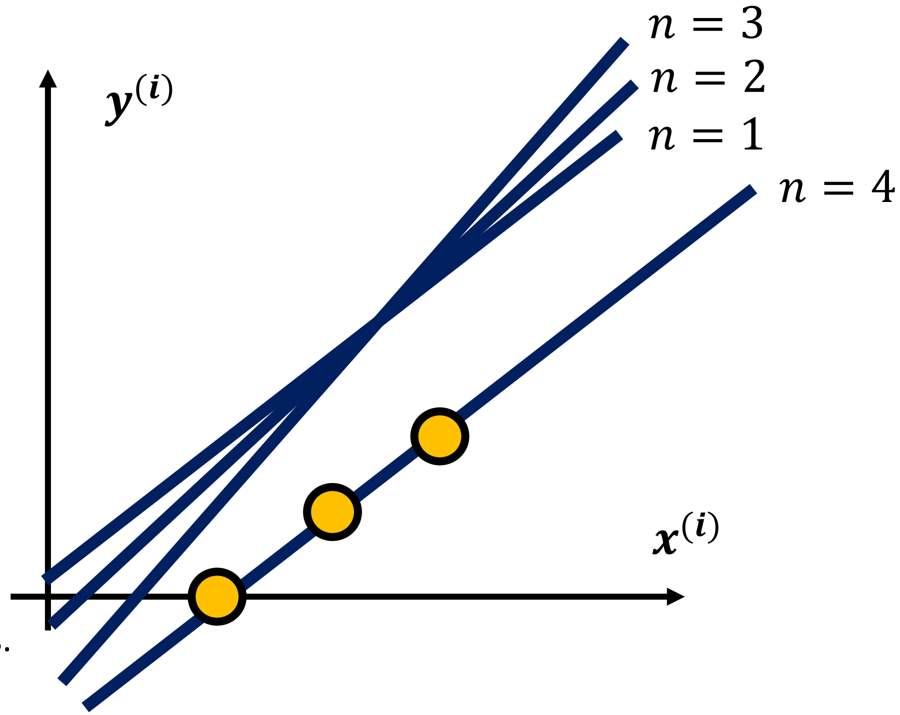
i	$x^{(i)}$	$y^{(i)}$
1	2	0
2	3	1
3	4	2
4	7	?

Itération $n = 4$

$$\theta_0^3 = \theta_0^2 - \frac{\alpha}{m}\sum_{i=1}^{m}\left(h_\theta\left(x^{(i)}\right) - y^{(i)}\right) = 1{,}45 - \frac{3{,}49 + 2{,}96 + 6{,}43}{12} = \mathbf{-2,52}$$

$$\theta_1^3 = \theta_1^2 - \frac{\alpha}{m}\sum_{i=1}^{m}\left(h_\theta\left(x^{(i)}\right) - y^{(i)}\right)x^{(i)} = 2{,}47 - \frac{6{,}98 + 8{,}88 + 25{,}72}{12} = \mathbf{1}$$

$$\boldsymbol{h_\theta\left(x^{(i)}\right) = -2{,}52 + x^{(i)}}$$

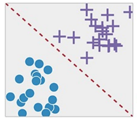


Application numérique pour

$$h_\theta\left(x^{(4)}\right) = 7*1-2{,}52 = \mathbf{4{,}48}$$

Remarque 1 : en continuant à itérer, on aurait plus de chiffres significatifs corrects.
Remarque 2 : en choisissant un α plus petit, on pourrait converger plus vite.

Régression logistique



classification

■ La **régression logistique** est un algorithme qui est utilisé pour résoudre les problèmes de **classification** sur des **variables discrètes**.
Il s'agit d'une analyse prédictive qui décrit les données et explique la relation entre les variables.

■ En fonction du nombre de catégories, la régression logistique peut être considérée comme :

▶ **Binomiale :** La variable cible ne peut avoir que 2 types possibles : "0" ou "1"
▶ **Multinomiale :** La variable cible peut avoir 3 types possibles ou plus qui ne sont pas ordonnés
Exemple: "maladie A" vs "maladie B" vs "maladie C".
▶ **Ordinale :** le résultat de test peut être classé dans les catégories suivant s variables cibles avec des catégories ordonnées.
Par exemple, : "très mauvais", "mauvais", "bon", "très bon".
Ici, chaque catégorie peut recevoir un score comme 0, 1, 2, 3.

EXEMPLES

Email spam ou pas spam
→ Régression logistique binomiale

Régression logistique binaire

classification

■ La **régression logistique binomiale** est appliquée à une variable d'entrée (x) où la variable de sortie (y) est une valeur discrète comprise entre 1 (oui) et 0 (non). Ce modèle peut être étendu pour modéliser plusieurs classes d'événements, par exemple pour déterminer si une image contient un chat, un chien, un lion, etc.

Approche naïve : faire une régression linéaire pour séparer les deux clusters

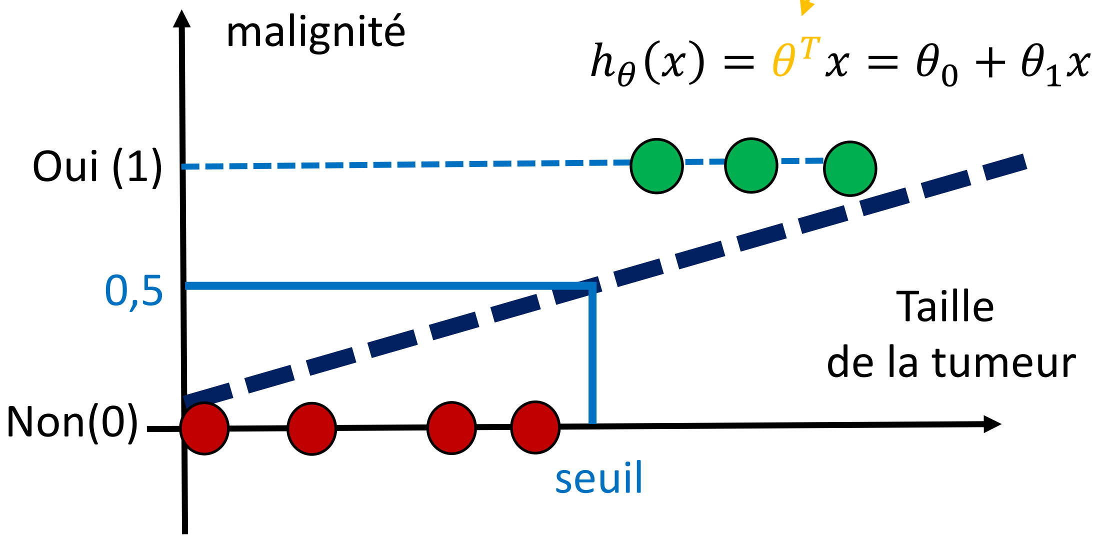
EXEMPLE Dire si une tumeur est maligne ou non en fonction de sa taille

Inconvénients :

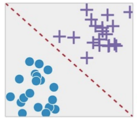
- Une régression linéaire donnerait des valeurs autres que 0 ou 1
- La régression linéaire est sensible aux points extrêmes : un point aberrant changerait le seuil.
- La pente de la régression linéaire change en fonction du nombre de points sur les deux plateaux (0 et 1)

$[\theta_0 \quad \theta_1]$

$$h_\theta(x) = \theta^T x = \theta_0 + \theta_1 x$$



Régression logistique binaire



classification

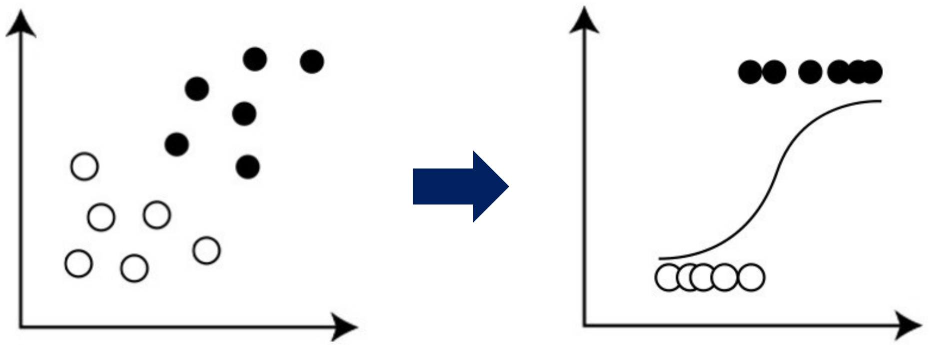
- On parle de **régression logistique** lorsque l'on utilise une fonction de prédiction de type logistique (= sigmoïde).
- L'hypothèse $h_\theta(x)$ produit un nombre entre 0 et 1, que l'on peut interpreter comme étant une probabilité.

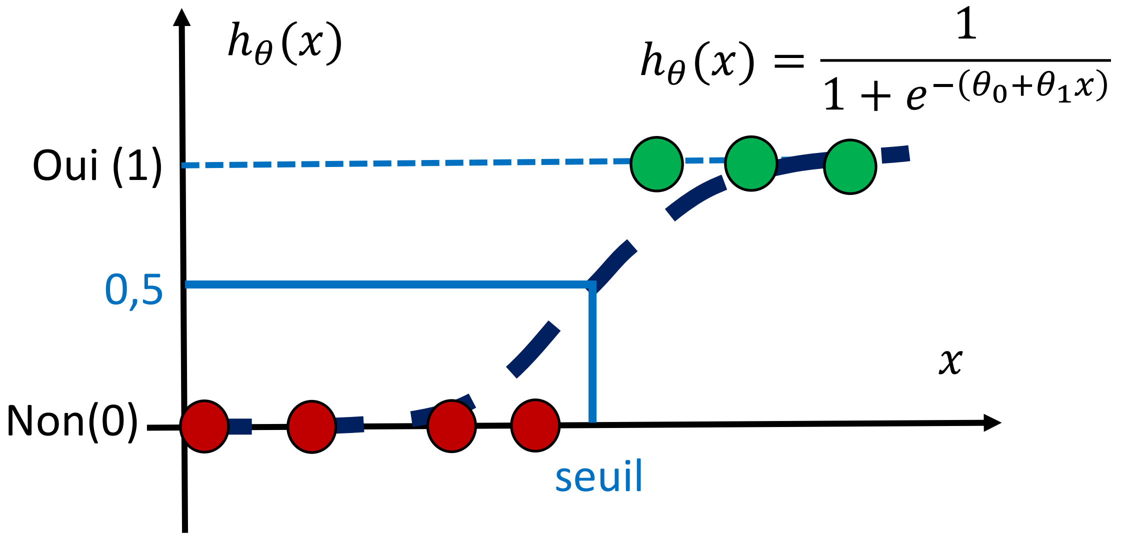
Fonction de prédiction logistique

$$h_\theta(x) = \frac{1}{1 + e^{-(\theta_0 + \theta_1 x)}}$$

Avantages :

Valeurs entre 0 et 1





Régression logistique binaire

classification

■ On peut alors écrire qu'il s'agit de la probabilité que $y = 1$ étant donné x, paramétré par θ :

Forme probabilistique de $h_\theta(x)$

$$h_\theta(x) = P(y = 1 | x\,;\theta)$$

Il s'agit alors d'une tâche de **classification** ($y = 0$ ou 1), et :

$$P(y = 1|x\,;\theta) + P(y = 0|x\,;\theta) = 1$$

$$\Leftrightarrow P(y = 0|x\,;\theta) = 1 - P(y = 1|x\,;\theta)$$

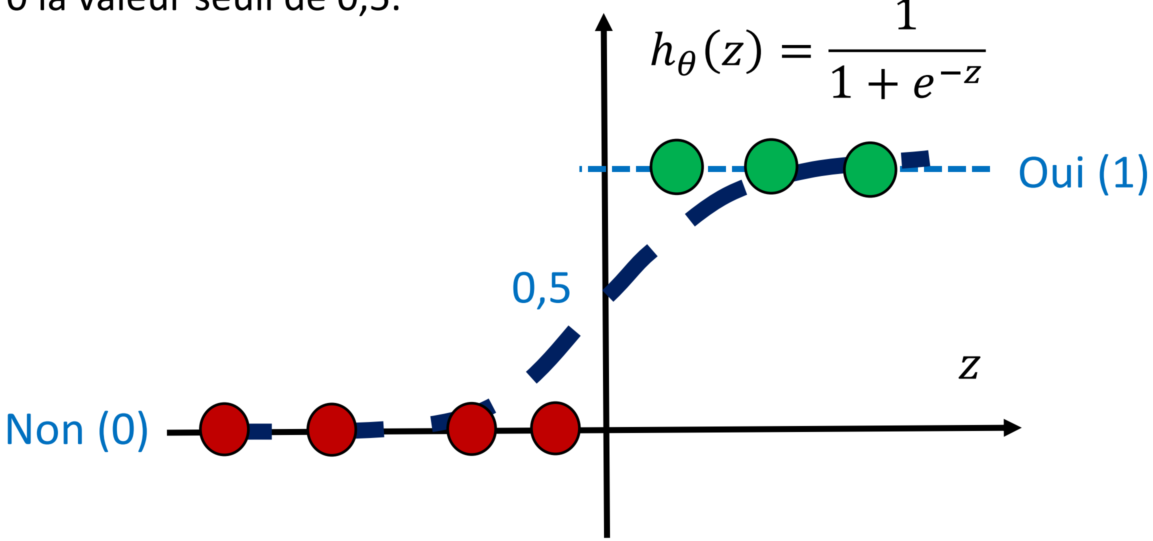
■ Lorsque la probabilité que y soit égale à 1 est supérieure <u>ou égale</u> à 0,5, on peut prédire que $y = 1$
Sinon, on prédit que $y = 0$

■ Il est alors courant de noter $\theta^T x = z$ et de centrer en 0 la valeur seuil de 0,5.

■ On a donc :

Si $z \geq 0$ alors $y = 1$

Si $z \leq 0$, alors $y = 0$



Régression logistique binaire

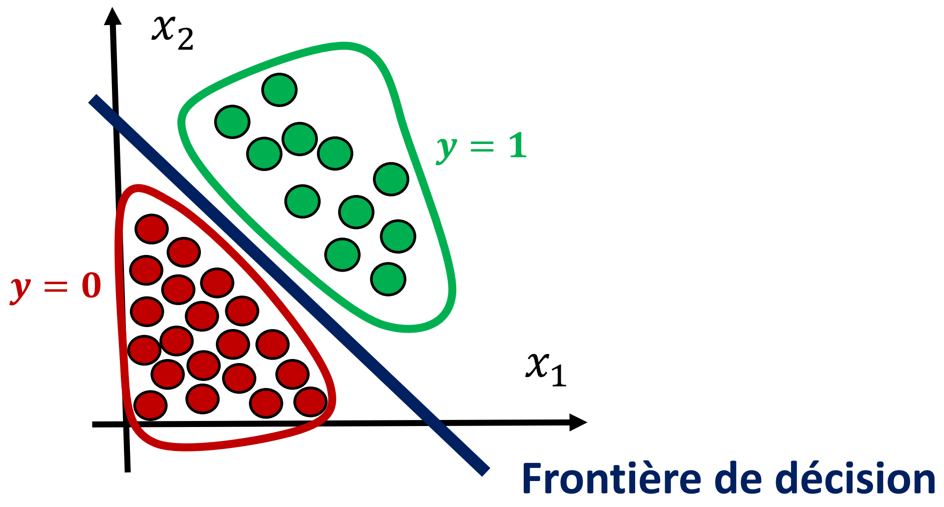
classification

EXEMPLE 1

On considère $h_\theta(x) = g(\theta_0 + \theta_1 x_1 + \theta_2 x_2) = g(\theta^T x)$
avec θ^T le vecteur ligne $[-3 \quad 1 \quad 1]$

On prédit alors $y = 1$ si $-3 + x_1 + x_2 \geq 0$
$\Leftrightarrow y = 1$ si $x_1 + x_2 \geq 3$

La droite $\boldsymbol{x_1 + x_2 = 3}$ est alors la **frontière de décision**.

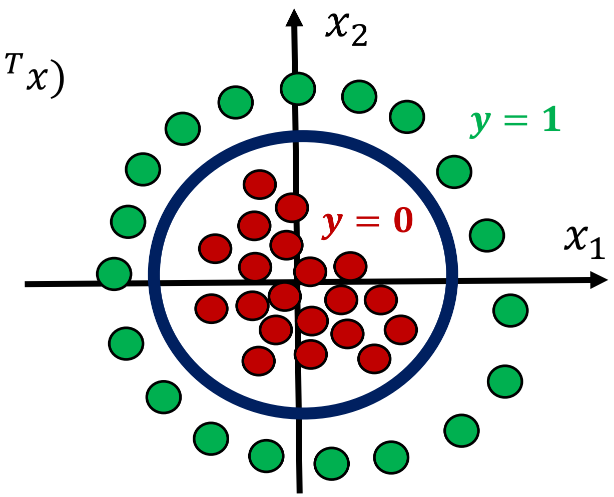


EXEMPLE 2

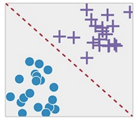
On considère $h_\theta(x) = g(\theta_0 + \theta_1 x_1 + \theta_2 x_2 + \theta_3 x_1^2 + \theta_4 x_2^2) = g(\theta^T x)$
avec θ^T le vecteur ligne $[-1 \quad 0 \quad 0 \quad 0 \quad 1 \quad 1]$

On prédit alors $y = 1$ si $x_1^2 + x_2^2 \geq 1$

La **frontière de décision** est donc $x_1^2 + x_2^2 = 1$



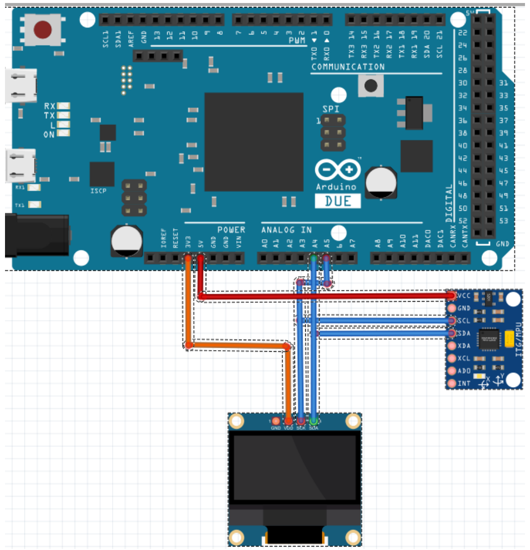
Régression logistique binaire



classification

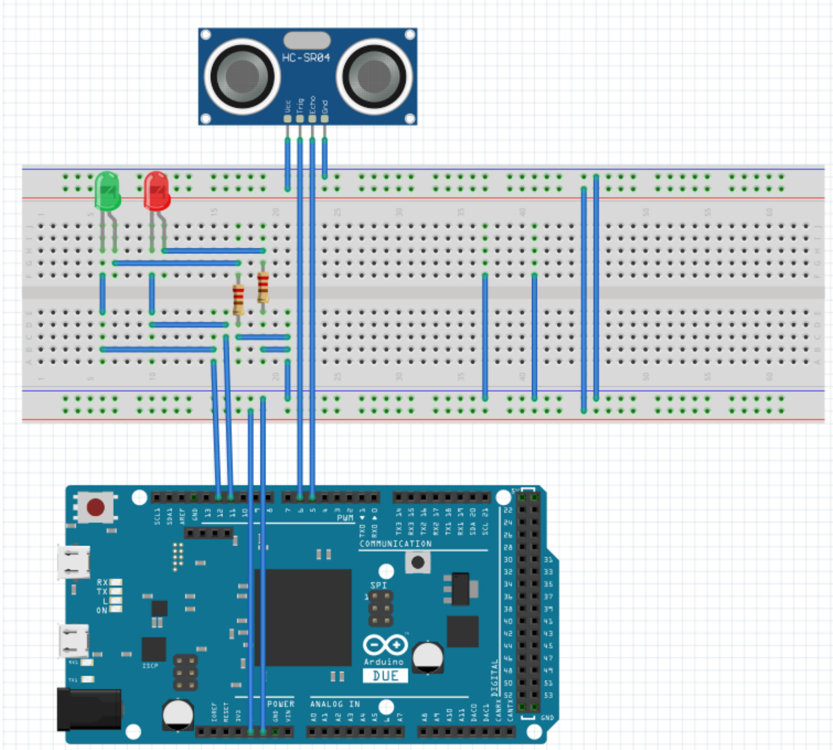
EXEMPLE 1

Prédire le mouvement d'une main en utilisant une centrale inertielle.

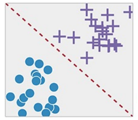


EXEMPLE 2

Prédire la trajectoire que doit suivre un robot en anticipant les prochaines directions.



Régression logistique binaire



classification

Algorithme

❶ Initialisation

$n \leftarrow 0;$
$\theta_0^0 \leftarrow 0;$
$\theta_1^0 \leftarrow 0;$
$m \coloneqq taille\ de\ vecteur\ data;$
erreur:=0;
lire vecteur x(1:m);
lire vecteur y (1:m);

❷ Entraînement

do
{

$$\theta_0^{n+1} := \theta_0^n - \frac{\alpha}{m}\sum_{i=1}^{m}\left(h_\theta\left(x^{(i)}\right) - y^{(i)}\right)h_\theta\left(x^{(i)}\right)\left(1 - h_\theta\left(x^{(i)}\right)\right);$$

$$\theta_1^{n+1} := \theta_1^n - \frac{\alpha}{m}\sum_{i=1}^{\mathrm{m}}\left(h_\theta\left(x^{(i)}\right) - y^{(i)}\right)h_\theta\left(x^{(i)}\right)\left(1 - h_\theta\left(x^{(i)}\right)\right)x^{(i)};$$

$$n \coloneqq n + 1;$$

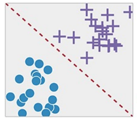
$$erreur = erreur + \left(h_\theta\left(x^{(i)}\right) - y^{(i)}\right)^2;$$

}
$while\ (erreur < 1e - 6)$
$\theta_0 := \theta_0^{n+1};$
$\theta_1 := \theta_1^{n+1};$

❸ Prédiction

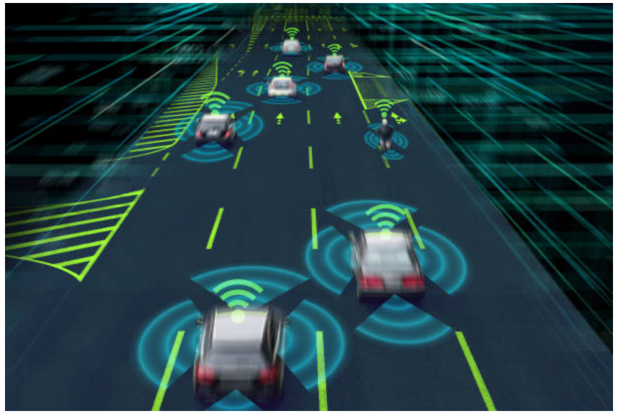
lire x(m+1);
$y(m+1) := \theta_1 x(m+1) + \theta_0$

Régression logistique : Exemple



classification

EXEMPLE Radar de sécurité pour la détection de piétons embarqué dans une voiture.



Le tableau suivant présente un relevé de distance entre les piétons et le parechoc avant de la voiture, et le message d'avertissement à remonter au conducteur :

i	$x^{(i)}$ $[m]$	$y^{(i)}$	**label**
1	-2	0,08	Très proche
2	-1	0,27	proche
3	0	0,5	loin
4	2	0,97	Très loin
5	1,5	?	?

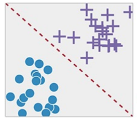
On émet l'hypothèse que la relation entre la distance entre le parechoc et le piéton, et l'intensité de l'avertisseur sonore n'est pas linéaire, mais suit la relation suivante :

$$h_\theta = \frac{1}{1 + e^{-(\theta_0 + \theta_1 x)}}$$

Calculer le couple de coefficients (θ_0, θ_1) permettant de prédire si le piéton est « très proche », « proche », « loin », ou « très loin » en fonction de la distance, et prédire si un piéton situé à 1,5 m est considéré comme « proche » ou « très proche ».

On prendra $\alpha = 0{,}25$ et $m = 4$

Régression logistique : Exemple



classification

EXEMPLE

i	$x^{(i)}\ [m]$	$y^{(i)}$	label
1	-2	0,08	Très proche
2	-1	0,27	proche
3	0	0,5	loin
4	2	0,97	Très loin
5	1,5	?	?

Avec les paramètres $\alpha = 0{,}25$ et $m = 4$, la descente de gradient donne :

$$\frac{\partial \theta_j}{\partial t} = -\alpha \frac{\partial}{\partial \theta_j} J(\theta) \text{ avec } \begin{cases} \dfrac{\partial}{\partial \theta_0} J(\theta) = -\dfrac{0{,}25}{4} \displaystyle\sum_{i=1}^{4} h_\theta\left(x^{(i)}\right) - y^{(i)} \\ \dfrac{\partial}{\partial \theta_1} J(\theta) = -\dfrac{0{,}25}{4} \displaystyle\sum_{i=1}^{4} \left(h_\theta\left(x^{(i)}\right) - y^{(i)}\right) x^{(i)} \end{cases}$$

Une approximation par différence finie donne :

$$\frac{\partial \theta_j}{\partial t} \approx \frac{\theta_j^{n+1} - \theta_j}{\alpha} \text{ d'où : } \begin{cases} \dfrac{\partial}{\partial \theta_0} J(\theta) = \theta_0^n - \dfrac{0{,}25}{3} \displaystyle\sum_{i=1}^{4} h_\theta\left(x^{(i)}\right) - y^{(i)} \\ \dfrac{\partial}{\partial \theta_1} J(\theta) = \theta_1^n - \dfrac{0{,}25}{3} \displaystyle\sum_{i=1}^{4} \left(h_\theta\left(x^{(i)}\right) - y^{(i)}\right) x^{(i)} \end{cases}$$

On fixe les conditions initiales : $\theta_0^0 = 0$ et $\theta_1^0 = 0$

$$h_\theta\left(x^{(i)}\right) = 0{,}5$$

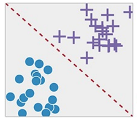
Itération $n = 1$

$$\theta_0^1 = -\frac{\alpha}{m} \sum_{i=1}^{m} h_\theta\left(x^{(i)}\right) - y^{(i)} = \frac{1}{4 \cdot 4} \sum_{i=1}^{4} 0{,}5 - y^{(i)} = \mathbf{0,011}$$

$$\theta_1^1 = -\frac{\alpha}{m} \sum_{i=1}^{m} \left(h_\theta\left(x^{(i)}\right) - y^{(i)}\right) x^{(i)} = \frac{1}{4 \cdot 4} \sum_{i=1}^{4} \left(0{,}5 - y^{(i)}\right) \cdot x^{(i)} = \mathbf{-0,125}$$

$$\boldsymbol{h_\theta(x) = \frac{1}{1 + e^{-\left(0{,}011 - 0{,}12 x^{(i)}\right)}}}$$

Régression logistique : Exemple



classification

Itération $n = 2$

$$\theta_0^2 = \theta_0^1 - \frac{\alpha}{m}\sum_{i=1}^{m} h_\theta\left(x^{(i)}\right) - y^{(i)} = 0{,}011 - \frac{0{,}35 + 0{,}19 - 0{,}0022 - 0{,}41}{4 \cdot 4} = \mathbf{0,003}$$

$$\theta_1^2 = \theta_1^1 - \frac{\alpha}{m}\sum_{i=1}^{m} \left(h_\theta\left(x^{(i)}\right) - y^{(i)}\right)x^{(i)} = -0{,}125 - \frac{-0{,}7 - 0{,}19 + 0 - 0{,}82}{4 \cdot 4} = \mathbf{-0,017}$$

$$\boldsymbol{h_\theta(x)} = \frac{\mathbf{1}}{\mathbf{1} + \boldsymbol{e}^{-\left(\mathbf{0{,}003 - 0{,}017}\boldsymbol{x}^{(i)}\right)}}$$

Itération $n = 3$

$$\theta_0^3 = \theta_0^2 - \frac{\alpha}{m}\sum_{i=1}^{m} h_\theta\left(x^{(i)}\right) - y^{(i)} = 0{,}003 - \frac{0{,}41 + 0{,}225 - 0{,}00075 - 0{,}46}{4 \cdot 4} = \mathbf{-0,008}$$

$$\theta_1^3 = \theta_1^2 - \frac{\alpha}{m}\sum_{i=1}^{m} \left(h_\theta\left(x^{(i)}\right) - y^{(i)}\right)x^{(i)} = -0{,}017 - \frac{-0{,}821 - 0{,}225 + 0 - 0{,}92}{4 \cdot 4} = \mathbf{0,105}$$

$$\boldsymbol{h_\theta(x)} = \frac{\mathbf{1}}{\mathbf{1} + \boldsymbol{e}^{-\left(\mathbf{-0{,}008 + 0{,}105}\boldsymbol{x}^{(i)}\right)}}$$

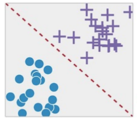
Itération $n = 4$

$$\theta_0^4 = \theta_0^3 - \frac{\alpha}{m}\sum_{i=1}^{m} h_\theta\left(x^{(i)}\right) - y^{(i)} = \mathbf{-0,21}$$

$$\theta_1^4 = \theta_1^3 - \frac{\alpha}{m}\sum_{i=1}^{m} \left(h_\theta\left(x^{(i)}\right) - y^{(i)}\right)x^{(i)} = \mathbf{0,245}$$

$$\boldsymbol{h_\theta(x)} = \frac{\mathbf{1}}{\mathbf{1} + \boldsymbol{e}^{-\left(\mathbf{-0{,}21 + 0{,}245}\boldsymbol{x}^{(i)}\right)}}$$

Régression logistique : Exemple



classification

Itération $n = 5$

$$\theta_0^5 = \theta_0^4 - \frac{\alpha}{m}\sum_{i=1}^{m} h_\theta(x^{(i)}) - y^{(i)} = 0{,}003 - \frac{0{,}41 + 0{,}225 - 0{,}00075 - 0{,}46}{4 \cdot 4} = \mathbf{-0,21}$$

$$\theta_1^5 = \theta_1^4 - \frac{\alpha}{m}\sum_{i=1}^{m} \left(h_\theta(x^{(i)}) - y^{(i)}\right)x^{(i)} = -0{,}017 - \frac{-0{,}821 - 0{,}225 + 0 - 0{,}92}{4 \cdot 4} = \mathbf{0,245}$$

$$\boldsymbol{h_\theta(x) = \frac{1}{1 + e^{-(-0{,}008 - 0{,}105x^{(i)})}}}$$

Itération $n = 6$

$$\theta_0^6 = \theta_0^5 - \frac{\alpha}{m}\sum_{i=1}^{m} h_\theta(x^{(i)}) - y^{(i)} = 0{,}21 - \frac{0{,}54 - 0{,}29 + 0{,}005 - 0{,}585}{4 \cdot 4} = \mathbf{-0,03}$$

$$\theta_1^6 = \theta_1^5 - \frac{\alpha}{m}\sum_{i=1}^{m} \left(h_\theta(x^{(i)}) - y^{(i)}\right)x^{(i)} = 0{,}245 - \frac{-1{,}09 - 0{,}29 + 0 - 1{,}17}{4 \cdot 4} = \mathbf{0,405}$$

$$\boldsymbol{h_\theta(x) = \frac{1}{1 + e^{-(-0{,}03 + 0{,}405x^{(i)})}}}$$

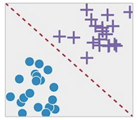
Itération $n = 20$

$$\theta_0^{20} = \theta_0^{19} - \frac{\alpha}{m}\sum_{i=1}^{m} h_\theta(x^{(i)}) - y^{(i)} = \mathbf{-0,21}$$

$$\theta_1^{20} = \theta_1^{19} - \frac{\alpha}{m}\sum_{i=1}^{m} \left(h_\theta(x^{(i)}) - y^{(i)}\right)x^{(i)} = \mathbf{0,245}$$

$$\boldsymbol{h_\theta(x) = \frac{1}{1 + e^{-(-0{,}21 + 0{,}245x^{(i)})}}}$$

Régression logistique : Exemple



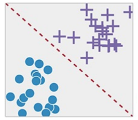
classification

$$h_\theta(-2) = \frac{1}{1 + e^{-(-0{,}21 + 0{,}245 \cdot (-2))}} = \mathbf{0{,}0759}$$

$$h_\theta(-1) = \frac{1}{1 + e^{-(-0{,}21 + 0{,}245 \cdot (-1))}} = \mathbf{0{,}2689}$$

$$h_\theta(-0{,}5) = \frac{1}{1 + e^{-(-0{,}21 + 0{,}245 \cdot (-0{,}5))}} = \mathbf{0{,}4378}$$

$$h_\theta(1{,}5) = \frac{1}{1 + e^{-(-0{,}21 + 0{,}245 \cdot 1{,}5)}} = \mathbf{0{,}9399}$$

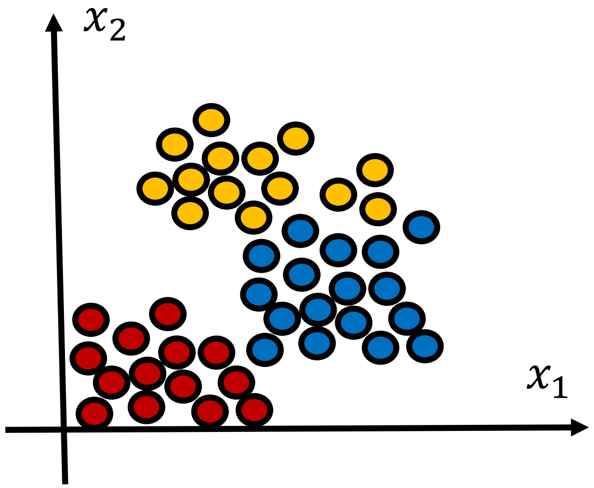


Classification multi-classe

■ Lorsque l'on a plus de deux classes à discriminer, on applique une classification « un contre tous » : pour n classes, on ramène le problème à n problèmes de classification binaire distincts.

■ Par exemple pour $n = 3$:

- Bleu (1) vs rouge et jaune (0) $h_\theta^1(x)$: $\boldsymbol{P(y = 1|x_1; \theta)}$
- Rouge (1) vs jaune et bleu (0) $h_\theta^1(x)$: $\boldsymbol{P(y = 1|x_2; \theta)}$
- Jaune (1) vs bleu et rouge (0) $h_\theta^1(x)$: $\boldsymbol{P(y = 1|x_3; \theta)}$



■ On entraine alors le classificateur de régression logistique $h_\theta^{(i)}(x)$ pour chaque classe i afin de prédire la probabilité que $y = i$

■ Sur une nouvelle entrée x, on choisit la classe i qui maximise la probabilité que $h_\theta^{(i)}(x) = 1$

Classification multi-classe

- Hypothèse :

$$h_\theta(x) = g(\theta_0 + \theta_1 x + \theta_2 x_2 + \theta_3 x_1^2 + \theta_4 x_2^2 + \theta_5 x_1 x_2 + \theta_6 x_1^3 x_2 + \theta_7 x_1 x_2^3 + \cdots)$$

Fonction coût pour une régression logistique régularisée

$$J(\theta) = \frac{1}{m}\left[\sum_{i=1}^{m} y^{(i)} \log\left(h_\theta\left(x^{(i)}\right)\right) + (1 - y^{(i)}) \log\left(1 - h_\theta\left(x^{(i)}\right)\right) + \frac{\lambda}{2}\sum_{j=1}^{n} \theta_j^2\right]$$

Algorithme

Répéter $\min_\theta J(\theta)$:

{

$$\theta_0 := \theta_0 - \alpha \frac{1}{m}\sum_{i=1}^{m}\left(h_\theta\left(x^{(i)}\right) - y^{(i)}\right)$$

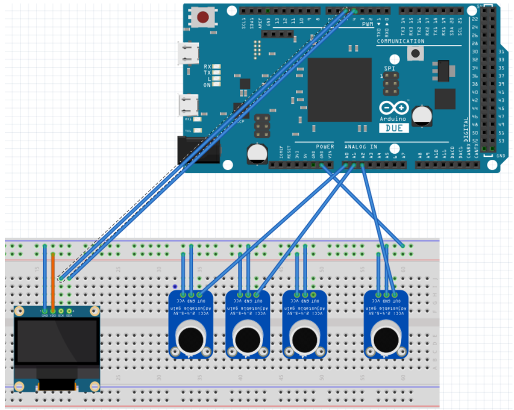
$$\theta_j := \theta_j - \alpha \left[\frac{1}{m}\sum_{i=1}^{m}\left(h_\theta\left(x^{(i)}\right) - y^{(i)}\right) x_j^{(i)} + \lambda \theta_j\right]$$

Mise à jour des θ_j

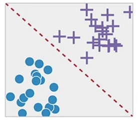
}

EXEMPLE

Classifier les voix dans une salle en utilisant plusieurs microphones.



Séparateur à Vaste Marge (1 / 6)



classification

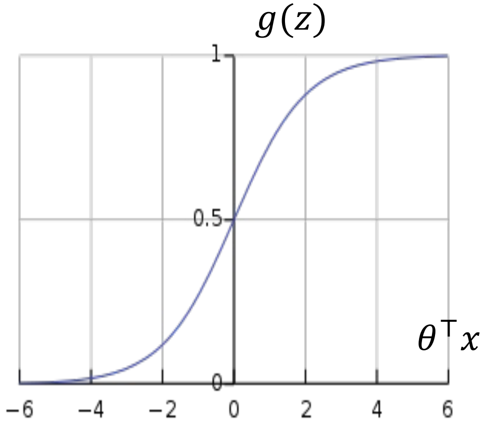
■ Le **Séparateur à vaste marge** (SVM) permet également de traiter les problèmes de classification supervisés en maximisant la marge (distance) entre les échantillons proches de la frontière de décision, appelés **vecteurs support**.

■ On utilise une fonction coût similaire à celle utilisée pour la régression logistique, mais avec quelques changements.

Avant changement :

$$h_\theta(x) = g(\theta^\top x) = \frac{1}{1 + e^{-\theta^\top x}}$$

$$\text{Prédire } y = \begin{cases} 1 & \text{si } h_\theta(x) \geq 0{,}5 \text{ si } \boldsymbol{\theta^\top x \gg 0} \text{ pour avoir } h_\theta(x) \approx 1 \\ 0 & \text{si } h_\theta(x) < 0{,}5 \text{ si } \boldsymbol{\theta^\top x \ll 0} \text{ pour avoir } h_\theta(x) \approx 0 \end{cases}$$



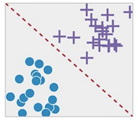
■ Afin de pouvoir optimiser les paramètres, il faut que la fonction coût $J(\theta)$ soit une **fonction convexe** : la descente de gradient doit converger vers un mimimul local.

Or cela peut ne pas être le cas ici (fonction non convexe → plusieurs optimums locaux et la descente de gradient peut restée bloquée dans un minimum local.)

■ Pour contourner ce problème, on utilise une fonction coût différente, qui cette fois-ci est convexe :

$$J(h_\theta(x), y) = \begin{cases} -log\big(h_\theta(x)\big) \text{ si } y = 1 \\ -log\big(1 - h_\theta(x)\big) \text{ si } y = 0 \end{cases}$$

Séparateur à Vaste Marge (2 / 6)



classification

Fonction coût convexe

$$J(\theta) = -\frac{1}{m}\sum_{i=1}^{m} y^{(i)} log\left(h_\theta\left(x^{(i)}\right)\right) + \left(1 - y^{(i)}\right) log\left(1 - h_\theta\left(x^{(i)}\right)\right)$$

- Il convient alors d'appliquer l'algorithme de descente de gradient sur $J(\theta)$ pour chaque paramètre :

Algorithme

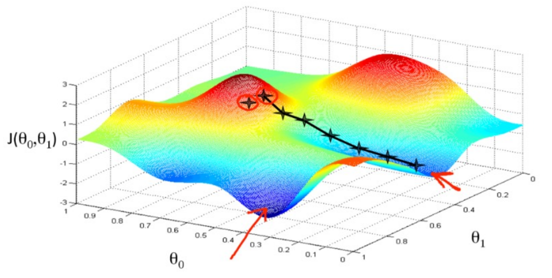
Répéter $\min_\theta J(\theta)$:

{

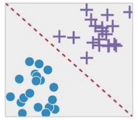
$$\theta_j^{n+1} = \theta_j^n - C\sum_{i=1}^{m} y^{(i)}\left(1 - h_\theta\left(x^{(i)}\right)\right) - \left(1 - y^{(i)}\right)h_\theta\left(x^{(i)}\right) + \theta_j^n$$

Mise à jour des θ_j

}



Séparateur à Vaste Marge (3 / 6)



classification

Fonction coût SVM

$$J(\theta) = C \sum_{i=1}^{m} y^{(i)} \log\left(h_\theta\left(x^{(i)}\right)\right) + \left(1 - y^{(i)}\right) log\left(1 - h_\theta\left(x^{(i)}\right)\right)$$

- Lorsque la valeur de C est très grande, il convient d'écrire $J(\theta)$ pour chaque paramètre :

Fonction coût SVM à grand paramètre *C*

$$J(\theta) = \begin{cases} \min\limits_{\theta} \left(\sum_{i=1}^{n} \theta_i^2 \right) \\ \text{sous contrainte :} \\ y = 1 \text{ si } \theta^T x^{(i)} \geq 1 \\ y = 0 \text{ si } \theta^T x^{(i)} \leq 1 \end{cases}$$

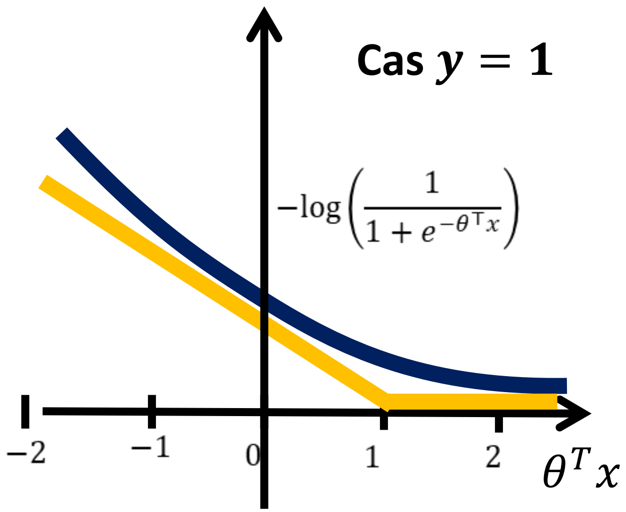
- On dit que les échantillons sont linéairement séparables si Les prédictions $y^{(i)} = \begin{cases} 1 \text{ si } \theta^T x^{(i)} \geq 1 \\ 0 \text{ si } \theta^T x^{(i)} \leq -1 \end{cases}$
- Les prédictions sont alors contrôlées par la largeur de la marge.

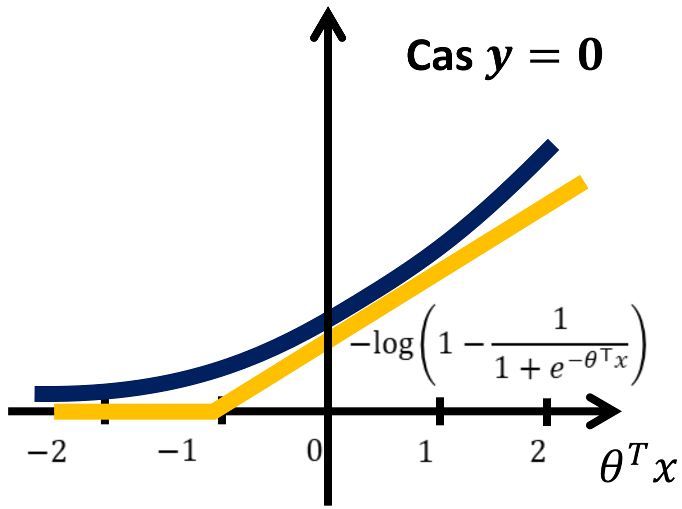
Séparateur à Vaste Marge (4 / 6)

classification

- Il convient également de rendre plus discriminant les conditions aux limites :

$$\text{Prédire } y = \begin{cases} 1 & \text{si } h_\theta(x) \geq 0{,}5 \text{ si } \theta^\top x \gg 0 \\ 0 & \text{si } h_\theta(x) < 0{,}5 \text{ si } \theta^\top x \ll 0 \end{cases} \Rightarrow \text{Prédire } y = \begin{cases} 1 & \text{si } h_\theta(x) \geq 0{,}5 \text{ si } \theta^\top x \geq 0 \\ 0 & \text{si } h_\theta(x) < 0{,}5 \text{ si } \theta^\top x \leq 0 \end{cases}$$





$$J(h_\theta(x), y) = -y \log(h_\theta(x)) - (1-y) \log(1 - h_\theta(x))$$

Séparateur à Vaste Marge (5 / 6)

classification

■ Il est nécessaire de représenter graphiquement l'ecart entre la prediction positif et négatif par une marge de sécurité notée *M* :

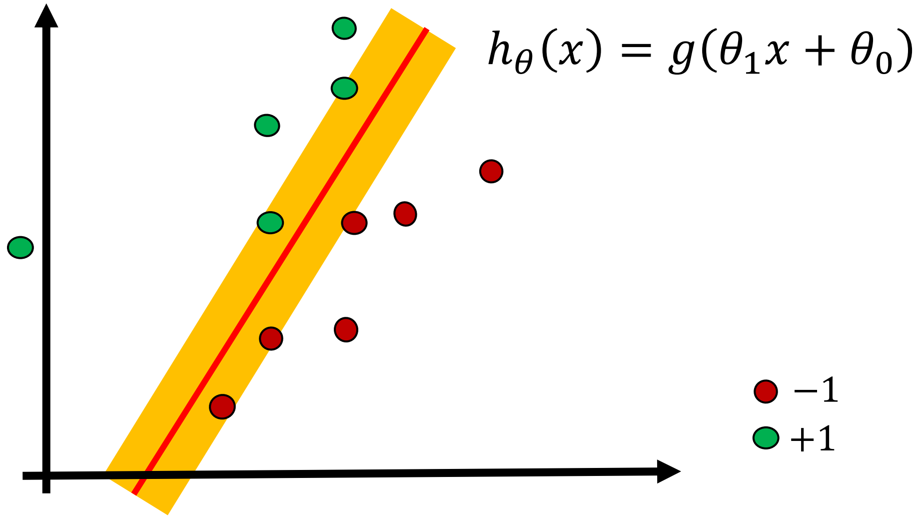
Marge de sécurité *M*

$$M = \frac{\theta_1 \cdot x^+ + \theta_0 - \theta_1 \cdot x^- + \theta_0}{|\theta_1|} = \frac{(x^+ - x^-) \cdot \theta_1}{|\theta_1|} = \frac{2}{|\theta_1|}$$

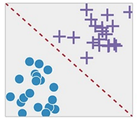
■ La marge *M* donnée par l'écart normalisé entre la frontière de prédiction positive et la frontière de prédiction négative est :

- $y = +1 = \theta_1 \cdot x^+ + \theta_0$
- $y = -1 = \theta_1 \cdot x^- + \theta_0$

■ On cherche à **maximiser** cette marge.



Séparateur à Vaste Marge (6 / 6)



- La fonction coût $J(\theta)$ permettant de maximiser la marge est dite **fonction de coût charnière**.
- L'apprentissage est similaire à celui qui a été vu précédemment.

Algorithme

$$h_\theta(x) = \frac{1}{1 + e^{-\theta^\top x}}$$

Répéter $\min_\theta J(\theta)$:

{

$$\theta_0 := \theta_0 - \alpha C \sum_{i=1}^{m} \left(h_\theta\left(x^{(i)}\right) - y^{(i)}\right)$$

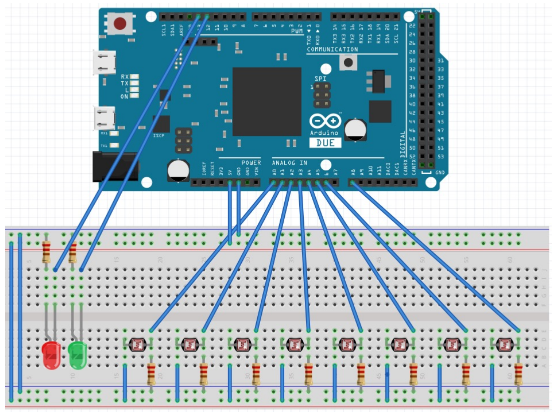
$$\theta_j := \theta_j - \alpha \left[C \sum_{i=1}^{m} \left(h_\theta\left(x^{(i)}\right) - y^{(i)}\right) x_j^{(i)} + \theta_j \right]$$

Mise à jour des θ_j

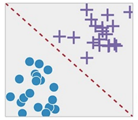
}

EXEMPLE

Contrôler l'eclairage d'une salle à l'aide de 2 capteurs au lieu de la salle et completer l'homogeneité d'eclaire par 6 autres capteurs repartis dans la salle



Astuce du noyau



- La SVM fait l'hypothèse que les groupes sont **linéairement séparables**, mais ce n'est pas toujours le cas en pratique.
- L'**astuce du noyau** permet de représenter la non-séparabilité entre les groupes. Elle consiste à :
  - plonger les données dans un nouvel espace appelé **espace de features** ;
  - appliquer une svm sur les features.
- Le terme **astuce** vient du fait que ce procédé ne nécessite pas de connaître explicitement les features: pour résoudre le problème de la SVM on a juste besoin de connaître le noyau associé au features. D'un point de vu formel un noyau est une fonction mathématique simple (il en existe plusieurs sortes avec des propriétés différentes…) :

Fonction coût SVM avec noyau

$$J(\theta) = C\sum_{i=1}^{m} y^{(i)}\log\left(h_\theta\left(x^{(i)}\right)\right) + \left(1-y^{(i)}\right)log\left(1-h_\theta\left(x^{(i)}\right)\right) + \frac{1}{2}\sum_{i=1}^{n}\theta_i^2$$

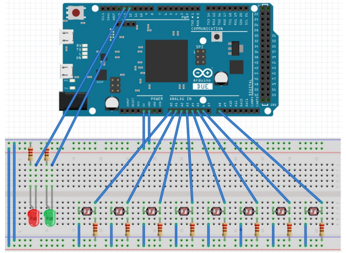
Fonction coût SVM avec noyau

$$J(\theta) = C\sum_{i=1}^{m} y^{(i)}J_1\left(\theta^T f^{(i)}\right) + \left(1-y^{(i)}\right)J_0\left(\theta^T f^{(i)}\right) + \frac{1}{2}\sum_{i=1}^{n}\theta_i^2$$

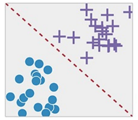
Exemple de features :
$f_1 = x_1,\ f_2 = x_2,\ f_3 = x_1 \cdot x_2$, etc.

EXEMPLE

Couplage Optique entre les 8 capteurs pour detection du bon electrage dans une sale.



Séparateur à Vaste Marge : Exemple

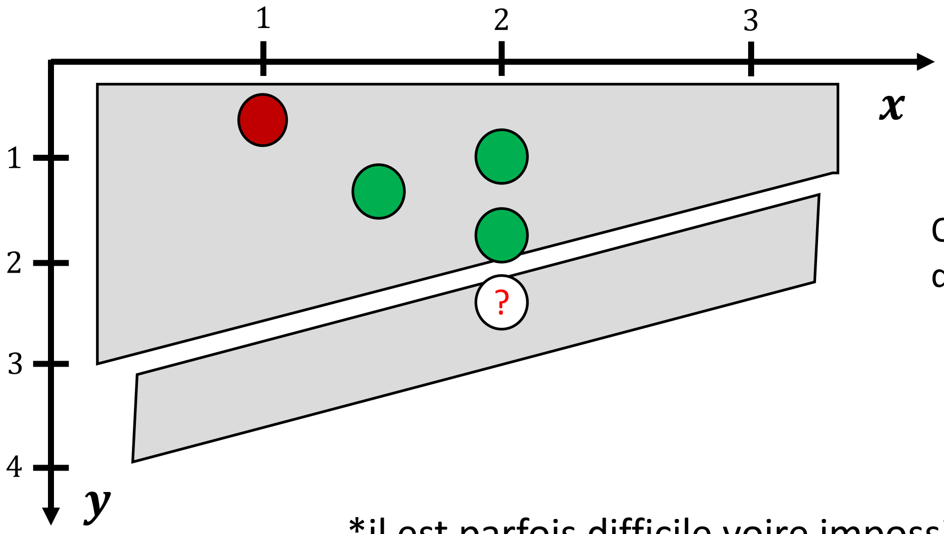


classification

EXEMPLE Classer les zones de l'aile d'un avion pour savoir si elles sont d'une température très différente (+1) ou non (-1) du fuselage de l'avion*.

Le tableau suivant présente pour différents couples de coordonnées $\left(x^{(i)}, y^{(i)}\right)$ un label correspondant au gradient de température entre le fuselage de l'avion et le point de mesure :

i	$x^{(i)}\ [m]$	$y^{(i)}\ [m]$	label
1	1	2	+1
2	1,5	1,5	+1
3	1	0,5	-1
4	2	2	+1
5	2,5	3	?

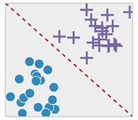


On émet l'hypothèse que que la température en fonction de l'éloignement du fuselage est une fonction non linéaire de la forme suivante :

$$h_\theta = \frac{1}{1 + e^{-(\theta_0 + \theta_1 x)}}$$

*il est parfois difficile voire impossible de tout modéliser ou de placer des capteurs à certains endroits.

Séparateur à Vaste Marge : Exemple



classification

On prend $\alpha = 0{,}25$ et $m = 4$

La descente de gradient donne :

$$\frac{\partial \theta_j}{\partial t} = -\frac{\partial}{\partial \theta_j} J(\theta) \text{ avec } \begin{cases} \dfrac{\partial}{\partial \theta_0} J(\theta) = -C\alpha \displaystyle\sum_{i=1}^{4} h_\theta\left(x^{(i)}\right) - y^{(i)} \\ \dfrac{\partial}{\partial \theta_0} J(\theta) = -C\alpha \displaystyle\sum_{i=1}^{4} \left(h_\theta\left(x^{(i)}\right) - y^{(i)}\right) x^{(i)} - \lambda \theta_1 \end{cases}$$

Itération 1

i	$x^{(i)}$ $[m]$	$y^{(i)}$ $[m]$	label
1	1	2	-1
2	1,5	1,5	-1
3	1	0,5	-1
4	2	2	+1
5	2,5	3	+1

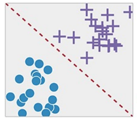
Itération 2

i	$x^{(i)}$ $[m]$	$y^{(i)}$ $[m]$	label
1	1	2	-1
2	1,5	1,5	-1
3	1	0,5	-1
4	2	2	+1
5	2,5	3	-1

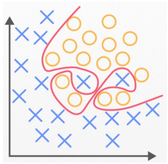
Itération 3

i	$x^{(i)}$ $[m]$	$y^{(i)}$ $[m]$	label
1	1	2	-1
2	1,5	1,5	-1
3	1	0,5	-1
4	2	2	+1
5	2,5	3	-1

Régression logistique vs SVM



classification

Régression logistique	Séparateur à vaste marge
• Sépare les données en plusieurs classes à l’aide d’une fonction coût au choix qui est minimisée.	• Maximise la marge (distance entre la frontière de décision et les vecteurs de support) afin de minimiser l’erreur de classification.
• Approche statistique	• Approche géométrique
• Adapté pour des données classifiées et indépendantes	• Adapté pour des données semi-structurées (texte, images, etc.)
• peut entrer dans un biais de surapprentissage 	• Pas de risque de surapprentissage

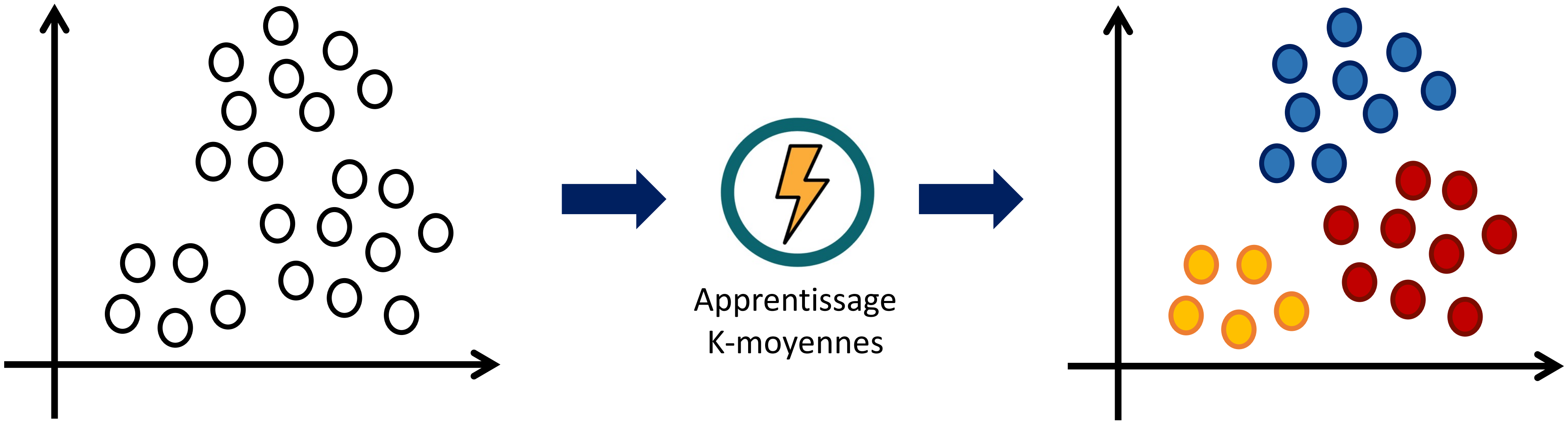
	Nb feature >> nb data	Peu de features Peu de données	Peu de features Beaucoup de données
RL	✓	✓	✓ créer des features intermédiaires
SVM	✓	✓ avec noyau	✓ créer des features intermédiaires

3 L’apprentissage non supervisé

K-moyennes

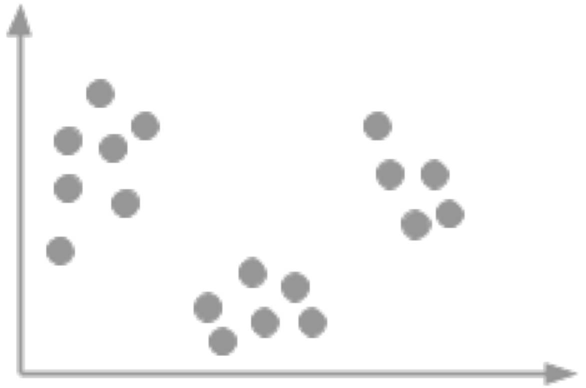
■ Dans ce cours, nous ne nous intéresserons qu'à une seule technique d'apprentissage machine non-supervisé : le **partitionnement en K-moyennes** (ou *k-means* en anglais).

■ Il s'agit de regrouper les points en k groupes, ou *clusters*, de façon à minimiser une fonction coût : la somme des carrés des distances entre les points et la moyenne des points de son cluster.



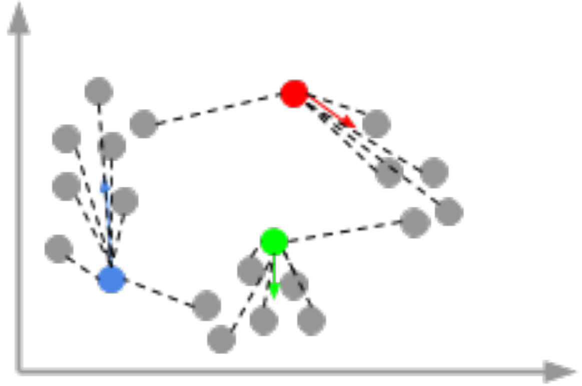
K-moyennes

- Étant donné :
  - un ensemble de n points $\{x^{(1)}, x^{(2)}, x^{(3)}, \cdots, x^{(m)}\}$
  - à partitionner en k clusters $S_k = \{S_1, S_2, \ldots, S_k\} (k \leq n)$
  - à partir de Mesures de distance/dissimilarité $\mathrm{d}\left(x^{(i)}, x^{(j)}\right)$

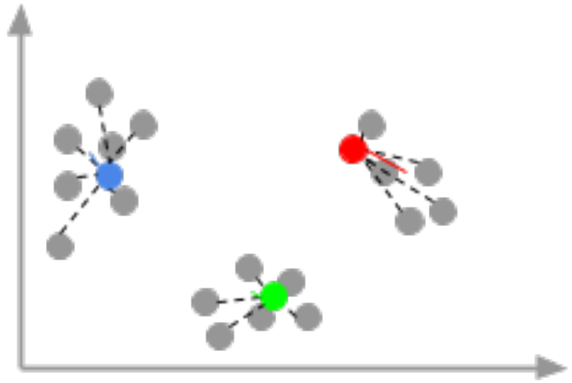


Données d'entrée

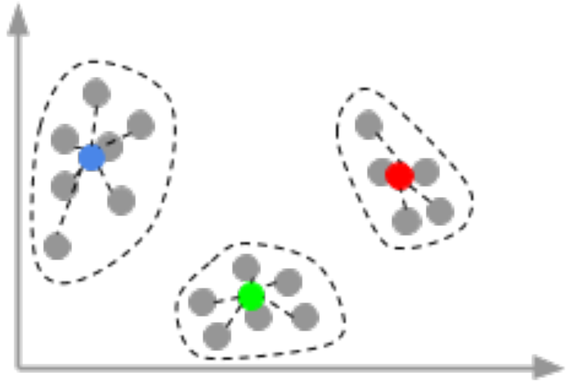
- On peut partitionner les données comme suit :
  - Par distances euclidiennes : **k-moyenne** ;
  - Par distances absolues : **K-médiane** ;
  - En minimisant le rayon maximum : **K-centre**.



❶ Initialisation

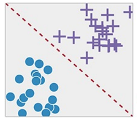


❷ Assignation



❸ Calcul des points moyens

K-moyennes



Démarche

❶ initialiser aléatoirement K centroïdes $\mu_1, \mu_2, \ldots \mu_K \in \mathbb{R}^n$

❷ Itérer :

Étape d'assignation des cluster

Pour $i = 1 : m$

$$J\left(\boldsymbol{c}^{(1)}, \cdots, \boldsymbol{c}^{(m)}, \mu_1, \cdots, \mu_K\right) = \frac{1}{m}\sum_{i=1}^{m}\left\|x^{(i)} - \mu_{c^{(i)}}\right\|^2$$

$c^{(i)}$ ≔ indice (1 -> K) du centroïde du cluster le plus proche de $x^{(i)}$

Étape de mise à jour des centroïdes

Pour j = $1 : K$

$$J\left(c^{(1)}, \cdots, c^{(m)}, \boldsymbol{\mu_1}, \cdots, \boldsymbol{\mu_K}\right) = \frac{1}{m}\sum_{i=1}^{m}\left\|x^{(i)} - \mu_{c^{(i)}}\right\|^2$$

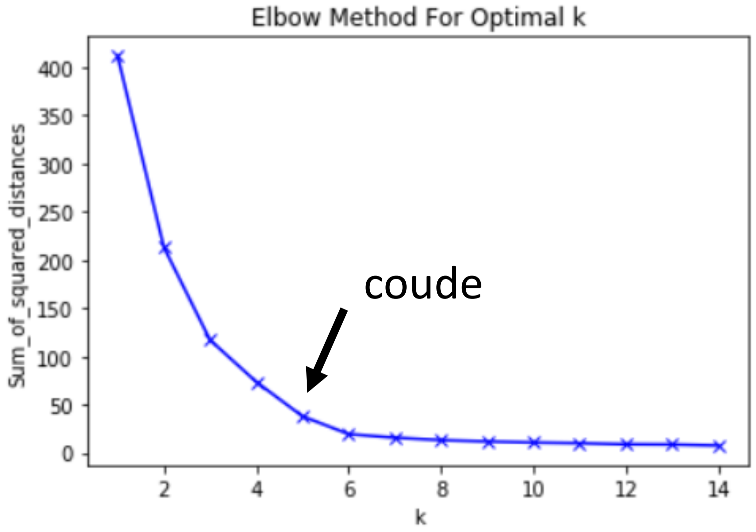
μ_k ≔ moyenne (mean) des points attribués au cluster$k$

K-moyennes

Choix du nombre de clusters K

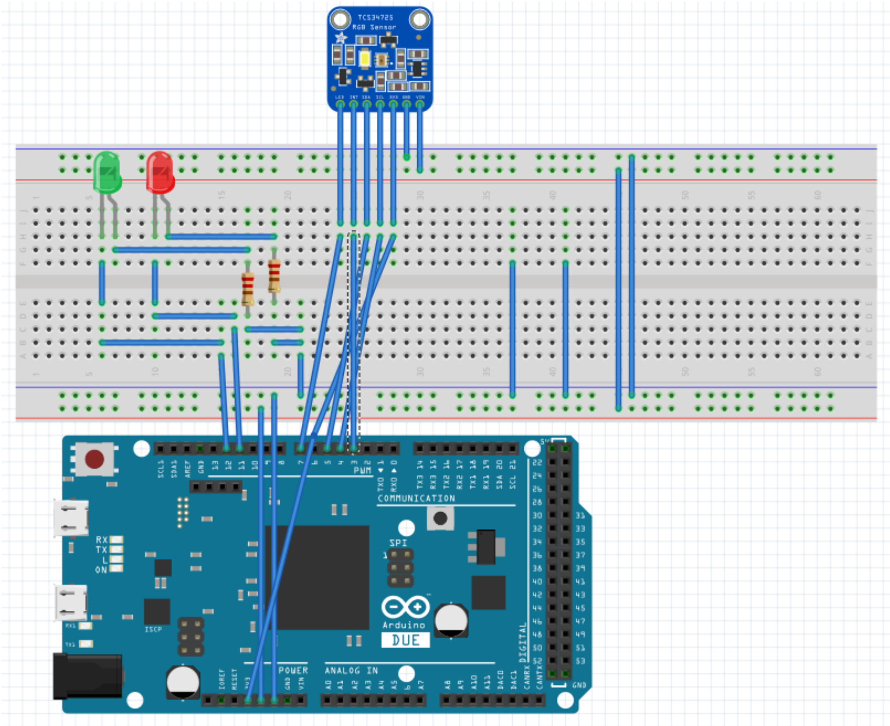
■ La fonction de coût diminue avec le nombre d'itérations.

■ L'optimal choisi est dans la région du coude : le moins de clusters possible pour un J minimisé.



EXEMPLE

Regrouper des fruits par leurs couleurs



K-moyennes

EXEMPLE Séparer les sources wifi/radiofréquences d'un module NRF24L en fonction de la puissance du signal reçu.

Un capteur NRF24L reçoit les signaux de 7 autres modules eux-mêmes connectés à des Arduino, avec des distances et des messages de natures différentes.

$\boldsymbol{i}$	1	2	3	4	5	6	7
$x^{(i)}\ [m]$	1	1,5	1	2	1	1,3	2,1
$y^{(i)}\ [m]$	2	1,2	0,5	2	3	2	4

→ Classer les sources en fonction de leurs puissance d'émission.

On émet l'hypothèse qu'il y a $K = 3$ centroïdes.
On initialise :

$$\mu_0 = \mu_1 = \mu_2 = 0$$

On assigne les clusters :

$$c^{(i)} = \frac{1}{7}\sum_{i=1}^{7}\left\|x^{(i)} - \mu_{c^{(i)}}\right\|^2$$

$\boldsymbol{i}$	1	2	3	4	5	6	7
$c^{(i)}\ [m]$	0,5	0,7	0,7	0,7	0,7	0,5	0,8

On met à jour les centroïdes

$$J\left(c^{(1)}, \cdots, c^{(m)}, \mu_1, \cdots, \mu_K\right) = \frac{1}{m}\sum_{i=1}^{m}\left\|x^{(i)} - \mu_{c^{(i)}}\right\|^2$$

Cours 4 : Introduction à l’apprentissage machine

■ L’apprentissage machine consiste à développer un modèle que l’on commence par entraîner à l’aide d’un algorithme d’optimisation en vue de minimiser les erreurs entre ce modèle et les données. Une fois entraîné, le modèle permet de faire des prédiction ou de classifier des données.

■ Nous avons vu deux méthodes d’apprentissage supervisé :

- De prédiction : la régression linéaire
- De classification : la régression logistique et le séparateur à vaste marge (SVM)

■ Et une méthode d’apprentissage non-supervisé :

- De classification : K-moyennes