

INTRODUCTION

Jérémie Cabessa
Laboratoire DAVID, UVSQ

INTELLIGENCE ARTIFICIELLE (IA)

- ▶ **L'intelligence artificielle (IA)** désigne l'ensemble des théories et techniques visant à simuler certains aspects de l'intelligence humaine par des systèmes informatiques.
- ▶ Ces aspects de l'intelligence comprennent entre autres: le raisonnement, l'apprentissage, la perception, etc.



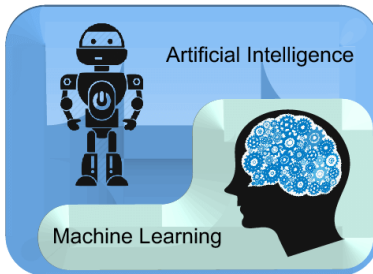
INTELLIGENCE ARTIFICIELLE (IA)

- ▶ L'**intelligence artificielle (IA)** désigne l'ensemble des théories et techniques visant à simuler certains aspects de l'intelligence humaine par des systèmes informatiques.
- ▶ Ces aspects de l'intelligence comprennent entre autres: le raisonnement, l'apprentissage, la perception, etc.



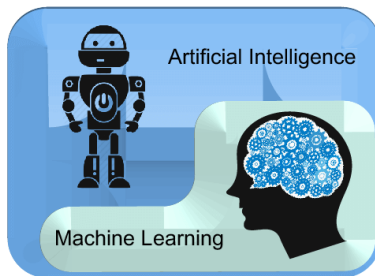
MACHINE LEARNING (ML)

- ▶ Le **machine learning (ML)** désigne l'ensemble des théories et techniques algorithmiques qui permettent d'apprendre à des systèmes informatiques à résoudre des problèmes, exécuter des tâches, etc., en se basant sur des données (data).
- ▶ Le machine learning (ML) est un sous-domaine de l'intelligence artificielle (IA).



MACHINE LEARNING (ML)

- ▶ Le **machine learning (ML)** désigne l'ensemble des théories et techniques algorithmiques qui permettent d'apprendre à des systèmes informatiques à résoudre des problèmes, exécuter des tâches, etc., en se basant sur des données (data).
- ▶ Le machine learning (ML) est un sous-domaine de l'intelligence artificielle (IA).



MACHINE LEARNING (ML)

- ▶ Le machine learning constitue un **changement de paradigme** dans la résolution de problèmes.
- ▶ **Exemple:** on aimerait créer un algorithme de reconnaissance d'images qui détecte les images contenant des fleurs.
 - ▶ Approche "hard-coded": se base sur notre compréhension du concept de "fleur", programmé en dur.
 - ▶ On code les formes, couleurs, etc., que peuvent avoir les différentes fleurs, et on crée un algorithme qui détecte les fleurs à partir de ces caractéristiques.
 - ▶ Approche "machine learning": se base sur des données d'exemples d'images de "fleurs" et de "non-fleurs" et l'algorithme apprend à détecter les fleurs à partir des données.

MACHINE LEARNING (ML)

- ▶ Le machine learning constitue un **changement de paradigme** dans la résolution de problèmes.
- ▶ **Exemple:** on aimerait créer un algorithme de reconnaissance d'images qui détecte les images contenant des fleurs.

- ▶ **Approche “hard-coded”:** se base sur notre compréhension du concept de “fleur”, programmé en dur.

On code les formes, couleurs, etc., que peuvent avoir les différentes fleurs, et on crée un algorithme qui détecte les fleurs à partir de ces caractéristiques.

- ▶ **Approche “machine learning”:** se base sur des data, i.e., des images de “fleurs” et de “non-fleurs” utilisées comme exemples.

On crée un algorithme qui apprend par lui-même à détecter les fleurs à partir des data.

MACHINE LEARNING (ML)

- ▶ Le machine learning constitue un **changement de paradigme** dans la résolution de problèmes.
- ▶ **Exemple:** on aimerait créer un algorithme de reconnaissance d'images qui détecte les images contenant des fleurs.

- ▶ **Approche “hard-coded”:** se base sur notre compréhension du concept de “fleur”, programmé en dur.

On code les formes, couleurs, etc., que peuvent avoir les différentes fleurs, et on crée un algorithme qui détecte les fleurs à partir de ces caractéristiques.

- ▶ **Approche “machine learning”:** se base sur des data, i.e., des images de “fleurs” et de “non-fleurs” utilisées comme exemples.

On crée un algorithme qui apprend par lui-même à détecter les fleurs à partir des data.

MACHINE LEARNING (ML)

- ▶ Le machine learning constitue un **changement de paradigme** dans la résolution de problèmes.
- ▶ **Exemple:** on aimerait créer un algorithme de reconnaissance d'images qui détecte les images contenant des fleurs.
 - ▶ **Approche “hard-coded”:** se base sur notre compréhension du concept de “fleur”, programmé en dur.

On code les formes, couleurs, etc., que peuvent avoir les différentes fleurs, et on crée un algorithme qui détecte les fleurs à partir de ces caractéristiques.
 - ▶ **Approche “machine learning”:** se base sur des data, i.e., des images de “fleurs” et de “non-fleurs” utilisées comme exemples.

On crée **un algorithme qui apprend par lui-même** à détecter les fleurs à partir des data.

APPRENTISSAGE SUPERVISÉ ET NON-SUPERVISÉ

- ▶ Deux types de méthodes d'apprentissage:
 - **apprentissage supervisé (supervised learning)**
 - **apprentissage non-supervisé (unsupervised learning)**

- ▶ **Apprentissage supervisé**

Les données contiennent des variables d'input et des variables d'output. Il s'agit de modéliser la relation entre les variables d'input et la variable d'output (prediction, classification).

- ▶ **Apprentissage non-supervisé**

Les données ne contiennent que des variables d'input, et pas de variable d'output. Il s'agit d'extraire des classes ou groupes de données présentant des caractéristiques communes (clustering).

APPRENTISSAGE SUPERVISÉ ET NON-SUPERVISÉ

- ▶ Deux types de méthodes d'apprentissage:
 - apprentissage supervisé (supervised learning)
 - apprentissage non-supervisé (unsupervised learning)

- ▶ **Apprentissage supervisé**

Les données contiennent des variables d'input et des variables d'output. Il s'agit de modéliser la relation entre les variables d'input et la variable d'output (prediction, classification).

- ▶ Apprentissage non-supervisé

Les données ne contiennent que des variables d'input, et pas de variable d'output. Il s'agit d'extraire des classes ou groupes de données présentant des caractéristiques communes (clustering).

APPRENTISSAGE SUPERVISÉ ET NON-SUPERVISÉ

- ▶ Deux types de méthodes d'apprentissage:
 - apprentissage supervisé (supervised learning)
 - apprentissage non-supervisé (unsupervised learning)

- ▶ **Apprentissage supervisé**

Les données contiennent des variables d'input et des variables d'output. Il s'agit de modéliser la relation entre les variables d'input et la variable d'output (prediction, classification).

- ▶ **Apprentissage non-supervisé**

Les données ne contiennent que des variables d'input, et pas de variable d'output. Il s'agit d'extraire des classes ou groupes de données présentant des caractéristiques communes (clustering).

APPRENTISSAGE SUPERVISÉ

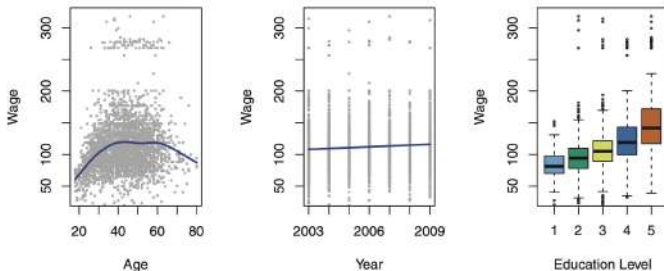


FIGURE 1.1. *Wage* data, which contains income survey information for males from the central Atlantic region of the United States. Left: *wage* as a function of *age*. On average, *wage* increases with *age* until about 60 years of age, at which point it begins to decline. Center: *wage* as a function of *year*. There is a slow but steady increase of approximately \$10,000 in the average *wage* between 2003 and 2009. Right: Boxplots displaying *wage* as a function of *education*, with 1 indicating the lowest level (no high school diploma) and 5 the highest level (an advanced graduate degree). On average, *wage* increases with the level of education.

Figure taken from [James et al., 2013].

APPRENTISSAGE SUPERVISÉ

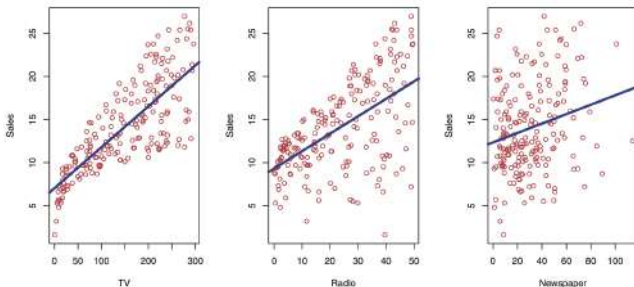


FIGURE 2.1. The Advertising data set. The plot displays `sales`, in thousands of units, as a function of `TV`, `radio`, and `newspaper` budgets, in thousands of dollars, for 200 different markets. In each plot we show the simple least squares fit of `sales` to that variable, as described in Chapter 3. In other words, each blue line represents a simple model that can be used to predict `sales` using `TV`, `radio`, and `newspaper`, respectively.

Figure taken from [James et al., 2013].

APPRENTISSAGE SUPERVISÉ

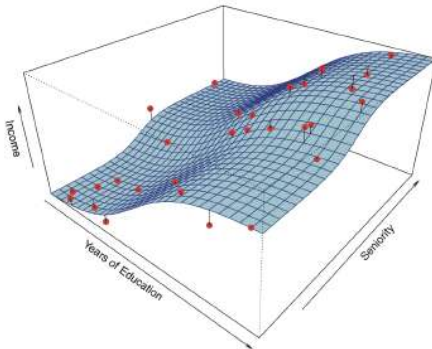


FIGURE 2.3. The plot displays **income** as a function of **years of education** and **seniority** in the **Income** data set. The blue surface represents the true underlying relationship between **income** and **years of education** and **seniority**, which is known since the data are simulated. The red dots indicate the observed values of these quantities for 30 individuals.

Figure taken from [James et al., 2013].

APPRENTISSAGE NON-SUPERVISÉ

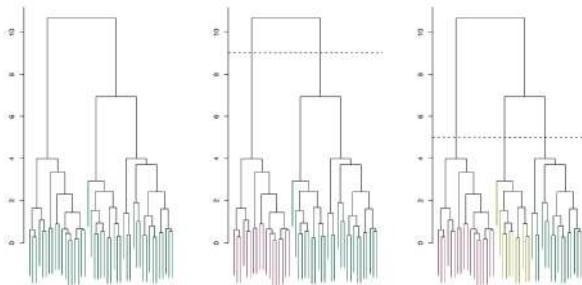


FIGURE 10.9. Left: dendrogram obtained from hierarchically clustering the data from Figure 10.8 with complete linkage and Euclidean distance. Center: the dendrogram from the left-hand panel, cut at a height of nine (indicated by the dashed line). This cut results in two distinct clusters, shown in different colors. Right: the dendrogram from the left-hand panel, now cut at a height of five. This cut results in three distinct clusters, shown in different colors. Note that the colors were not used in clustering, but are simply used for display purposes in this figure.

Figure taken from [James et al., 2013].

RÉGRESSION ET CLASSIFICATION

- ▶ Dans le cadre de l'**apprentissage supervisé**, on distingue deux types de méthodes:
- ▶ Méthodes de régression
La variable d'output est **quantitative**.
- ▶ Méthodes de classification
La variable d'output est **qualitative**.

RÉGRESSION ET CLASSIFICATION

- ▶ Dans le cadre de l'**apprentissage supervisé**, on distingue deux types de méthodes:
- ▶ **Méthodes de régression**
La variable d'output est **quantitative**.
- ▶ Méthodes de classification
La variable d'output est **qualitative**.

RÉGRESSION ET CLASSIFICATION

- ▶ Dans le cadre de l'**apprentissage supervisé**, on distingue deux types de méthodes:
- ▶ **Méthodes de régression**
La variable d'output est **quantitative**.
- ▶ **Méthodes de classification**
La variable d'output est **qualitative**.

RÉGRESSION

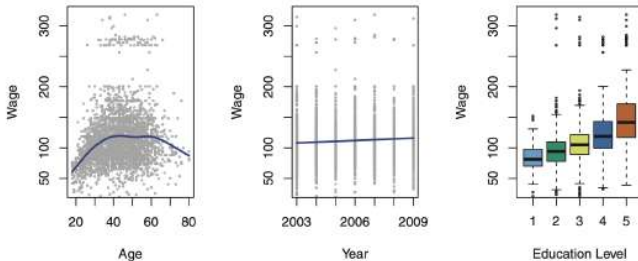


FIGURE 1.1. Wage data, which contains income survey information for males from the central Atlantic region of the United States. Left: wage as a function of age. On average, wage increases with age until about 60 years of age, at which point it begins to decline. Center: wage as a function of year. There is a slow but steady increase of approximately \$10,000 in the average wage between 2003 and 2009. Right: Boxplots displaying wage as a function of education, with 1 indicating the lowest level (no high school diploma) and 5 the highest level (an advanced graduate degree). On average, wage increases with the level of education.

Figure taken from [James et al., 2013].

RÉGRESSION

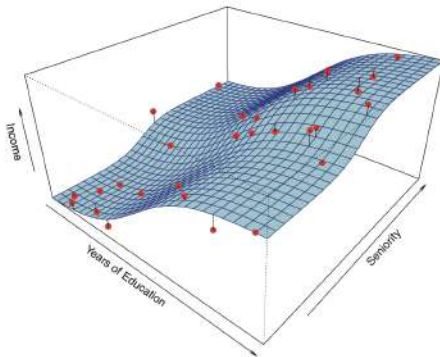


FIGURE 2.3. The plot displays **income** as a function of **years of education** and **seniority** in the **Income** data set. The blue surface represents the true underlying relationship between **income** and **years of education** and **seniority**, which is known since the data are simulated. The red dots indicate the observed values of these quantities for 30 individuals.

Figure taken from [James et al., 2013].

CLASSIFICATION

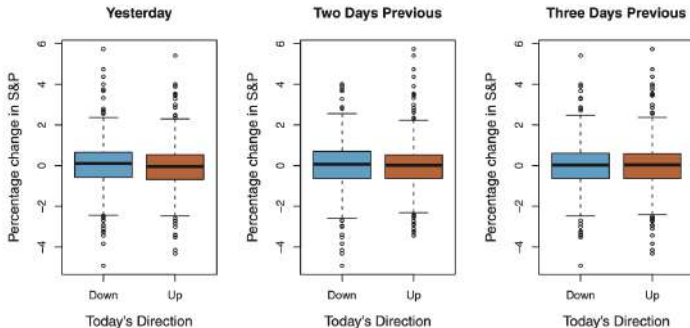


FIGURE 1.2. Left: *Boxplots of the previous day's percentage change in the S&P index for the days for which the market increased or decreased, obtained from the Smarket data.* Center and Right: *Same as left panel, but the percentage changes for 2 and 3 days previous are shown.*

Figure taken from [James et al., 2013].

CLASSIFICATION

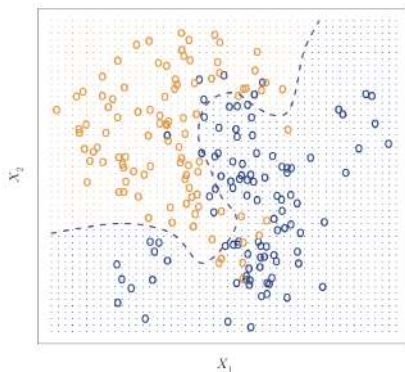


FIGURE 2.13. A simulated data set consisting of 100 observations in each of two groups, indicated in blue and in orange. The purple dashed line represents the Bayes decision boundary. The orange background grid indicates the region in which a test observation will be assigned to the orange class, and the blue background grid indicates the region in which a test observation will be assigned to the blue class.

Figure taken from [James et al., 2013].

BIBLIOGRAPHIE

Most images taken from [James et al., 2013].



James, G., Witten, D., Hastie, T., and Tibshirani, R. (2013). *An Introduction to Statistical Learning: with Applications in R*, volume 103 of *Springer Texts in Statistics*. Springer, New York.