Cours : Machine Learning

Chapitre 2 : Apprentissage supervisé

Dr. Asma Ouertatani

Université Centrale

October 7, 2024

Plan du Chapitre II

Les deux cadres d'apprentissage principaux

Apprentissage supervisé :

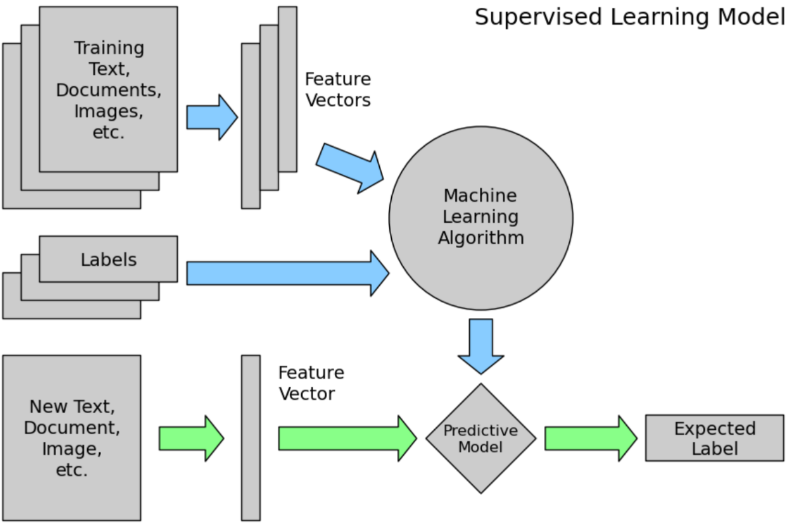
- **données** : observations $\{(x_i, y_i)\}$, $i = 1, \ldots, n$.
  - descripteurs / variables explicatives + variable d'intérêt
- **objectif(s)** : prédiction
  - (+ compréhension du lien entre X et Y)

Apprentissage non-supervisé :

- **données** : observations $\{x_i\}$, $i = 1, \ldots, n$.
  - pas de variable à expliquer
- **objectif** : identifier des "structures" dans les données
  - moins clairement formalisé que le cadre supervisé

Apprentissage supervisé

Apprentissage supervisé - principe



Terminologie de l'apprentissage supervisé

- Variables et étiquettes
- Observations et jeux de données
- Terminologie alternative

Terminologie de l'apprentissage supervisé

Le machine learning, étant issu de plusieurs disciplines et champs d'application, utilise des termes variés pour désigner les mêmes concepts. Voici quelques correspondances courantes :

- **Variables :**
  - Aussi appelées : descripteurs, attributs, prédicteurs, ou caractéristiques.
  - En anglais : variables, descriptors, attributes, predictors, or features.
- **Observations :**
  - Aussi appelées : exemples, échantillons, ou points du jeu de données.
  - En anglais : samples or data points.
- **Étiquettes :**
  - Aussi appelées : variables cibles.
  - En anglais : labels, targets, or outcomes.

Exemple de terminologie

Variables (Features)

- Exemple : Dans un modèle de prédiction du prix d'une maison, les variables peuvent être :
  - La surface de la maison (en m²)
  - Le nombre de chambres
  - La localisation géographique
  - L'année de construction

Exemple de terminologie

Observations (Samples)

- Exemple : Chaque observation représente une maison individuelle :
  - Surface : 100 m²
  - Chambres : 3
  - Localisation : Paris
  - Année de construction : 1995

Étiquettes (Labels)

- Exemple : Le prix de vente d'une maison est l'étiquette à prédire :
  - Maison : 100 m², 3 chambres, Paris
  - Étiquette (prix de vente) : 350 000 €

Définition de l'Apprentissage Supervisé

Définition : L'apprentissage supervisé consiste à apprendre une fonction f qui associe un espace d'entrées X à un espace de sorties Y.

- X : l'ensemble des entrées (exemples, observations)
- Y : l'ensemble des sorties (étiquettes, valeurs à prédire)

Objectif : Trouver une fonction $f : X \to Y$ telle que $f(\mathbf{x}) \approx y$, en utilisant un ensemble d'apprentissage :

Données : On dispose d'un ensemble d'entraînement composé de n exemples :

$$D = \{(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \ldots, (\mathbf{x}_n, y_n)\}$$

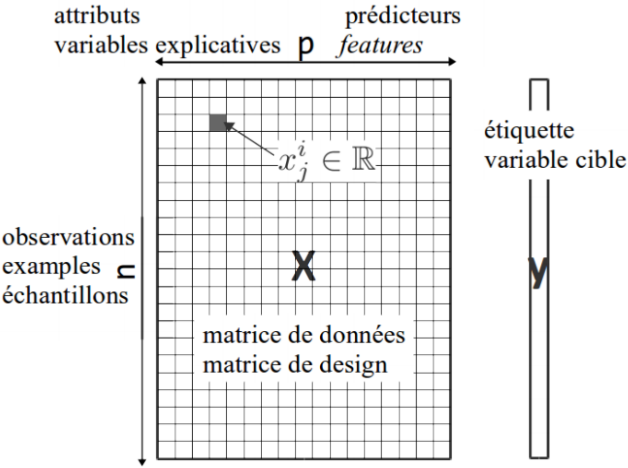
Représentation des Données

- **Ensemble des observations :** $X = \mathbb{R}^p$ est constitué de vecteurs à p dimensions.
    - p : Nombre de caractéristiques (ou variables).
- **Matrice de données :** $X \in \mathbb{R}^{n \times p}$
    - n : Nombre d'observations (ou échantillons).
    - X_{ij} : Représente la valeur de la j-ème caractéristique de la i-ème observation.

Par exemple, pour $n = 3$ et $p = 2$, la matrice de données pourrait ressembler à :

$$X = \begin{pmatrix} x_{11} & x_{12} \\ x_{21} & x_{22} \\ x_{31} & x_{32} \end{pmatrix}$$

Exemple de Représentation des Données



Exemple : Matrice de données

Exemple : Prédiction du prix de vente d'une maison

Observation	Surface (m²)	Chambres	Localisation	Prix (label)
1	100	3	Paris	350 000 €
2	80	2	Lyon	220 000 €
3	120	4	Marseille	280 000 €
4	95	3	Lille	260 000 €

Table: Matrice de données avec observations, variables et étiquettes (label)

$$\begin{bmatrix} 100 & 3 & 1 & 350000 \\ 80 & 2 & 2 & 220000 \\ 120 & 4 & 3 & 280000 \\ 95 & 3 & 4 & 260000 \end{bmatrix}$$

Exercice : Création d'une Matrice de Données

- **Caractéristiques des objets** :
  - Couleur : Rouge, Vert, Jaune
  - Taille : Petite, Moyenne, Grande
  - Poids : en grammes (g)
- **Liste d'observations** :
  - Fruit 1 : Rouge, petite, 150g
  - Fruit 2 : Jaune, grande, 300g
  - Fruit 3 : Vert, moyenne, 200g
  - Fruit 4 : Rouge, moyenne, 180g
  - Fruit 5 : Vert, grande, 250g
- **Tâche** :
  - Créez une matrice de données avec chaque ligne représentant un fruit.
  - Encodez les couleurs et tailles numériquement :
    - Couleur : Rouge = 1, Vert = 2, Jaune = 3
    - Taille : Petite = 1, Moyenne = 2, Grande = 3

Correction : Matrice de Données

Fruit	Couleur (1=R, 2=V, 3=J)	Taille (1=P, 2=Mo, 3=G)	Poids (g)
Fruit 1	1	1	150
Fruit 2	3	3	300
Fruit 3	2	2	200
Fruit 4	1	2	180
Fruit 5	2	3	250

Voici la matrice représentant les caractéristiques des fruits :

$$\begin{bmatrix} 1 & 1 & 150 \\ 3 & 3 & 300 \\ 2 & 2 & 200 \\ 1 & 2 & 180 \\ 2 & 3 & 250 \end{bmatrix}$$

Apprentissage supervisé - formalisation

On dispose d'un échantillon $\{(x_i, y_i)\}, i = 1, \ldots, n$:

- des **observations** $x_i \in \mathcal{X}$,
- des **réponses associées** $y_i \in \mathcal{Y}$.

Typiquement :

- $\mathcal{X} = \mathbb{R}^p$: on parle de *vecteurs de descripteurs* (*features, attributes, input variables*),
- Si $\mathcal{Y} = \mathbb{R}$, on parle de *régression*,
- Si $\mathcal{Y} = \{1, \ldots, K\}$, on parle de *classification*,
- Si $\mathcal{Y} = \{0, 1\}$, on parle de *classification binaire*

Classification Binaire

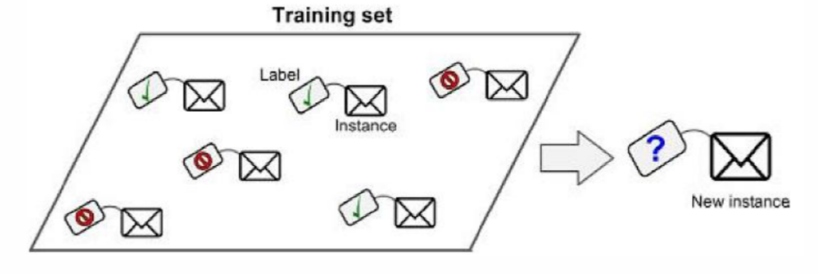
Définition 1.2 : Un problème d'apprentissage supervisé dans lequel l'espace des étiquettes est binaire, autrement dit $Y = \{0, 1\}$, est appelé un problème de classification binaire.

Exemples :

- Identifier si un email est un spam ou non.
- Identifier si un tableau a été peint par Picasso ou non.
- Identifier si une image contient ou non une girafe.
- Identifier si une molécule peut ou non traiter la dépression.
- Identifier si une transaction financière est frauduleuse ou non.

Classification - Illustration

Classification :



- $\mathcal{X} = \{\text{e-mails}\}$
- $\mathcal{Y} = \{0, 1\}$ (ici : spams / non-spams)

Classification Multi-classe

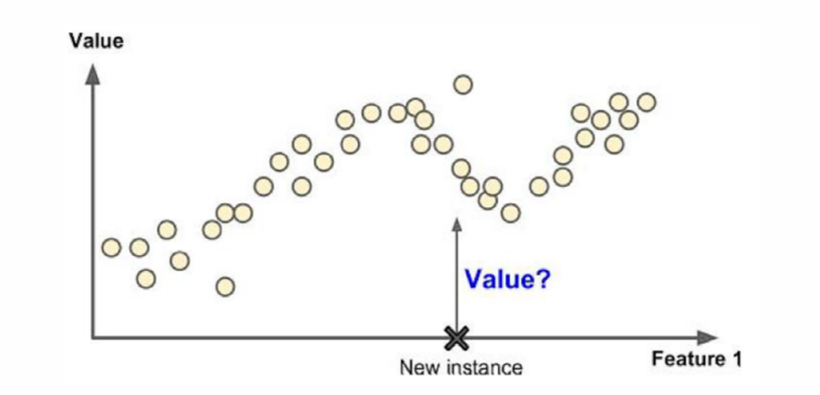
Définition 1.3 : Un problème d'apprentissage supervisé dans lequel l'espace des étiquettes est discret et fini, autrement dit $Y = \{1, 2, \ldots, C\}$ avec C étant le nombre de classes, est appelé un problème de classification multi-classe.

Exemples :

- Identifier en quelle langue un texte est écrit.
- Identifier lequel des 10 chiffres arabes est un chiffre manuscrit.
- Identifier l'expression d'un visage parmi une liste prédéfinie de possibilités (colère, tristesse, joie, etc.).
- Identifier à quelle espèce appartient une plante.
- Identifier les objets présents sur une photographie.

Régression - Illustration

Régression :



- $\mathcal{X} = \{\mathbb{R}\}$ (Feature 1)
- $\mathcal{Y} = \{1, \ldots, K\}$ (Value)

Exercice

Pour chaque scénario ci-dessous, identifiez s'il s'agit d'un problème de : de régression ou un problème de classification Binaire ou bien Multi-Classe :

1. Prédire la température d'une ville demain
2. Classer des photos en fonction de l'animal : chien, chat ou oiseau
3. Détecter si un email est un spam ou non
4. Estimer le prix de vente d'une maison
5. Prédire si un patient souffre de diabète (oui ou non)
6. Classer des images de chiffres manuscrits (de 0 à 9)

Correction

- **1. Prédire la température d'une ville demain** Type de problème : **Régression**
  On prédit une valeur continue (la température en degrés).
- **2. Classer des photos en fonction de l'animal : chien, chat ou oiseau** Type de problème : **Classification Multi-Classe**
  Il y a plusieurs classes possibles (chien, chat, oiseau).
- **3. Détecter si un email est un spam ou non** Type de problème : **Classification Binaire**
  Deux classes possibles : spam ou non-spam.

Correction

- **4. Estimer le prix de vente d'une maison** Type de problème : **Régression**
  On prédit une valeur continue (le prix en euros).
- **5. Prédire si un patient souffre de diabète (oui ou non)** Type de problème : **Classification Binaire**
  Deux classes : diabétique ou non-diabétique.
- **6. Classer des images de chiffres manuscrits (de 0 à 9)** Type de problème : **Classification Multi-Classe**
  Dix classes possibles (chiffres de 0 à 9).

Apprentissage supervisé - formalisation

Données d'entrée : échantillon $\{(x_i, y_i)\}_{i=1,\ldots,n} \in \mathcal{X} \times \mathcal{Y}$.

Objectif : apprendre une fonction $f : \mathcal{X} \to \mathcal{Y}$ permettant de **prédire** la réponse associée à une **nouvelle observation**.

Objectif :

- Apprendre une fonction $f : \mathcal{X} \to \mathcal{Y}$ qui minimise l'erreur de prédiction.
- L'erreur est mesurée à l'aide d'une fonction de perte $L(f(x), y)$.

Critère : une **fonction de perte** L (pour "*loss*") mesurant l'erreur de prédiction entre y et $f(x)$.

Qu'est-ce que le Risque Empirique ?

Définition : Le risque empirique est la moyenne des erreurs de prédiction d'un modèle sur un ensemble d'apprentissage.

- Représente la performance du modèle sur les données dont il a appris.
- Formule :

$$R_{\text{emp}}(f) = \frac{1}{N} \sum_{i=1}^{N} L(f(x_i), y_i)$$

où L est la fonction de perte, (x_i, y_i) sont les exemples, et N est le nombre total d'exemples.

Rôle du Risque Empirique

Importance dans le Processus d'Apprentissage :

- Mesure la capacité du modèle à faire des prédictions sur les données d'entraînement.
- Utilisé pour ajuster les paramètres du modèle afin de réduire l'erreur.
- Permet de comparer différents modèles.

Risque Empirique

- Ce concept est fondamental en apprentissage supervisé, car il définit notre objectif
- apprendre une fonction qui fait de bonnes prédictions.

Exemples de **Fonctions de perte :**

- Régression : **Erreur quadratique**

$$L(f(x), y) = (f(x) - y)^2$$

- Classification : **Entropie croisée**

$$L(f(x), y) = -y \log(f(x)) - (1 - y) \log(1 - f(x))$$

Erreur Quadratique

- L'erreur quadratique est une fonction de perte fréquemment utilisée dans les problèmes de **régression** supervisée.
- Elle mesure l'écart entre les prédictions du modèle et les valeurs réelles, en mettant d'avantage l'accent sur les erreurs plus importantes.

Définition de l'Erreur Quadratique

L'erreur quadratique est définie comme :

$$L(y, f(x)) = (y - f(x))^2$$

- y : la valeur réelle ou vraie (étiquette).
- $f(x)$: la prédiction faite par le modèle pour l'entrée x.
- $L(y, f(x))$: la fonction de perte qui quantifie l'erreur en calculant le carré de la différence entre y et $f(x)$.

Interprétation de l'Erreur Quadratique

- La différence entre la valeur réelle y et la prédiction $f(x)$ est appelée **erreur de prédiction**.
- En élevant cette différence au carré, l'erreur quadratique donne plus de poids aux **grandes erreurs**.
- Si $f(x)$ est très proche de y, l'erreur quadratique est petite. Si $f(x)$ est très éloigné de y, l'erreur est grande.

Utilisation dans l'Apprentissage Supervisé

Dans un cadre d'apprentissage supervisé, l'objectif est de minimiser la fonction de perte sur l'ensemble des données d'apprentissage.

Supposons que l'on dispose d'un ensemble d'apprentissage $D = \{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\}$, où n est le nombre d'observations.

$$MSE = \frac{1}{n} \sum_{i=1}^{n} (y_i - f(x_i))^2$$

Le modèle ajuste la fonction f pour minimiser cette erreur sur l'ensemble des données.

Exemple Pratique

Supposons que vous développiez un modèle pour prédire la valeur d'une maison en fonction de plusieurs caractéristiques (superficie, nombre de chambres, etc.).

Si la vraie valeur y_i est de 300 000 € et que votre modèle prédit $f(x_i) = 290000$ €, l'erreur quadratique serait :

$$L(300000, 290000) = (300000 - 290000)^2 = 100000000$$

L'objectif de l'apprentissage est de minimiser cette erreur pour toutes les maisons de l'ensemble.

Exemple de Calcul du Risque Empirique

Illustration :

- Considérons un ensemble d'apprentissage avec 3 points : $(x_1, y_1), (x_2, y_2), (x_3, y_3)$
- Fonction de perte : Erreur quadratique.
- Calcul du risque empirique :

$$R_{\text{emp}}(f) = \frac{1}{3}\left(L(f(x_1), y_1) + L(f(x_2), y_2) + L(f(x_3), y_3)\right)$$

Entropie croisée

L'entropie croisée est une mesure de la différence entre deux distributions de probabilité :

- La distribution réelle des classes (étiquettes)
- La distribution prédites par le modèle

Elle quantifie combien d'information est nécessaire pour décrire la distribution réelle en utilisant la distribution prédite.

Formule - Classification Binaire

Pour une classification binaire, l'entropie croisée $H(p, q)$ est donnée par :

$$H(p, q) = -\left(p \log(q) + (1 - p) \log(1 - q)\right) \tag{1}$$

où :

- p : Probabilité réelle de la classe (1 si l'échantillon appartient à la classe positive, 0 sinon)
- q : Probabilité prédite par le modèle pour la classe positive

Formule - Classification Multi-Classe

Pour une classification multi-classe, l'entropie croisée devient :

$$H(p,q) = -\sum_{i=1}^{C} p_i \log(q_i) \tag{2}$$

où :

- C : Nombre total de classes
- p_i : Probabilité réelle de la classe i
- q_i : Probabilité prédite pour la classe i

Exemple

Considérons un modèle de classification binaire :

- Vraie étiquette : $p = 1$ (classe positive)
- Probabilité prédite : $q = 0.9$

Calculons l'entropie croisée :

$$\begin{aligned} H(p, q) &= -\left(1 \log(0.9) + (1-1)\log(1-0.9)\right) \\ &= -\log(0.9) \approx 0.1054 \end{aligned}$$

Pour $p = 0$ (classe négative) et $q = 0.1$:

$$\begin{aligned} H(p, q) &= -\left(0 \log(0.1) + (1-0)\log(1-0.1)\right) \\ &= -\log(0.9) \approx 0.1054 \end{aligned}$$

Avantages

- **Différenciabilité** : Fonction différentiable, adaptée à l'optimisation par des algorithmes de gradient.
- **Interprétabilité** : Une valeur nulle indique des prédictions parfaites.

Conclusion

L'entropie croisée est essentielle en apprentissage supervisé pour la classification, permettant d'évaluer et d'optimiser la performance des modèles.

Énoncé de l'exercice

Nous cherchons à prédire si un utilisateur est abonné ($y = 1$) ou non abonné ($y = 0$) à un service.
L'ensemble de données d'entraînement contient $N = 4$ observations avec les prédictions probabilistes q_i fournies par le modèle et les étiquettes réelles p_i.

i	p_i (réel)	q_i (prédiction probabiliste)
1	1	0.8
2	0	0.3
3	1	0.6
4	0	0.2

Question 1 :
Calculez la fonction d'entropie croisée pour cet ensemble de données.
Question 2 :
Calculez le risque empirique du modèle créé.

Correction

1.Fonction d'entropie croisée

La fonction d'entropie croisée pour une observation est définie comme :

$$L(p_i, q_i) = -\left(p_i \log(q_i) + (1 - p_i)\log(1 - q_i)\right)$$

Le risque empirique pour un ensemble de N observations est :

$$R_{\text{emp}}(h) = \frac{1}{N}\sum_{i=1}^{N} L(p_i, q_i)$$

Correction

2.Calcul du risque empirique

Calculons $L(p_i, q_i)$ pour chaque observation :

- Pour $i = 1$: $L(1, 0.8) = -\log(0.8) = 0.223$
- Pour $i = 2$: $L(0, 0.3) = -\log(0.7) = 0.357$
- Pour $i = 3$: $L(1, 0.6) = -\log(0.6) = 0.511$
- Pour $i = 4$: $L(0, 0.2) = -\log(0.8) = 0.223$

Correction

Résultat du calcul

Le risque empirique est alors donné par :

$$R_{\text{emp}}(h) = \frac{1}{4} \times (0.223 + 0.357 + 0.511 + 0.223) = 0.3285$$

Énoncé de l'exercice 2

On cherche à prédire la valeur d'une variable cible continue y à partir d'un ensemble d'observations.
On dispose de $N = 4$ observations dans l'ensemble de données d'entraînement avec les prédictions q_i données par le modèle et les valeurs réelles p_i.

i	p_i (réel)	q_i (prédiction)
1	2.5	2.7
2	0.0	-0.1
3	3.6	3.4
4	1.2	1.5

Question 1 :
Calculez la fonction d'erreur quadratique pour cet ensemble de données.
Question 2 :
Calculez le risque empirique a modèle créé.

Erreur quadratique

La fonction d'erreur quadratique pour une observation est définie comme :

$$L(p_i, q_i) = (p_i - q_i)^2$$

Le risque empirique pour un ensemble de N observations est donné par :

$$R_{\text{emp}}(h) = \frac{1}{N} \sum_{i=1}^{N} L(p_i, q_i)$$

Calcul du risque empirique

Calculons $L(p_i, q_i)$ pour chaque observation :

- Pour $i = 1$: $L(2.5, 2.7) = (2.5 - 2.7)^2 = 0.04$
- Pour $i = 2$: $L(0.0, -0.1) = (0.0 + 0.1)^2 = 0.01$
- Pour $i = 3$: $L(3.6, 3.4) = (3.6 - 3.4)^2 = 0.04$
- Pour $i = 4$: $L(1.2, 1.5) = (1.2 - 1.5)^2 = 0.09$

Résultat du calcul

Le risque empirique est alors donné par :

$$R_{\text{emp}}(h) = \frac{1}{4} \times (0.04 + 0.01 + 0.04 + 0.09) = \frac{1}{4} \times 0.09 = 0.045$$

Le risque empirique pour cet ensemble de données est $R_{\text{emp}}(h) = 0.045$.

Minimisation du Risque Empirique (ERM)

- La **minimisation du risque empirique** (ERM) est une méthode clé en apprentissage supervisé.
- L'objectif est de trouver la fonction f dans une classe de fonctions $\mathcal{F}$ qui minimise le risque empirique sur les données d'entraînement.
- Cela permet de choisir un modèle qui s'ajuste bien aux données observées.

Formule du Risque Empirique

- Le risque empirique $R_{\text{emp}}(f)$ est défini par la formule suivante :

$$R_{\text{emp}}(f) = \frac{1}{N} \sum_{i=1}^{N} L(f(x_i), y_i)$$

où :

  - L est la fonction de perte,
  - (x_i, y_i) sont les exemples d'entraînement,
  - N est le nombre total d'exemples.

Problème d'Optimisation Associé

- L'objectif d'ERM est de résoudre le problème d'optimisation suivant :

$$f_{\mathcal{F}} = \arg\min_{f \in \mathcal{F}} R_{\text{emp}}(f)$$

- Cela implique de rechercher la fonction f qui minimise le risque empirique, permettant ainsi de généraliser les performances du modèle sur des données non observées.
- Il est important de prendre en compte des techniques de régularisation pour éviter le sur-apprentissage.