Chapitre I
Introduction à l'apprentissage supervisé

Cycle pluridisciplinaire d'études supérieures
Université Paris sciences et lettres

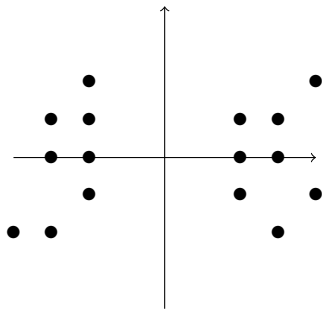
Nous introduisons dans ce chapitre le cadre et le vocabulaire de l'apprentissage supervisé. Donnons de ce dernier une première présentation informelle : il s'agit d'*apprendre* à *prédire*. La tâche de prédiction en question consiste à prédire une variable de sortie en fonction d'une variable d'entrée. L'*apprentissage* se fait sur la base d'exemples donnés, qui sont des couples constitués d'une valeur d'entrée et d'une valeur de sortie (aussi appelée étiquette). Dans le tableau suivant, les quatre premières lignes correspondent aux exemples d'apprentissage, et la dernière ligne à un exemple à prédire.

Entrée	Sortie
1	2
4	8
50	100
10	20
3	?

Ici, la fonction de prédiction que l'on peut *apprendre* à partir des exemples d'apprentissage, et à laquelle chacun pense, est la fonction $f(x) = 2x$, qui donne donc la prédiction 6 pour l'exemple à prédire.

En apprentissage non-supervisé en revanche, les exemples d'apprentissage ne possèdent pas de variable de sortie. Par exemple, on peut avoir pour exemples

d'apprentissage un ensemble de points dans $\mathbb{R}^2$, et avoir pour but de classer ces points en deux groupes, nommés 1 et 2.



Un fonction de prédiction apprise pourrait alors être :

$$f(x_1, x_2) = \begin{cases} 1 & \text{si } x_1 < 0 \\ 2 & \text{si } x_1 \geqslant 0. \end{cases}$$

L'apprentissage non-supervisé ne sera pas abordé dans ce cours.

I. CADRE

On se donne $\mathcal{X}$ un ensemble d'*entrées* (aussi appelées *input*, *variables explicatives*, ou encore *feature vector* lorsque $\mathcal{X} = \mathbb{R}^d$), et $\mathcal{Y}$ un ensemble de *sorties* (aussi appelées *étiquette* ou *label*).

Lorsque $\mathcal{X}$ est défini par un produit cartésien (par exemple $\mathcal{X} = \mathbb{R} \times \{a, b, c\} \times \mathbb{N}$), les composantes des entrées sont appelées *variables explicatives*. Une variable explicative est dite *quantitative* lorsqu'il s'agit d'un nombre, et est dite *catégorielle* si elle prend ses valeurs dans un ensemble fini dont les éléments ne sont pas interprétés comme des quantités.

Lorsque $\mathcal{Y} = \mathbb{R}$ (resp. $\mathcal{Y}$ est un ensemble fini), on dit qu'il s'agit d'un cadre de *régression* (resp. de *classification*).

DÉFINITION. — *Échantillon.* — On appelle *échantillon* une famille de la forme

$$S = ((x_1, y_1), \dots, (x_n, y_n)) \in (\mathcal{X} \times \mathcal{Y})^n,$$

où $n \geqslant 1$ est un entier. On pourra écrire de façon abrégée : $S = (x_i, y_i)_{i \in [n]}$. On note $\mathcal{S}(\mathcal{X}, \mathcal{Y}) = \bigcup_{n \geqslant 1} (\mathcal{X} \times \mathcal{Y})^n$ l'ensemble de tous les échantillons.

Soit $S_{\text{train}} = (x_i, y_i)_{i\in[n]}$ un échantillon dit *d'apprentissage*. On suppose que les $(x_i, y_i)_{i\in[n]}$ sont des variables aléatoires i.i.d. suivant une distribution (inconnue) P sur $\mathcal{X} \times \mathcal{Y}$.

DÉFINITION. — *Fonction de prédiction.* — Une fonction $f : \mathcal{X} \to \mathcal{Y}$ est appelée *fonction de prédiction* (ou *prédicteur*, *estimateur*, *modèle* ou encore *hypothèse*). On note $\mathcal{F}(\mathcal{X}, \mathcal{Y})$ l'ensemble des prédicteurs. Si $\mathcal{Y} = \mathbb{R}$ (resp. si $\mathcal{Y}$ est fini), f est aussi appelé *régresseur* (resp. *classifieur*).

La démarche de l'apprentissage supervisé consiste à utiliser l'échantillon d'apprentissage S_{train} pour construire un bon prédicteur, et ce, à l'aide d'un algorithme d'apprentissage.

DÉFINITION. — *Algorithme d'apprentissage.* — Un *algorithme d'apprentissage* est une application

$$\begin{array}{ccc} \mathcal{S}(\mathcal{X}, \mathcal{Y}) & \longrightarrow & \mathcal{F}(\mathcal{X}, \mathcal{Y}) \\ S_{\text{train}} & \longmapsto & \hat{f}. \end{array}$$

On note $\mathcal{A}(\mathcal{X}, \mathcal{Y})$ l'ensemble des algorithmes d'apprentissage.

EXEMPLE. — *Prédiction du prix des appartements.* — On considère des données se présentant sous la forme suivante où les quatre premières colonnes correspondent à l'entrée (variables explicatives) et la dernière colonne à la sortie (variable à prédire).

SURFACE	NB. PIÈCES	ARRONDISSE$^{\text{T}}$	ÉTAGE	PRIX
92	4	14	9	910000
60	3	15	3	630000
⋮	⋮	⋮	⋮	⋮
124	5	7	5	?

L'ensemble d'entrées peut s'écrire :

$$\mathcal{X} = \mathbb{R} \times \mathbb{N}^* \times \{1, \dots, 20\} \times \mathbb{N}.$$

Parmi ces 4 variables explicatives, la troisième (l'arrondissement) est une variable dite *catégorielle*, c'est-à-dire qu'elle prend ses valeurs dans un ensemble fini, tandis que les trois autres sont des variables dites *quantitatives*, c'est-à-dire qu'il s'agit de valeurs numériques [1]. L'ensemble de sorties peut s'écrire $\mathcal{Y} = \mathbb{R}$. Il s'agit d'une variable à prédire quantitative, c'est donc un cadre de régression.

1. Il se trouve que l'arrondissement est également un nombre, mais sa valeur ne s'interprète pas comme une quantité.

Exemple. — *Prédiction des résultats de football.* — On considère des données dont toutes les variables sont catégorielles. C'est en particulier le cas de la variable à prédire, il s'agit donc d'un cadre de classification.

Éq. Domicile	Éq. Extérieur	Résultat éq. domicile
PSG	OM	Victoire
OM	Dijon	Nul
Lyon	PSG	Défaite
⋮	⋮	⋮
PSG	Dijon	?

On a donc les ensembles d'entrées et de sortie qui s'écrivent $\mathcal{X} = \{\text{PSG, OM, ...}\}^2$ et $\mathcal{Y} = \{\text{Victoire, Nul, Défaite}\}$.

2. Mesure de la qualité d'un prédicteur : risque et risque empirique

Informellement, la qualité d'un prédicteur est sa capacité à prédire correctement sur de nouvelles données, c'est-à-dire des exemples n'appartenant pas à l'échantillon d'apprentissage.

Donnons rapidement quelques rappels de probabilités. Soit $n \geqslant 1$ un entier et $Z_1, \dots, Z_n$ des variables aléatoires réelles i.i.d. d'espérance $\mu \in \mathbb{R}$. Alors, on définit comme suit la moyenne (dite *empirique*) des $(Z_i)_{1 \leqslant i \leqslant n}$:

$$\bar{Z}_n = \frac{1}{n} \sum_{i=1}^{n} Z_i.$$

La moyenne empirique est un estimateur de μ dans le sens où elle converge vers μ presque-sûrement (lorsque $n \to +\infty$) en vertu de la loi forte des grands nombres.

On se donne les objets suivants :

- $S_{\text{train}} = (x_i, y_i)_{i \in [n]} \in \mathcal{S}(\mathcal{X}, \mathcal{Y})$ un échantillon d'apprentissage dont les exemples (x_i, y_i) sont tirés de façon i.i.d. selon une loi inconnue P,
- $S_{\text{test}} = (x'_i, y'_i)_{i \in [n']} \in \mathcal{S}(\mathcal{X}, \mathcal{Y})$ un échantillon dit *de test*, dont les exemples sont tirés de façon i.i.d. selon la même loi P, et indépendamment de S_{train},
- $\hat{f} \in \mathcal{F}(\mathcal{X}, \mathcal{Y})$ un prédicteur construit à l'aide de S_{train},
- $\ell \colon \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$ une *fonction de perte* ($\ell(y, y')$ mesure la *perte* encourue lorsqu'on prédit y' alors que la sortie réelle est y).

Exemple. — *Fonctions de perte.* — 1) En régression, la fonction

$$\ell(y, y') = (y - y')^2$$

est appelée *perte quadratique* (ou encore *perte des moindres carrés*).

2) En classification *binaire* (autrement dit lorsque $\mathrm{Card}\, \mathrm{Y} = 2$), la fonction :

$$\ell(y, y') = \begin{cases} 0 & \text{si } y = y' \\ 1 & \text{si } y \neq y' \end{cases}$$

est appelée *perte 0–1*.

Définition. — *Risque.* — Le risque d'un prédicteur $\widehat{f} \in \mathscr{F}(\mathscr{X}, \mathscr{Y})$ est défini par :

$$\mathrm{R}(\widehat{f}) = \mathbb{E}_{(\mathrm{X},\mathrm{Y})\sim \mathrm{P}} \left[\ell(\mathrm{Y}, \widehat{f}(\mathrm{X})) \right].$$

Le risque est une mesure de la qualité du prédicteur. La loi P étant inconnue, on ne peut pas calculer $\mathrm{R}(\widehat{f})$ directement.

Définition. — *Erreurs d'apprentissage et de test.* — (i) *L'erreur d'apprentissage* d'un prédicteur $\widehat{f}$ est définie par :

$$\varepsilon_{\text{train}} = \frac{1}{n} \sum_{i=1}^{n} \ell(y_i, \widehat{f}(x_i)).$$

Autrement dit, il s'agit du risque empirique sur l'échantillon d'apprentissage $\mathrm{S}_{\text{train}}$.

(ii) *L'erreur de test* d'un prédicteur $\widehat{f}$ est définie par :

$$\varepsilon_{\text{test}} = \frac{1}{n} \sum_{i=1}^{n'} \ell(y'_i, \widehat{f}(x'_i)).$$

Autrement dit, il s'agit du risque empirique sur l'échantillon de test S_{test}.

L'erreur d'apprentissage $\varepsilon_{\text{train}}$ n'est pas un bon estimateur du risque $\mathrm{R}(\widehat{f})$; en effet, l'estimateur $\widehat{f}$ a été construit à l'aide de l'échantillon d'apprentissage $\mathrm{S}_{\text{train}}$. En revanche, l'erreur de test $\varepsilon_{\text{test}}$ est bien un estimateur du risque. L'erreur d'apprentissage est parfois beaucoup plus petit que l'erreur de test. On parle alors de *sur-apprentissage* (*overfitting* en anglais) ; l'intuition correspondante étant qu'il

est beaucoup plus facile de bien prédire sur des données déjà observées (celles de S_{train}) que sur des données jamais observées (celles de S_{test}).

Pour un même problème, on considère parfois différentes fonctions de perte. Une fonction de perte ℓ' différente de celle initialement considérée est appelée une *métrique*, et un risque empirique associé est appelé un *score*.

3. Exemple d'algorithme : minimisation du risque empirique

On présente ici la minimisation du risque empirique, qui est un exemple de base d'algorithme d'apprentissage.

Notation. — Soit $f : A \to \mathbb{R}$ une fonction. Si $f(a_*) = \min_{a \in A} f(a)$, a_* est appelé un *minimiseur* de f. On note $\operatorname{Arg\,min}_{a \in A} f(a)$ l'ensemble des minimiseur de f. On note $\arg\min_{a \in A} f(a)$ un minimiseur quelconque de f (s'il en existe).

Définition. — *Minimisation du risque empirique.* — Soit un ensemble de prédicteurs $\mathcal{F} \subset \mathcal{F}(\mathcal{X}, \mathcal{Y})$, appelée *classe* de prédicteurs, et $S_{train} = (x_i, y_i)_{i \in [n]} \in \mathcal{S}(\mathcal{X}, \mathcal{Y})$ un échantillon d'apprentissage. Le prédicteur donné par *minimisation du risque empirique* (ERM) est défini par :

$$\hat{f} = \arg\min_{f \in \mathcal{F}} \left\{ \frac{1}{n} \sum_{i=1}^{n} \ell(y_i, f(x_i)) \right\}.$$

L'algorithme d'apprentissage correspondant est donc l'application :

$$(x_i, y_i)_{i \in [n]} \mapsto \arg\min_{f \in \mathcal{F}} \left\{ \frac{1}{n} \sum_{i=1}^{n} \ell(y_i, f(x_i)) \right\}.$$

Nous verrons dans les chapitres suivants qu'un grand nombre d'algorithmes importants sont des variantes de la minimisation du risque empirique.