Chapitre 2 : Modèle Supervisé

24 février 2026

Plan du chapitre

Principe de l'apprentissage supervisé

But : Prédire l'étiquette de nouveaux exemples à partir d'une base d'entraînement étiquetée (ex : images de chats/chiens, ECG sains/malades).

Les Ingrédients

- **Espace des caractéristiques** $\mathscr{X}$ **:** Représentation de l'objet (souvent un vecteur $\mathbf{x} \in \mathbb{R}^p$). *Synonymes : attributs, descripteurs, features.*
- **Espace des étiquettes** $\mathscr{Y}$ **:** La variable cible à prédire. *Synonymes : labels, targets.*
  - Discret (fini) $\rightarrow$ **Classification** (ex : $\{-1, +1\}$).
  - Continu $\rightarrow$ **Régression**.

Attention : La sélection des "bonnes" caractéristiques est cruciale. L'algorithme ne voit **que** ce vecteur $\mathbf{x}$, pas l'objet réel.

Le Prédicteur optimal de Bayes

Si l'on connaissait parfaitement la loi de distribution des données, l'apprentissage serait inutile ! On utiliserait le **prédicteur de Bayes**.

Principe

Il choisit l'étiquette dont la **probabilité a posteriori** est maximale.
Ce choix minimise mathématiquement la probabilité d'erreur.

La réalité du Machine Learning

Les lois de probabilité (la distribution $\mathbb{P}$) sont **inconnues**.
La seule chose connue est en ensemble de données d'apprentissage.
Le ML consiste à estimer cette règle optimale à partir des données.

Optimalité : Le Risque $0/1$

Soit $X \in \mathbb{R}^d$ et $Y \in \{-1, +1\}$.
Le risque $0/1$ (probabilité d'erreur) d'un classifieur h est :

$$\mathscr{R}(h) = \mathbb{P}(h(X) \neq Y) = \mathbb{E}[\mathbb{1}_{h(X) \neq Y}]$$

Si h et Y sont à valeurs dans $\{-1, 1\}$, il est proportionnel à l'erreur quadratique :

$$\mathscr{R}(h) = \frac{1}{4}\mathbb{E}\left[(Y - h(X))^2\right]$$

Optimalité : Définition du prédicteur de Bayes

Définissons la probabilité conditionnelle a posteriori :

$$\eta(x) = \mathbb{P}(Y = +1 \mid X = x)$$

Le prédicteur de Bayes h^* assigne la classe la plus probable :

$$h^*(x) = \begin{cases} +1 & \text{si } \eta(x) \geq 1/2 \\ -1 & \text{si } \eta(x) < 1/2 \end{cases}$$

Théorème : h^* minimise le risque $\mathscr{R}(h)$.

Éléments de preuve : Le risque conditionnel en un point x est :

$$r(h, x) = \mathbb{P}(h(x) \neq Y \mid X = x)$$

$$r(h, x) = \mathbb{1}_{h(x)=+1}(1 - \eta(x)) + \mathbb{1}_{h(x)=-1}\eta(x)$$

Optimalité : Calcul de l'excès de risque

Comparons un classifieur quelconque h avec h^* au point x :

$$D(x) = r(h, x) - r(h^*, x)$$

- Si $h(x) = h^*(x)$, alors $D(x) = 0$.
- Si $h^*(x) = +1$ et $h(x) = -1$ (donc $\eta(x) \geq 1/2$) :
  $D(x) = \eta(x) - (1 - \eta(x)) = 2\eta(x) - 1 \geq 0$.
- Si $h^*(x) = -1$ et $h(x) = +1$ (donc $\eta(x) < 1/2$) :
  $D(x) = (1 - \eta(x)) - \eta(x) = 1 - 2\eta(x) > 0$.

Dans tous les cas :

$$r(h, x) - r(h^*, x) = |2\eta(x) - 1| \cdot \mathbb{1}_{h(x) \neq h^*(x)} \geq 0$$

Lien entre Classification et Régression (1/2)

Comment relier l'erreur à la fonction de régression $f^*(x) = \mathbb{E}[Y \mid X = x]$?
En utilisant $\mathbb{1}_{h(x) \neq y} = \frac{1 - y \cdot h(x)}{2}$, le risque devient :

$$\mathscr{R}(h) = \frac{1}{2} - \frac{1}{2}\mathbb{E}[Y \cdot h(X)]$$

Minimiser le risque revient à **maximiser** $\mathbb{E}[Y \cdot h(X)]$.
En conditionnant par X :

$$\mathbb{E}[Y \cdot h(X) \mid X] = h(X) \cdot \mathbb{E}[Y \mid X] = h(X) \cdot f^*(X)$$

Pour maximiser $h(x) \cdot f^*(x)$ en chaque point x, il faut que $h(x)$ ait le même signe que $f^*(x)$. D'où le classifieur optimal :

$$h^\star(x) = \text{sign}(f^\star(x))$$

Lien entre Classification et Régression (2/2)

Preuve par l'excès de risque :

$$\mathscr{R}(h) - \mathscr{R}(h^\star) = \frac{1}{2}\mathbb{E}_X\Big[f^\star(X)h^\star(X) - f^\star(X)h(X)\Big]$$

Comme $h^\star(X) = \text{sign}(f^\star(X))$, on a $f^\star(X)h^*(X) = |f^*(X)|$:

$$\mathscr{R}(h) - \mathscr{R}(h^\star) = \frac{1}{2}\mathbb{E}_X\Big[\underbrace{|f^\star(X)| - h(X)f^\star(X)}_{\Delta(X)}\Big]$$

Pour tout réel a et signe s, on a $|a| \geq s \cdot a$. Donc $\Delta(X) \geq 0$.
L'espérance d'une variable positive est positive :

$$\mathscr{R}(h) \geq \mathscr{R}(h^\star)$$

Risque

La qualité d'un classifieur h est mesurée par un risque, une quantité théorique qui ne dépend pas des données. Ce risque est aussi appelé risque vrai, théorique ou de généralisation.

Risque théorique

C'est la probabilité de se tromper sur une **nouvelle** donnée tirée selon la vraie distribution $\mathbb{P}$:

$$\mathscr{R}(h) = \mathbb{P}\left(h(\mathbf{x}) \neq f(\mathbf{x})\right) = \mathbb{E}[\mathbb{1}_{h(\mathbf{x})\neq f(\mathbf{x})}]$$

Pour expliciter sa dépendance en $\mathbb{P}$, nous le notons parfois $\mathscr{L}_{\mathbb{P}}(h)$.

Le Problème fondamental :

Ce risque théorique est impossible à calculer en pratique car la **loi** $\mathbb{P}$ est **inconnue** : il faut l'estimer.

Ensemble d'entraînement et modèle

Ensemble d'entraînement $\mathscr{S}$ (Training Set)

L'ensemble $\mathscr{S}$ est composé d'exemples x_i dotés d'une étiquette y_i :

$$\mathscr{S} = \{(\mathbf{x}_1, y_1), \cdots (\mathbf{x}_n, y_n)\}$$

Les données sont supposées tirées **i.i.d.** selon une loi $\mathbb{P}$.
Cet ensemble est souvent décomposé en une matrice de n caractéristiques et un vecteur de n étiquettes.

L'espace des hypothèses $\mathscr{H}$

L'algorithme cherche une fonction $h \in \mathscr{H}$ telle que $h : \mathscr{X} \to \mathscr{Y}$.
Il est indispensable de restreindre $\mathscr{H}$ à un sous-ensemble pertinent (ex : modèles linéaires, arbres) plutôt que de chercher parmi toutes les fonctions possibles.

Risque Empirique

Puisqu'on ne peut pas calculer le risque théorique, on l'estime à partir des données de $\mathscr{S}$.

Le Risque Empirique 0/1

C'est la proportion d'erreurs commises par h **sur l'ensemble d'entraînement** :

$$\widehat{\mathscr{R}}(h) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_{h(\mathbf{x}_i) \neq y_i}$$

Pour expliciter sa dépendance en $\mathscr{S}$, nous le notons parfois $L_S(h)$.

Minimisation du Risque Empirique (MRE)

Un algorithme d'apprentissage recherche un minimiseur du risque empirique :

$$h_S = \arg\min L_S(h)$$

Il prend comme entrée un ensemble d'apprentissage $\mathscr{S}$ et produit en sortie un classifieur, c'est-à-dire une fonction $h : \mathscr{X} \to \mathscr{Y}$.

Le danger du MRE : Le Surapprentissage (Overfitting)

Minimiser aveuglément $L_S(h)$ est dangereux.
Exemple du classifieur "par cœur" :

$$h_S(\mathbf{x}) = \begin{cases} y_i & \text{si } \mathbf{x} = \mathbf{x}_i \in \mathscr{S} \\ 0 & \text{sinon} \end{cases}$$

- Son risque empirique est **parfait** : $L_S(h_S) = 0$.
- Mais face à une donnée inédite, il prévoira toujours 0 ! Son Risque Vrai sera désastreux.

Régularisation

Pour éviter cela, on fait de la **MRE Régularisée** : on restreint drastiquement l'espace $\mathscr{H}$ *avant* de voir les données pour empêcher l'algorithme d'apprendre "par cœur".

Illustration du Surapprentissage