Machine Learning

Part I
Bases du Machine Learning

1 Notations

- $\mathcal{X}$: espace des entrées.
- $\mathcal{Y}$: espace des sorties (exemples : $\{-1, 1\}$, $\mathbb{R}$).
- $S = \{(x_i, y_i)\}_{1 \leq i \leq m}$.
- $h : \mathcal{X} \longrightarrow \mathcal{Y}$: modèle de prédiction.
- $\mathcal{H}$: ensemble des modèles de prédiction.
- $L : \mathcal{H} \times (\mathcal{X} \times \mathcal{Y}) \longrightarrow \mathbb{R}$: fonction de perte.
- $$\begin{array}{rccll} R_D : & \mathcal{H} & \longrightarrow & \mathbb{R} & \text{risque empirique} \\ & h & \longmapsto & \mathbb{E}_{(x,y)\sim D}[L(h(x), y)] & \end{array}$$

Exemples :

- $L(h(x), y) = \mathbb{1}_{h(x) \neq y}$, $R_D(h) = \mathbb{P}_{(x,y)\sim D}[h(x) \neq y]$.
- $L(h(x), y) = (h(x) - y)^2$, $R_D(h) = \mathbb{E}_{(x,y)\sim D}[(h(x) - y)^2]$.

$$h_S = \underset{h \in \mathcal{H}}{\operatorname{argmin}}\ \hat{R}_S(h) \quad \text{avec} \quad \hat{R}_S(h) \text{ estimateur de } R_D(h) \tag{1}$$

Comment analyser $R_D(\hat{h}_S)$?

2 Courbes ROC

- h_S : seuil score.
- score : $\mathcal{X} \longrightarrow \mathbb{R}$.
- $$\begin{array}{rccl} \text{seuil} : & \mathbb{R} & \longrightarrow & \mathcal{Y} \\ & x & \longmapsto & \begin{cases} 1 & \text{si } x < \mu \\ -1 & \text{si } x \geq \mu \end{cases} \end{array}$$

Exemple :

x	x_1	x_2	x_3	x_4	x_5	x_6
y	1	-1	1	1	-1	-1
score(x)	0.99	0.95	0.51	0.45	0.10	0.01

Définition - Sensibilité et spécificité

On pose : $\mathcal{T}^+ = \{(x_i, y_i) \in S \text{ tels que } y_i = 1\}$
et : $\mathcal{T}^- = \{(x_i, y_i) \in S \text{ tels que } y_i = -1\}$

Sensibilité : $\frac{1}{|\mathcal{T}^+|} \sum_{(x_i,y_i)\in\mathcal{T}^+} \mathbb{1}_{h_S(x_i)=1}$
Spécificité : $\frac{1}{|\mathcal{T}^-|} \sum_{(x_i,y_i)\in\mathcal{T}^-} \mathbb{1}_{h_S(x_i)=-1}$

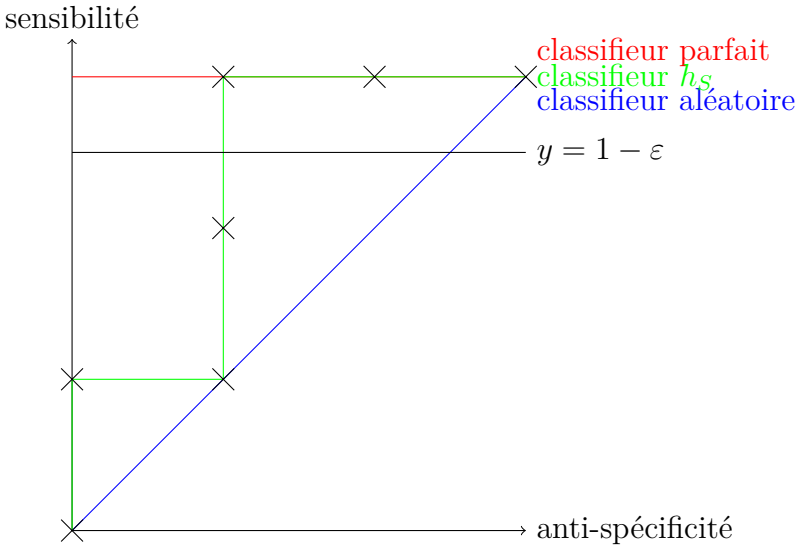
Remarque :

- Sensibilité : taux de vrais positifs.
- Spécificité : taux de vrais négatifs.

Remarque :
Anti-spécificité (ou taux de faux positifs) : $\frac{1}{|\mathcal{T}^-|} \sum_{(x_i,y_i)\in\mathcal{T}^-} \mathbb{1}_{h_S(x_i)=1}$

Exemple :

μ	> 0.99	$]0.95, 0.99]$	$]0.51, 0.95]$	$]0.45, 0.51]$	$]0.10, 0.45]$	$]0.01, 0.10]$
Sensibilité	0	1/3	1/3	2/3	1	1
Anti-spécificité	0	0	1/3	1/3	1/3	2/3



Aire sous la courbe :

- $\text{AUC}_{\text{classifieur parfait}} = 1$.
- $\text{AUC}_{\text{classifieur aléatoire}} = 0.5$.
- $\text{AUC}_{h_S} = 7/9$.