Chapitre 16 : Métriques d'Évaluation

23 février 2026

Plan du chapitre

La Matrice de Confusion

Précision, Rappel et F-Score

Courbe ROC et AUC

Au-delà du Risque Empirique

Le risque empirique $L_S(h)$ (l'erreur sur les données d'entraînement) ne reflète pas la **capacité de généralisation**. On évalue donc le modèle sur un ensemble de test inédit.

Le Paradoxe de l'Exactitude (Accuracy Paradox)

Si on se contente du taux d'erreur global, on devient aveugle face aux classes déséquilibrées. *Exemple :* Détection d'une maladie touchant 1% de la population. Un classifieur trivial (qui prédit toujours "Sain") aura 99% d'exactitude, mais 0% d'utilité médicale !

Il nous faut des métriques plus fines pour décomposer les types d'erreurs.

La Matrice de Confusion

Pour une classification binaire ($\mathcal{Y} = \{-1, +1\}$), on catégorise les prédictions en 4 ensembles :

Vérité y	**Prédiction Positive** $h(\mathbf{x}) = +1$	**Prédiction Négative** $h(\mathbf{x}) = -1$
Vérité $y = +1$	Vrais Positifs (VP)	Faux Négatifs (FN) *Erreur Type II*
	Faux Positifs (FP) *Erreur Type I*	Vrais Négatifs (VN)

1. Précision et Rappel

Précision (Valeur Prédictive Positive)

"Parmi toutes mes alarmes, combien étaient justifiées ?"

$$\text{Précision} = \frac{\text{VP}}{\text{VP} + \text{FP}}$$

Crucial quand les Faux Positifs coûtent cher (ex : spam).

Rappel (Sensibilité ou Taux de Vrais Positifs)

"Parmi tous les vrais cas existants, combien ai-je réussi à trouver ?"

$$\text{Rappel} = \frac{\text{VP}}{\text{VP} + \text{FN}}$$

Crucial quand les Faux Négatifs sont dangereux (ex : cancer, crash d'avion).

Le Compromis Précision-Rappel

Théorème empirique : Précision et Rappel sont généralement inversement proportionnels.

- Si le modèle est **très sévère** (ne déclenche l'alarme que s'il est 100% sûr) $\Longrightarrow$ FP chute, FN explose. La Précision monte, le Rappel s'effondre.
- Si le modèle est **très laxiste** (déclenche l'alarme au moindre doute) $\Longrightarrow$ FN chute, FP explose. Le Rappel monte, la Précision s'effondre.

Comment résumer ces deux forces contradictoires en un seul chiffre ?

2. Le F_1-Score (Moyenne Harmonique)

On utilise le F_1-**Score**, qui est la **moyenne harmonique** de la Précision (P) et du Rappel (R) :

$$F_1 = 2 \cdot \frac{P \cdot R}{P + R} = \frac{2\text{VP}}{2\text{VP} + \text{FP} + \text{FN}}$$

Pourquoi la Moyenne Harmonique ?

Contrairement à la moyenne arithmétique ($(P + R)/2$), la moyenne harmonique **pénalise lourdement les valeurs extrêmes**.
Exemple : Si $P = 1.0$ et $R = 0.0$:

- Moyenne arithmétique $= 0.5$ (Biaisé ! Le modèle est inutile).
- Moyenne harmonique $= 0.0$ (Reflète la vraie utilité).

Généralisation : Le F_β-Score

En pratique, la Précision n'a pas toujours la même importance financière ou humaine que le Rappel.
On définit le F_β-Score, une moyenne harmonique pondérée :

$$F_\beta = (1 + \beta^2) \frac{P \cdot R}{(\beta^2 \cdot P) + R}$$

- $\beta = 1$: P et R ont la même importance (F_1-Score).
- $\beta = 2$: Le Rappel est considéré 2 fois plus important que la Précision (F_2-Score, souvent utilisé en diagnostic médical).
- $\beta = 0.5$: La Précision est 2 fois plus importante.

Du score au seuil

Beaucoup de classifieurs (SVM, Régression Logistique, Réseaux de Neurones) ne sortent pas une classe binaire stricte, mais un **score de confiance** $\hbar(\mathbf{x}) \in \mathbb{R}$.
La décision finale se fait via un **seuil** t :

$$h(\mathbf{x}) = \text{sign}(\hbar(\mathbf{x}) - t)$$

Le F1-Score et la Matrice de Confusion ne mesurent la performance que pour **un seul seuil** t **donné**. Comment évaluer la qualité intrinsèque du score $\hbar(\mathbf{x})$ indépendamment de t ?

La Courbe ROC

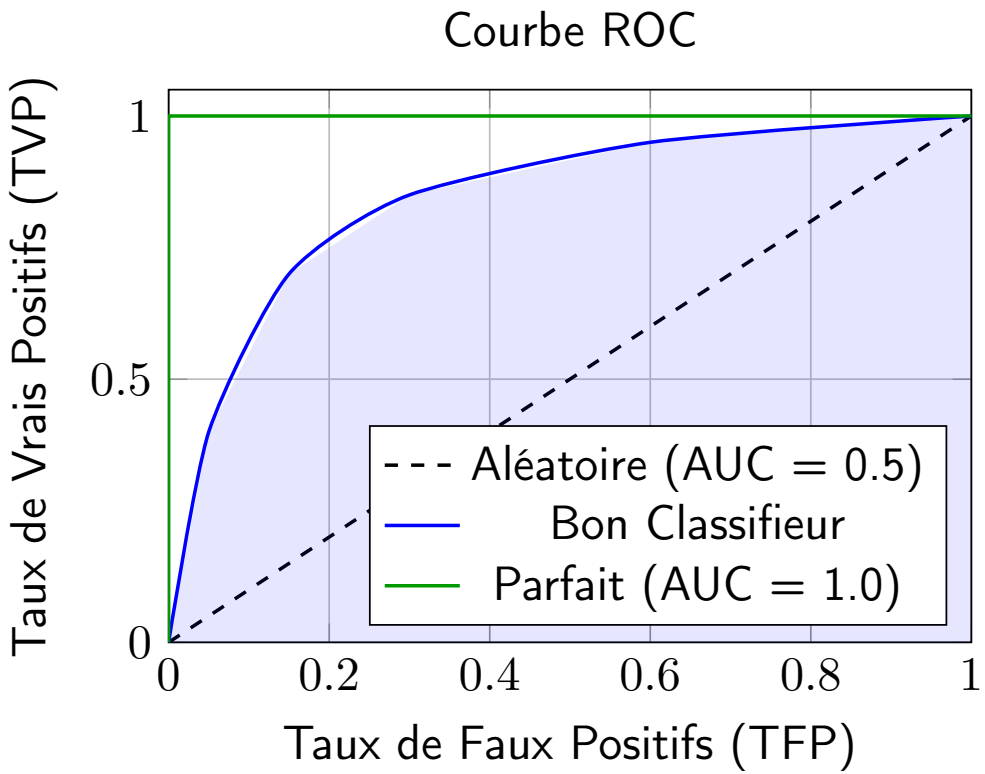
La courbe ROC (*Receiver Operating Characteristic*) trace la performance pour **tous les seuils** $t \in [-\infty, +\infty]$.

- **Axe Y (Sensibilité) :** Taux de Vrais Positifs (TVP) = Rappel $= \frac{\text{VP}}{\text{VP+FN}}$
- **Axe X (Antispécificité) :** Taux de Faux Positifs (TFP) $= \frac{\text{FP}}{\text{FP+VN}}$

On fait varier le seuil t :

- $t = +\infty$: Le modèle prédit tout Négatif. (0, 0).
- $t = -\infty$: Le modèle prédit tout Positif. (1, 1).
- Le but est de "tendre" la courbe vers le coin supérieur gauche (0, 1).

Visualisation de la Courbe ROC



AUC : Interprétation Probabiliste

L'**AUC** (*Area Under the Curve*) est l'intégrale (l'aire) sous la courbe ROC. Elle résume la performance du score globalement (entre 0.5 et 1.0).

Propriété statistique fondamentale (Wilcoxon-Mann-Whitney)

L'AUC est mathématiquement égale à la probabilité qu'un exemple positif tiré au hasard obtienne un score de confiance strictement supérieur à un exemple négatif tiré au hasard.

$$\text{AUC} = \mathbb{P}\left(\hbar(\mathbf{x}_+) > \hbar(\mathbf{x}_-)\right)$$

C'est une excellente métrique, robuste au déséquilibre des classes, car elle évalue uniquement la capacité du modèle à **trier** (ordonnancer) correctement les données, peu importe où le seuil de décision final sera placé.