Sélection de Modèles

Introduction

- **Objectif** : Apprendre une fonction $f : X \to Y$ capable de prédire les labels y sur des données non vues.
- Quel modèle choisir ?
- Comment évaluer sa capacité à généraliser ?

Principes de l'Apprentissage Statistique

Fonction de loss

La fonction de loss $\ell(Y, f(X))$ évalue la distance entre la prédiction $f(x)$ et la vérité y.
Exemple de loss 0-1 (classification binaire) :

$$\ell(y, f(x)) = \begin{cases} 0 & \text{si } yf(x) > 0 \\ 1 & \text{sinon} \end{cases}$$

Risque Empirique et Généralisation

- **Risque réel** :
$$R(f) = \mathbb{E}_{X,Y}[\ell(Y, f(X))]$$
- **Risque empirique** :
$$R_n(f) = \frac{1}{n} \sum_{i=1}^{n} \ell(y_i, f(x_i))$$

Régularisation pour Contrôler la Complexité

Pour éviter le sur-apprentissage (overfitting), on régularise le risque empirique :

$$\min_f \frac{1}{n} \sum_{i=1}^{n} \ell(y_i, f(x_i)) + \lambda \Omega(f)$$

où λ est un paramètre de régularisation qui contrôle la complexité du modèle et $\Omega(f)$ une fonction de régularisation.

Évaluation de la Qualité d'un Modèle

Mesures de Performance

- **Matrice de confusion** utile pour les pb à classes déséquilibrés :

Prédit/ Réel	Positif	Négatif
Positif	TP	FP
Négatif	FN	TN

- **Indicateurs** :

$$\text{Précision} = \frac{\text{TP}}{\text{TP} + \text{FP}}$$
$$\text{Rappel (Sensibilité)} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$
$$\text{F-mesure} = 2\frac{\text{Précision} \cdot \text{Rappel}}{\text{Précision} + \text{Rappel}}$$

Courbe ROC et AUC

- **ROC** : TPR (*True Positive Rate*) contre FPR (*False Positive Rate*).
- **AUC** : Aire sous la courbe ROC. L'AUC est comprise entre 0 et 1.

Hugo Tondenier

Sélection de Modèle

Principe

- **Objectif** : Choisir un modèle f dans une famille F qui présente les meilleures performances.
- Exemples de modèles :
  - K-NN : choix de K
  - SVM : réglage du paramètre C et choix du noyau

Méthodologie Pratique

1. Diviser les données en deux ensembles : D_{train_val} et D_{test}.
2. Sélectionner le modèle optimal :
   - Diviser D_{train_val} en D_{train} et D_{val}.
   - Entraîner chaque modèle sur D_{train} et évaluer sa performance sur D_{val}.
3. Tester le modèle optimal sélectionné sur D_{test}.

Validation Croisée K-Fold

Pour des ensembles de données petits :

- Diviser D_{train} en K sous-ensembles.
- Entraîner le modèle sur $K-1$ sous-ensembles et le tester sur le sous-ensemble restant.
- Répéter l'opération pour chaque sous-ensemble et calculer la moyenne des performances.