Ecole d'été de Santé Publique et d'Epidémiologie de Bicêtre

M.Sedki, JP Teglas

7. Introduction au Machine Learning en recherche biomédicale *Du 23 juin au 27 juillet 2025*

Objectifs

L'objectif du cours est d'apporter une introduction par l'exemple aux principes et aux outils du *Machine Learning* en recherche biomédicale.
Nous allons aborder la problématique dite d'apprentissage supervisé qui vise à prédire une variable cible à partir d'une ou plusieurs autres variables dites explicatives. Nous distinguerons deux cas pour la variable à prédire. L'un est le problème de la régression, où la variable à prédire est quantitative telle que le coût d'un traitement, un taux de mortalité par ville, la concentration en rétinol plasmatique ou la pression artérielle. Dans le problème de la classification, la variable à prédire prend un nombre fini de valeurs ou modalités telles que "survécu" ou "mort" ou le type de cancer d'un échantillon de tissu.

L'objectif des méthodes rassemblées sous le nom *Machine Learning* (apprentissage statistique en français) est de construire un modèle à partir de données observées, dites données d'apprentissage, qui sera utilisé pour prédire la partie réponse de cas dits tests pour lesquels nous n'observons que les variables explicatives. Il est nécessaire de prédire avec précision les cas tests, mais aussi comprendre quelles variables explicatives affectent le résultat et comment, et aussi d'évaluer la qualité des prédictions.

Il est facile d'appliquer un algorithme pour répondre à l'objectif de prédiction et de nos jours, on peut simplement lancer un logiciel, néanmoins il est important et parfois difficile de comprendre à quel point la méthode fonctionne réellement. Nous essayerons de détailler le fonctionnement d'un ensemble de familles de modèles à travers une série d'exemples d'application sur jeux de données réelles avec le logiciel R.

Pré-requis

-Notions de base de probabilité et statistique
-Une pratique régulière de la programmation avec R est indispensable.

Programme

-Formalisme de la régression et de la classification.
-Problématique du choix du meilleur modèle dans une famille de modèles et validation croisée
-Arbres de décision : arbres de régression et classification.
-Réseaux de neurones pour la classification d'images : introduction et illustration sur un exemple
- Si le temps le permet : présentation d'une architecture neuronale de type LLM.

Déroulement

Chaque partie du programme sera accompagnée d'un TP R sur un jeu de données réel issu de la recherche biomédicale. Les deux dernières parties sont uniquement consacrées à la présentation des architectures neuronales avec illustration sur des exemples calculés par le formateur en amont sur des ressources informatiques adaptées (GPU et serveur).