Stage M2 : Modélisation statistique pour la dynamique physiologique des plantes via des approches d'apprentissage profond

Contexte

Les modèles statistiques de la dynamique de la physiologie des plantes visent à décrire et analyser les processus sous-jacents au développement des plantes. Dans sa formulation la plus simple, le modèle est un modèle à effets mixtes non linéaire dans lequel un nombre limité de paramètres contrôlent la forme et la dynamique du trait physiologique considéré (par exemple la croissance de la plante). Ces paramètres peuvent à leur tour être décrits comme des variables aléatoires dont les distributions dépendent d'un ensemble (éventuellement de grande dimension) de cofacteurs explicatifs. L'inférence de tels modèles est complexe, et nécessite généralement l'utilisation de techniques d'échantillonnage telles que SAEM ou MCEM [1,2].

Sujet

L'objectif du projet est de développer une stratégie d'inférence alternative basée sur les approches récemment développées pour l'estimation des paramètres en apprentissage profond (différentiation automatique, batch gradient descent) pour ajuster des modèles à effets mixtes non linéaires de pointe pour la dynamique physiologique. À cette fin, l'inférence sera présentée comme un problème d'optimisation où la cible est un réseau neuronal, et où la fonction de perte sera choisie en fonction de la nature de la caractéristique physiologique (continue ou discrète). La méthode sera mise en œuvre à l'aide de bibliothèques dédiées au calcul numérique haute performance et à l'optimisation en Python (PyTorch, Jax) ou en R (torch).

Les modèles seront appliqués aux données collectées dans le projet ANR G2WAS. Les données comprennent 250 variétés de raisin qui ont été phénotypées de manière dynamique pendant 3 semaines pour la production de biomasse végétative par imagerie à la plateforme PhenoArch. Chaque variété a subi 3 scénarios hydriques différents (de "bien arrosé" à "stress hydrique sévère"). Un ensemble d'environ 60 000 marqueurs génétiques sera utilisé comme variables explicatives dans le modèle statistique pour expliquer et prédire la dynamique de la physiologie des plantes.

Compétences requises

Ce stage s'adresse à un ou une étudiant.e de Master 2 dans l'un de ces domaines : informatique, statistiques ou apprentissage automatique. Une maîtrise de la programmation en Python et/ou en R, une expérience de travail avec de grands ensembles de données, ainsi qu'un intérêt pour les applications en biologie, et plus particulièrement en génétique sont requis.

Environnement de travail

Le stagiaire sera financé par le projet ANR G2WAS. Vous travaillerez dans l'équipe SOLsTIS de l'unité MIA Paris-Saclay, située à AgroParisTech (Palaiseau), sous la supervision de Tristan Mary-Huard, Laure Sansonnet et Julien Chiquet, et en étroite collaboration avec Vincent Segura et Timothée Flutre pour les aspects de physiologie des plantes et de génétique.

La durée du stage envisagée et de 5 ou 6 mois, avec une date de début comprise entre février et avril 2024 suivant la disponibilité du futur.e stagiaire.

Contact

Les candidat.e.s intéressé.e.s doivent postuler en envoyant un CV et une lettre de motivation à :
tristan.mary-huard@agroparistech.fr
laure.sansonnet@agroparistech.fr
julien.chiquet@inrae.fr

Références

[1] Kuhn & Lavielle (2005). Maximum likelihood estimation in nonlinear mixed effects model, *Comput. Stat. and Data Analysis*.

[2] Liu & Wu (2007). Simultaneous inference for semiparametric nonlinear mixed-effects models with covariate measurement errors and missing responses, *Biometrics*.