Développement de méthodes d'apprentissage « Label Probability Learning » pour décoder l'épitranscriptome

Proposition de stage niveau M1 ou M2 (printemps 2027)

Unité MIAT, INRAE Toulouse

Pour postuler envoyer CV, lettre de motivation et dernier relevé de notes à nathalie.vialaneix@inrae.fr

Problématique du stage

On s'intéresse aux problèmes d'apprentissage de type « apprentissage de probabilités d'étiquettes » (*Label Probability Learning*, [2]). Dans ce type de problèmes, on cherche à entraîner un prédicteur, Φ, permettant de prédire une variable aléatoire Y (binaire, $\in \{0,1\}$) à partir d'une variable aléatoire X ($\in \mathbb{R}^d$) mais, contrairement au cadre classique, on ne dispose pas d'un échantillon d'apprentissage $(x_i, y_i)_{i=1,\dots,n}$ mais d'un ensemble de « sacs » disjoints, $(\mathcal{S}_j)_{j=1,\dots,p}$ pour lesquels on dispose de :

- la mesure de X pour chacun des éléments du sac : $(x_i)_{i\in\mathcal{S}_j}$;
- la mesure de la moyenne de Y pour le sac : $m_j = \sum_{i\in\mathcal{S}_j} y_i$, la valeur de y_i elle-même restant inconnue.

Un cadre un peu plus complexe décrit dans [3, 5, 1] considère que la mesure m_j est partiellement inconnue (mesurée de manière bruitée), par exemple par contamination mutuelle des échantillons positifs ($y_i = 1$) et des échantillons négatifs ($y_i = 0$), comme dans [3].

L'objectif de ce stage sera d'étudier et/ou implémenter et tester ce cadre d'apprentissage pour l'analyse de données biologiques complexes se présentant sous la forme d'une mesure d'un courant ionique (données de séquençage direct de l'ARN par technologie Nanopore).

Travail à effectuer

Selon l'appétence de l'étudiant-e recruté-e, le stage pourra consister à réaliser :

- une synthèse de la bibliographie de ce domaine ;
- une implémentation d'une ou deux stratégies d'apprentissage ;
- le test de celle(s)-ci sur des données issues du projet DEMETER (données de séquençage direct de l'ARN par technologie Nanopore). L'objectif de ces tests est d'étudier les modifications de l'ARN et leur implication dans les processus de résistance au stress hydrique chez le tournesol.

En savoir plus : focus sur l'épitranscriptomique...

Plus de 170 modifications chimiques sont actuellement répertoriées dans les ARN, qui ont un impact sur le fonctionnement des organismes vivants, notamment pour une adaptation rapide aux stress de l'environnement [4]. Le séquençage d'ARN « direct » par la technologie Nanopore constitue une approche prometteuse pour identifier ces modifications. L'approche produit, en effet, des données sous la forme d'un courant ionique influencé par les différents types de modifications.

Toutefois, le problème d'apprentissage est très particulier car, la technique de séquençage étant destructive de la molécule d'ARN, il n'existe pas de bases d'apprentissage d'ARN natifs (« naturels ») mais des techniques expérimentales orthogonales permettent d'obtenir une estimation de la probabilité de modification pour chaque position (ou site) des ARN.

Le sujet de stage pourrait donner lieu à une poursuite en thèse en statistique / machine learning (financement en cours d'obtention et/ou soutien pour préparer le concours de l'ED MITT sinon).

Conditions du stage

durée 4 à 6 mois

localisation Unité MIA-T, INRAE Toulouse (Castanet-Tolosan)

rémunération taux légal

encadrement Nathalie Vialaneix (INRAE, MIAT, Toulouse) http://www.nathalievialaneix.eu, Christine Gaspin (INRAE, MIAT, Toulouse), Malvina Marku (INRAE, MIAT, Toulouse) et Benjamin Charlier (INRAE, MIAT, Toulouse)
Contact : nathalie.vialaneix@inrae.fr

Modalité de candidature Envoyer CV, lettre de motivation et relevé de notes de l'année précédente ou en cours à nathalie.vialaneix@inrae.fr

Profil recherché

— Master 1 ou 4ème/5ème année d'école d'ingénieur en statistique et/ou machine learning ;
— bonne compréhension du formalisme mathématique ;
— aisance en programmation ; des bases dans l'utilisation de pytorch seraient un plus ;
— bonne maîtrise de l'anglais ;
— un intérêt pour les approches d'IA et de deep learning ;
— les connaissances en biologie ne sont pas nécessaires mais un intérêt pour ce domaine d'application est nécessaire.

Références

[1] Shreyas Havaldar, Navodita Sharma, Shubhi Sareen, Karthikeyan Shanmugam, and Aravindan Raghuveer. Learning from label proportions : bootstrapping supervised learners via belief propagation. Preprint arXiv :2310.08056, 2023.

[2] Novi Quadrianto, Alex J. Smola, Tibério S. Caetano, and Quoc V. Le. Estimating labels from label proportions. *Journal of Machine Learning Research*, 10(82), 2009.

[3] Clayton Scott and Jianxin Zhang. Learning from label proportions : A mutual contamination framework. In *In Proceedings of 33th Conference on Neural Information Processing Systems (NeurIPS 2020)*, 2020.

[4] Dominik Sordyl, Etienne Boileau, Agata Bernat, Satyabrata Maiti, Sunandan Mukherjee, S. Naeim Moafinejad, Masoud Amiri Farsani, Anastasiya Shavina, Andrea Cappannini, Giada Agostini, Silvestro G. Conticello, Filip Stefaniak, Christoph Dieterich, Elżbieta Purta, and Janusz M. Bujnicki. MODOMICS : a database of RNA modifications and related information. 2025 update and 20th anniversary. *Nucleic Acids Research*, 54(D1) :D219–D225, 2026.

[5] Jianxin Zhang, Yutong Wang, and Clayton Scott. Learning from label proportions by learning with label noise. In *In Proceedings of 36th Conference on Neural Information Processing Systems (NeurIPS 2022)*, 2022.