

Auto Machine Learning en temps réel : des modèles de classification plus intelligents pour les flux de données

Ubiquity AI – Bordeaux – 6 mois

Contexte

La croissance exponentielle des données produites par l'économie numérique et la dépendance au stockage de données posent des problèmes majeurs : coût, sécurisation et mise en conformité, risque d'approvisionnement, impact écologique.

Alors que les GAFAMs dominent toujours plus l'économie numérique, outre notre dépendance technologique à ces acteurs, ils ne devraient pas assurer seuls la prise en compte des enjeux de maîtrise de l'empreinte écologique de l'IA, de la protection des données personnelles et sensibles, des biais algorithmiques, de la démocratisation de son usage dans notre société.

“We have the people, we have the compute, we have the data, we have everything.”
Satya Nadella, CEO de Microsoft

Les entreprises et les collectivités veulent extraire de la connaissance à partir des données à leur disposition pour créer de la valeur et innover, mais sous le capot, cette création de valeur peut être difficile à construire et orchestrer. L'IA décuple à la fois les attentes et les coûts de cette création de valeur.

Une toute nouvelle approche algorithmique, le stream learning ou machine learning en temps-réel[1], promet de grandement simplifier l'intégration des modèles d'IA en offrant aux entreprises et collectivités des modèles jamais obsolètes, avec des coûts de production réduits, tout en prenant naturellement en compte les enjeux de protection des données personnelles et d'impact écologique. Cette innovation de rupture permet d'apprendre en continu, de s'adapter aux changements en temps-réel, directement à partir des flux de données pris à leur source, sans recours au stockage.

Le stream learning, cependant, est un nouveau courant de recherche qui n'a pas encore la maturité industrielle du machine learning, et l'objectif de ce stage est de contribuer à son développement.

À propos d'Ubiquity AI

Ubiquity AI est une deeptech bordelaise, essaimage du LIPN-CNRS de l'Université Sorbonne Paris Nord avec plus de 20 ans de recherche fondamentale et 10 ans de R&D travaillant sur cette nouvelle génération de système d'IA. Sa mission est de démocratiser l'intelligence artificielle en la rendant accessible grâce au stream learning qui est la forme la plus pure et naturelle d'apprentissage.

L'objectif

L'objectif de ce stage est de réaliser le tout premier benchmark à grande échelle des algorithmes de classification en stream learning, et de proposer un algorithme basé sur la sélection de modèles capable de rivaliser avec l'état de l'art sur un large éventail de jeux de données.

Un stage similaire est proposé pour un autre paradigme de machine learning : la régression.

Les tâches principales

- Bibliographie sur les algorithmes de classification en stream learning
- Étudier les algorithmes de classification les plus populaires en stream learning et comprendre leurs hyperparamètres
- Créer un répertoire de jeux de données divers pour la classification
- Réaliser un benchmark de ces algorithmes de classification sur les jeux de données sélectionnés en utilisant les métriques pertinentes (Accuracy, AUC-ROC, Temps, Mémoire)
- Concevoir et tester un algorithme de classification compétitif basé sur la sélection de modèles[2]

Ce que vous allez apprendre

- Les bases du paradigme du stream learning.
- Faire de la bibliographie scientifique sur un sujet particulier "la classification en stream learning".
- Collecter des jeux de données pertinents.
- Appliquer des tests statistiques pour comparer les performances des algorithmes.
- Construire des algorithmes de classification compétitifs en utilisant la sélection de modèles.

- Utiliser et étendre les bibliothèques de stream learning les plus populaires en Python (River[3], CappyMoa[4])

Plus de détails

- **Lieu** : Bordeaux centre-ville.
- **Période** : Février-Mars à Août-Septembre 2025 (6 mois).
- **Formation** : Étudiant(e) en Master 2 Recherche spécialisation IA (ou équivalent ingénieur), à la recherche d'un stage de fin d'études (6 mois).
- **Gratification** : 1000€ brut par mois
- **Candidature** : Envoyez votre CV à l'adresse suivante : subhy.albakour@ubiquity.ai.
- **La suite** : Le stage peut être suivi par un poste en CDI d'ingénieur R&D en machine learning temps-réel au sein de notre équipe R&D. Pour ce CDI, il est attendu un très bon niveau de programmation en Java, en algorithmique, une connaissance de l'optimisation software/hardware, et un réel intérêt d'apprendre et de maîtriser les architectures de machine learning distribuées de haute performance.

Références

- [1] Albert Bifet, Ricard Gavaldà, Geoffrey Holmes, and Bernhard Pfahringer, Machine Learning for Data Streams: with Practical Examples in MOA. Cambridge, Massachusetts: The MIT Press, 2018.
- [2] Y. Ho and D. Pepyne, "Simple explanation of the no-free-lunch theorem and its implications," Journal of Optimization Theory and Applications, vol. 115, pp. 549–570, Dec 2002.
- [3] J. Montiel, M. Halford, S. M. Mastelini, G. Bolmier, R. Sourty, R. Vaysse, A. Zouitine, H. M. Gomes, J. Read, T. Abdessalem, and A. Bifet, "River: machine learning for streaming data in python," Journal of Machine Learning Research, vol. 22, no. 1, pp. 4945–4952, 2021.
- [4] Gomes, Heitor Murilo, et al. "CappyMOA: efficient machine learning for data streams in python." arXiv preprint arXiv:2502.07432 (2025).