Data science for economists

ECTS : 3

Langue du cours : Français

Volume horaire : 21

Description du contenu de l'enseignement :

Introduction

- Les différentes étapes d'un projet de science des données

- Le vocabulaire de science de données et ses correspondances avec les statistiques classiques

- Différents types de question et de modélisation

Infrastructure et gestion de projets de science des données

- Outils de contrôle de version des données et modèles

- Outils de manipulation d'un grand jeu de données

- Récupérer des données, entraîner et déployer un modèle sur Google Cloud Platform

- Outils de développement et assistance par IA.

Outils d'apprentissage automatique

- Détection d'anomalies

- Sélection de variable et régression pénalisée

- Gestion des données manquantes

- Modélisation semi-supervisée

- Récupération et traitement de données alternatives : exemple de web scraping

Application d'apprentissage profond en économie

- Introduction aux réseaux de neurones

- Traitement automatique du langage naturel

- Vision par ordinateur

Compétences à acquérir :

On distinguera deux blocs d'acquisition de compétences l'un correspondant à ce qu'on peut appeler le ML Engineering, l'autre correspondant plutôt au ML Ops.

Dans le premier bloc, nous étudierons des méthodes issues de la science des données venant compléter celles vues habituellement en statistiques et économétrie de la détection des anomalies par isolation forest au traitement des données non structurées (images ou textes), en passant par la gestion des valeurs manquantes avec les librairies de gradient boosting. Il s'agira d'ouvrir la boîte noire de certaines méthodes pour mieux comprendre leur fonctionnement, leurs limites et leur intérêt pour des questions économiques.

Dans le second bloc, nous verrons les outils de base pour développer un projet de science des données automatisé de bout en bout de la récupération des données, à la mise en production d'un modèle en passant par la gestion des versions. Cette partie sera particulièrement importante pour exploiter des modèles ou des données qui ne seraient pas exploitables sur la plupart des PC. L'automatisation et la mise en production de modèles sont utiles pour des besoins de mise à jour régulière (par exemple prévision/analyse de conjoncture mensuelle) ou encore le traitement automatisé d'un flux de requête ou de données (détection d'anomalie, traitement automatique de communiqués, scoring de crédit…).

Pré-requis recommandés

- Maîtrise des outils statistiques et économétriques de L3 et M1.

- Connaissance de base d'un langage de programmation ou statistique (R ou Python).

Mode de contrôle des connaissances :

Examen écrit

Coefficient : 3.00

Bibliographie, lectures recommandées :

Aruoba S. B., Drechsel T. (2024), Identifying Monetary Policy Shocks: A Natural Language Approach. https://econweb.umd.edu/~drechsel/papers/Aruoba_Drechsel.pdf

d'Aspremont A., Ben Arous S., Bricongne J-C., Lietti B., Meunier B. (2024) Satellites Turn "Concrete": Tracking Cement with Satellite Data and Neural Networks. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4712741

Banquet, A. et al. (2022), Monitoring land use in cities using satellite imagery and deep learning, OECD Regional Development Papers, No. 28, OECD Publishing, Paris, https://doi.org/10.1787/dc8e85d5-en.

Blum A., Mitchell T. (1998) Combining labeled and unlabeled data with co-training https://www.cs.cmu.edu/~avrim/Papers/cotrain.pdf

Dixon M. F. , Halperin I. , Bilokon P. (2020) Machine Learning in Finance, From Theory to Practice https://link.springer.com/book/10.1007/978-3-030-41068-1

Gaillac C., L'Hour J. (2023) Machine Learning pour l'économétrie,https://www.economica.fr/machine-learning-pour-leconometrie-c2x40149680

Han, S., Hu, X., Huang, H., Jiang, M., & Zhao, Y. (2022). Adbench: Anomaly detection benchmark. arXiv preprint arXiv:2206.09426.

IMF (2025), Nowcasting Global Trade from Space. https://www.imf.org/en/Publications/WP/Issues/2025/05/16/Nowcasting-Global-Trade-from-Space-566957

Lu, Sha, Lin Liu, Jiuyong Li, Thuc Duy Le, et Jixue Liu. « Dependency-Based Anomaly Detection: Framework, Methods and Benchmark ». arXiv, 12 novembre 2020.

Pfeifer M., Marohl V. P. (2023) CentralBankRoBERTa: A fine-tuned large language model for central bank communications. https://www.sciencedirect.com/science/article/pii/S2405918823000302

Zhao Z., Hryniewicki M. (2019) Xgbod: improving supervised outlier detection with unsupervised representation learning. In 2018 International Joint Conference on Neural Networks (IJCNN) (pp. 1-8). IEEE.