Data Science in Practice - MAP670 - M2 Data &AI - IP Paris - 2022 - 2023

Lecturers: Mariam BARRY
Contacts: mariam.barry@polytechnique.edu,

Exam = 10 min of oral presentation of projects on - Date to be defined:
https://docs.google.com/document/d/1N2aI1hsIBU4juocpqLKXJiX9hkjKux7a/edit?usp=sharing&ouid=112457993494032970404&rtpof=true&sd=true

River RoadMap :
https://www.notion.so/maxhalford/d1e86fcdf21e4deda16eedab2b3361fb?v=503f44740b8b44a99a961aa96e9e46e1

The expected deliverable for the projects are :

1. Put the code Source .py and HTML in Zip file (**Moodle**) and available on your **Github**
2. Presentation PowerPoint & Results (5" min): Problem, Solution, Results
3. Demo (Proof that the code compiles and everything run) – 5 min
4. Send an email to mariam.barry@polytechnique.edu with groupe ID - Student Names in object + Github link, slides and put all the materials (code source & slides)

Option 1: Implement an online learning method from a research paper

The goal of this project is to build **teams of 3 people** to work together on an open-source implementation of streaming algorithms that can be used with river in Python or streaming application using Apache Kafka. You can choose one of listed paper or propose your own research paper on online learning or data stream processing to implement on river.

The deliveries of the project are:

1. Source code following river guidelines:
    a. https://github.com/online-ml/river
2. **Documentation following RIVER guidelines:**
    a. **https://github.com/online-ml/river/blob/main/CONTRIBUTING.md**
3. Presentation slides with experimental results

Refer https://riverml.xyz/latest/ for additional guidelines specific to river

Note: Teams whose implementation displays high-quality may be invited to contribute their code to the project on GitHub. This contribution is voluntary and unrelated to the grading of your project, however code and projects with documentation following the official guidelines will be taken into account for a bonus https://github.com/online-ml/river/blob/main/CONTRIBUTING.md

River Github https://github.com/online-ml/river or website https://riverml.xyz/latest/

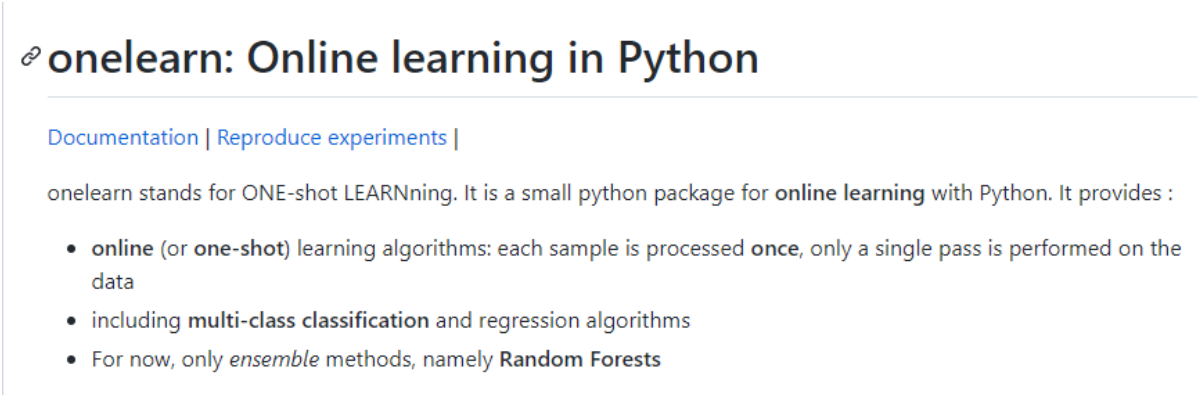
- Project 1 : NEW - CONTRIBUTING IN METHODS FROM RIVER ROADMAP TO GET YOUR CODE INTEGRATED IN THE OFFFICIAL API

You have the opportunity to contribute in one of the online learning methods and ROADMAP of new models that will be integrated official in the open source package (cf my email following my discussion with the maintainer of the package).

1.1 Choose some methods or papers from the **official RoadMap** https://www.notion.so/maxhalford/d1e86fcdf21e4deda16eedab2b3361fb?v=503f44740b8b44a99a961aa96e9e46e1

1.2 Implement a **Graph learning method (Deep Walk)** https://www.notion.so/Graph-learning-b31e3b1c7e624401ae6eb0a9ebb089a1

1.3 Integrate **onelearn in river**
https://github.com/onelearn/onelearn



onelearn: Online learning in Python

Documentation | Reproduce experiments |

onelearn stands for ONE-shot LEARNning. It is a small python package for **online learning** with Python. It provides :

- **online** (or **one-shot**) learning algorithms: each sample is processed **once**, only a single pass is performed on the data
- including **multi-class classification** and regression algorithms
- For now, only *ensemble* methods, namely **Random Forests**

1.4 Implement Conformal prediction Another nice thing to contribute would be conformal predictions: they're a robust way to produce confidence intervals, and work naturally online.

https://towardsdatascience.com/conformal-prediction-4775e78b47b6
https://github.com/valeman/awesome-conformal-prediction

NB : If we have more than 10 students working on methods from RIVER Roadmap (1.1, 1.2 or 1.3), I will tray to get a meeting with the main maintainer of the package to answer you technical questions / issues on river. https://maxhalford.github.io/

Project 2 : Fast Anomaly Detection in Multiple Multi-Dimensional Data Streams

Multiple multi-dimensional data streams are ubiquitous in the modern world, such as IoT applications, GIS applications and social networks. Detecting anomalies in such data streams in real-time is an important and challenging task. It is able to provide valuable information from data and then assists decision-making. However, existing approaches for anomaly detection in multi-dimensional data streams do not properly consider the correlations among multiple multi-dimensional streams. Moreover, for multi-dimensional streaming data, online detection speed is often an important concern. In this paper, we propose a fast yet effective anomaly detection approach in multiple multi-dimensional data streams. This is based on a combination of ideas, i.e., stream pre-processing, locality sensitive hashing and dynamic isolation forest.

>> Email me at mariam.barry@polytechnique.edu to get the PDF of the paper. https://ieeexplore.ieee.org/document/9006354

Project 3 : Streaming Graph Neural Networks via Continual Learning

Paper (PDF) https://arxiv.org/pdf/2009.10951.pdf

Graph neural networks (GNNs) have achieved strong performance in various applications. In the real world, network data is usually formed in a streaming fashion. The distributions of patterns that refer to neighborhood information of nodes may shift over time. The GNN model needs to learn the new patterns that cannot yet be captured. But learning incrementally leads to the catastrophic forgetting problem that historical knowledge is overwritten by newly learned knowledge. Therefore, it is important to train GNN model to learn new patterns and maintain existing patterns simultaneously, which few works focus on. In this paper, we propose a streaming GNN model based on continual learning so that the model is trained incrementally and up-to-date node representations can be obtained at each time step. Firstly, we design an approximation algorithm to detect new coming patterns efficiently based on information propagation. Secondly, we combine two perspectives of data replaying and model regularization for existing pattern consolidation. Specially, a hierarchy-importance sampling strategy for nodes is designed and a weighted

regularization term for GNN parameters is derived, achieving greater stability and generalization of knowledge consolidation.

Project 4 : ROLAND: Graph Learning Framework for Dynamic Graphs

Paper (PDF) https://dl.acm.org/doi/abs/10.1145/3534678.3539300

Graph Neural Networks (GNNs) have been successfully applied to many real-world static graphs. However, the success of static graphs has not fully translated to dynamic graphs due to the limitations in model design, evaluation settings, and training strategies. Concretely, existing dynamic GNNs do not incorporate state-of-the-art designs from static GNNs, which limits their performance. Current evaluation settings for dynamic GNNs do not fully reflect the evolving nature of dynamic graphs. Finally, commonly used training methods for dynamic GNNs are not scalable. Here we propose ROLAND, an effective graph representation learning framework for real-world dynamic graphs. At its core, the ROLAND framework can help researchers easily repurpose any static GNN to dynamic graphs. Our insight is to view the node embeddings at different GNN layers as hierarchical node states and then recurrently update them over time. We then introduce a live-update evaluation setting for dynamic graphs that mimics real-world use cases, where GNNs are making predictions and being updated on a rolling basis. Finally, we propose a scalable and efficient training approach for dynamic GNNs via incremental training and meta-learning. We conduct experiments over eight different dynamic graph datasets on future link prediction tasks.

Option 2 : Extend an existing code or method to fit in river framework and do a benchmark (comparison) of methods with open data and public anomaly datasets

Project 5: MSTreams - MStream: Fast Anomaly Detection in Multi-Aspect Streams -

The work aims to define a streaming multi-aspect data anomaly detection framework, termed MSTREAM which can detect unusual group anomalies as they occur, in a dynamic manner. MSTREAM has the following properties: (a) it detects anomalies in multi-aspect data including both categorical and numeric attributes; (b) it is online, thus processing each record in constant time and constant memory; (c) it can capture the correlation between multiple aspects of the data.

The code source is available partially in Python and C++: https://github.com/Stream-AD/MStream

https://riverml.xyz/latest/api/tree/HoeffdingAdaptiveTreeRegressor/

Project 6 : Adapt existing python code of anomaly detection methods in the river framework (Online Anomaly Detection Package)
PySAD provides methods for online/sequential anomaly detection, i.e. anomaly detection on streaming data, where the model updates itself as a new instance arrives.

Code source on anomaly detection methods :
Github https://github.com/selimfirat/pysad

Anomaly detection datasets: http://odds.cs.stonybrook.edu/

Your task is to re-adapt the code of PySOD framework to River framework using method leran_one and predict_one and benchmark the different methods using public datasets.

Please, make a team of 2-3 students and fill the table below to insert your wishes for Project Attributions

Groupe ID - **Option Number**	Teams Members (2 to 3)	Project Bid (at least 3)	Project Assignation (TBC)

Groupe ID - **Option Number**	Teams Members (2 to 3)	Project Bid (at least 3)	Project Assignation (TBC)

Each group have 10min of Presentation and Demo + 3-5 min for Q&A

Presentation should include the following

- Problematic / Problem Statement
- Data sources & streams ingestion or Feature Engineering with Kafka
- Models implemented / research papers analysis or Stream Processing
- Experiments & Application results (online ML on river vs batch on scikit-learn))
- Conclusion
- Made the code source & Data set available on private / public Github