Intership or engineering position : Contribution to a python library for temporal sequences analysis

1. Supervision

Co-supervisor: Thomas Guyet
Research group: AIstroSight research team (Inria, UCBL)
Contact: thomas.guyet@inria.fr

Co-supervisor: Etienne Audureau
Research group: CePIA (APHP/UPEC)
Contact: etienne.audureau@aphp.fr

Co-supervisor: Mike Rye
Research group: AIstroSight research team (Inria, UCBL)
Contact: jan-michael.rye@inria.fr

2. Context

The AIRACLES Chair (https://www.bernoulli-lab.fr/project/chaire-ai-racles/) focuses on developing methods and tools for analyzing care pathways. In particular, it is interested in COVID patients to describe and understand their care pathways within the APHP, according to different waves. Data collected by information systems (Electronic Health Records) provide access to rich information on hospital stays for a very large population of hospitalized patients.

This information constitutes their care pathway and is modeled by a temporal sequence of cares, which is a collection of timestamped events. Within the context of the AIRACLES project, we are interested in the longitudinal description of patients. Hence, the temporal information is an important dimension of the data, but it requires data analytics tools that can handle this information. For instance, clustering patients based on their care pathways would require adapting classical clustering approaches to specifically handle the temporal dimension of the data.

There is no standard way to address the problem of timed sequence clustering. Many different techniques have been proposed in the literature. Depending on the analysis to be conducted and on the data characteristics, the analyst may be interested in testing different approaches and choosing the most suitable one. Unfortunately, this would require a lot of effort to compare different implementations coming from different researchers/developers.

This motivates the need to develop a Python library dedicated to the analysis of timed sequences. The objective of this library is to gather a collection of data science tools that can be used to analyze timed sequences. The choice of the Python language fits the current practices in the field of data science but also the environment provided for analyzing EHR data at APHP. This is motivated by the success of the development of libraries such as:

- TraMineR[1], which proposes different methods to analyze state sequences but with the R language.

1 http://traminer.unige.ch/

- tslearn[2] or aeon[3], which are Python libraries for the analysis of time series, but time series data are slightly different from timed sequences.

The success of these time series libraries is based on the quality of the implementations and the wide involvement of the community to integrate their algorithms. The project that we propose aims to offer a well-conceived and well-implemented library that will integrate a few algorithms for timed sequence analysis and, in a second step, to open it to integrate contributions from different teams.

3. Objectives

The objective of the engineer will be to contribute to the conception and development of a Python library dedicated to the analysis of time sequences.

1. You will collaborate with supervisors to propose the overall architecture of the library (defining main packages and organization) and plan the development phases of the project.
2. You will implement the structure of the collaborative development project based on Inria GitLab tools (including documentation generation, testing, continuous integration, etc.).
3. You will implement and test a data structure for timed sequences, including visualization tools and import/export functionalities. This data structure must be flexible, compatible with other storage facilities (such as pandas dataframes), and provide efficient accessibility functionalities.
4. You will select and implement clustering methods for timed sequences. Clustering will be the primary class of methods to be implemented, starting with metric-based clustering techniques.
5. You will contribute to experimenting with the library on real studies of care pathways in collaboration with data scientists at APHP.

4. Profile

You are an engineer with good knowledge in Python library development and good practices for contributing to collaborative development projects (Git, GitLab, continuous integration, documentation generation, package design, etc.). You have from 0 (M2 interns) to 10 years of experience in Python development.

Practical skills with libraries such as Pandas, NumPy, Matplotlib, or Seaborn are expected. Additionally, some knowledge of advanced libraries like scikit-learn, SciPy, and Arrow is considered a plus.

You are interested in contributing to the development of an impactful Python library and value clean, shareable, tested, and well-documented code. Your background knowledge in data science is also welcomed, as you may need to implement algorithms described in scientific articles or research code.

No specific background in timed sequence or care pathways analysis is required.

2 https://tslearn.readthedocs.io/en/stable/
3 https://www.aeon-toolkit.org/en/stable/

5. Important information

- The position is open for hiring as soon as possible and will end by the end of 2024 (the exact length will be determined).
- The Engineer's location will preferably be in Lyon (in the Inria Team AIstroSight), but a location in Paris could be discussed
- The employer will be Fondation APHP
- Interns from engineering school or master 2 in computer science are welcome to apply to this position (a combined internship + few months will be proposed)
- Salary amount depends on your experience

6. Application

To apply to this position, we invite you to send the following documents by email to the three supervisors :

- a CV
- a motivation letter that precises 1) your own professional objectives and how this project could contribute to it ; 2) your interest in the proposed project.
- a link to your github (or source code you developped by your own)
- recommendation letter(s) or contact(s) from your past-experience
- information about your intership dates if your are intern

Applications would be studied on the fly and candidates that would fit the need would be invited for an interview. The position will be closed as soon as we found the right person.