Investigating informative missingness in time series laboratory data

Project Supervisor: Christel Sirocchi

Multivariate time series data collected in practical applications, such as laboratory measurements and vital signs recorded in healthcare settings, are typically sparse and irregularly spaced, reflecting clinicians' decisions to perform specific tests based on a patient's condition. In time series analysis, irregular time series are often temporally discretised into regular sequences, with missing values introduced for unobserved features. Notably, patterns of missing values in time series have been observed to correlate with target outcomes in several clinical prediction tasks. In some cases, models trained solely on missingness patterns have demonstrated predictive performance comparable to models trained on measured feature values. Therefore, the study of unobserved patterns in time series and their predictive potential represents a promising area in clinical machine learning.

This project aims to investigate the predictive value of missingness patterns in cancer research, using laboratory measurements from the MIMIC-IV dataset to predict rare forms of cancer and side effects of chemotherapy treatments. The study will begin by assessing whether missingness patterns alone provide meaningful predictive signals, then identify key predictive features and temporal patterns, and finally investigate how these patterns vary across diseases, treatment stages, and demographics.

The proposed project should address this research question through the following steps:

1. Install and explore the MIMIC-IV database, familiarise with its structure and content, and write SQL queries to extract relevant patient cohorts for different tasks.
2. Engineer features that capture missingness patterns (i.e., features indicating when a measurement was taken without including its value).
3. Train a variety of machine learning models to assess the predictive power of these representations and compare them to models trained on measured values.
4. Analyse missingness patterns associated with different diseases and their progression over time.
5. Investigate differences in missingness patterns across various demographics (e.g., age, gender, ethnicity, insurance type) to identify potential biases in test prescription.

Requirements

To complete this project, the following prerequisites are recommended:

- Knowledge of Python and machine learning concepts.
- Familiarity with UNIX environments and command line usage.
- Willingness to independently dive into the ML literature and test new approaches.

Depending on the results, continuation of the project in the context of a MSc thesis is possible.

References

Che, Z., Purushotham, S., Cho, K., Sontag, D., and Liu, Y. (2018). Recurrent neural networks for multivariate time series with missing values. *Scientific reports*, 8(1):6085.

Giesa, N., Akgül, M., Boie, S. D., and Balzer, F. (2024). Gru-d characterizes age-specific temporal missingness in mimic-iv. *arXiv preprint arXiv:2410.05350*.