Development of machine learning models to process Electronic Health Records- Explainable Models

Lee Lok Hang Toby

2431180L

Proposal

Motivation

Publicly available large scale Electronic Health records (EHR) is an important resource to developing a robust clinical decision system. However, these data are often complex, irregularly sampled, and with a lot of missing values, making the use of this data a challenging task.

Aims

This project aims to compare the effects of preprocessing methods for Electronic Health record using the MIMIC-III database. This includes choosing what information to extract from EHR, develop strategies to reshape data meaningfully, develop imputation strategies for missing values, and developing a state-of-the-art machine learning model to predict in-hospital mortality with combinations of different preprocessing strategies.

Progress

- *Created a local version on the MIMIC-III database*
- Extracted useful information from database using SQL queries
- Preprocessed and reshaped the data
- Applied different imputation methods to fill in missing values
- Trained machine learning models to predict in-hospital mortality
- Compared the performance of using different models, with a few different imputation strategies
- Started applying Deep learning inference models

Problems and risks

Problems

- *Setting up the local postgre database raised errors, tricky fixes are applied to get it to work*
- Panda's version difference caused the extraction notebook to not work properly
- Part of the extraction scripts is from previous projects, might want to alter it so it fits the purpose of this project better.

Risks

- A lot of inference models, different models might work differently with different imputation strategies. **Mitigation:** Will stick to a only a few models, to be decided at the start of next semester
- The effect in the performance for different imputation strategies could be similar. **Mitigation:** Will do background research to try back up and help evaluate the results.

Plan

Semester 2

- Week 1-2: Deep learning for inference
  - Deliverable: Complete deep learning models for predicting in hospital mortality
- Week 2-4: Deep learning for imputation
  - Deliverable: Complete MCMC imputation, and imputation with deep learning
- Week 5-6: Compare and collect results of different imputation methods
  - Deliverable: Start comparing the results between the combinations of different methods in the dissertation
- Week 7-8: Code cleanup, write up GitHub page, and start dissertation
  - Deliverable: Clean up code, write up GitHub readme, wiki page for replicating the results and finish dissertation introduction and background
- Week 7-9: Evaluate and explain different preprocessing strategies used
  - Deliverable: Explain preprocessing strategies used in
  - Compare and evaluate performances of inference models using different combination of preprocessing strategies

- Week 8-10: Write up dissertation
    - Deliverable: First draft submitted to supervisor two weeks before final deadline