CS 760: Machine Learning Spring 2022

PROJECT DESCRIPTION

INSTRUCTOR: DANIEL L. PIMENTEL-ALARCÓN

DO NOT POLLUTE! AVOID PRINTING, OR PRINT 2-SIDED MULTIPAGE.

1 Introduction

This project will challenge you to use the tools learnt on this course (and ideally others) to analyze a real dataset related to *Opportunistic Cardiometabolic Screening*. The **main idea** is to exploit data collected for a specific purpose, and use it for an alternative purpose. For example, using a thoracic X-ray that was collected to treat a broken rib to predict an osteoporotic fracture later on. Specifically, your **goal** is to predict *clinical outcomes* (e.g., death, Alzheimer's, or cancer) using incidental data that typically goes unused/underuse – this includes clinical data and computerized tomography (CT) data.

2 The Dataset

The dataset that you will be using was made available by Perry Pickhardt (Department of Radiology, UW-Madison), who will collaborate designing, supervising and evaluating this project. Without his explicit approval, you are not allowed share this dataset with anyone outside this course, and you may only use it for the purposes of this course. This data will be available to you through Canvas, and it contains the following information about a collection of patients:

- Clinical Data
  - Cols A-C: anonymized Case ID info
  - Col D: Clinical F/U interval [days from CT]
  - Cols E-J: pt BMI, sex, age (at time 0=CT date), smoking/drinking hx)
  - Col K: FRS = Framingham Risk Score (multivariable 10-yr cardiovascular risk score)
  - Cols L-M: FRAX = Fracture risk assessment score (multivariable 10-yr risk for all & hip fx)
  - Col N: Metabolic Syndrome (Y/N/blank=unknown) ? really more of an outcome
- Clinical Outcomes
  - Col P: Death
  - Cols Q-V: Cardiovascular events w/ dates (CVD=stroke, Heart failure, MI=heart attack; any=positive)
  - Col W-X: T2 Diabetes (if dx)
  - Cols Y-AH: Pathologic/osteoporotic fracture w/ date (any=positive; femoral=hip fx)
  - Cols AI-AJ: Alzheimer's Dx

  - Cols AK-AN: Cancer Dx's
  - (Col N: Metabolic Syndrome; could be considered an outocome)
- Computerized Tomography Data
  - Col AP: Bone measure/BMD (L1 HU)
  - Cols AQ-AU: Fat measures (total/visceral/subcutaneous; V/S ratio; all total body X-section)
  - Cols AV-AX: Muscle measures (HU/Area/SMI)
  - Col AY: Aortic Calcification (Ag)
  - Col AZ: Liver fat (HU)

3 General Goals

Use and compare *at least* two methods from this class for the following purposes:

- Predict adverse clinical outcomes using CT data. The main outcome, which you must include in your predictions, is death. All other outcomes are optional but highly encouraged.
- Explore how predictions improve when we use Clinical data in addition to CT data.
- Derive a patient's *biological age* (i.e., relative to actual chronological age) using CT data. (The date of the CT scan is *day zero*, and all other dates are now in days relative to the date of CT).

4 Deliverables

You may work in teams of no more than 3 people. Each team must:

- Send an **email** to the Instructor *and* the TA's indicating the team members. The email must be sent no later than March 30^{th}, and it must have **CS760 Spring 2022 Project Team** as subject (so it is easy to track). Please send only one email per team, cc-ing all team members.

In addition, each team must submit through Canvas no later than May 12^{th} a **.zip** file containing:

- **Source code** or reference to **github** repository with all the code required to replicate experiments and results.
- **Slides** for a 5-minute *flash* talk, which you will present at the end of the semester. Timing will be strictly enforced, and scheduling will be specified by April 15^{th}.
- All **.tex** source files and **.pdf** of a report, in the *Project_Template.tex* format, provided at `https://danielpimentel.github.io/teaching.html`. Your report should be no more than 8 pages long, excluding references and appendix. I *suggest* that your report includes the following sections:
  - Abstract. Summary of the entire project.
  - Introduction. Describe main ideas and goals.
  - Related/Similar work.

- Dataset. Details about where you got it from, who is $\mathbf{x}$, who is y, what is N, what is D, etc.
- Approach. Details of your approach: data cleaning and preprocessing (e.g., strategy to mitigate missing values), method, algorithms, packages, etc.
- Results. Description of experiments, comparisons, and results (tables, plots, etc.)
- Conclusions and Future Work.
- References.

5 Evaluation

You will be evaluated on:

1. (25%) Novelty.
2. (25%) Correctness.
3. (25%) Reproducibility.
4. (25%) Clarity at presenting your ideas, methodology, data, results, etc.