Machine Learning Nanodegree Capstone Project Proposal: Mortality Predictor using MIMIC-III Database

Maxim V. Ivanov

I. BACKGROUND

Over the last decade, there has been a rising interest in adopting large medical databases to store patient records, data from clinical trials as well as medical research data. The ongoing big-data revolution in health care is largely driven by the benefits it brings to patients, physicians and stakeholders. In particular, a traditional approach where physicians used their judgement to make treatment decisions is being replaced by a data-oriented algorithmic approaches. In order for these algorithms to work most effectively, aggregation of individual medical data sets into big-data platforms is of outmost importance. A representative example of a large medical database is the Medical Information Mart for Intensive Care (MIMIC-III) database.[1] The MIMIC-III database is freely available and has been widely used in building predictive models as well as in epidemiological and educational studies.[2,3] Among various models that can be built using the information available in MIMIC-III is the mortality predictor of the patients admitted to intensive care units (ICUs). Availability of accurate models trained on health care data, including mortality prediction, is critically important as it provide health providers with useful review of patient health upon their admission to the hospital, at the bedside and at the discharge. In this proposal, we outline strategies on how the mortality predictor model will be constructed, including creation of training set, feature engineering, choice of models and metrics.

II. PROBLEM STATEMENT

One of the central issues in the data-oriented medical research is reproducibility of medical studies. As Johnson and co-workers put it in their recent work on the reproducibility in critical care research:[4] *"Our results demonstrate that, in spite of best efforts, reproducing cohorts using textual descriptions of patient selection criteria is difficult... More than this, we believe that the public dissemination of open source code is central to facilitating iterative improvement in*

the field." Therefore, the goal of this project is twofold: on one hand, we will demonstrate how machine learning algorithms can be used to build accurate models using health care data on the example of mortality predictor model and, on the other hand, we will focus on the reproducibility of our work. To this end, we will employ the MIMIC Code Repository—a centralized location of concepts related to the critical care research. This will ensure that the training set contains features that represent a domain knowledge and enable reproducibility of the work.

III. DATASET

Here we use MIMIC-III v1.4 database, a freely-available dataset containing deidentified information on more than 60,000 stays in ICUs at the Beth Israel Deaconess Medical Center in Boston, Massachusetts between 2001 and 2012. The database includes information on patients' demographics, vital measurements, laboratory tests, medications, procedures, caregiver notes as well as survival data. In order to access the database, a researcher must complete a course that includes Health Insurance Portability and Accountability Act (HIPAA) requirements and sign an agreement that outlines appropriate data usage and security standards.

IV. SOLUTION STATEMENT

MIMIC-III database contains a detailed demographics and medical data on each patient who stayed at the ICUs and, as such, requires domain knowledge in order to perform feature engineering that is relevant for mortality prediction. To this end, we will utilize MIMIC Code Repository that provides detailed descriptions on how various relevant concepts are defined and extracted from the database. For example, severity of illness score measures the extent of organ system derangement and provides a quantitative assessment of the patient's acuity. There are 5 severity of illness scores implemented in the MIMIC Code Repository, and these scores as well as other quantitative measures available in the repository will be used as features in the training set.

Performance of various models with different complexities will be tested and compared with the available benchmark models. We will start with simple linear classifier model and move on to more complex models based on the ensemble methods and neural networks. As the

benchmark model we will use a recent study by Johnson et al where mortality predictor was built using gradient boosting and logistic regression methods.[4]

To evaluate peformance of our models we will rely on the Area Under the Receiver Operator Characteristic Curve (AUROC)—a common metric in the mortality prediction research. AUROC is equal to the probability that a classifier will rank a randomly chosen positive instance higher than a randomly chosen negative one.

V. PROJECT DESIGN

First, access to MIMIC-III database will be obtained by completing a CITI "Data or Specimens Only Research" course. Once access is granted, we will obtain the database using the protocols available in Amazon Web Services. The protocols include access to the Amazon SageMaker notebook instance where the models will be developed and deployed. Using Amazon Athena, we will execute standard SQL queries in order to access required information from MIMIC-III.

Before proceeding to the model development, we will perform exploratory analysis in order to develop a better understanding of the data. We will first examine demographics data that includes features like patients' age, weight, height, marital status, etc. We will then move towards exploration of the medical data obtained during the first 24 hours, including heart rate, blood pressure, respiration rate, body temperature, etc. Finally, using the SQL queries provided in the MIMIC Code Repository we will access concepts that are related to the patient survival. These concepts include severity of illness scores, organ dysfunction scores, treatments, sepsis and others.

Once the training and test data are obtained, we will examine the performance of the linear classifier to predict whether a patient will survive at ICU given its demographics and medical data. We will use the LinearLearner estimator available in Amazon SageMaker. We will perform hyperparameter tuning and try optimize different metrics using binary_classifier_model_selection_criteria keyword. We will then explore performance of custom models such as XGBoost within scikit-learn module and custom PyTorch classifier. To evaluate performance of the models developed here with those of the available benchmark models we will use AUROC as a metric.

[1] Johnson, Alistair EW; Pollard, Tom J; Shen, Lu; Li-wei, H Lehman; Feng, Mengling; Ghassemi, Mohammad; Moody, Benjamin; Szolovits, Peter; Celi, Leo Anthony; Mark, Roger G MIMIC-III, a freely accessible critical care database *Scientific data* **2016**, *3*, 160035.

[2] Hsu, Douglas J; Feng, Mengling; Kothari, Rishi; Zhou, Hufeng; Chen, Kenneth P; Celi, Leo A The association between indwelling arterial catheters and mortality in hemodynamically stable patients with respiratory failure: a propensity score analysis *Chest* **2015**, *148*, 1470–1476.

[3] Sun, JX; Reisner, AT; Saeed, M; Mark, RG Estimating cardiac output from arterial blood pressure-waveforms: a critical evaluation using the MIMIC II database *Computers in Cardiology* **2005**, page 295.

[4] Johnson, Alistair EW; Pollard, Tom J; Mark, Roger G Reproducibility in critical care: a mortality prediction case study *Machine Learning for Healthcare Conference* **2017**, page 361.