Sprint 2 Plan

Understanding Healthcare Data

Sprint Completion Date:

Revision 1.1, Date: 2/28/2020

Goal: Develop a pipeline to extract a set of features from a large set of synthetic patient data, convert the data to one-hot and embedded formats then run a CNN vs. RNN training on it, in order to determine the factors that lead to a patient being rehospitalized.

User Stories

User Story 1: *As a programmer, I want to be able to extract a chosen set of features from a data set, and convert it into a format able to be fed into a neural network.*

Task 1: Extract data from the provided synthea data stored in json format. (8 points)

- Convert the data features we're looking for from the json files into pandas dataframes.

Task 2: Convert the extracted data into a csv format suitable for use in a neural network. (12 points)

- Take the data from the pandas dataframes and get a list of all snomed codes. Then take those snomed codes and insert them into a csv file showing the conditions that have affected each patient two years prior to their first rehospitalization, month by month.

Total for user story: (20 story points)

User Story 2: *As a patient, I want to be able to predict whether or not I will be re-hospitalized for CHF in the future.*

Task 3: Create a CNN in PyTorch (10 points)

- Take processed data, treat it as an image, and use it to train and test the CNN.
- Split data 50/50 for training and testing
- Create a 3D CNN

Task 4: Create an RNN in PyTorch (10 points)

- Take processed data, treat it a sequence of readings, and use it to train and test the RNN.
- Split data 50/50 for training and testing

Initial Task Assignment:

Shayan Shaikh - task 1, 2

Cassidy Norfleet - task 1, 2

Brendan Reilly-Langer task 1, 2

Aman Prasad - task 3, 4
Perry Yang - task 3, 4
Harshitha Arul Murugan - task 3, 4