Sprint 3 Plan

Understanding Healthcare Data
Sprint Completion Date: 3/20/2020
Revision 1.2, Date: 3/13/2020

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 filter patients based on a time restriction on rehospitalization.*

Task 1: Download 100k dataset and load patients with CHF. (5 points)

- Sift through 100k patients to find only patients with CHF and apply a data pipeline to create an image of patient records.

Task 2: Convert the extracted data into a csv format suitable for use in a neural network. (9 points)

- Use the previously created data pipeline and adapt to handle converting data for large amounts of patients.

Total for user story: (14 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: Improve the CNN in PyTorch (10 points)

- Increase kernel width
- Play around with different functions
- Achieve over 90% accuracy
- Use train_test_split from scikit-learn library

Task 4: Create a correct RNN in PyTorch (10 points)

- Change model to loop over each row in the matrix
- Achieve over 90% accuracy
- Use train_test_split from scikit-learn library

Task 5: Create an SVM in Python(10 points)

- Create and train an SVM using scikit-learn in order to create a baseline

Initial Task Assignment:
Shayan Shaikh - task 1, 2

Cassidy Norfleet - task 1, 2
Brendan Reilly-Langer task 1, 2
Aman Prasad - task 3, 4, 5
Perry Yang - task 3, 4, 5
Harshitha Arul Murugan - task 3, 4. 5