Domain Background

There has been serious advancements in the usage of artificial intelligence in healthcare. Some go as far as to say that we are in midst of the Fourth Industrial Revolution, and a branch of industry that is heavily impacted is healthcare. With the help of artificial intelligence, healthcare and medicine could improve and optimize patient routes and treatment plans, keep better record of a patient's medical history, actively screen using preventative methodologies and detect anomalies and diseases more accurately than any human could.

To demonstrate this merging of artificial intelligence and healthcare, IBM Watson for oncology has been used in various hospitals around the world to suggest cancer treatment plans. A lot of research is currently ongoing to apply machine learning to cancer detection, and so far there have been some success in these studies. Deep learning has been used to classify skin cancer [1], and various contests and research has been performed on developing kernels for classifying Lung Cancer [2].

Problem Statement

Diabetes is a serious problem. Currently, approximately 422 million people in the world have diabetes (type I and type II), and approximately 1.5 million people die due to a form of diabetes every year (an additional 2.2 million deaths were cause by higher-than-optimal levels of blood glucose). Patients, however, with Diabetes can live long and healthy lives if Diabetes is detected and well-managed. However, early diagnosis and intervention is key in living a healthy life with a diabetic condition. Diabetes, if left mismanaged can lead to blindness, amputation and kidney failure [3]. So it is imperative to really keep a good track of diabetes patients from the hospitals' standpoints.

Also with the recent legislature, hospitals are subject to a fine (to be more specific they face a penalty from Medicare) if they discharge a patient and they are readmitted within a 30 day window [4]. So, focusing on retaining the patient if they are likely to be readmitted within the next month is beneficial not just for the patient but also the healthcare institutions. This is an area where machine learning can help. If a kernel can be developed to predict the whether or not a patient will be readmitted in the near future with attributes including medical records, this would save patient lives, as well as levy a lesser financial burden on hospitals. I hence propose experimenting on predicting readmission with a dataset provided by the UCI Machine Learning repository.

Datasets and Inputs

The dataset I wish to investigate, which was collected over the course of 10 years of clinical care at 130 hospitals in the US, consists of patient chart data which include

demographic data such as race, gender, and age, medical record data such as number of lab tests performed on the patient, number of procedures, number of medications administered during the visit, diagnosis values including glucose serum test result, A1c test result, and if there had been a change in the prescription of medication during the encounter. Finally, there is data on whether or not the patient was readmitted or not (categorical in ranges <30 days, > 30 days, and no readmission). A full list of features is provided below.

The dataset has 101765 rows of data with 50 columns as features. The data has a roughly even distribution of the label (readmittance) which sits between 40-50% for 1 – readmitted (including both cases for 30 day readmission and over 30 day readmission), and 50-60% for 0 - not readmitted.

This data is very in line with the problem statement in that it provides plenty of data on the hospital's interaction with a patient with diabetes and data on if the patient was readmitted or not. So it is quite fitting to answer the question of readmission post treatment for a serious diabetic condition.

Solution Statement

I plan on measuring the fitness of various ML models to address this problem of successfully classifying readmission. I hypothesize a logistic regression model may be the best fit for this type of dataset, since I will be classifying the readmission result to a binary one (readmitted / not readmitted), and the number of features are decently sized. My aim for this study is to investigate if utilizing a kernel would be beneficial in predicting readmission that outperforms a simple guess like a flip of a coin.

Benchmark Model

A good baseline to compare is comparing to a random guess with 50% accuracy, as there are no official statistic available that point to how accurate a doctor's educated guess is, but existing studies currently have achieved accuracies ranging from 73% to 63% depending on which pre-existing condition the patient had before an encounter at the hospital [5]. We can compare the results of this study to the bare minimum at 50%, and compare how strong it is relative to other existing literature which lie in the 60-70% range.

Evaluation Metrics

The evaluation metric will primarily be the classification accuracy of the test data. Another metric may be false negative error. It is in both the hospital and patients interest for the patient to not be readmitted. Hence, it is ideal if less patients are predicted not to be readmitted but actually are.

Project Design

I will be mostly doing the work in a Jupyter Notebook using Python 2.7.

Libraries that will most likely be used: NumPy, Pandas, Matplotlib, Sklearn, and potentially other visualization tools

Workflow:

1. First gain a better understanding of the data (data type, average values etc)

2. Assess whether or not some features are relevant. (feature selection) This will involve potentially viewing the correlation between features

3. Split test and training data

4. Fit models (I will be starting with Logistic Regression, SVM, and Random Forest)

5. Assess feature importance and refine parameters for the models

6. Cross validation

7. Grid search to tune parameters for models

8. Evaluate performance

References

[1] https://www.nature.com/articles/nature21056.epdf?referrer_access_token=shuy6MM06sXL_Up3K_gZSNRgN0jAjWel9jnR3ZoTv0NXpMHRAJy8Qn10ys2O4tuPQzN7VitPjMSrm-_eh79EcBa7E8gqt-pkDzYFswnz9xZ5-KXIIz-q1vIeEhxcYyFWeH_8pHiygm9DWyu_08hysBauLS8xsLhzVFYyEQLpF_WX-ahz18wIB3jPkFPChRFtTi9ijo_NvI_2omWmkk3kqEQksepx_FTvN9qjvf1eJLbOCa1YFcF3VQMzgK_045RJI9DcyV26TC53xZgESNhDjDF0lYBgzJemLOn6qKw1Xpc%3D&tracking_referrer=www.technologyreview.com

[2] https://www.kaggle.com/c/data-science-bowl-2017

[3] http://www.who.int/features/factfiles/diabetes/en/

[4] https://www.cms.gov/Medicare/Medicare-Fee-for-Service-Payment/AcuteInpatientPPS/FY2016-IPPS-Final-Rule-Home-Page-Items/FY2016-IPPS-Final-Rule-Tables.html

[5] http://www.sciencedirect.com/science/article/pii/S1532046415000969