Predicting Diagnoses Using Patient EMRS

Gaurab Banerjee

Background

Electronic medical records (EMRs) are the main form of storing patient information in clinical settings. Previous work has used EMRs to establish links between hospital stay duration and mortality[1]. In this project, I predict the diagnosis of the patient and compare different prediction methods in their accuracy. My models output categorical code probabilities.

Data Collection

The data is from the MIMIC-III[2] (Medical Information Mart for Intensive Care III) dataset: aggregated EMRs from Beth Israel Deaconess Medical Center's ICU. I merged on SUBJECT_ID as the common pivot column.

I one-hot encoded all of the categorical variables while using a label binarizer to clean the ICD9_CODE (diagnosis code) column for use as my output set. I split the dataset 80/20 for train and test in all of these methods.

Features

SUBJECT_ID	ETHNICITY	ICD9_CODE	LABEL	212	HEF
1	Cuban	280	heart rate	80	0

Fig. 1 Sample example of inputs and the output, ICD9_CODE.

Feature	Description
ROW_ID	Unique Data Linker
SUBJECT_ID	Admission/Patient ID
HADM_ID	Admission ID
ADMITTIME	Admission Time
DISCHTIME	Discharge Time
DEATHTIME	Time of Death
ADMISSION_TYPE	Elective, Urgent, Newborn, Emergency
ADMISSION_LOCATION	Pre-admit location: Categorical Var.
DISCHARGE_LOCATION	Categorical Var.
INSURANCE	Categorical Var.
LANGUAGE	Categorical Var.
RELIGION	Categorical Var.
MARITAL_STATUS	Categorical Var.
ETHNICITY	Categorical Var.
EDREGTIME	ED Entry
EDOUTTIME	ED Out
DIAGNOSIS	Free text notes
HOSPITAL_EXPIRE_FLAG	Did patient survive to discharge?
HAS_CHARTEVENTS_DATA	Chart populated?
SEQ_NUM	Priority order of ICD9
ICD9_CODE	Patient diagnosis

Models

This was modeled as a reflex-based problem. Since I had several columns of related data, it was evident that these would need to be the features and the output would be the ICD9_CODES or the columns that I wanted to predict. Specifically, I used three primary types of classifiers and each classifier with different feature sets.

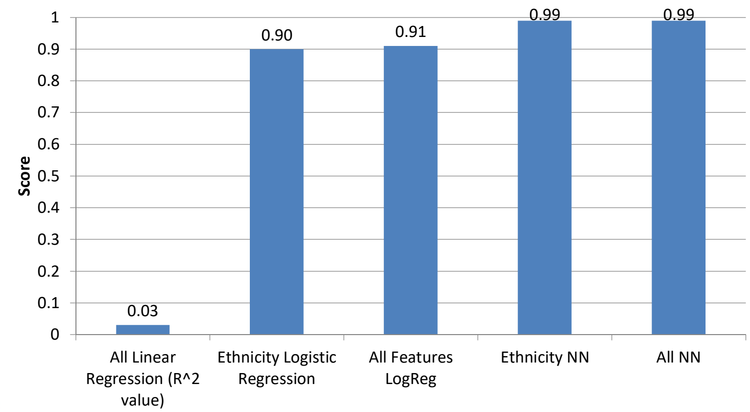
Linear Regression
This is a basic regressor model where the feature vector is given and a dot function is performed with a weight vector to classify using a score.

$$log\frac{p(x)}{1-p(x)} = \beta_0 + x \cdot \beta$$

Logistic Regression
This is a classification model which uses the function identified below. Y=1 when the probability is greater than 0.5.

Multilayer Perceptron Classifier
This is composed of hidden and input layers feeding into a softmax layer. I used hidden layer sizes (5,2)

Results and Discussion



Key Takeaways

- Linear Regression is not optimal
- Due to the high scores of the NN and logistic regression, especially when comparing a single feature to all features, there is likely overfitting occuring
- Why Overfitting?: One theory is that while there are nearly 1.1 million patients in this data, the feature vectors for many of the ICD9_CODES are extremely sparse while others are overrepresented

Future Research

- Use NLP on the free text diagnoses and see if mortality can be predicted
- Use NLP on free text notes to predict ICD9_CODE

References

1. Pirracchio, R., Petersen, M. L., Carone, M., Rigon, M. R., Chevret, S., \& van der Laan, M. J. (2015). Mortality prediction in intensive care units with the Super ICU Learner Algorithm (SICULA): a population-based study. The Lancet Respiratory Medicine, 3(1), 42-52.
2. MIMIC-III, a freely accessible critical care database. Johnson AEW, Pollard TJ, Shen L, Lehman L, Feng M, Ghassemi M, Moody B, Szolovits P, Celi LA, and Mark RG. Scientific Data (2016). DOI: 10.1038/sdata.2016.35. Available from: www.nature.com/articles/sdata201635
3. Previous work done by me