Team 40
Project Topic: Mortality Prediction in ICU
Project Students:

1. Vincent La (vla6@gatech.edu)
2. Avinash Ananthakrishnan (aanantha8@gatech.edu)

Motivation

Why is this problem important? Why do we care of this problem?

Accurate knowledge of patient's clinical state is critical in a clinical setting. Patients in an intensive care unit are particularly in a critical state. Thus, if we can have better accurate mortality prediction within an ICU, we can use this to better understand which patients should be prioritized for clinical care and how we can better allocate resources. Within an ICU, there are many measurements being made from from various devices, but we know that there are both problems in terms of false positives and false negatives. While false negatives might seem to be more problematic, being too sensitive and introducing a lot of false positives also has problems as well. Too many false positives results in information fatigue and providers lose trust in signal as it's hard to separate from noise. By improving mortality prediction using data from EMRs and other clinically-relevant data, we can provide better predictions to improve efficiency and quality of care.

Literature Survey

What kind of approaches have been used? What are the related works?

Siontis et. al (2011) provide an empirical review of methodologies around mortality models. In this paper, the authors use Medline to identify studies published in 2009 that assessed AUC of predicting all cause mortality. What they find is that most studies at the time predicted mortality with modest accuracy, with large variability across populations, which makes sense since for since very high risk populations it may be easier to predict mortality, but for lower risk general populations, the class imbalance is very high with relatively few patients actually dying. A full list of the studies that the authors review and the approaches they use are documented in the paper (linked above). However, we also copy a portion of that table below as an illustrative example:

Disease Clinical Condition	Predictive Model	Variables Used
Cardiovascular Disease	Acute Kidney Injury Network (AKIN)	serum creatinine criteria or urine output criteria

Critical Illness	Acute Lung Injury (ALI) score	chest X-ray, hypoxemia, PEEP, compliance
Cardiovascular disease	Acute Myocardial Infarction in Switzerland (AMIS) model	age, Killip class, systolic blood pressure, heart rate, prehospital cardiopulmonary resuscitation, history of heart failure, history of cardiovascular disease
Critical illness Gastroenterology-related Malignancies Infectious disease Other	Acute Physiology And Chronic Health Evaluation (APACHE) II	temperature, mean arterial pressure, heart rate, respiratory rate, oxygenation or PaO2 , arterial pH, serum sodium, serum potassium, serum creatinine, hematocrit, white blood cell (WBC) count, Clasgow Coma Score

Next, there are a few papers that actually use the MIMIC III data for mortality prediction using ICU data. In Unfolding Physiological State: Mortality Model in Intensive Care Units (Ghassemi et. al), the authors use Latent Variables Models to decompose free-text hospital notes into meaningful features to predict patient mortality. For the actual prediction, they use linear SVM in various setting (e.g. patient first arrives at hospital, In hospital mortality, 30 day mortality, etc.) and they largely find that using hospital notes improved AUCs.

Approach

How will you solve the problem?

What we will do in this project is to try to reproduce and improve models motivated by Unfolding Physiological State: Mortality Model in Intensive Care Units (Ghassemi et. al). In particular, what we find interesting is the use of free-text hospital notes to better predict mortality. This is because in many cases, there are already many mortality models, built both on Business Rules, and machine learning to predict mortality. Perhaps the biggest methodology to improve on existing mortality models is to use unstructured data, to see if there is any additional signal from notes and other text that we don’t already get from structured data using EHRs.

To be more specific, we will use the same dataset as indicated in the Project guidelines, MIMIC III. This data set has ~45,000 patients, ~60,000 Hospital Admissions, ~200,000 Clinical notes, and ~6000 mortality events. We will then divide into 80% training, 20% test set. Similar to what is presented in the paper, we will first extract clinical baseline features, including age, sex, and SAPS-II score, from the database for every patient. We will then perform topic modeling, perhaps LDA, to do natural language processing on the notes to extract meaningful features.

In terms of evaluating our model, we will predict on multiple iterations of the outcome variable, just as the paper does. For example, to evaluate 24 hour mortality predictions, in the training set, we will create a 24 hour prediction window and using the data that is transformed into features, we will measure AUC of mortality where outcome variable is mortality within 24 hours. We will then test on held-out test data set to report final test AUC.

Data

Describe the dataset you use. e.g. descriptive stats, preliminary results(play with sample data as the minimum requirement), etc. If you'd like to do MIMIC related project like NLP, start to request data from MIT ASAP, ***see more instructions @524 @501***

As described above, we will use the MIMIC III dataset which contains ~45,000 patients, ~60,000 Hospital Admissions, ~200,000 Clinical notes, and ~6000 mortality events.

Experimental Setup

Describe your project environment. e.g. software stack you use, on which hardware spec, etc.

In terms of software stack, we will use HIVE to store and process the data. We will use Pig to implement models and Hadoop for MapReduce jobs. On hardware specs, we will run on a Docker container on a Macbook Pro (2015)

- Processor: 2.2 GHz Intel Core i7
- Memory: 16 GB

The Docker Container is the same as was recommended in the Homeworks using Bigbox: https://github.com/yuikns/bigbox.