CS 3750 final projects

The final projects for the class uses MIMIC III database (https://www.nature.com/articles/sdata201635). MIIMIC-III is a database of Electronic Health Records (EHRs) for patients admitted to ICUs of Beth Israel Deaconess Medical Center in Boston, Ma, which is one of the teaching hospitals in Boston area. We already used MIMIC-III for the midterm project. In that case we used a subset of data in the database based on lab results, medication administrations and procedures events for a subset of patients. Your final project is more open ended and you can utilize information in other tables in the database, such as diagnosis, vital signs, patient demographics etc.

MIMIC III

Briefly, MIMIC data consists of different kinds of events and readings. For example, labs represent the readings of specific test values obtained from the sample (e.g. lab tests on the blood sample). Medications represent administrations of the different medications in time and their dosages. Vital signs are readings of the most critical values reflecting body functions such as blood pressures, heart rate, respiratory rates, or temperatures. Procedures cover surgeries and other important procedures. In addition, the records include diagnosis, admission and discharge information for each patient. Finally, there are text reports written by clinicians that reflect the status of the patient. These are referred to as progress notes (or progress reports). Other text-based entries in the database include various reports interpreting some investigative procedures such as X-ray, CT. For a list of different types of information recorded in MIMIC-III dataset please see: https://www.nature.com/articles/sdata201635/tables/3

Possible ML problems to look at

The objective of the final project is to try and test some of the methodologies used for modeling lower dimensional embeddings, time series and sequences, and deep neural nets or their combinations for solving various machine learning problems on the MIMIC data. What follows are examples of two important problems one can define on MIMIC III data:

Event prediction. One type of problem very important in clinical practice is adverse event prediction. Basically, given a current time t we want to make prediction of an adverse event in the future given the information only up to time t. The future can be limited to a fixed or a variable time window T, that is, we predict the occurrence of an event in between (t,t+T]. Example of prediction problems may include:

- Discharge from ICU (can we accurately predict within a day, two days, or next 6 hours that the patient will discharged from the ICU to other hospital unit)
- Death prediction (can we predict within a day, or next 12 hours that the patient will die).
- Sepsis prediction. Can we predict that the patient will develop or is developing sepsis? We can use administration of the intravenous antibiotics as a proxy to detection of sepsis).
- Acute kidney injury. Can we predict the patient will develop or is developing serious kidney problem where the kidneys stopped to function properly (we can measure this indirectly in terms of high creatinine lab values).

Time series modeling. Another problem is related to the development and learning of a model that lets us predict the patient states (or various observations) in the future. For example, we can consider a sequence of sets of daily observations defined by a collection of observed lab values, medications, etc

and we want to build a model that cab predict observations we expect to see in the next day or in two days. We can also restrict the predictions to just one or a few observations, say prediction of blood pressures, or values of some labs, and we can also change the time discretization defining the sequences and observations (say 1 day, or six hours intervals). The past events can be based on events or can take into account the real-valued time series.

Possible connection to the midterm project

Patient state abstractions. One of the important lessons we learned from the midterm project is that a lower dimensional representation of patient events in EHR obtained via SVD, LSA or CBOW can help us to predict some aspects of the patient case. These types of representations can be extended also to temporal representations that work in time. That is, we can view events that occurred on a specific day to define the patient state for just that day. Defining a lower dimensional embedding of these daily events can help us to model the evolutions or dynamics of a patient state in time and connect these lower dimensional representations using a time series model such Autoregressive Process, Markov process etc. Such models can provide important context for both the prediction of lower level events (administration of the new medication next day) or target adverse events such a prediction of a sepsis, an acute kidney injury or a death. An alternative is to define the patient state dynamics in terms of a hidden state space that is defined directly by the time series models like HMM, LDS, or LSTM, or by combining the two ideas, e.g. by using an HMM, or LDS where observations are defined by a lower dimensional embeddings defined say by the SVD.

Text based abstractions. In the midterm project we focused our attention on lab, medication, and procedure events recorded in the EHR. However, in EHR we have other types of information, including free text entries. One important type of text-based information are progress notes that represent physicians' summaries of the patient for that day. So this text information can be informative of the current patient state and one can use it to define the abstraction of the patient state for that day and also feed it to a time series model defined upon such an abstraction.

Project rules

The final project is a group project with the maximum group size of three. Smaller groups are OK. Each group should select a leader who similarly to midterm project can communicate with the instructor on organizational issues.

Data access. All students in the class should be able to see the mimic_iii database when you log in to the CS database server. You have only read privileges to this database. Please report ASAP any problem with the access to the database to the instructor. In addition, each group will be allocated a group specific database space which you can use to store auxiliary data, results, models etc.

Project timeline

- Noon, Tuesday, April 7, 2020. Forming of the groups, selection of the leaders. A brief description of the chosen project.
- Online project presentations, Thursday, April 23, 2020
- Final project reports due at noon on Friday, April 24, 2020.