CSC 591/791: ML-User
Assigned Project 1
Fall, 2019

This is the assigned project on missing data handling in healthcare domain. Please read and follow the following instructions.

- **Important Dates (all due at 11:45PM on specified dates)**:
  - Baseline (3D-MICE) code and instruction and results on training data available: **Aug 30th**
  - Test data available: **Sep 23rd**
  - Submit your imputed test data and code: **Sep 25th**
  - Leaderboard is available: **Sep 26th**.
  - Submit your project presentation: **Sep 27th, Classroom** (Details later)
  - Submit your project report: **Sep 27th** (Details later)
- **Submissions**:
  - Data and code: submit only a **single zip file** named "AP1_G[group#].zip" containing your imputed test data in a folder named "test_imputed_G[group#] with same format as the original test data, your codes, and a readme file with running instructions.
  - Report and presentation (both **PDF only**): clearly list your team ID, each team member's **names and Unity IDs** at the top of your submission.
- Only submit **one single** pdf file or zip file per group.

Problem Definition: This project involves missing data imputation for longitudinal ICU laboratory test data. A key challenge in clinical data mining is that most clinical datasets contain missing data. Since many commonly-used machine learning algorithms require complete datasets with no missing data, clinical analytic approaches often entail an imputation procedure to "fill-in" missing data. However, although most clinical datasets contain a temporal component, most commonly used imputation methods do not directly accommodate longitudinal, time-series-based clinical data.

Task: In this project you will develop machine learning models to impute missingness in a clinical dataset of longitudinal multivariable laboratory test results. You are free to design your own algorithms or exploring existing ones. The dataset consists of test results for 13 commonly measured analytes (clinical laboratory tests). To evaluate imputation method performance, data are randomly masked for selected test results at varying time points. Your task is to impute these masked results alongside results missing from the original data. The benchmarking performance metric is evaluated by comparing predicted results to actual results for masked data points only.

Dataset: Our training dataset is a synthetic dataset of 6,000 patients. Each patient's history contains a varying number of records stored in each row, and the values for 13 clinical lab tests are stored in columns. Each record is associated with a time stamp (minutes from arrival/admission) and time interval between two consecutive records are irregular.
You will be provided with two versions of training samples, where each folder will contain records of patients stored in [pid].csv format. "train_with_missing" includes data with original and random missingness, and "train_groundtrurh" includes the original values we eliminated to be used as ground truth for evaluation purposes. "pts.tr.csv" and "naidx.csv" include patient IDs and the index of masked value per each patient, respectively. These two files are needed when running the baseline. Later, you will be provided with a test dataset of ∼ 2,200 patients with original and masked missingness called "test_with_missing".

Baseline: A baseline model called 3D-MICE is provided to compare your results against this model. The goal in this project is to come up with an algorithm that can beat 3D-MICE. 3D-MICE is developed in R language and the readme file includes instructions on how to run it. Note that time complexity of this algorithm is not efficient to run on all training samples on a personal computer, and takes ∼ 18 hours to run. Therefore, in addition to the code, we have provided the performance of this model on the training set for you to have a baseline performance to compare against. Feel free to run this algorithm on smaller subsets of training data, but be advised that this might result in poor performance.

Evaluation: Evaluation is to be run on held-out test data. Please make sure your training model(s) are ready before Sep 23rd so you can directly evaluate your trained model(s) on the test dataset that contains ∼ 2,200 samples with randomly masked missingness. Note that you only have two days to run your trained model(s) on the test set and submit the imputed test data, **in the same format**, along with your codes and readme file. The ground truth for test set will be the measure to evaluate your methods and will not be accessible for you. Later, we will use your imputed test results and compare it with test ground truth values and rank the groups based on their performance. The evaluation metric that will be used in this project is normalized root mean square deviation (nRMSD) calculated by the following formula:

$$nRMSD(a) = \sqrt{\frac{\Sigma_{p,i} I_{p,a,i} (\frac{|X_{p,a,i} - Y_{p,a,i}|}{range(Y_{p,a})})^2}{\Sigma_{p,i} I_{p,a,i}}} \tag{1}$$

where $X_{p,a}$ represents test result predictions for analyte a of patient p, $Y_{p,a}$ represents the actual measured values of this analyte, and we can index corresponding values from $X_{p,a}$ and $Y_{p,a}$ using an index i based on the order of records. Function $I(.)$ is an indicator of whether this value was a randomly masked missing or not. This metric is evaluated per analyte.

Report: In your report you need to include the following:

- Preliminary statistics of the dataset
- Complete description of the method you used, along with the motivation to use, preprocessing steps, and any important detail or observation in applying your algorithm

- Report your training nRMSD performance
- Copy your rank and testing nRMSD performance from the Leaderboard.