CSE6250: Big Data Analytics in Healthcare
Homework 1

Jimeng Sun

Deadline: Jan 21, 2018 Anytime on Earth
(i.e., 7AM Jan 22, 2018, Eastern Time)

- You can use 2 days of late submission throughout the semester **ONLY for homeworks**
- Discussion is encouraged, but each student must write his/her own answers and explicitly mention any collaborators.
- Each student is expected to respect and follow the GT Honor Code.
- Please type the submission with LaTeX or Microsoft Word. We **will not** accept hand written submissions.
- Please **do not** change the filenames and function definitions in the skeleton code provided, as this will cause the test scripts to fail and subsequently no points will be awarded. Built-in modules of python and the following libraries - pandas, numpy, scipy, scikit-learn can be used.

Overview

Preparing the data, computing basic statistics and constructing simple models are essential steps for various data analysis work. In this homework, you will use clinical data as raw input to perform **Mortality Prediction**. For this homework, **Python** programming will be required. See the attached skeleton code as a start-point for the programming questions. Also, you need to make a single PDF file (*homework1_answer.pdf*) of a compiled document for non-programming questions.

It is your responsibility to make sure that all code and other deliverables are in the correct format and that your submission compiles and runs. We will not manually check your code. Thus non-runnable code will directly lead to 0 score.

1 CITI Certification [10 points]

Georgia Tech policy requires that all personnel involved with human subjects research must pass a training course before doing so. This requirement encompasses all types of interactions with human subjects, including the analysis of data.

Follow the steps below to complete the CITI certification and attach your two certifications.

1. Go to https://www.citiprogram.org/
2. Login via Single Sign On (SSO), which will allow to login using your Georgia Tech username and password
3. Select Georgia Institute of Technology as the authentication provider
4. Once logged in, under Georgia Institute of Technology courses, click on "Add Course or Update Learner Groups"
5. Now you will have three main courses to select. You will check the box for "Human Subjects Research"
6. Click next, then you will select the radio button "NO, I have NOT completed the basic course"
7. Now, you will see three learner groups. You are required to complete Group 1 and Group 2. Let us start with Group 1 (select Group 1) and click next
8. Good Clinical Practice is not required so select "N/A", then click next
9. Health Information Privacy and Security (HIPS) is required, click "CITI Health Information Privacy and Security (HIPS) for Biomedical Research Investigators"
10. Now under Georgia Tech courses you will have "Group 1 Biomedical research Investigators and Key Personnel" listed as incomplete. You will have to go through every tutorial in that course and complete a quiz for each.
11. Once you have completed and passed Group 1, repeat the steps above to complete Group 2 (Social / Behavioral Research Investigators and Key Personnel)

Solution. Sample certification displayed in Figure 1.

□

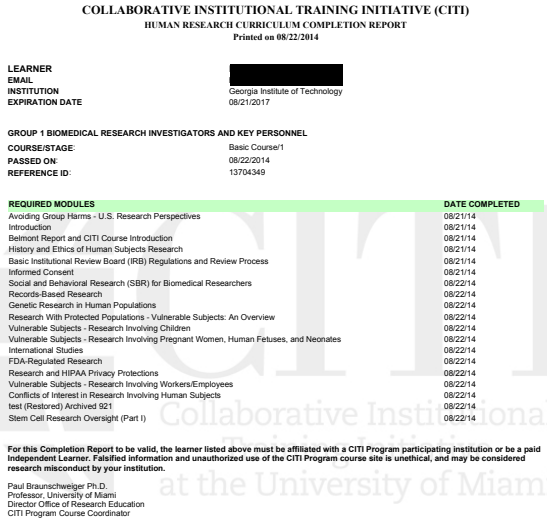
About Raw Data

Navigate to *homework1/data/train*. There are three CSV files which will be the input data in this assignment.

The data provided in *events.csv* are event sequences. Each line of this file consists of a tuple with the format *(patient_id, event_id, event_description, timestamp, value)*.

For example,

```
1053,DIAG319049,Acute respiratory failure,2924-10-08,1.0
1053,DIAG197320,Acute renal failure syndrome,2924-10-08,1.0
1053,DRUG19122121,Insulin,2924-10-08,1.0
```



COLLABORATIVE INSTITUTIONAL TRAINING INITIATIVE (CITI)
HUMAN RESEARCH CURRICULUM COMPLETION REPORT
Printed on 08/22/2014

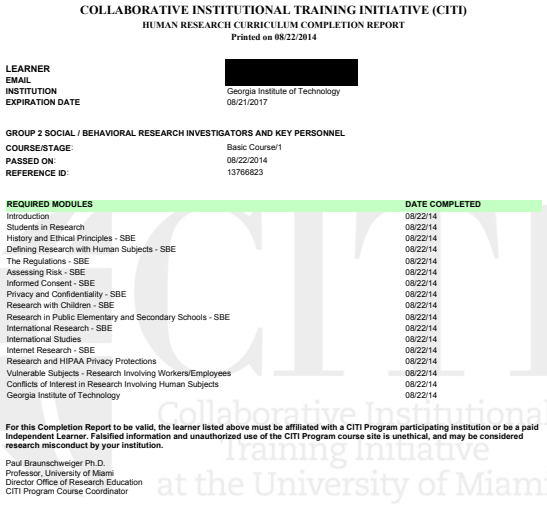
LEARNER [redacted]
EMAIL
INSTITUTION Georgia Institute of Technology
EXPIRATION DATE 08/21/2017

GROUP 1 BIOMEDICAL RESEARCH INVESTIGATORS AND KEY PERSONNEL
COURSE/STAGE: Basic Course/1
PASSED ON: 08/22/2014
REFERENCE ID: 13704349

REQUIRED MODULES	DATE COMPLETED
Avoiding Group Harms - U.S. Research Perspectives	08/21/14
Introduction	08/21/14
Belmont Report and CITI Course Introduction	08/21/14
History and Ethics of Human Subjects Research	08/21/14
Basic Institutional Review Board (IRB) Regulations and Review Process	08/21/14
Informed Consent	08/21/14
Social and Behavioral Research (SBR) for Biomedical Researchers	08/22/14
Records-Based Research	08/22/14
Genetic Research in Human Populations	08/22/14
Research With Protected Populations - Vulnerable Subjects: An Overview	08/22/14
Vulnerable Subjects - Research Involving Children	08/22/14
Vulnerable Subjects - Research Involving Pregnant Women, Human Fetuses, and Neonates	08/22/14
International Studies	08/22/14
FDA-Regulated Research	08/22/14
Research and HIPAA Privacy Protections	08/22/14
Vulnerable Subjects - Research Involving Workers/Employees	08/22/14
Conflicts of Interest in Research Involving Human Subjects	08/22/14
test (Restored) Archived 921	08/22/14
Stem Cell Research Oversight (Part I)	08/22/14

For this Completion Report to be valid, the learner listed above must be affiliated with a CITI Program participating institution or be a paid Independent Learner. Falsified information and unauthorized use of the CITI Program course site is unethical, and may be considered research misconduct by your institution.

Paul Braunschweiger Ph.D.
Professor, University of Miami
Director Office of Research Education
CITI Program Course Coordinator



COLLABORATIVE INSTITUTIONAL TRAINING INITIATIVE (CITI)
HUMAN RESEARCH CURRICULUM COMPLETION REPORT
Printed on 08/22/2014

LEARNER [redacted]
EMAIL
INSTITUTION Georgia Institute of Technology
EXPIRATION DATE 08/21/2017

GROUP 2 SOCIAL / BEHAVIORAL RESEARCH INVESTIGATORS AND KEY PERSONNEL
COURSE/STAGE: Basic Course/1
PASSED ON: 08/22/2014
REFERENCE ID: 13766823

REQUIRED MODULES	DATE COMPLETED
Introduction	08/22/14
Students in Research	08/22/14
History and Ethical Principles - SBE	08/22/14
Defining Research with Human Subjects - SBE	08/22/14
The Regulations - SBE	08/22/14
Assessing Risk - SBE	08/22/14
Informed Consent - SBE	08/22/14
Privacy and Confidentiality - SBE	08/22/14
Research with Children - SBE	08/22/14
Research in Public Elementary and Secondary Schools - SBE	08/22/14
International Research - SBE	08/22/14
International Studies	08/22/14
Internet Research - SBE	08/22/14
Research and HIPAA Privacy Protections	08/22/14
Vulnerable Subjects - Research Involving Workers/Employees	08/22/14
Conflicts of Interest in Research Involving Human Subjects	08/22/14
Georgia Institute of Technology	08/22/14

For this Completion Report to be valid, the learner listed above must be affiliated with a CITI Program participating institution or be a paid Independent Learner. Falsified information and unauthorized use of the CITI Program course site is unethical, and may be considered research misconduct by your institution.

Paul Braunschweiger Ph.D.
Professor, University of Miami
Director Office of Research Education
CITI Program Course Coordinator

(a) Group 1 (b) Group 2

Figure 1: Sample CITI Certification

```
1053,DRUG19122121,Insulin,2924-10-11,1.0
1053,LAB3026361,Erythrocytes in Blood,2924-10-08,3.000
1053,LAB3026361,Erythrocytes in Blood,2924-10-08,3.690
1053,LAB3026361,Erythrocytes in Blood,2924-10-09,3.240
1053,LAB3026361,Erythrocytes in Blood,2924-10-10,3.470
```

- **patient_id**: Identifies the patients in order to differentiate them from others. For example, the patient in the example above has patient id 1053.
- **event_id**: Encodes all the clinical events that a patient has had. For example, DRUG19122121 means that a drug with RxNorm code as 19122121 was prescribed to the patient. DIAG319049 means the patient was diagnosed of disease with SNOMED code of 319049 and LAB3026361 means that the laboratory test with a LOINC code of 3026361 was performed on the patient.
- **event_description**: Shows the description of the event. For example, DIAG319049 is the code for Acute respiratory failure and DRUG19122121 is the code for Insulin.
- **timestamp**: Indicates the date at which the event happened. Here the timestamp is not a real date but a shifted date for protecting privacy of patients.
- **value**: Contains the value associated to an event. See Table 1 for the detailed description.

event type	sample event_id	value meaning	example
diagnostic code	DIAG319049	diagnosed with a certain disease, value always be 1.0	1.0
drug consumption	DRUG19122121	prescribed a certain medication, value will always be 1.0	1.0
laboratory test	LAB3026361	test conducted on a patient and its value	3.690

Table 1: Event sequence value explanation

The data provided in *mortality_events.csv* contains the patient ids of only the deceased people. They are in the form of a tuple with the format *(patient_id, timestamp, label)*. For example,

```
37,3265-12-31,1
40,3202-11-11,1
```

The timestamp indicates the death date of a deceased person and a label of 1 indicates death. Patients that are not mentioned in this file are considered alive.

The *event_feature_map.csv* is a map from an event_id (SNOMED, LOINC and RxNorm) to an integer index. This file contains *(int_id, event_id)* pairs for all event ids.

Python and dependencies

In this homework, we will work on Python 2.7 environment. If you do not have a python distribution installed yet, we recommend installing Anaconda (or miniconda) with Python 2.7. We provide *homework1/environment.yml* which contains a list of libraries needed to set environment for this homework. You can use it to create a copy of conda 'environment' (http://conda.pydata.org/docs/using/envs.html#use-environment-from-file). If you already have your own Python development environment (it should be Python 2.7), please refer to this file to find necessary libraries.

Running the tests

Test cases are provided for every module in this assignment and operate on a subset of the data. To run a test, execute the following commands from the base folder i.e. homework1. If any of the test cases fail, an error will be shown. For example to test the statistics computed, the following command should be executed:

nosetests tests/test_statistics.py

A single test can also be run using this syntax:

nosetests tests/$<$ *filename* $>$**:**$<$ *test_method* $>$

2 Descriptive Statistics [10 points]

Before starting analytic modeling, it is a good practice to get descriptive statistics of the input raw data. In this section, you need to write code that computes various metrics on the data described previously. A skeleton code is provided to you as a starting point.

The definition of terms used in the result table are described below:

- **Event count**: Number of events recorded for a given patient. Note that every line in the input file is an event.
- **Encounter count**: Count of unique dates on which a given patient visited the ICU. All the events - DIAG, LAB and DRUG - should be considered as ICU visiting events.
- **Record length**: Duration (in number of days) between the first event and last event for a given patient.
  For example, if the first event is on 2014-05-21 and the last event is on 2014-05-24, the duration is 3 days. If a patient has only one event, the duration is 0.

a. Complete *src/event_statistics.py* to implement the required statistics.

Please be aware that **you are NOT allowed to change the filename and any existing function declarations.** Only numpy, scipy, scikit-learn and other built-in modules of python will be available for you to use. The use of *pandas* library is suggested.

b. Use *events.csv* and *mortality_events.csv* provided in **data/train** as input and fill Table 2 with actual values. Include this table in *homework1_answer.pdf*

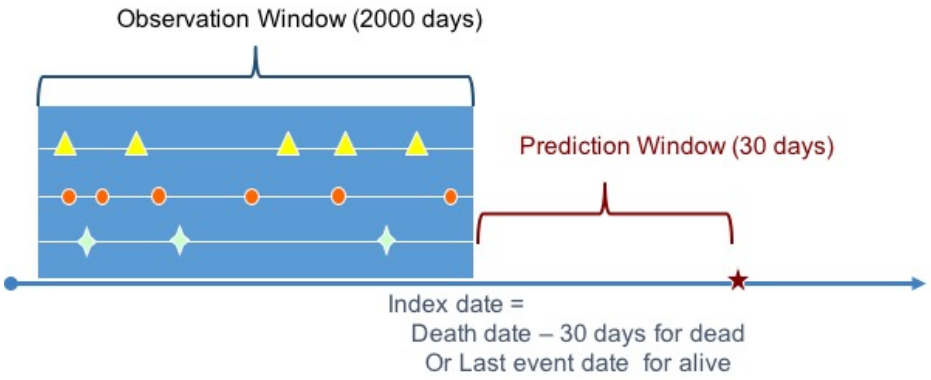
Deliverable: src/event_statistics.py, homework1_answer.pdf [10 points]

3 Feature construction [30 points]

It is a common practice to convert raw data into a standard data format before running real machine learning models. In this section, you will work on *src/etl.py* file and implement the necessary python functions in this script. You will work with *events.csv*, *mortality_events.csv* and *event_feature_map.csv* files provided in **data/train** folder. The use of *pandas* library in this question is recommended. Listed below are a few concepts you need to know before beginning feature construction (for details please refer to lectures).

Metric	Deceased patients	Alive patients	Function to complete
Event Count 1. Average Event Count 2. Max Event Count 3. Min Event Count			event_count_metrics
Encounter Count 1. Average Encounter Count 2. Max Encounter Count 3. Min Encounter Count			encounter_count_metrics
Record Length 1. Average Record Length 2. Max Record Length 3. Min Record Length			record_length_metrics

Table 2: Descriptive statistics for alive and dead patients



- Observation Window: The time interval you will use to identify relevant events. Only events present in this window should be included while constructing feature vectors. The size of observation window is 2000 days.
- Prediction Window: A fixed time interval that is to be used to make the prediction. Events in this interval should not be included in constructing feature vectors. The size of prediction window is 30 days.
- Index date: The day on which mortality is to be predicted. Index date is evaluated as follows:
  - For deceased patients: Index date is 30 days prior to the death date (timestamp field) in *data/train/mortality_events.csv*.
  - For alive patients: Index date is the last event date in *data/train/events.csv* for each alive patient.

Step - a. Compute the index date [8 points]
Use the definition provided above to compute the index date for all patients. Complete the method *calculate_index_date* provided in *src/etl.py* .

Deliverable: *src/etl.py, deliverables/etl_index_dates.csv*

Step - b. Filter events [5 points]
Consider an observation window (2000 days) and prediction window (30 days). Remove the events that occur outside the observation window. Complete the method *filter_events* provided in *src/etl.py* .

Deliverable: *src/etl.py*, deliverables/etl_filtered_events.csv

c. Aggregate events [10 points]
To create features suitable for machine learning, we will need to aggregate the events for each patient as follows:

- **sum** values for diagnostics and medication events (i.e. event_id starting with DIAG and DRUG).
- **count** occurences for lab events (i.e. event_id starting with LAB).

Each event type will become a feature and we will directly use event_id as feature name. For example, given below raw event sequence for a patient,

```
1053,DIAG319049,Acute respiratory failure,2924-10-08,1.0
1053,DIAG197320,Acute renal failure syndrome,2924-10-08,1.0
1053,DRUG19122121,Insulin,2924-10-08,1.0
1053,DRUG19122121,Insulin,2924-10-11,1.0
1053,LAB3026361,Erythrocytes in Blood,2924-10-08,3.000
1053,LAB3026361,Erythrocytes in Blood,2924-10-08,3.690
1053,LAB3026361,Erythrocytes in Blood,2924-10-09,3.240
1053,LAB3026361,Erythrocytes in Blood,2924-10-10,3.470
```

We can get feature value pairs(*event_id*, *value*) for this patient with ID *1053* as

```
(DIAG319049, 1.0)
(DIAG197320, 1.0)
(DRUG19122121, 2.0)
(LAB3026361, 4)
```

You will notice there are certain events with no entries in the values column. Handle these missing values by removing all events with null values while constructing the features. Next, replace each *event_id* with the *feature_id* provided in *data/train/event_feature_map.csv*

```
(708, 1.0)
(306, 1.0)
(2475, 2.0)
(3030, 3.35)
```

Further, in machine learning algorithm like logistic regression, it is important to normalize different features into the same scale using an approach like min-max normalization (hint: $min(x_i)$ maps to 0 and $max(x_i)$ 1 for feature x_i). Complete the method *aggregate_events* provided in *src/etl.py* .

Deliverable: *src/etl.py* and deliverables/etl_aggregated_events.csv

d. Save in SVMLight format [7 points]
If the dimensionality of a feature vector is large but the feature vector is sparse (i.e. it has only a few nonzero elements), sparse representation should be employed. In this problem you will use the provided data for each patient to construct a feature vector and represent the feature vector in SVMLight format.

```
<line> .=. <target> <feature>:<value> <feature>:<value> #<info>
<target> .=. +1 | -1 | 0 | <float>
<feature> .=. <integer> | qid
<value> .=. <float>
<info> .=. <string>
```

For example, the feature vector in SVMLight format will look like:

```
1 2:0.5 3:0.12 10:0.9 2000:0.3
0 4:1.0 78:0.6 1009:0.2
1 33:0.1 34:0.98 1000:0.8 3300:0.2
1 34:0.1 389:0.32
```

where, 1 or 0 will indicate whether the patient is alive or dead i.e. the label and it will be followed by a series of feature-value pairs sorted by the feature index (idx) value.

Deliverable: *src/etl.py*, deliverables/featured_svmlight.train and deliverables/features.train

4 Predictive Modeling [45 points]

4.1 Model Creation [15 points]

In the previous section, you constructed feature vectors for patients to be used as training data in various predictive models (classifiers). Now you will use this training data (deliverables/featured_svmlight.train) in 3 predictive models.

a. Implement Logistic Regression, SVM and Decision Tree. Skeleton code is provided in *src/models_partb.py, src/models_partc.py.*

b. Report all the performance metrics on the training data (deliverables/featured_svmlight.train). Skeleton code is provided in *src/models_partb.py.* You will evaluate and report the performance of your predictive models based on the metrics listed in Table 3. Include this table in *homework1_answer.pdf*

Model	Accuracy	AUC	Precision	Recall	F-Score
Logistic Regression					
SVM					
Decision Tree					

Table 3: Model performance on training data

c. Evaluate your predictive models on a separate test data that is provided to you in *data/features_svmlight.validate* (binary labels are provided in that svmlight file as the first field). Skeleton code is provided in *src/models_partc.py.* You will evaluate and report the performance of your predictive models based on the metrics listed in Table 4. Include this table in *homework1_answer.pdf*

Model	Accuracy	AUC	Precision	Recall	F-Score
Logistic Regression					
SVM					
Decision Tree					

Table 4: Model performance on test data

d. Based on the performance metrics on training and test data, please propose some strategies to improve the test performance and also provide the justification for your recommendation. For example, the strategies can be "gather more training data" or "do parameter tuning more on the algorithms".

Deliverable: src/models_partb.py, src/models_partc.py, homework1_answer.pdf [15 points]

4.2 Model Validation [10 points]

In order to fully utilize the available data and obtain more reliable results, machine learning practitioners use cross-validation to evaluate and improve their predictive models. In this section, you will demonstrate using two cross-validation strategies against Logistic Regression.

- K-Fold: Divide all the data into k groups of samples. Each time $\frac{1}{k}$ samples will be used as test data and the remaining samples as training data.
- Randomized (Shuffle Split): Iteratively random shuffle the whole dataset and use top specific percentage of data as training and the rest as test.

a. Implement the two strategies in *src/cross.py*. Use k=5 for K-Fold. Use a test data percentage of 0.2 and 5 for the number of iterations for Randomized.

b. Report the average Accuracy and AUC in Table 5. Include this table in *homework1_answer.pdf*

CV strategy	Accuracy	AUC
K-Fold		
Randomized		

Table 5: Cross Validation

NOTE: You will use the features that you constructed in Section 3 as the entire dataset for this problem.

Deliverable: src/cross.py, homework1_answer.pdf [10 points]

4.3 Self Model Creation [15 points]

In this part, you will choose your own predictive model and set of features to attempt to obtain better performance compared to what you just worked on. You are advised to try out different things to improve the performance of your predictive model. One may try out new features, or use feature selection techniques to reduce the feature set, or tune the parameters of the predictive model or try ensemble techniques. However, one **must not** change the observation window and prediction window.

You should use the data available in *data/train* to construct features and train your predictive model. It is advisable to use cross validation and AUC as the metric to determine the relative performance of your model. Your final model will be evaluated on a (blind) test set for which the labels are unknown to you. The events in the observation window corresponding to the test patients are available in *data/test/events.csv*. If you are

using the same features as in Section 3 for the test set, you may use the feature map in *data/test/event_feature_map.csv*.

a. Implement your predictive model in *src/my_model.py*. You are free to use your own features and predictive model. Please ensure that your predicted labels are either 0 or 1. Report your features of the test patients in *deliverables/test_features.txt*. Submit your predictions in the form of a csv file (patient_id, predicted label) in *deliverables/my_predictions.csv*

b. Write a short paragraph on your best predictive model (based on cross validation and AUC) and the other models that you tried. What was the rationale behind your approach? Did your model perform better than in the previous section? Include this in *homework1_answer.pdf*

Deliverable 1: deliverables/test_features.txt
(Refer to the skeleton code for the required format)
Deliverable 2: src/my_model.py
Deliverable 3: deliverables/my_predictions.csv

4.4 Kaggle [5 + up to 5 bonus points]

Kaggle is a platform for predictive modelling and analytics competitions. Submit your *deliverables/my_predictions.csv* to a kaggle competition created specifically for this assignment to compete with your fellow classmates. A *gatech.edu* email address is required to participate; follow the sign up directions using your university email address or if you already have an account, change the email on your existing account via your profile settings so you can participate. Make sure your display name (not necessarily your username) matches either your actual full name or your GT account username so your kaggle submission can be linked back to T-Square.

Evaluation criteria is AUC. The predicted label is a soft label, which represents the possibility of mortality for each patient you predict. The label range is between 0 and 1. 50% of the data is used for the public leaderboard where you can receive feedback on your model. The final private leaderboard will use the remaining 50% of the data and will determine your final class ranking. More specific details can be found on the kaggle competition website.

Score at least 0.6 AUC to receive 5 points of credit. Additional bonus points can be received for your performance according to the following:

- Top 10%: 5 bonus points
- Top 15%: 4 bonus points
- Top 20%: 3 bonus points
- Top 25%: 2 bonus points
- Top 30%: 1 bonus point

Percentages are based on the entire class size, not just those who submit to kaggle.

5 Submission [5 points]

The folder structure of your submission should be as below. You can use the *tree* command to dispay and verify the folder structure is as seen below. **All modified src code should be in src folder. You should not change the given function names in source code. All other unrelated files will be discarded during testing and you will get ZERO score for Submission part.**

```
<your gtid>-<your gt account>-hw1
|-- homework1_answer.pdf
|-- src
   |-- event_statistics.py
   |-- etl.py
   |-- models_partb.py
   |-- models_partc.py
   |-- cross.py
   |-- utils.py
   \-- my_model.py
|-- deliverables
   |-- etl_index_dates.csv
   |-- etl_filtered_events.csv
   |-- etl_aggregated_events.csv
   |-- features_svmlight.train
   |-- features.train
   |-- test_features.txt
   \-- my_predictions.csv
```

Create a tar archive of the folder above with the following command and submit the tar file.

```
tar -czvf <your gtid>-<your gt account>-hw1.tar.gz \
  <your gtid>-<your gt account>-hw1
```

Example submission: 901234567-gburdell3-hw1.tar.gz