CSE8803: Big Data Analytics in Healthcare
Homework 1

Deadline: 11:55 PM EST, February 10, 2015

- 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 GT Honor Code.
- Please type the submission with LaTeXor Microsoft Word. We don't accept hand written submission.

Overview

Accurate knowledge of a patient's disease state is critical. Electronic monitoring systems and health records provide rich information for making prediction. In this homework, you will use MIMIC2 ICU clinical data to predict the 1-year after discharge mortality of patients as an indicator of patient's severity.

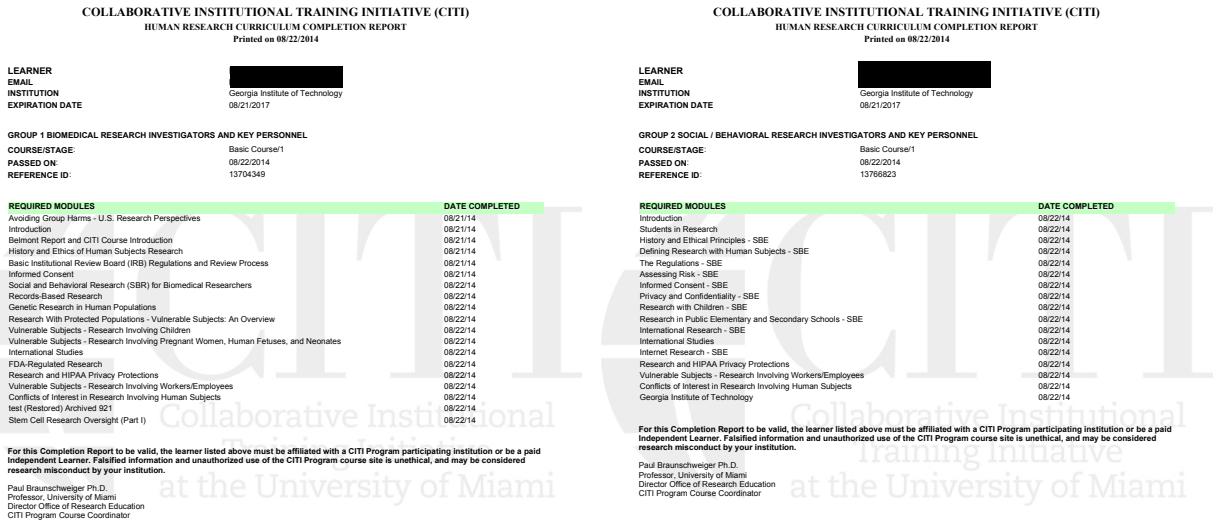
1 Certification [10 points]

Dealing with medical datasets like MIMIC2, you are required to complete an online training course to make sure you know the basic ethics and rules. Follow the steps below to complete the certification and attach your certification.

1. Go to https://www.citiprogram.org/
2. Login via SSO (Single Sign On). SSO 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 not required so select "N/A", click next
10. Select "RCR for engineering"
11. 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.
12. Once you completed and passed Group 1, repeats the steps above to complete Group 2 (Social / Behavioral Research Investigators and Key Personnel)

Solution. Sample certification displayed in Figure 1.



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

LEARNER [redacted]
EMAIL [redacted]
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

(a) Group 1

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

LEARNER [redacted]
EMAIL [redacted]
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

(b) Group 2

Figure 1: Sample CITI Certification

□

2 Logistic Regression [20 points]

With a set of historical healthcare data, you could train a Logistic Regression classifier to make prediction. A training set D is composed of $\{(\mathbf{x}_i, y_i)\}_1^N$, where $y_i \in \{0, 1\}$ is the

label, and $\mathbf{x}_i \in \mathbf{R}^d$ is the feature vector of the i-th patient. In logistic regression we have $p(y = 1|\mathbf{x}, \mathbf{w}) = \sigma(\mathbf{w}^T\mathbf{x})$, where $\sigma(t) = \frac{1}{1+\exp -t}$ is the sigmod function.

a. Show the negative *log*-likelihood for D, simplify it as much as you can [5 points]

b. Show the gradient of the negative *log*-likelihood in terms of $\mathbf{w}$ [5 points]

c. Show the Hessian of the negative *log*-likelihood [5 points]

d. In order to use gradient descent to find best $\mathbf{w}$, you need to make sure the negative *log*-likelihood is convex. Proof Hessian in question c is positive definite [5 points]

3 Stochastic Gradient Descent [20 points]

In previous problem, you studied Logistic Regression learning in batch style. In this problem, you will approach the problem with another method. Suppose your system continuously collects patient data and predicts their severity using **Logistic Regression**. When patient data vector $\mathbf{x}$ arrive your system, the system need to predict whether he/she is in severe condition(predicted label $\hat{y} \in \{0, 1\}$) and need immediate care or not. The prediction will be delivered to physician and physician will look into the patient. Finally, the physician will provide a feedback(real label $y \in \{0, 1\}$) to your system, so that the system could be updated to make better prediction in the future. In this problem, you are going to derive the equation behind this setting.

a. Show the log likelihood l of a $(\mathbf{x}_t, y_t)$ pair. [5 points]

b. Show how to update the coefficient vector $\mathbf{w}_t$ when you get patient feature vector $\mathbf{x}_t$ and physician feedback label y_t at time t using $\mathbf{w}_{t-1}$(suppose learning rate η is given). ?[5 points]

c. What's the time complexity of the update rule from b if $\mathbf{x}_t$ is very sparse? [2 points]

d. Briefly explain the consequence of too large η and too small η. [3 points]

e. Show how to update $\mathbf{w}_t$ under the penalty of L2 norm regularization. In other words, update $\mathbf{w}_t$ according to $l - \mu\|\mathbf{w}\|^2$, where μ is a constant. What's the time complexity? [5 points]

f. When you use L2 norm, you will find each time when you get a new $(\mathbf{x}_t, y_t)$ you need to update every element of vector $\mathbf{w}_t$ even when $\mathbf{x}_t$ only have very few non-zero elements. Show how to update $\mathbf{w}_t$ lazily(hint: update i-th element of $\mathbf{w}_t$, $\mathbf{w}_{ti}$, only when i-th element of $\mathbf{x}_t$, $\mathbf{x}_{ti}$, is non-zero, and you have to read this paper a paper). [Extra 5 points]

4 Programming [50 points]

First, follow the instructions to install Cloudera VM if you haven't done that yet. Then you need to download data from AWS S3. We have made the files public, so that you could download them as follows

```
wget http://s3.amazonaws.com/cse8803/ICD9EVENT.csv
wget http://s3.amazonaws.com/cse8803/MORTALITYEVENT.csv
wget http://s3.amazonaws.com/cse8803/ICD9_FEATURE_MAP.csv
wget http://s3.amazonaws.com/cse8803/training.data
wget http://s3.amazonaws.com/cse8803/testing.data
```

4.1 Transform data [10 points]

It's a common practice to convert raw data into some common data format before conducting real machine learning study. If the dimensionality of a feature vector is large but the feature vector is sparse (i.e., only a few nonzero elements), sparse representation should be employed. In this problem you will use ICD9 diagnostic code for each patient to construct the feature vector and represent the feature vector in SVMLight format.

You will use diagnostics events of patients. Specifically, each patient is associated with a list of events, and you need to aggregate the same events that satisfy certain criteria into one feature. Some concept you need to know (for details please refer to lectures):

- Diagnosis date: The day at which event of interest you are to predict. In this homework, you are predicting whether a patient is deceased or not at specific date.
- Observation Window: The time interval you will use to screen events.
- Prediction Window: A fixed time interval you will use to make prediction.

Suppose a patient was deceased in 2010, while the observation window is 5 years and prediction window is 2 years, you will use diagnostic events between 2004 and 2008 to predict the mortality.

You will work on following files in *pig* folder

- **etl.pig**: Modify this scripts as you need. The sample implementation doesn't consider observation window and prediction window.
- **utils.py**: Implement necessary python User Defined Function (UDF) in this file (optional).

You will need to use following data files you downloaded from S3:

- **MORTALITYEVENT.csv** contains the target label which is one-year after discharge mortality indicator. Each line contains a tuple $(p, label, date)$, where p is patient id, label is 0 (alive) or 1 (deceased) and *date* is the date after 1 year of discharge.

- **ICD9EVENT.csv**: Each line contains a tuple $(p, icd9\_code, date)$, which indicates patient p was diagnosed with $icd9\_code$ at a given $date$. Because a patient can visit the hospital several times and diagnosed with the same ICD9 code, you will need to aggregate $(p, code)$ using COUNT (a pig command).
- **ICD9_FEATURE_MAP.csv**: You need to map a feature (ICD9 code in this problem) from its name to an index. This file contains $(index, name)$ pairs for ICD9 codes you will use.

Further, in machine learning algorithm like logistic regression, it is important to normalize different feature into the same scale. Modify **etl.pig** to consider observation window(1000 days) and prediction window(365 days). Then, scale feature values using min-max normalizationthis approach. Please do not change the input/output location and format. (hint: $min(x_i)$ is zero in this problem)

Run your pig script in local mode, you will need the output for the next question.

```
pig -x local etl.pig
```

Deliverable: pig/etl.pig and pig/utils.py [10 points]

4.2 SGD Logistic Regression [20 points]

In this question, you are going to implement your own Logistic Regression classifier in python using the equations you derived in question **3.e**. To help you start, a skeleton of code will be provided. Find related files in *lr* folder. You will train and test a classifier by running

1. cat path/to/training/data | python train.py -f 5676
2. cat path/to/testing/data | python test.py

You will use the two sets of training/testing data. One comes from output of previous problem. Another pair of training/testing data set is downloaded from S3 (training.data and testing.data).

To better understand the performance of your classifier, you will need to use standard metrics like AUC. A script file named **install-conda.sh** is provided to help you install necessary modules for drawing ROC curve. The script is tested in Cloudera VM. You may need to modify it if you want to install that somewhere else. Restart the terminal after installation.

a. Update the **lrsgd.py** file. You are allowed to add extra methods, but please make sure the existing method names and parameters remain unchanged. Use standard modules of Python 2.7 only, as we will not guarantee the availability of any third party modules while testing your code. [14 points]

b. Show the ROC curve generated by test.py in this writing report for different learning rates η and regulization parameters μ combination and briefly explain the result. [3 points per dataset]

[Extra 5 points] Implement using result of question **3.f**, and show the speed up. You could use a larger data set *s3://cse8803/train.large.data*, and the number of features is 299135.

4.3 Hadoop [20 points]

In this problem, you are going to train multiple logistic regression classifiers using your implementation of previous problem with Hadoop in parallel. The pseudo code of Mapper and Reducer are listed as Algo 1 and Algo 2 respectively. Find related files in *lr* folder.

input : Ratio of sampling r, number of models M, input pair (k, v)
output: key-value pairs

```
1 for i ← 1 to M do
2 |   m ←Random;
3 |   if m < r then
4 |   |   Emit (i, v)
5 |   end
6 end
```

Algorithm 1: Map function

input : (k, v)
output: Trained model

```
1 Fit model on v;
```

Algorithm 2: Reduce function

You need to copy *training.data* into HDFS using command line.

```
hdfs dfs -mkdir hw1
hdfs dfs -mkdir hw1/training
hdfs dfs -put training.data hw1/training/
```

a. complete the **mapper.py** according to pseudo code. [6 points]
You could train 5 ensembles by invoking

```
hadoop jar \
   /usr/lib/hadoop-mapreduce/hadoop-streaming.jar \
  -D mapreduce.job.reduces=5    \
  -files lr \
  -mapper "python lr/mapper.py -n 5 -r 0.4" \
  -reducer "python lr/reducer.py -f 5676" \
  -input hw1/training \
  -output hw1/models
```

Notice that you could apply other parameters to reducer. To test the performance of ensembles, copy the trained models to local via

```
hdfs dfs -get hw1/models
```

b. Complete the **testensemble.py** to generate the ROC curve. [6 points]

```
cat testing.data | python lr/testensemble.py -m models
```

c. Compare the performance with that of previous problem. [8 points]

4.4 Submission

The folder structure of your submission should be as below. You could display fold structure using *tree* command. All other unrelated files will be discarded during testing.

```
<your gtid>-<your gt account>-hw1
|-- homework1answer.pdf
|-- lr
|   |-- lrsgd.py
|   |-- mapper.py
|   |-- testensemble.py
\-- pig
    |-- etl.pig
    \-- utils.py
```

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
```