Assignment III – CS 8803 Data Analytics for Well-Being

Topic	Linking Academic Performance with Sensed Health Related Data
Grade	Max 60 points; 10% of overall grade (late policy applies)
Due	May 3, 2016, 2:05pm Eastern Time
What to hand in	Code and a report with answers to the different questions
Where to submit	T-Square
Useful resource	Python's statsmodels library (http://statsmodels.sourceforge.net/) Python's scikit-learn library (http://scikit-learn.org/stable/) [You can also use your favorite other library, tool or software]

Goal

This third and final assignment will test your understanding and skills toward analyzing mobile health sensing based data, relating to health and well-being. Our dataset is a sample from the publicly available dataset made available by the authors of the StudentLife project [1]. The main task is to build and fit a linear regression model of academic performance using a variety of self-reported and sensing data at the per individual level.

Data

The enclosed folder contains a sub-folder called "GroundTruth" that includes a csv file on grades. Use the overall grade (field "gpa all") of each individual (indicated in the "uid" field) as the ground truth label of each individual in the regression model (the dependent variable).

The enclosed folder then contains two more folders – "Surveys" and "SensingData". The data inside these folders will be used as features in the regression model (the independent variables). Detailed description of these data are presented below.

The "Surveys" folder has csv files, each corresponding to one standardized mental health questionnaire – they are the Flourishing Scale, Loneliness Scale, PANAS, Perceived Stress Scale, and PHQ-9. Each line in each of the csv files is an individual (indicated in the "uid" field); there are several columns starting with the third column, each of which corresponds to the response given by an individual on a particular question of the survey. Some of these responses are numeric, some are categorical; in case of the latter please assign categorical values to discrete numbers (e.g., in the Perceived Stress Scale, assign 0 = Never, 1 = Almost Never, 2 = Sometimes, 3 = Fairly Often, 4 = Very Often). The field "type" in the csv files indicates when the survey was taken – the StudentLife folks took these survey responses from individuals at the start of the study ("pre") and at the end ("post"). For the purposes of using these survey responses as features in the regression model, use the average value of each question across the pre and post responses of an individual.

The second features related folder is "SensingData" and it contains several subfolders, each corresponding to a type of sensed data of the individuals in the StudentLife study.

i) The "CallLog" subfolder includes a csv file each for data of an individual. Each row indicates information on a call placed from the mobile phone of the individual – timestamps are in Unix timestamp format and correspond to Eastern Standard Time.
ii) The "Conversations" file also includes csv files for each individual. Within each file (that corresponds to an individual), there are two fields: conversation start timestamp and conversation end timestamp. For example, a row could indicate that the individual was around a conversation

from Unix timestamp 1364425656 to Unix time stamp 1364425727. Each row is a separate and unique conversation the individual was involved in during the period of the study.

iii) The "PhoneLight" folder includes a csv file for each individual, as above. Within each file (that corresponds to an individual), each row indicates a record of when the phone was at a dark environment for a significant long time (>=1 hour). There are two fields in each data file: start timestamp and end timestamp, indicating the period of time the phone was present in a dark environment during the entire period of the study.

iv) The "PhoneLock" folder data files includes records of when the phone was locked for a significant long time (>=1 hour). There are two fields in each data file (corresponding to an individual): start timestamp and end timestamp, indicating the duration when the phone was locked during the entire period of the study.

v) The "Activity" folder contains data files of information around various forms of physical activity each individual was involved in during the period of the study. Each row in each csv file indicates the timestamps of specific activities recorded for an individual (using the phone's accelerometer), where the activity type ("activity inference" field) is recorded as a numeric value with the following meanings: 0 = Stationary; 1 = Walking; 2 = Running; = Unknown.

vi) The "Wifi_Location" folder includes csv files for the individuals, and consists of information about the individuals' rough on-campus location (in the Dartmouth Campus). Each row in a file records the time of an individual's presence in a certain location ("location" field) spanning the entire duration of the study.

For further information about the data, please refer to the StudentLife project page: http://studentlife.cs.dartmouth.edu/dataset.html

Task A

(30 points; 6 points for each feature category) Using the survey and sensed data above, construct the following feature or independent variable categories for the linear regression framework:

i) *Mental Well-being:* The responses to each question of each mental health survey (inside "Surveys" folder) would be a feature value for an individual. The survey feature vector will thus have length corresponding to the sum of the questions in all of the five surveys.

ii) *Social Engagement:* For an individual, this feature vector will consist of: the total number of calls made in the call log file of the individual (inside "CallLog" folder); the total number of conversations (in the "Conversations" folder); the mean duration of the conversations; and the standard deviation of the duration of conversations over the entire period of the study.

iii) *Mobility:* For an individual, use the data in the "Wifi_Location" folder to get the following feature vector: total number of locations collected; and the number of unique locations collected over the entire period of the study.

iv) *Physical Activity:* For an individual, use the data inside the "Activity" folder to construct the following feature vector: mode (most frequent) activity; and the proportion of activity that is running/walking spanning the entire period of the study.

v) *Phone (Non)-Use:* For an individual, use the data inside the "PhoneLight" and "PhoneLock" folders to construct the following feature vector of phone (non)-use: mean duration when phone was in a dark environment; standard deviation of the duration when phone was in a dark environment; mean phone lock duration; and the standard deviation of phone lock duration, spanning the entire period of the study.

In your submission, include five tsv or csv files: one corresponding to each feature category type. Within each file, each line would correspond to the feature vector of each individual, with rows as individuals and columns the various feature values. Treat all missing feature values as 0's.

Task B

(20 points) Use the above feature categories to fit five different linear regression models, each of which predicts the academic performance (or "gpa all" field inside the grades file in "GroundTruth"). Let us represent the matrix of feature vectors of individuals (of each category) as X, and the gpa grades as y. Thereafter, use the statsmodels Python package, and obtain model fit and results:

```python
import statsmodels.api as sm
model = sm.OLS(y, X)
results = model.fit()
print(results.summary())
```

In your assignment submission, include the five model outputs given by the above function summary() for the five feature categories. Here is a link to details of the above code snippet: http://statsmodels.sourceforge.net/devel/examples/notebooks/generated/ols.html

Task C

(10 points) Present a discussion (1 page) of the performance of the above model fits based on the "Adj. R-squared" metric given by the summary() function. Here, lower values are better – 0 indicates the model fit for the particular feature category was the poorest, 1 indicates it perfectly fit the gpa grade data, and values between 0 and 1 indicate everything in between. Specifically discuss your rationale behind why one model (i.e., one feature category type) performs better than the other.

References

1. Wang, Rui, Fanglin Chen, Zhenyu Chen, Tianxing Li, Gabriella Harari, Stefanie Tignor, Xia Zhou, Dror Ben-Zeev, and Andrew T. Campbell. "StudentLife: Assessing Mental Health, Academic Performance and Behavioral Trends of College Students using Smartphones." In *Proceedings of the ACM Conference on Ubiquitous Computing*. 2014.