*Probability and Statistics for Data Science   
Stony Brook University  
CSE594 - Spring 2016*

Assignment 3: Prediction and Clustering

Assigned: 4/11/2016; Due: 4/21/2016, 3:45pm EST

**PART A. Team Project Proposal (50 points).**

2 pages (strictly enforced) describing the motivation and plan for your research project. Specifically this document must include:

* Title, Authors (team members)
* **Motivation** for the project (why would someone care?) -- 2 to 3 paragraphs
* Description of the **approach** for the task you are tackling -- 2 to 3 paragraphs
* Description of how you are utilizing each of the **course 3 big course topics** (i.e. Probability Theory, Discovery, Prediction or Clustering). -- 1 to 2 sentences on each.
* A **results figure** mockup with a caption. What is going to be your main result? How will you communicate the result? (e.g. bar plots, tag clouds, scatter plot, density plots...etc..).
* A **timeline** for when you hope to have 3 to 4 key objectives completed.

**Only one team member needs to submit Part A.** There will be a separate submission section on blackboard for this.

**PART B. Programming Prediction and Clustering (50 points)   
-- independent project.**

**Dataset:** 2015 County Health Rankings (same as assignment 1)

<http://www.countyhealthrankings.org/sites/default/files/2015%20CHR%20Analytic%20Data.csv>

A list of important factors related to health of members of communities.

**NOTE:** Besides counties,this csv contains state totals. Such rows should be filtered out when working with the data, so that you only work with county data.

**Your code** should be self-contained in a file called **“A3\_FULLNAME.py”** which can be run in python 2.7 by typing: python a2\_studentid.py “2015 CHR Analytic Data.csv”.

**Python libraries permitted:** default file IO libraries, csvreader, pandas, numpy, scipy.stats, sklearn.   
**Academic Integrity:** As with all assignments (sans the team project), although you may discuss concepts with others, you must work independently and insure your work and code is not visible to any classmates.

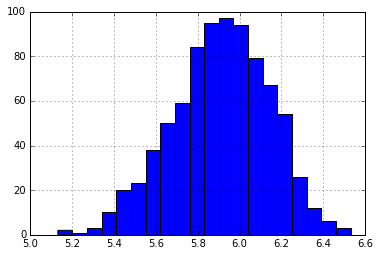
**1. Read in the CSV,** and limit the rows and columns as follows (5 points):

1. Only counties with at least 30,000 people according to “2011 population estimate Value”.
2. Only columns ending in “Value” (plus the county code). To check, here is the complete list:  
   ['COUNTYCODE’, 'Premature death Value', 'Poor or fair health Value', 'Poor physical health days Value', 'Poor mental health days Value', 'Low birthweight Value', 'Adult smoking Value', 'Adult obesity Value', 'Food environment index Value', 'Physical inactivity Value', 'Access to exercise opportunities Value', 'Excessive drinking Value', 'Alcohol-impaired driving deaths Value', 'Sexually transmitted infections Value', 'Teen births Value', 'Uninsured Value', 'Primary care physicians Value', 'Dentists Value', 'Mental health providers Value', 'Preventable hospital stays Value', 'Diabetic screening Value', 'Mammography screening Value', 'High school graduation Value', 'Some college Value', 'Unemployment Value', 'Children in poverty Value', 'Income inequality Value', 'Children in single-parent households Value', 'Social associations Value', 'Violent crime Value', 'Injury deaths Value', 'Air pollution - particulate matter Value', 'Drinking water violations Value', 'Severe housing problems Value', 'Driving alone to work Value', 'Long commute - driving alone Value', '2011 population estimate Value', 'Population that is not proficient in English Value', 'Population living in a rural area Value', 'Diabetes Value', 'HIV prevalence rate Value', 'Premature age-adjusted mortality Value', 'Infant mortality Value', 'Child mortality Value', 'Food insecurity Value', 'Limited access to healthy foods Value', 'Motor vehicle crash deaths Value', 'Drug poisoning deaths Value', 'Uninsured adults Value', 'Uninsured children Value', 'Health care costs Value', 'Could not see doctor due to cost Value', 'Other primary care providers Value', 'Median household income Value', 'Children eligible for free lunch Value', 'Homicide rate Value']  
   (COUNTYCODE is the unique id for each row).
3. Only counties with non-nan values for each value column. (keep zeros though)  
     
   **Output:** print the total number of counties:

“1. TOTAL NUMBER OF COUNTIES: 823“

**2. Create a new column,** ‘log\_paamv’ which is the log transform of 'Premature age-adjusted mortality Value'. This will be the outcome, response variable to predict. (5 points)

**Output:** produce a histogram to a png

“2. log\_paamv HISTOGRAM: 2histogram.png”  


**3. Predict “log\_paamv” (*y*)** using all “Value” variables **except REMOVED\_COLUMNS (listed below)** as your predictors/features (*X*). Implement 10-fold cross-validation, and use multiple linear regression, **without regularization**, to predict “log\_paamv”. You may use statistical or other machine learning packages to help, except for implemented the 10-fold cross validation (make sure regularization is effectively “turned off”). To check your answer versus below, you must standardize the data. (10 points)  
REMOVED\_COLUMNS = ['COUNTYCODE', 'log\_paamv', 'Premature age-adjusted mortality Value', 'Premature death Value', 'Uninsured adults Value', 'Teen births Value', 'Food insecurity Value', 'Physical inactivity Value', 'Adult smoking Value', 'Injury deaths Value', 'Motor vehicle crash deaths Value', 'Drug poisoning deaths Value', 'Child mortality Value', 'Uninsured Value']

**Output:** print the mean squared error of the predictions across the 10 folds:

“3. Non-regularized Linear Regression MSE: 0.124”

**4. Run PCA** over all of the predictors from 3 (all columns of *X*) and all counties (i.e. no nfold cross-validation). (10 points)   
*Hint:* sklearn.PCA requires that X is standardized in order for “explained\_variance\_ratio\_” to be correct.   
 **Output:** Print the **percentage of** variance explained for each of the first 3 components.

“4. Percentage variance explained of first three components: [0.281, 0.139, 0.079]”

**5. Run regularized predictions** using: (a) principal component regression (i.e. use PCA on *X* and then linear regression fitting the components to *y*), (b) L2 regularized linear regression, and (c) L1 regularized linear regression to predict y from X. Use 10-fold cross-validation, making sure to use the development set to (a) choose the best number of components, and (b,c) choose an ideal penalty for regularization. (20 points)

**Output:** Print the 10-fold test MSEs:  
 “5. a) principal components regression mse: XXX.XXX”

“5. b) L2 regularized mse: 0.121”

“5. c) L1 regularized mse: 0.122”

**Testing.** We will remove a random subset of 5 to 15 counties from the csv and test on that. Please make sure your code works for such versions of the csv.

**Hints:**

* Implement 10-fold cross-validation as a method that you can call for both 3 and 5, that either produces folds or runs specified models across folds.
* You may use sklearn or scipy.stats for regularized linear regression implementations and PCA.
* Don’t expect your MSEs to be exactly as above. If they are more than 5% larger or smaller then make sure you are standardizing the variables before step 3.
* You may use any code in the assignment 1 or 2 key (i.e. for reading the csv).
* If you use pandas dataframes, you can use “drop” to remove the columns specified above.
* If you are not managing to get mses close to the above, check that you are not selected a “boundary alpha”during cross-validation. You might need to try alphas larger or smaller. Powers of 10 ranging from 10-5 to 105 will often work.

**Further Learning:** To see a larger effect of regularization, limit your data to ~100 counties.