Machine Learning for Cities - Homework assignment 1

(Ravi Shroff, 3/31/2016)
DUE on 4/12/2016 by 11:59pm EST

This homework assignment concerns NYPD police stops in New York City in the years 2011-2012 where the suspected crime was CPW (criminal possession of a weapon). You can read more about these stops at
https://en.wikipedia.org/wiki/Terry_stop
https://en.wikipedia.org/wiki/Stop-and-frisk_in_New_York_City
(we will be working with this dataset for the next homework assignment as well).
You may use any language/software package of your choice to complete this assignment. Note that Python's scikit-learn package has a module called sklearn.cluster (http://scikit-learn.org/stable/modules/classes.html#module-sklearn.cluster) and documentation (http://scikit-learn.org/stable/modules/clustering.html#clustering) that should get you started.

1) You are given 1,069 stop records in the file 'original_with_duplicates_noid.csv', but unfortunately technical errors have resulted in some exact duplicate records. You know there are around 1,000 unique stops in the file (there may be slightly more or slightly fewer).
    a) Use a clustering technique of your choice to determine how many unique records there are.
    b) Check your work by looking for row-wise duplicates using any technique of your choice (this should be straightforward). How well did your clustering technique perform?
    c) Write a paragraph explaining what you did in parts a) and b)
2) You are given 1,100 stop records in the file 'original_with_errors_noid.csv', but unfortunately technical errors have resulted in some almost-duplicate records (that is, there are some records that are duplicates of others except for slight variations in two features). You know there are around 1,000 unique stops in the file (there may be slightly more or slightly fewer).
    a) Use a clustering technique of your choice to determine how many unique records you think there should be.
    b) Which are the two features where errors were introduced?
    c) Write a paragraph explaining what you did in parts a) and b)
3) You are given records for all recorded CPW stops made in New York City during 2012 in the file 'cpw_stops_2012.csv'. Each record includes the lat/long of the stop and the month, day, and time period (this feature has six values, where each corresponds to a four-hour time period. For example, time.period = 1 means the stop occurred between midnight and 4am, and time.period = 2 means the stop occurred between 4am and 8am). Explore "hot-spots" of CPW stops by applying clustering methods based on space, or space and time. In particular,
    a) Apply at least two different clustering algorithms to the data

b) Plot some clusters on a map of New York City

c) Write at least two paragraphs explaining which clustering methods you chose and the parameters you used (and why you used those parameters). Do your results make sense? **[This question is deliberately open-ended. You will be evaluated on whether or not you made an effort to understand the data, successfully applied different clustering algorithms, and made at least one plot]**