TTE 6608: Algorithms and Models for Smart Cities
Assignment #1
Due Date: 09/21/2017
Total Points 100

Instructions

- Upload your source code and written report in Webcourse before the due date. No late submissions will be accepted.
- Python is the preferred language for solving the programming problems. If you want to use R or Matlab then please talk to me first.
- If you use Python, you may upload your Jupyter Notebooks.
- Program files should be named after the problem (e.g. solution to problem 1 should be problem1.ipynb etc). Include detailed instructions on how to run your code.

Problem 1 (30 pt) Download the taxi cab data of the New York City for the month of October 2012 from here http://www.nyc.gov/html/tlc/html/about/trip_record_data.shtml. Plot the data points of daily taxi trips for each day between October 24 to October 30 (7 days). You should use the datashader tool. We recommended a tutorial on datashader in lecture 4 (see at the end of Data_exploration_visualization.ipynb).

In addition to visualizing the taxi cab data, separately plot the following items:

- A distribution of the number of daily trips within this period.
- The pdf and cdf of trip distance, taxi fare and tips
- Plot a scatter matrix of all the reasonable variables (exclude variables such as vendor_id, datetime, and coordinates)

Briefly describe the most interesting findings. Hurricane Sandy made landfall on October 29, 2012; did you find any interesting trends in taxicab data due to Sandy?

Problem 2 (10 pt) Let X be a random variable which denotes if a patient has AIDS and let $T1$ and $T2$ be the outcomes of two clinical tests with the following error profiles:

$p(t_1\|x)$	$x = HIV-$	$x = HIV+$
$t_1 = HIV-$	0.91	0.07
$t_1 = HIV+$	0.09	0.93

$p(t_2\|x)$	$x = HIV-$	$x = HIV+$
$t_2 = HIV-$	0.98	0.01
$t_2 = HIV+$	0.02	0.99

Compute

1. $p(x = +|t_1 = +, t_2 = +)$
2. $p(x = +|t_1 = +, t_2 = -)$
3. $p(x = +|t_1 = -, t_2 = +)$
4. $p(x = +|t_1 = -, t_2 = -)$

You may assume that the prior probability $p(x = +)$ is 0.005

Problem 3 (30 pt) You can read about the ECML/PAKDD discovery challenge 2006 which dealt with email spam detection here: http://www.ecmlpkdd2006.org/challenge.html. Your task is to download the dataset for task A from http://www.ecmlpkdd2006.org/data_task_a.zip. **Implement** and train a Naive Bayes classifier using the data found in task_a_labeled_train.tf file. You can test the performance of your classifier on the data found in task_a_u00_tune.tf . In your report, briefly describe how you approached the problem, what results you obtained, what practical difficulties you faced, and how you overcame these difficulties.

Problem 4 (30 pt) You can read about the bikesharing dataset here https://archive.ics.uci.edu/ml/datasets/bike+sharing+dataset. This dataset records the number bikes rented in Capital bikeshare system in Washington D.C. and many other weather and environmental attributes.

Download the dataset. Using the daily data from day.csv file, create a label based on the attribute "cnt" into three categories {low, medium, high} demand based on some threshold values according to your judgment. Your task will be to create a nearest neighbor classifier for such a dataset using **scikit-learn** and evaluate your classifier's performance. In your report, briefly describe on how you approached the problem, what interesting results you obtained, what practical difficulties you faced, and how you overcame these difficulties.