APSTA-GE 2047 - Fall 2018
Instructor: Ravi Shroff

HOMEWORK ASSIGNMENT 2

Instructions:

- Please submit this assignment on NYU classes by **11:00am on Thursday, 10/11/2018** (the beginning of Week 6 lecture). It is worth 100 pts.
- Work in groups of 2-3, turn in one assignment per group, and indicate group members. *Work together on all parts of the assignment.*
- A submission should include either:
    - a pdf document of all written work and any R scripts you wrote, or
    - one or more RMarkdown files.
- All scripts should be clearly written, commented, and self-contained so that I can easily run them to reproduce your analysis. Each of the two questions below contains more details on what you should submit.
- You will be graded on completeness, writing and visualization quality, and effort/creativity.
- You have two weeks for this assignment, so please start early, and talk to me if you need help!

1) **Logistic regression applied to voting [50 pts]**

A) Download and explore the exit poll data, poll_data.tsv, available at NYU classes in the web page for this assignment.

B) Build a logistic regression model to estimate people's probability of voting for Obama in the 2008 presidential election using all the features in the dataset. *If you are not using RMarkdown, you should provide separate written answers in to parts i), ii), and iv) as a pdf, as well as an R script of your code.*

i) List the coefficients for each age group and gender; if your model does not fit all these coefficients, explain why not. **[5 pts]**

ii) Provide a summary of the model. What is your interpretation of these values? **[5 pts]**

iii) Convert the probabilistic predictions for each individual into binary predictions, based on the candidate they are most likely to vote for. Compute accuracy, precision, and recall for your predictions. Don't use any package that automatically computes these values! **[5 pts]**

iv) Repeat step iii), but now convert each individual's prediction to a binary prediction for Obama only if the individual's probability of voting for Obama is at least 70%. What differences do you see in accuracy, precision, and recall compared to step iii)? **[5 pts]**

C) Not everyone votes for major party candidates in elections, so a binary prediction isn't always the best approach for predicting votes. Download and explore the revised exit poll data, poll_data_full.tsv, available at NYU classes in the web page for this assignment (this dataset includes individuals who voted 'other'). *If you are not using RMarkdown, you should provide separate written answers in to parts i), iii), and iv) as a pdf, as well as an R script of your code.*

i) Using this revised exit poll data, build a binary logistic regression model to predict whether an individual voted for a major-party candidate in the 2008 elections. Make a histogram of the resulting predicted probabilities using ggplot2. **[10 pts]**

ii) Filter the revised exit poll data to only individuals who actually voted for major party candidates. On this subset, build a binary logistic regression model to predict whether an individual voted for Obama. This gives an estimate of
Pr(voted Obama | voted major party candidate). **[10 pts]**

iii) Using the model from step ii), generate estimates of
Pr(voted Obama | voted major party candidate) for *every* individual in the revised exit poll data, and make a histogram of the resulting predicted probabilities using ggplot2. **[5 pts]**

iv) Use the models from steps i) and ii) to compute, for each individual, the probability that the individual votes for:

a) Obama
b) McCain
c) 'Other'

Generate categorical predictions for each individual based on these probabilities, and report the accuracy of your classifier. **[5 pts]**

2) **Tweet classification using Naive Bayes [50 pts]**

There is a theory that whether or not a tweet from the @realDonaldTrump twitter account is written by the president or his staff can be determined by the device from which the tweet was sent (Donald uses an Android phone, and his staff uses an iPhone). For this problem, you are going to use the Naive Bayes classification technique to see how well you can classify Trump's tweets based on the tweet text and timestamp (but without device information). The performance of your classifier will be judged using a hidden test dataset. *You should turn in one R script for this question, called trump_classifier.R; more details are in the instructions for part D, below.*

A) Download trump_data.tsv from NYU classes in the web page for this assignment. This file contains a sample of 1,240 tweets from @realDonaldTrump, and has three columns:
source (Trump or Staff), time_posted (EST), and text (full text of each tweet). Create a script called trump_classifier.R and put your work below in this script.

B) Clean and organize the data for binary classification. Determine what features you want to use: think about how to use the time and full tweet information to generate features (dealing with capitalization, punctuation, hashtags, etc). Note that we will spend more time on text data later in the course; don't feel obligated to go crazy here, but create at least a few features that you think might be helpful in classification. You can get some ideas from the link above. **[15 pts]**

C) Divide your data into a training set (80% of the data) and a test set (20% of the data); that is, randomly shuffle your cleaned data from part B), and use 80% of this data to train the model, and reserve 20% to check performance. Implement the Naive Bayes classifier using this package in R. Feel free to explore various options (e.g., Laplace smoothing). **[20 pts]**

D) Select a single model, and save the options you used to generate it in an R script. You will submit a single R script, trump_classifier.R which should:
-load the training data, trump_data.tsv
-load a test dataset called trump_hidden_test_set.csv
-clean the training data and test data
-fit your model on the training data
-apply the fitted model to generate predictions on the test dataset, and save these predictions as a single csv file, called predictions.csv, which consists of one column of 0's and 1's (1 for Trump, 0 for staff), with one prediction for each row in the test dataset.

To evaluate this part of the assignment, I will take your script, put it in my md_and_ml folder along with trump_data.tsv and trump_hidden_test_set.csv, and run the following command from the shell:
Rscript trump_classifier.R
It should output predictions.csv, which I will use to judge the accuracy, precision, and recall of your model.

You can find an example test dataset also called trump_hidden_test_set.csv on the NYU classes page for this assignment; you can use this to make sure your scripts will work with the evaluation process (note that the example test dataset is *not* the same dataset that will be used to evaluate your assignment). **[15 pts]**