Problem Set 4

Due April 2, 10:00 AM (Before Class)

Instructions

1. The following questions should each be answered within an R script. Be sure to provide many comments in the script to facilitate grading. Undocumented code will not be graded.
2. Work on git. Fork the repository found at https://github.com/domlockett/PDS-PS3 and add your code, committing and pushing frequently. Use meaningful commit messages – these may affect your grade.
3. You may work in teams, but each student should develop their own R script. To be clear, there should be no copy and paste. Each keystroke in the assignment should be your own.
4. If you have any questions regarding the Problem Set, contact the TAs or use their office hours.
5. For students new to programming, this may take a while. Get started.
6. You will need to install `ggplot2` and `dplyr` to complete this dataset.

Sample Statistics

1. Load the following data: http://politicaldatascience.com/PDS/Datasets/GSS-data.csv. \

   The variable `poleff11` asks participants to rate their level of agreement with the statement "People like me don't have any say about what the government does" (see the codebook for more information on all variables in this dataset at: http://politicaldatascience.com/PDS/Datasets/gss_codebook.csv).

   - Convert this variable into a numeric where higher values indicate higher levels of political efficacy (1- strongly agrees with the statement; 5- strongly disagrees with the statment) and all other values ('Cant choose' etc.) become NA's.
   - What is the proportion of individuals from the entire sample who feel as though they "have a say in the government?""
   - Using a sample of **25** from this dataset. What is the average proportion who feel as thought hey have a say?
   - Pull a random sample of 25 from the `poleff11` data and calculate the mean for this outcome. Now repeat this process 500 times and store these values in a variable called `trials_25`.
   - Now create a variable called `trials_100` where we do 500 trials with n=100 instead of 25.
   - Draw a histogram of the **sampling distribution** for the two trials (n=25 vs. n=100) you just conducted. Give the plots meaningful titles and axis labels. Save these plots in your repository.
   - What notable difference occur when we use a larger sample size in our trials?

Supervised Learning

1. Load the following data: http://politicaldatascience.com/PDS/Datasets/SenateForecast/PollingCandidateData92-16.csv.

   This is is data for *incumbents* running for re-election to the US Senate.

- *Poll Percentage.of.Vote.won.x* is the percentage of the vote the candidate won.

- The other variabels are mostly self-explanatory or have been used before in class.

- However, this datset differes in that it is organized at the poll level. That is, there is one row for each poll of each senate race.

- So there are some new variables including: the polling firm, the starting date of the poll, the "days left" until Eleciton Day, sample size, and the `numberSupport` (the number of respondents in that poll who indicated they supported the incumbent candidate.)
- There is also a `win` variable that indicates whether the incumbent candidate won the election.

2. Re-organize the data so it is a the election level (as opposed to the poll level).
   - This means you will have to figure out how to reduce the polling data into a summary statistic.

   - You might try to do this a couple of different ways based on sample size and date of the poll for use later.
3. Randomly select 20 percent of your data to use as a "validation sample" to assess the quality of your model. You will use this division of the data in the rest of the problems below.
4. Using the *Poll Percentage.of.Vote.won.x* variable, create at least two linear regression models to predict vote share for incumbents.

   - You are free to do this any way you want, but you must assess the quality of your model using cross-validation.

   - Train your model on your "training" data (80% of the data) and test on on the "test" data.
   - Provide an appropriate summary statistic for your compeing models using only the validation set. (Meaning: what is your out-of-sample performance?)
5. Now, using the `win` variable as your outcome, create at least 3 classification models. You should again assess each model on your "validation" set using appropriate methods. You must fit at least one of each:
   - linear classifier
   - random forest model
   - K-nearest neighbors
6. Now you are going to assess your classifiers using the 2018 election.
   - Most of the data you need is here: http://politicaldatascience.com/PDS/Datasets/SenateForecast/PollingCandidateData18.csv.
   - BUT, this dataset is missing (a) the final outcome and (b) a lot of the polling data.
   - Scrape the election results and polls from ballotpedia.org.

   - This does not need to be perfect, but should demonstrate the basic skills covered on webscraping.
   - Assess how each of your classifiers performs for 2018 using appropriate metrics.