Homework 8 Completion

Sheridan Grant

Must be uploaded to Canvas under "Homework 8 Completion" by
Thursday, May 21 at 11:59pm

Instructions

You may submit a .txt, .pdf, or .doc/.docx file for this assignment. This completion homework is to get you started on the final project. You've got 3 days to do it, and I expect it will take you 3 hours. You must submit a project proposal in 3 sections:

1. Either the name of 1 partner from the class that you'll work with, or "I will work alone." **If you decide to work with a partner, you should both submit the same document, just with the name at the top and the name of the partner switched. Note bullet 4, below.** Recognize that the quantity of work required will increase by about 50% if you choose to work with a partner, i.e. that you'll be writing 700-1000 words instead of 500-700 words and that your "interview" with me will be 9 minutes instead of 6.

2. Your first proposed data set, with a link so that I can access the data. See data set requirements and suggestions, below. Describe the data set in 2-3 sentences (where it comes from and how it was collected; what it contains). In 1-2 sentences, what broad scientific question do you propose to answer? Provide three *specific* scientific hypotheses to test or questions to answer:

   (a) Scientific question 1

   (b) Scientific question 2

   (c) Scientific question 3

3. Repeat the previous bullet and sub-bullets (i.e., provide a second data set with description and proposed scientific questions).

4. **If you chose to work with a partner:** repeat bullet 2 for a 3rd data set.

I'll look through your proposals and make sure everyone has fairly different and reasonable data sets. If you pick something to similar to a data set from class, I won't let you use it. If you want to use a COVID-19 data set that has lots of new, interesting variables and you propose analyses we haven't done in class, I'll allow it.

Project and Data Requirements

A full project description will be posted in the next week as I finalize the requirements. This should be enough to inform you about selecting a data set and partner:

- Data set, in tidy form, (number of rows) $\times$ (number of columns) ≥ 1000. Remember how hard it will be to properly model data with very few observations.
- Data set contains at least 3 of the following data types: continuous, binary, categorical (non-binary factor), integer, date/time.
- 500-700 written words (not including plots, tables, code) if you work alone, 700-1000 with a partner.
- During final exam period (or other time if you're in a far-away time zone or have a major conflict), you'll have a 6-minute (alone) or 9-minute (with partner) "interview" with me. I'll ask you (and your partner, if you have one) about your project to make sure you understand what you did and can explain the basics clearly.
- Grade breakdown will be 20% each of: interview, writing (clear, comprehensive, organized), modeling (appropriate assumptions, implementation, interpretation), visualization and exploratory analysis, and code (readability, correctness, usability).
- Use at least 1 regression model we learned about in class (linear or logistic regression) and 1 kind of model we didn't learn about.

Suggested sources:

- Kaggle
- UCI Machine Learning Repository
- FiveThirtyEight (politics, sports)
- Washington Post

Your best bet is to pick a data set you can analyze well. If you try to do something super hard ("artificially intelligent machine learning algorithms for natural language processing" or something) and do a bad job, you'll fare worse than if you do a good job at something more straightforward (see example, below). I'm familiar with the most popular of the data sets on these sites from my own stats classes, it's easy to look up popular analyses on the internet, and many analyses posted on the internet are bad. So while you may look at other's analyses for inspiration or information about the data set, don't expect to get credit for them.

Example

Here's my example, for the Johns Hopkins COVID-19 data.

1. I will work alone.

2. Johns Hopkins COVID-19 data (source). 13 variables with almost 1 million observations. Aggregated COVID-19 daily cumulative number of confirmed cases and number of deaths across the world, broken down by country, sometimes province/state, sometimes county/municipality. Additionally contains latitude/longitude information. I propose to study the rate of increase in confirmed cases early in the pandemic, trying to understand which factors affect the growth rate. I will supplement these data with information about the countries or states I study.

   (a) Is population of a country related to the length of time it takes for the number of new cases per day to begin decreasing steadily? I will investigate for different definitions of "decreasing steadily."

   (b) Is the political party of the governor of US states statistically significantly related to the fraction of the state's population that has coronavirus 30 days after the first case appears in that state? Is it related to the estimated IFR (Infection Fatality Rate) in that state?

   (c) Is it possible to predict the number of cases a US state will see in the future using historical data from other states that are "further along" in the pandemic? I will research time series analysis and use various lags and autoregression modeling techniques, comparing via test prediction accuracy.