Homework 5: Self-Organizing Maps

MACS 30100: Perspectives on Computational Modeling
University of Chicago

Overview

For each of the following prompts, produce responses *with* code in-line. While you are encouraged to stage and draft your problem set solutions using any files, code, and data you'd like within the private repo for the assignment, *only the final, rendered PDF with responses and code in-line will be graded.*

A Computational Social Science Problem

This final problem set will be a bit different. I am interested in how you address a social science problem with a computational social science workflow given very general guidelines. The goal in this final problem set, then, is to evaluate your ability to take a general set of instructions, and develop a technically correct solution that allows for substantively insightful inferences. This marries the *computational* with the *social science* parts of the program and course, informed by some functional programming skills you have developed.

Your task in this final problem set is to *design and implement a well-rounded self-organizing map analysis to mine public opinion data on 14 questions from the 2016 ANES.* Using these data we've encountered a bit to this point, you will develop your own solution, which requires selection of packages you think are best for completing the task, whether covered in class or not.

You will be evaluated on your ability to accomplish a task using both new and more familiar tools. On the substantive side, you will be mining the ANES data for evidence of whether there are likely to be *partisan* differences in public opinion, where *public opinion* is defined here as responses to 14 survey questions on salient social issues. For simplicity, you may treat "partisan" as three levels, including the two major US political parties and all others (Democrat, Republican, and Other).

As I am giving you only a general set of prompts to guide your process, the following rubric in addition to a technical set of solutions will be used to grade this problem set:

Preprocessing (10 points)	**Exploration** (15 points)	**Modeling** (25 points)	**Validation** (25 points)	**Writing & Programming** (25 points)
Designed and implemented appropriate techniques to get the data in a form that is usable for analysis	Explored the space fully (numeric, viz, etc.) prior to fitting models or training algorithms, resulting in a clear rendering of the space	Fit the correct model with all hyperparameters tuned appropriately, and *all* decisions throughout sufficiently defended	Validated results appropriately, and clearly dug deep and beyond the main model to defend and explain recovered patterns	Proper writing with excellent grammar (spelling, etc.) and thorough responses; elegant and *replicable* code

The Social Issue Questions

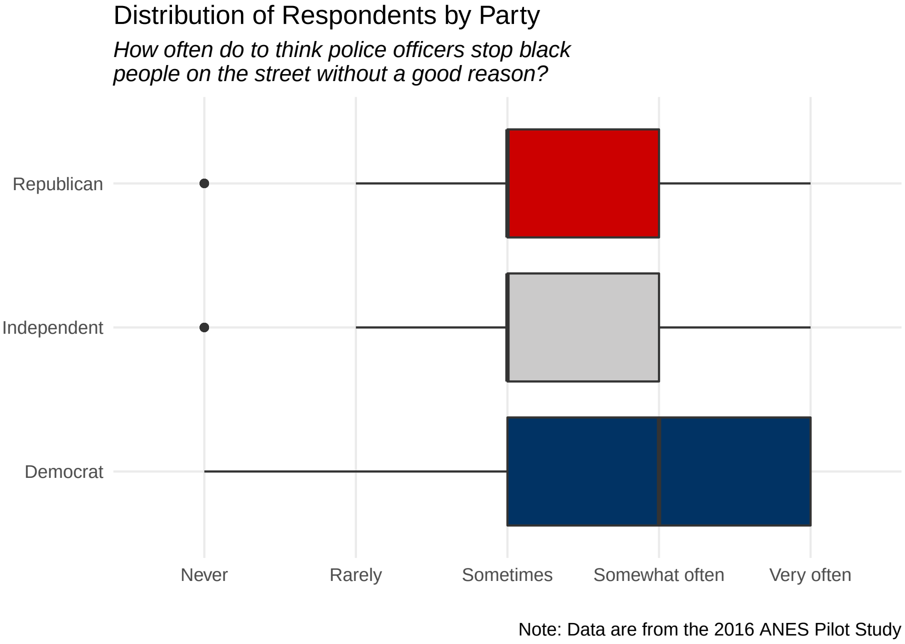
Below are the 14 social issue questions along with scales and variable code names to be used in the analysis. Question wording and response categories were copied and pasted from the ANES 2016 Pilot Study Questionnaire.

- `vaccine` - “Do you favor, oppose, or neither favor nor oppose requiring children to be vaccinated in order to attend public schools?” (7 point from favor a great deal (1) to oppose a great deal (7))
- `autism` - “How likely or unlikely is it that vaccines cause autism?” (6 point from Extremely likely (1) to Extremely unlikely (6))
- `birthright_b` - “Do you favor, oppose, or neither favor nor oppose children of unauthorized immigrants automatically getting citizenship if they are born in this country?” (7 point from Favor a great deal (1) to Oppose a great deal (7))
- `forceblack` - “How often do you think police officers use more force than is necessary under the circumstances when dealing with BLACK people?” (5 point from Never (1) to Very often (5))
- `forcewhite` - “How often do you think police officers use more force than is necessary under the circumstances when dealing with WHITE people?” (5 point from Never (1) to Very often (5))
- `stopblack` - “How often do to think police officers stop BLACK people on the street without a good reason?” (5 point from Never (1) to Very often (5))
- `stopwhite` - “How often do to think police officers stop WHITE people on the street without a good reason?” (5 point from Never (1) to Very often (5))
- `freetrade` - “Do you favor, oppose, or neither favor nor oppose the U.S. making free trade agreements with other countries?” (7 point from Favor a great deal (1) to Oppose a great deal (7))
- `aa3` - “Do you favor, oppose, or neither favor nor oppose allowing universities to increase the number of underrepresented minority students studying at their schools by considering race along with other factors when choosing students?” (7 point from Favor a great deal (1) to Oppose a great deal (7))
- `warmdo` - "Do you think the federal government should be doing more about rising temperatures, should be doing less, or is it currently doing the right amount? (7 point from Should be doing a great deal more (1) to Should be doing a great deal less (7))
- `finwell` - “Do you think people’s ability to improve their financial well-being is now better, worse, or the same as it was 20 years ago?” (7 point from A great deal better (1) to A great deal worse (7))
- `childcare` - “Do you favor an increase, decrease, or no change in government spending to help working parents pay for CHILD CARE when they can’t pay for it all themselves?” (7 point from Increase a great deal (1) to Decrease a great deal (7))
- `healthspend` - “Do you favor an increase, decrease, or no change in government spending to help people pay for HEALTH INSURANCE when they can’t pay for it all themselves?” (7 point from Increase a great deal (1) to Decrease a great deal (7))
- `minwage` - “Should the minimum wage be raised, kept the same, lowered but not eliminated, or eliminated altogether?” (4 point from Raised [1], Kept the same [2], Lowered [3], Eliminated [4])

The Task

Here are the prompts to guide your task. Again, **there is no single way this problem set should be executed**. Simply do your best, leveraging all tools and techniques we have covered, and most importantly defend ***all*** choices you make throughout the process so you can at least earn partial credit where appropriate

1. Read in the 2016 ANES data we have been using (`anes_2016.csv`), and create a subset of the data containing *at least* the main 14 questions/features (from above) as these are the core of the analysis.
2. Preprocess and clean the data.
3. Explore the data using any approach(es) or tool(s) you think best, such as feature-level correlations, boxplots, scatterplots, density plots, etc. For example,



4. Construct and present a self-organizing map (SOM) of the *question space*. **Think carefully about the scale of response categories, as these vary across questions.** Also, remember to tune the relevant hyperparameters appropriately. You might consider the `kohonen` package in R (though there are many others), or the `minisom` package in Python. A response to this may include creating grids, fitting models, and creating (multiple) visualizations of the results.
5. Comment on the results thus far as it relates to the main goal of the task. In other words, did you uncover evidence of partisan differences in the data? Did you not? Regardless, why do you think you got the results you did? What are some other substantive patterns you detected?
6. To validate the SOM results, fit a k-means algorithm to the data and plot respondents' political party affiliations as well as their cluster assignments from the k-means fit. Discuss the results. E.g., Do you see evidence of partisan differences across the groups? Do you not? How do you know?
7. Taken with the SOM results, do the k-means results show similar or different patterns? Or is it unclear? Discuss both models in relation to each other for a full, well-rounded validation. *Note*: you are encouraged to think of and implement other ways to validate your results. You are welcome to go back to class notes, Google, etc.