APSTA-GE 2094 / APSY-GE 2524: PS 7

Klint Kanopka

In this assignment, you will explore a dataset of your choice. If you're looking for data, the Item Response Warehouse is a great place to start. There are also a large number of interesting datasets available at the NYC OpenData Portal. While most of what we do in this course is, in some way, related to item response data, **you are not technically required to use item response data**. Feel free to explore text data or some other suitably high dimensional dataset. **You may work in a group of up to three. Each member should submit an identical report on Brightspace for grading.**

This assignment has you applying some technique from the first half of the course (PCA, factor analysis, some flavor of IRT, or topic modeling) and some technique from the second half of the course (some flavor of clustering, Latent Class Analysis, or a Diagnostic Classification Model), so be sure that you have enough rows and columns to do something interesting.

Some (very general) ideas of project setups might include:

- Starting from a dataset that has individuals responding to cognitive items with demographic information, use factor analysis to look for multidimensionality and generate factor scores for individuals. Name the factors. Cluster individuals based on their factor scores and name the clusters. Then, use the demographic data to describe who is in each cluster.
- From a dataset that includes text documents with covariate metadata, fit a topic model. Cluster documents based on their topic profiles. Use the metadata to predict cluster membership.
- Using data with cognitive items, survey items, and demographics, use IRT to summarize the cognitive items and find θ scores for individuals. Use Latent Class Analysis to group respondents together. With demographics and ability scores, interpret and describe the Latent Classes.

Your submission should be in the form of a `.pdf` a **maximum** of twenty (20) pages, including figures but not counting references and an appendix of code. Code should not be inline with the text. Figures should be referenced in the text and have suitable captions. You may organize the narrative how you see fit, but structuring it as an academic article or blog post is a good place to start.

Grading Criteria

These requirements are designed to identify the mininmal components of a successful project, not constrain the work you do.

Part 1: Introduction (40pts total)

In this section, your job is to introduce your project and research questions, as well as describe the data you're using. This section should begin with a (brief) literature review that provides some background on the context and questions you'll be addressing. Then, it should contain visual and written descriptions of the distributions of the variables, information from any codebooks you use, and may include the text of items. Your job here is to give the reader a clear picture of questions you are asking and the data that you are using to answer them.

- 10pts for literature review
- 10pts for clear research questions
- 10pts for visual descriptions of the data
- 10pts for written descriptions of the data

Part 2: Working with "item responses" (30pts total)

In this section, you'll employ (at least) one of PCA, factor analysis, item response theory, or topic modeling to explore your data. This section should have a brief description of the method you're using, why you've selected it, and tables and/or figures displaying your results. Key to this section is not just using the method, but interpreting the results and connecting them back to your questions from Part 1.

- 10pts for implementing the method and justifying its use
- 10pts for visualizations related to the results
- 10pts for written descriptions and interpretations of the results

Part 3: Clustering (30pts total)

As in Part 2, you'll employ (at least) one of clustering, latent class analysis, or diagonostic classification models. This section should have a description and rationale for the method you choose, along with the results. Remember to interpret the results and connect them back to your questions from Part 1.

- 10pts for implementing the method and justifying its use
- 10pts for visualizations related to the results
- 10pts for written descriptions and interpretations of the results