Homework 4: Unsupervised Learning

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.*

Dimension Reduction

Conceptual Problems

1. (5 points) Compute the total variance from the following PCA output.

	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	PC9	PC10
Standard deviation	3.55	2.41	1.82	1.31	1.05	0.86	0.81	0.79	0.72	0.70
Variance	3.45	3.10	1.75	0.98	0.64	0.33	0.31	0.30	0.09	0.05

2. (10 points) Make a *manual* scree plot based on these results. That is, *no* canned functions or packages (e.g., `factoextra`).

3. (10 points) Based on your results in the previous question, how many PCs would you suggest characterize these data well? That is, what would the dimensionality of your new reduced data space be?

4. (10 points) Calculate the Euclidean distance between each of the following observations, i, and some observation at 0 (i.e., x_0) in 4-dimensional space $\forall X \in \{1, 2, 3, 4\}$.

i	X_1	X_2	X_3	X_4	Euclidean Distance
1	2	2	3	1	...
2	1	1	-2	2	...
3	1	-2	-2	-1	...
4	3	3	2	2	...
5	-3	2	-1	1	...

An Applied Problem

For the following applied problem, use the 2019 American National Election Study (ANES) Pilot survey data. These data include, among many other features, a battery of 35 feeling thermometers, which are questions with answers ranging from 1 to 100 for how respondents "rate" some topic (e.g., *How would you rate Obama?* or *How would you rate Japan?*). See the documentation and more detail here.

To make your lives a bit easier, I have preprocessed the data for you, including: 1) feature engineering (via kNN) for missing data, and 2) reduction of the feature space to include only the 35 feeling thermometers and a feature for the respondent's party affiliation (`democrat`), where 1 = Democrat and 0 = non-Democrat (which could be Republican, Independent, or decline to say).

5. (10 points) Fit a PCA model on all 35 feeling thermometers from the 2019 ANES, but be careful to *not* include the party affiliation feature.
6. (20 points) Plot the feature contributions from each of the feeling thermometers in the first two dimensions (i.e., PC1 and PC2). Describe the patterns, groupings, and structure of the lower-dimensional projections in *substantive* terms.

Clustering

A Conceptual Problem

7. (10 points) What are the two properties required for a *hard* partitional solution, and when thus relaxed, give a *soft* partitional clustering solution? Be sure to answer this both formally (with mathematical notation) and substantively (with words). Then, give an example or two of each and how they relate to these two central properties of clustering.

An Applied Problem

In this applied problem, you will again use the 2019 ANES data, but this time to explore the clustering solution from fitting a fuzzy c-means (FCM) algorithm to all feeling thermometers. As with the dimension reduction exercise, derive a clustering solution using *only* the feeling thermometers. The idea here is to explore whether attitudes on these issues, countries, and people map onto natural groupings between major American political parties.

8. (5 points) Load and scale the ANES *feeling thermometer* data.
9. (5 points) Fit an FCM algorithm to the scaled data initialized at $k = 2$, driven by the assumption that party affiliation (Democrat or non-Democrat) underlies these data.
10. (15 points) Visualize the cluster scores from your FCM solution plotted over the range of feelings toward `Trump` and `Obama`, with data points colored by cluster assignment and also labeled by the respondent's true party affiliation (the `democrat` feature). As party wasn't included in your clustering solution, what can you conclude based on these patterns? Is there a grouping pattern among observations along a partisan dimension, or isn't there? Do respondents group in expected ways (e.g., Trump supporters to the right and Obama supporters to the left)? Do cluster assignments align with the true party affiliation or not? How would you evaluate the effectiveness of FCM for this type of task?