Application Project - Requirements

INFO 432 - Drexel University

August 7, 2024

Instructions: The purpose of this assignment is to test your ability to apply the multivariate data analysis methods covered in this course.

All individuals will be required to submit a write-up of their approach in a PDF formatted '{first_name}_{last_name}_Final_Solutions.pdf', along with a zip file of code used in the format '{first_name}_{last_name}_Final_Code.zip'.

1. **Problem 1: Selecting a Dataset (20 points)**

   Research and select a dataset in an application domain of interest to you. The requirements for the selection are that the data must have greated than 20 features (variables) and 200 samples.

   For this assignment, you'll submit two parts. First will be a proposal for why the dataset is of interest, a description of the variables, and some initial hypotheses. This should be a one-page summary of what you intend to find out from the dataset.

   For the second piece, you'll submit an R script to perform exploratory data analysis. This EDA should be explained in your final write-up, but is not required for the proposal step.

2. **Problem 2: Dimensional Reduction (40 points)** In this problem, we will fit models to learn lower-dimensional representations of the text. We have looked at three main methods:

   - Principal Component Analysis (PCA)
   - Canonical Correlation Analysis (CCA)
   - Factor analysis

   Choose two of the above as they are applicable to your dataset and find the 'best fitting' model for the data. Explain which original features contribute to the new low-dimensional features and what these new features may be capturing about the data set.

3. **Problem 3: Fitting Unsupervised Clustering (40 points)**

   As the majority of problems do not have readily available labels associated with each sample, we'll utilize unsupervised clustering techniques to pick up on underlying structures in our dataset. Fit two clustering models of your choosing to both the original variables and to the low-dimensional representations you learned in the prior problem. Which feature sets provide the best clustering structure? How easy are the clustering results to explain? Develop a two page narrative about your experience in working with this data set and if you were able to confirm or refute any of the hypotheses developed in Problem 1.