Project 4
Unsupervised Learning and Deep Learning
Sociology 273L: Computational Social Science

1 Introduction

In this project, you will use unsupervised learning and deep learning techniques to explore a large dataset and prepare it for prediction. You will examine data taken from the National Health and Nutrition Examination Survey (NHANES). NHANES is a nationally representative survey that combines interviews with medical examinations to collect about 200 variables about 5000 people each year. The data you are working with is drawn from the 2013-2014 version of this study.

You have been tasked with using this dataset to study social determinants of public health outcomes. NHANES contains data about social determinants such as family income, education, occupation etc., and health outcomes related to conditions like diabetes, heart disease, osteoporosis, etc. Some of these data are collected via survey, and others from medical exams. You will specifically try to predict the variable, "HSD010" which corresponds to the question:

Would you say your/SP's health in general is:

- a. Excellent
- b. Very good
- c. Good
- d. Fair
- e. Poor

2 Data Description and Preprocessing

The NHANES data has already been combined and preprocessed for you. The overall dataset, "nhanes.csv" was combined from datasets covering demographics, diet, lab results, examination results, and a questionnaire. There was significant missingness in some of these datasets, so they were cleaned for the purposes of this project.

Your ultimate goal is to predict individuals' self-reported health condition ('HSD010'). Respondents can say they are feeling "excellent," "very good," "good," "fair," or "poor." You may either predict HSD010's original five categories, or convert it to a binary where "excellent," "very good," and "good" are "good" and "fair" and "poor" are "poor" health.

Before turning to prediction however, you you should first explore and summarize the data with some unsupervised learning techniques. Specifically, you will be solving two problems. First, you will tackle the curse of dimensionality — meaning you have many features in this dataset that can rapidly increase the overall complexity of a model. Second, you will address the issue of grouping different observations to aid in downstream classification. You might think of these unsupervised learning procedures as minimizing the complexity introduced by your *columns* and the noise introduced by your *rows*.

Consider the following boxplots that visualize the relationship between the ratio of family income to the federal poverty line and self-reported health condition:

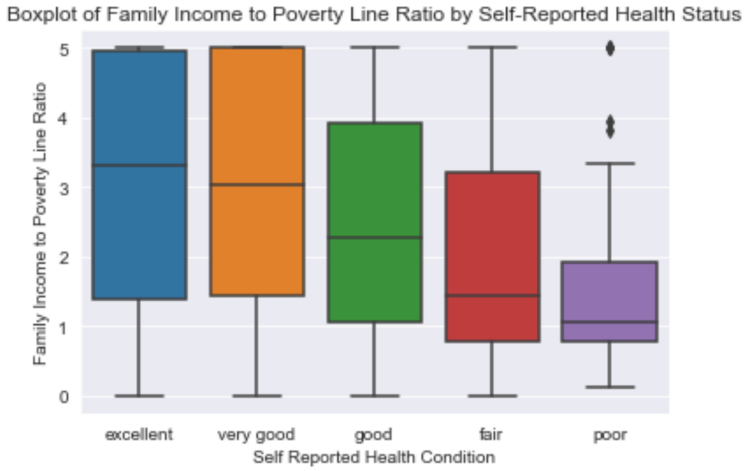

Figure 1: Boxplot of Income-Poverty Ratio and Health Condition

There are some clear patterns here — reported health condition declines the closer a family is to the poverty line. However, lots of features in the dataset probably correlate with both family income and health condition. One of your tasks will be to simplify the nearly 250 features in the dataset into fewer features so you better predict self-reported health condition.

It's also possible that while income strongly correlates with self-reported health condition, it does not necessarily correlate with actual health condition. Consider Figure 3 which shows the relationship between body mass index (BMI) and the family income-poverty line ratio. BMI is a strong predictor of other health conditions like diabetes, heart disease, obesity, etc. and yet does not have a clear relationship with family income. Figuring out how to automatically

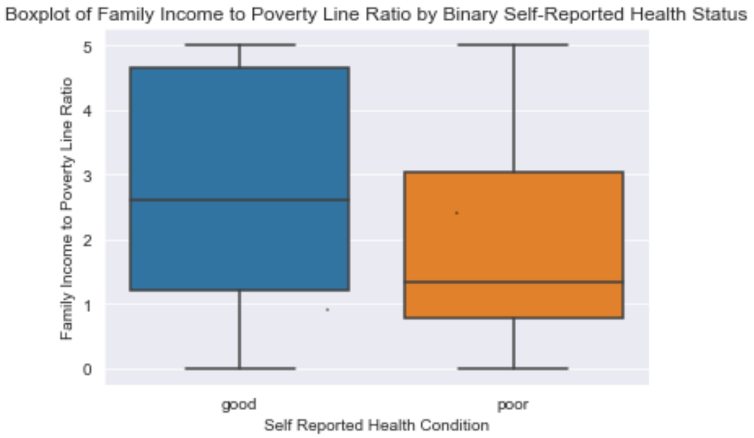

Figure 2: Boxplot of Income-Poverty Ratio and Binary Health Condition

group data points together to better predict health condition might be better than making judgments about these proxies.

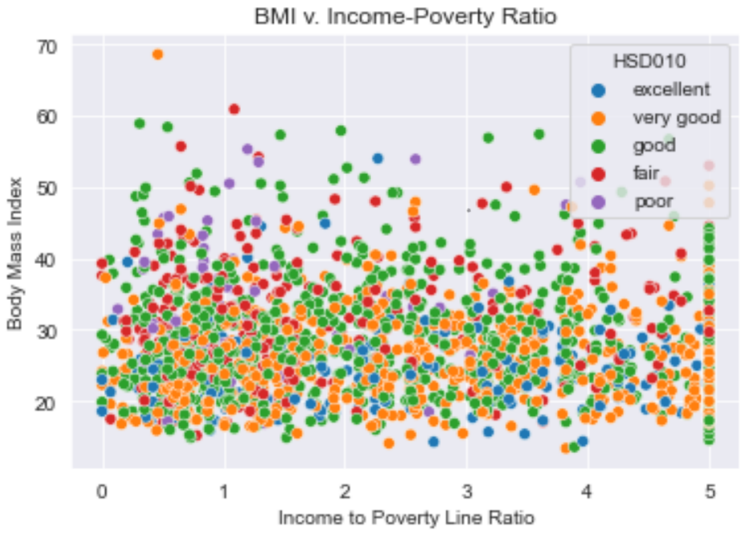

Figure 3: BMI v. Income to Poverty Ratio

3 Principal Component Analysis

Conduct a Principal Component Analysis (PCA) of the nhanes data. The data have already been prepared for you, so you can work directly on nhanes_scaled. Be sure to do the following:

- Choose the number of components and provide 1-2 sentences about your choice of the number of components.
- Plot a barplot of the variation explained by each component. Hint: look at the attributes associated with your model.
- Choose how many components you will use to fit a supervised learning model and provide 1-2 sentences to explain that choice.
- Plot a 2D scatterplot of the first two components and provide 1-2 sentences analyzing the plot.

4 Clustering

Next, cluster your observations. Be sure to do the following:

- Choose a clustering algorithm and explain it in 1-2 sentences.
- Cluster the nhanes data (X). Detail any choice you need to make with regards to number of clusters, and how you arrived at that choice. For instance, you might use the elbow method if you choose k-means.
- Plot your clusters on top of BMI v. Income Poverty Ratio Plot. Describe what you see in 1-2 sentences.
- Retrain the clustering algorithm, but this time use your PCA results instead of the original dataframe. Plot the clusters on top of the 2D PCA scatterplot from the previous step. Describe your results in 1-2 sentences.

5 Neural Network

Now we are ready to predict! Do the following:

- Choose either HSD010 or HSD010_binary as your target outcome.
- Train a neural network using the original features. Much of the code to train a basic neural net has been set up for you, but you will need to fill in a couple of missing pieces.
- Train a neural network using only your PCA components as features.
- Train a neural network using your PCA components and the predicted class membership from your clustering algorithm as features.

- Compare and contrast how well each algorithm did. Which featurization technique would you pick and why?

6 Discussion Question

1. In your own words, what is the difference between PCA and clustering?
2. Did you notice any advantages to combining PCA and clustering? If so, what do you think they were? If not, why do you think you didn't see any gains from this combination?
3. How can unsupervised techniques help with downstream supervised learning tasks when working with "big data?"