STAT430 – Unsupervised Learning - Final Project – 100 Points

Due: Tuesday, May 11 by 8am CST on Compass.

Main Goal of Analysis

The main goal of your final project report and presentation, is to tell a compelling story based on the unsupervised learning data analyses you will perform on a dataset that you have selected. If you do not want to select a dataset, there are two contained in this zip file and discussed at the end of this document that you can use.

There are several places you can go to to find interesting datasets, but here are some places you can start.

https://archive.ics.uci.edu/ml/datasets.php

https://www.kaggle.com/datasets

To receive full credit, you should follow the steps and answer the questions given in this document for your project. However, if you think that there are additional questions or analyses that would add additional insights to your overall research goal, you're more than welcome to pursue these in addition to what is stipulated in this document.

Project Format

This project will have two components.

Project Report [60 points]

Deadline: Tuesday, May 11 by 8am CST on Compass.

Format: Jupyter notebook.

Should contain: Everything stipulated in the Project Specifications discussed below. This should look like a clean data analysis report that you would theoretically submit to an employer (not a homework assignment), so please use appropriate and well-formatted markdown headings to present your report.

Project Presentation [40 points]

You will have two options for "submitting" your presentation for this project.

Presentation Style Option 1: ***On*** **Tuesday, May 11 by 8am-10am CST** we will set aside two hours for you to present your findings "live" on Zoom. If you do it this way, your classmates and I will be able to give you feedback and ask you questions in real-time. However, once the presentation is over, you won't be expected to answer any follow up questions (unless you want to). I will post this presentation session on Compass. If you plan on presenting live, be sure to sign up here. https://docs.google.com/spreadsheets/d/1Ax3rMMY1BNSkGhUpodmf2wu6wr4oDksoErKBmuFUyM4/edit?usp=sharing

Presentation Style Option 2: ***By* Tuesday, May 11 by 8am-10am CST** you and your team can record yourselves giving your presentation and then submit the video on Compass along with your report. I will then post your videos on Compass for your classmates and me to watch. If you choose to present this way, your classmates and/or I *may* ask you questions about your work on Piazza. Part of your grade for the presentation will be answering any questions that are asked on Piazza. You will only be expected to answer questions that are asked within 24 hours of your video being posted.

Format and Other Specifications for Presentation:

- You can work by yourself on this project, or in groups of up to 3. *(My preference is that you work in groups).* If you're looking for teammates you can try posting on Piazza to see if anyone else was working on a similar dataset as you.
- Ideally, keep your presentation within 7-10 minutes long.
- Every member of your group must present some part of the presentation to get full credit.
- See attached presentation rubric for what you should present and how you will be graded.

What to do *first* if you're choosing your own dataset.

Choose a Dataset (if you decide to pick your own)

Choose a dataset that has at least 6 attributes (not including the pre-assigned class labels if there are any) and at least 150 rows.

Data Cleaning

Make sure you conduct any necessary data cleaning. If your dataset needed to be cleaned, be sure to discuss what was done to the dataset as well as explain any data cleaning decisions that you made.

Dataset Suitability Pre-Analysis

In this project you will be asked to apply at least two of the unsupervised learning algorithms that we have learned in this class to your chosen dataset. Two of these must be clustering algorithms. Thus, before proceeding with further analysis, you should do the following.

- First, test whether your dataset is clusterable. We have learned at least two methods that test this in this class. **Make sure you test whether the data is clusterable on the *scaled data as well (ie. each attribute has been mean-subtracted and standard deviation-divided).*** Given the results of these methods and what you know about the dataset, think about whether the scaled or unscaled dataset would be useful to perform your unsupervised learning analyses on.
- Next, you want to make sure that you *know of* at least two clustering algorithms that will cluster this particular type of dataset. (You are not constrained to the clustering algorithms that we have learned in this class. However, ensuring that there are at least two clustering algorithms that we have learned in this class that will cluster your dataset can be a useful backup just in case the work in this project takes longer than you expected.)

Dataset Options (if you don't want to pick your own)

1. **Country Data**: The country_data.csv contained in the zip file contains the following statistics for 167 countries:

a. child_mort: Death of children under 5 years of age per 1000 live births

b. exports: Exports of goods and services. Given as %age of the Total GDP

c. health: Total health spending as %age of Total GDP

d. imports: Imports of goods and services. Given as %age of the Total GDP

e. Income: Net income per person

f. Inflation: The measurement of the annual growth rate of the Total GDP

g. life_expec: The average number of years a new born child would live if the current mortality patterns are to remain the same.

h. total_fer: The number of children that would be born to each woman if the current age-fertility rates remain the same.

i. gdpp: The GDP per capita. Calculated as the Total GDP divided by the total population.

Unfortunately, the year(s) that the data was collected and the data collection process were not given for this dataset, so do proceed with caution with your interpretation of your analyses given this.

2. **Clothing Image Dataset** The fashion_mnist_sample.csv in the zip file contains a representation of a random sample of 2000 gray-scale images of 10 different types of clothing items (T-shirt/top, Trousers, Pullover, Dress, Coat, Sandal, Shirt, Sneaker, Bag, and Ankle boot). Each of the images is a 28-by-28 pixel, and has been flattened into an array of 784 pixels. *(To visualize a given image in this dataset, you can reference the code for when we explored the MNIST dataset).*

   You can read more about how these fashion items were generated here: https://github.com/zalandoresearch/fashion-mnist.

Project Report

Your report should include the analyses, code, and explanations detailed in each of the following section.

1. Introduction and Dataset Research

Motivation: After picking a suitable dataset (or deciding to use the one I gave you) and citing at least three resources, describe the motivation for why someone would want to conduct an unsupervised learning analysis on this dataset or a dataset of this type. These resources could be news articles, scholarly publications, additional data, etc. Give a brief explanation of the significance of each resource. Your resources should be cited in a standard citation format (ie. MLA, APA, Chicago, etc).

Dataset Information: If this is a dataset you chose from somewhere else, be sure to look up and discuss how the dataset was collected and any type of preprocessing that was conducted on the dataset that you will be using.

2. Preliminary Exploratory Data Analysis (EDA)

Before using any unsupervised learning algorithms, learn more about your dataset using basic exploratory data analysis.

1. For your numerical attributes, calculate basic summary statistics about each attribute.
2. For any categorical attributes (including the pre-assigned class labels, if your dataset has any) count up the number of observations of each type.
3. Determine if there exist are any strong pairwise relationships between at least one pair of attributes in your dataset and visualize these relationships.
4. From your analyses conducted here, should you scale your dataset? Explain why or why not. If you choose to scale, then do so.

3. Pre-Analysis Questions

Use the methods that we have learned in this class to answer the following questions. If the answer to some of these questions is not clear-cut, explain why.

1. Is the dataset clusterable? (The answer to this should be yes). Explain why.
2. If so, describe the underlying clustering structure of the data.
    a. Approximately how many underlying clusters does the data have?
    b. What are the shapes of the underlying clusters?
    c. Are the clusters balanced in size?
    d. Do any of the clusters that you identified overlap with each other?

4. Algorithm Selection Motivation

Next, using your background research, your findings from your preliminary EDA, and/or your findings from your pre-analysis questions explain why you chose these two (or more) unsupervised learning algorithms to use on your dataset.

5. Run these Two (or More) Algorithms on your Dataset

Perform your two (or more) selected algorithms on your dataset.

Results Presentation

Present and discuss the results from each of these algorithms to the reader of your report in an insightful way that relates back to your original motivation for performing the unsupervised learning analysis.

For instance:

- If your clustering is a hard assignment, you can color-code your t-SNE plot with the cluster labels.
- If your clustering algorithm is a hierarchical clustering algorithm, give the dendrogram and explain the nested relationships.
- If your clustering is a fuzzy clustering (or has cluster membership scores), you can plot K *(# of clusters) t-sne* plots, and color code each plot by the cluster membership score for the kth clusters.

Input Parameters

For any algorithms that require input parameters (most do), make sure you explain (and show the work for) why you selected these particular parameters. If you choose to use multiple sets of input parameters, make sure you discuss and show how the results of your algorithm changed when you tried these different input parameters.

6. Post-Analysis Questions [Do this for Each of your Clustering Algorithm Results]

1. **Separation and Cohesion:** For the clustering returned by your clustering algorithms assess and discuss the cohesion of each of the clusters. Are there are any objects that have poor cohesion with their assigned cluster?
    a. Note: If you used a hierarchical clustering algorithm, select (at least one) of the clusterings returned by the algorithm. If you used a clustering algorithm with a cluster membership score, create a hard partition by assigning each object to the cluster that it most belongs to.
2. **Cluster Attributes**: For each of the clusters in each of your clusterings, what attributes (and attribute relationships) standout for objects assigned to this dataset? (There are many ways that we learned how to do this in this class, you can pick one).
3. ***If* your dataset had pre-assigned class labels [Supervised Learning Evaluation]**
   For each of the two (or more) clusterings returned by your clustering algorithms, do the following.
    i. Calculate the following and interpret the result.
        1. Adjusted RAND Index between the clustering and the class labels.
        2. Homogeneity score between the clustering and the class labels.
        3. Completeness score between the clustering and the class labels.
    ii. Color code the points in your t-sne plot by cluster labels and code the "style" of the marker with your class labels. Then interpret this plot. Did what you observe in this plot corroborate what you calculated in your adjusted rand index, the homogeneity score, and the completeness score?

4. ***If* your dataset DID NOT have pre-assigned class labels [Cluster Distance]**
    a. Which clusters are closer to eachother than others?

7. Analysis Summary

1. **Algorithm Comparison Summary:** Given your research motivation you discussed for this dataset, are the results and performance from just one of these algorithms categorically better than all the others? Explain why or why not. (Show code and visualizations to help explain why).
2. **Insights Summary:** Summarize the insights you found with all of your analyses and relate how these insights might be useful towards your research motivation.

IMPORTANT TO READ - Things to Remember to Do in your Analysis

- **Non-Deterministic Algorithms/Methods** For any non-deterministic algorithm or method in your analysis, make sure that you run the algorithm/method using ***multiple different random states*** *(for instance, random_states 1000,1001,…,1004).*
    - You should use a random state (in general) for these type of algorithms so that your results do not change if you have to re-run the algorithm and when you write your report.
    - You should use multiple random states in attempt to ensure that the results that you are seeing with this non-deterministic algorithm/method are consistent.
    - When you run your algorithm/method with multiple random state, be sure to **comment on any variability of your results and any differences that you see**.

Report Grading Rubric (60 points)

Components of the Report	Points
Dataset meets size specifications	2
Data cleaning: * code correctness * explained well * correct decisions	**2**
Dataset research - motivation: * cites three resources * explains the significance of each resource well * correct citation format	**4**
Dataset research - information: * discusses dataset collection * discusses data preprocessing	**2**
Preliminary EDA: * basic summary stats for categorical/numerical variables * pairwise relationships * scaling (explained why or why not)	**4**
Pre-analysis questions * is dataset clusterable (correct explanations) * clustering structure (correct explanations) - how many clusters - cluster shapes - clusters balanced in size - overlapping clusters?	**5**
Algorithm Selection Motivation * well explained and correct	**4**
Running the algorithms * correctness * presented/discussed in a way that maps back to research goal	**10**
Algorithm Parameters Selection * correctness * explains why these were selected (correct explanation) * discusses how the results changed with different parameter values	**4**
Separation and Cohesion (clusters and objects) * correctness * explanation	**3**

Cluster Attributes * correctness * explanation * maps back to research goal	**4**
Supervised Learning Evaluation * Adjusted rand (correctness, explanation) * Homogeneity score (correctness, explanation) * Completeness score (correctness, explanation) * Corroboration with t-sne plot (correctness, explanation)	**4 (or na)**
Cluster Distance * correctness, explanation	**4 (or na)**
Comparing clustering performance/results * correctness * explains well * maps back onto research motivation	**4**
Insights summary * correctness * explains well * maps back onto research motivation	**4**
Uses Multiple Random states * does this for non-deterministic algorithms	**1**
Professionalism/tidiness of report	**3**

STAT430 Project Presentation Rubric (40 points)

Team Members: ____

SLIDES / 25

Content (15) – You should present *some* content on each of these topics

- (1) Intro/Conclusion
- (2) Presentation of data research
- (2) Presentation of *some* EDA
- (2) Answered pre-analysis questions
- (2) Explained algorithm selection motivation
- (2) Presented algorithm results
- (2) Answered post analysis questions
- (2) Presented analysis summary

Correctness (5)

- Analyses are appropriate for the data, results are interpreted correctly.

Layout (5)

- Content is well organized, fonts are easy to read.
- Slides are engaging and not too wordy.

PRESENTATION / 15

Narrative / Motivation (5)

- Explain why they chose their dataset, and why conducting unsupervised learning analysis is meaningful.
- Explain how their findings relate back to the research goal/motivation.

Preparedness (5)

- Team members understand the material, they are not reading directly from a notecard or script.

Presentation (5)

- All team members speak and present some portion of the material.
- Team members speak loud enough for everyone to hear