STAT430 – Unsupervised Learning - Final Project – 100 Points

Due: Wednesday, December 15 by 8am CST on Compass.

Main Goal of Analysis

The main goal of this project, is to tell a compelling story based on the unsupervised learning analyses you will perform on a dataset. **You can work in groups of up to 3 people. Or you can work by yourself.**

- **If you work with a group of 3, you must do at least 25% of the work in order to get full credit.**
- **If you work with a group of 2, you must do at least 33% of the work in order to get full credit.**

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.

In addition to being graded for correctness and completion, this project will be graded on a qualitative basis. Qualitatively, we will be looking for the following things.

- **Clarity about Analyses, Algorithms, and Data Choices**
  - Someone who has taken this class should be able to read through your report and/or watch your presentation and easily be able to do the following.
    - Replicate what you did in your analyses.
    - Know why you made the choices that you did in your analyses.
- **Clarity about Motivation (ie. the "so what?") of your Analyses**
  - *Beginning of the Report and Presentation:*
    - Someone who is **about to** read your report and watch your presentation should be able to clearly answer the questions.
      - *"Why should I (or someone else) care about the report that I am about to read/listen to?"*
      - *"What research questions do they intend to answer?"*
      - *"How do these research questions relate to their motivation?"*
    - Therefore, in the introduction of your report and presentation you should make this clear.
  - *Middle of the Report and Presentation:*
    - While **in the middle of** your report and presentation, your audience should be able to clearly answer the question.
      - *"How do each of these analyses/algorithms/data choices that they're making/using tie back into the overarching motivation of this whole analysis?"*
    - Therefore, for each new analysis/model/algorithm/data choice that you make, you should explain this and make it clear to your audience.
  - *End of the Report and Presentation:*
    - Someone who has **just finished** reading your report and watching your presentation should be able to clearly answer the questions:

- *"Why should I (or someone else) care about the analysis that I just read/listened to?"*
- *"Did their analyses and conclusions answer the research questions that they stated at the beginning of the report/presentation? If so, how? What were the answers to these research questions?"*
- *"How would the results/answers to these research questions be useful to someone?"*

- Therefore, in the conclusion of your report and presentation you should make this clear.

Intended Audience/Reader of your Project

The intended audience of your report/presentations should be someone who taken this STAT430 class. **Theoretically, you should be able to send/present your report to one of your classmates (who is not on your team), and they should be able to understand everything that you did and the claims that you are making.**

Project Format

This project will have three components.

Project Report [70 pt]

Deadline: Wednesday, December 15 by 8am CST on Compass.

Should contain: Everything stipulated in the **Project Report Specifications** discussed below.

Format:

- Jupyter notebook.
- This should look like a clean data analysis report that you would theoretically submit to an employer (not a homework assignment). Thus, at the very least, your report should have:
  - a title
  - headings for each of your sections
  - You should write paragraphs and in complete sentences.
- You can use and modify the attached project **project_template.ipynb** file as a template for this report if you'd like. You can add and delete as many cells as you'd like in this file.

Graded:

- See "Project Report Specifications" section below for point breakdown.

Project Presentation [26 pt]

Presentation Dates: ***On*** **Wednesday, December 15 8am-10am CST** we will set aside two hours for you to present your findings "live." Based on the survey results, we may have these presentations **in-person** (in our CIF Building classroom) or on **Zoom**. Please fill out the survey if you have a preference.

Format:

- Ideally, keep your presentation within 7-10 minutes long.
- You must present some part of the presentation (if you're in a group) in order to get full presentation credit.
- Presentation should be presented in **slides** (not the Jupyter notebook).

Graded:

- See attached **presentation rubric** for what you should present and how you will be graded.

Student Summarization for Another Group Presentation [4 pts]

- **Deadline: Wednesday, December 15 by 11:59pm CST on Github.**
- **Purpose**:
  - For presenters:
    - The purpose of this final part of the project **for the presenters** is to give the presenting teams constructive feedback on how clearly they were able to communicate and answer their research questions with their analyses and how well they were able to motivate their research to a peer.
  - For listeners:
    - The purpose of this final part of the project **for the listeners** is to gain practice being able to extract the most important parts of an oral research presentation.
- **Steps**:
  - On the day of the presentations, you (as an individual) will be randomly assigned to another group presentation.
  - After watching this group's presentation, you should fill out the **"Student Summarization of Presentation"** document and submit it individually on Compass..
  - The group that you summarized in this report will be able to see the constructive feedback and your summarization.
  - If you are unclear about how to answer the questions in this document, you are encouraged to reach out to the group that you were assigned to for clarification.
- **Graded**:
  - For completeness

Dataset Options

You can choose your own dataset or you can choose from one of the two supplied datasets below. The csvs for each of these datasets are located in the same folder that this document is in. There is more information about each of these datasets below.

Choosing your Own Dataset

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

https://www.kaggle.com/datasets

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

If you decide to choose your own dataset, it must meet the following specifications.

Dataset Size Specifications

Your dataset should have:

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

Data Cleaning and Scaling

Before moving onto to checking whether the dataset is clusterable below, you should think about any type of cleaning and scaling that would need to be done in order to create an insightful analysis. Do this cleaning and scaling before moving on to the clusterability check.

Clusterability and Cluster Algorithm Fit Specifications

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**.
  - You should apply the t-SNE algorithm on your **scaled** and/or **unscaled** dataset (depending on what you intend to use).
  - If the t-SNE algorithm suggests that there is a clustering structure, your dataset has passed this check.
- **Clustering Algorithm Suitability**
  - Next, you want to make sure that you *know of* at least two clustering algorithms that will able to cluster this particular type of dataset. For instance, if this is a numerical, structure dataset, then you know many clustering algorithms that can take this dataset as input. (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 choose your own)

1. **2016 U.S. Primary Data**: The election.csv contained in the zip file contains the voting information for all of the U.S. counties in the 2016 U.S. presidential election. For each county, the percentage and number of votes that went to each primary presidential candidate is listed. This dataset can be joined with other available U.S. county datasets if that is something that interest you. More information on this dataset can be found here. https://corgis-edu.github.io/corgis/csv/election/

2. **notMNIST Letter Image Dataset** In the notMNIST_sample500.csv in the zip file, 50 different font types were randomly selected. Then, for each of these font types, 28-by-28 pixel images of the first ten letters (ie. "A", "B",…, "J") were collected. Thus, this dataset is comprised of the 500 images of letters, each with 784 pixel values. I collected a random sample of the images that were originally collected and constructed here: http://yaroslavvb.blogspot.com/2011/09/notmnist-dataset.html.

Project Report

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

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. 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. Data Cleaning and Data Manipulation

If you perform any type of data cleaning and/or manipulation, show this cleaning and data manipulation process. Also, discuss what you did here in this section and why you did it.

3. Basic Descriptive Analytics

Before using any unsupervised learning algorithms, you should learn more about your dataset by performing some basic descriptive analytics.

1. **If your dataset is a structured dataset (ie. not image, audio, time-series data etc.), do the following.**
    i. For your numerical attributes, calculate basic summary statistics about each attribute.
    ii. For any categorical attributes (including the pre-assigned class labels, if your dataset has any) count up the number of observations of each type.
    iii. Determine if there exist are any strong pairwise relationships between at least one pair of attributes in your dataset and visualize these relationships.

2. **If your dataset is an image dataset, do the following.**
    i. <u>If your dataset has pre-assigned class labels:</u>
        1. Visualize the first few images of each type of class-label.
        2. Discuss how much image variability each of the classes has, and what image elements are different.
    ii. <u>If your dataset DOES NOT have pre-assigned class labels:</u>
        1. Visualize a random sample of images from this dataset.
        2. Discuss how much image variability the image in your dataset have, and what image elements are different.

4. Scaling Decisions

From your analyses conducted here, discuss whether you should scale the dataset or not. Explain why or why not. If you choose to scale, then do so in this section here.

5. Clusterability and Clustering Structure 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. **Does your analysis suggest that the dataset clusterable?** (The answer to this should be yes). Explain why.

2. **Describe the Underlying Clustering Structure of the Dataset**
   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?

6. Algorithm Selection Motivation

Next, using your research goals (section 1), your background research (section 1), your findings from your descriptive analytics section (section 3), and/or your findings from your clusterability section (section 5) explain why you chose the two (or more) unsupervised learning algorithms to use intend to use on your dataset.

7. Clustering the Dataset and Post-Cluster Analysis for <u>Algorithm 1</u>

7.1. <u>Parameter Selection</u>

Select the parameters that you intend to use for this clustering algorithm. Explain and show your 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.

7.2. <u>Clustering Algorithm</u>

Cluster the dataset with this algorithm and the parameters that you chose. Make sure to use a random state for non-deterministic algorithms.

7.3. <u>Clustering Algorithm Results Presentation</u>

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.

7.4. Assessing Clustering Result Separation and Cohesion

For the clustering returned by your clustering algorithm **create a silhouette plot** and **calculate the average silhouette score.** Discuss the separation and cohesion of a.) each of the clusters and b.) the overall clustering. Are there are any objects that have poor cohesion with their assigned cluster? Explain.

Note:

- If you used a hierarchical clustering algorithm, select the "best" clustering from the list of nested clusterings returned by the algorithm.
- If you used an algorithm which returns a fuzzy clustering, create a hard partition by assigning each object to the cluster that it most belongs to.

7.5. Additional Analysis

- ***If* your dataset had pre-assigned class labels [Supervised Learning Evaluation]**
  - Calculate the following and interpret the result.
    - Adjusted RAND Index between the clustering and the class labels.
    - Homogeneity score between the clustering and the class labels.
    - Completeness score between the clustering and the class labels.
  - 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?

- ***If* your dataset DID NOT have pre-assigned class labels [Cluster Distance]**
  - Use a **cluster-sorted similarity matrix** to determine which clusters are closer to eachother than others.

7.6. Describing Each of the Clusters

Finally, describe what type of attribute values and attribute relationships characterize each of the resulting clusters in your final clustering. You can choose at least one of these options (or pick multiple options to learn more).

- Option 1 (don't use this if your dataset is an image dataset):
  - Create a side-by-side boxplots visualization for each numerical attribute in your dataset (where each cluster label is given a boxplot).

- Create a side-by-side barplot visualization for each categorical attribute in your dataset (where each cluster label appears on the x-axis).
- Use these plots to **thoroughly** describe which type of attribute **values** characterize each of the resulting clusters in your final clustering.

- Option 2 (if you used a prototype-based clustering algorithm):
  - If you clustering algorithm is a prototype-based clustering algorithm, display (visualize if it's an image dataset) and compare each of the prototypes of the clusters.
  - Use these prototypes to **thoroughly** describe which type of attribute **values** characterize each of the resulting clusters in your final clustering.
- Option 3: (If you used NMF)
  - If you used NMF to cluster the rows of a dataset, display (visualize if it's an image datset) the rows of H.
  - Use these rows of H to **thoroughly** describe which type of attribute **values** characterize each of the resulting clusters in your final clustering.
- Option 4:
  - Perform PCA on each of the K resulting cluster observations.
  - Use each of these K PCA results to **thoroughly** describe which type of attribute **relationships** characterize each of the resulting clusters in your final clustering.

8. Clustering the Dataset and Post-Cluster Analysis for Algorithm 2

(ie. Do everything you did for clustering algorithm 1 and do it again for clustering algorithm 2).

8.1. Parameter Selection

Select the parameters that you intend to use for this clustering algorithm. Explain and show your 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.

8.2. Clustering Algorithm

Cluster the dataset with this algorithm and the parameters that you chose. Make sure to use a random state for non-deterministic algorithms.

8.3. Clustering Algorithm 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.

8.4. Assessing Clustering Result Separation and Cohesion

For the clustering returned by your clustering algorithm **create a silhouette plot** and **calculate the average silhouette score.** Discuss the separation and cohesion of a.) each of the clusters and b.) the overall clustering. Are there are any objects that have poor cohesion with their assigned cluster? Explain.

Note:

- If you used a hierarchical clustering algorithm, select the "best" clustering from the list of nested clusterings returned by the algorithm.
- If you used an algorithm which returns a fuzzy clustering, create a hard partition by assigning each object to the cluster that it most belongs to.

8.5. Additional Analysis

- ***If* your dataset had pre-assigned class labels [Supervised Learning Evaluation]**
  - Calculate the following and interpret the result.
    - Adjusted RAND Index between the clustering and the class labels.
    - Homogeneity score between the clustering and the class labels.
    - Completeness score between the clustering and the class labels.
  - 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?

- ***If* your dataset DID NOT have pre-assigned class labels [Cluster Distance]**
  - Use a **cluster-sorted similarity matrix** to determine which clusters are closer to eachother than others.

8.6. Describing Each of the Clusters

Finally, describe what type of attribute values and attribute relationships characterize each of the resulting clusters in your final clustering. You can choose at least one of these options (or pick multiple options to learn more).

- Option 1 (don't use this if your dataset is an image dataset):
  - Create a side-by-side boxplots visualization for each numerical attribute in your dataset (where each cluster label is given a boxplot).
  - Create a side-by-side barplot visualization for each categorical attribute in your dataset (where each cluster label appears on the x-axis).
  - Use these plots to **thoroughly** describe which type of attribute **values** characterize each of the resulting clusters in your final clustering.
- Option 2 (if you used a prototype-based clustering algorithm:
  - If you clustering algorithm is a prototype-based clustering algorithm, display (visualize if it's an image dataset) and compare each of the prototypes of the clusters.
  - Use these prototypes to **thoroughly** describe which type of attribute **values** characterize each of the resulting clusters in your final clustering.
- Option 3: (If you used NMF)
  - If you used NMF to cluster the rows of a dataset, display (visualize if it's an image datset) the rows of H.
  - Use these rows of H to **thoroughly** describe which type of attribute **values** characterize each of the resulting clusters in your final clustering.
- Option 4:
  - Perform PCA on each of the K resulting cluster observations.
  - Use each of these K PCA results to **thoroughly** describe which type of attribute **relationships** characterize each of the resulting clusters in your final clustering.

9. Analysis Summary and Conclusion

1. Algorithm Comparison Summary

i. Algorithm Performance

Given your research goals and motivation stated at the beginning of your analysis, compare and contrast the performance of your two clustering algorithms.

ii. Algorithm Results:

Compare and contrast the results of your two clustering algorithms (ie. what you discovered about the clusters in 7.6 and 8.6)

2. Conclusion and Insights Summary

Summarize the insights you found with all of your analyses and relate how these insights might be useful towards your research motivation.

10. Group Contribution Report

If you worked in a group, list the contributions of each team member.

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	1
(Section 2) Data cleaning: *** code correctness** *** explained well** *** correct decisions**	**2**
(Section 1) Dataset research - motivation: *** explains motivation well** *** cites three resources** *** correct citation format**	**4**
(Section 1) Dataset research - information: *** discusses dataset collection** *** discusses data preprocessing**	**2**
(Section 3) Descriptive Analytics *** does what is asked** *** correctness**	**4**
(Section 4) Scaling *** discusses why they scaled or did not scale** *** correctness**	**2**
(Section 5) Clusterability and Clustering Structure *** is dataset clusterable (correct explanations and code)** *** clustering structure (correct explanations and code)** **- how many clusters** **- cluster shapes** **- clusters balanced in size** **- overlapping clusters?**	**5**
(Section 6) Algorithm Selection Motivation *** well explained and correct**	**4**
(Section 7.1.) Parameter Selection *** correctness and logical explanations**	**3**
(Section 7.2.) Clustering Algorithm *** correctness**	**2**
(Section 7.3.) Clustering Algorithm Results Presentation *** correctness** *** clear and insighful**	**2**

(Section 7.4.) Assessing Clustering Result Separation and Cohesion *** correctness, correct interpretations**	**2**
(Section 7.5.) Additional Analysis *** correctness, correct interpretations**	**3**
(Section 7.6.) Describing Each of the Clusters *** correctness, correct interpretations**	**5**
(Section 8.1.) Parameter Selection *** correctness and logical explanations**	**3**
(Section 8.2.) Clustering Algorithm *** correctness**	**2**
(Section 8.3.) Clustering Algorithm Results Presentation *** correctness** *** clear and insighful**	**2**
(Section 8.4.) Assessing Clustering Result Separation and Cohesion *** correctness, correct interpretations**	**2**
(Section 8.5.) Additional Analysis *** correctness, correct interpretations**	**3**
(Section 8.6.) Describing Each of the Clusters *** correctness, correct interpretations**	**5**
(Section 9.1) Comparing clustering performance/results *** correctness** *** explains well** *** maps back onto research motivation**	**4**
(Section 9.2) 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 *** writes in complete sentences** *** titles/headings**	**3**
SUM	70

STAT430 Project Presentation Rubric (26 points)

Team Members: _____

SLIDES / 20

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 (2)

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

Layout (3)

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

PRESENTATION / 6

Narrative / Motivation (3)

- 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.

Presentation (3)

- All team members speak and present some portion of the material.
- Team members speak loud enough for everyone to hear
- Team members understand the material, they are not reading directly from a notecard or script.