Project 2 - Clustering

CS (STAT) 5525 **Total Points: 39**

Instructions and Experiments

Note: Please read the entire project description before you begin. The goal of this project is to analyze the performance of clustering algorithms on several synthetic and real-world data sets. This will be done in the following steps:

- First, you will explore the data sets.
- Next, you will perform a series of experiments on which you will be asked to answer a series of questions. For these experiments, you will be running a python Jupyter notebook.
- Third, you will compile your answers in the form of a report.

Python Jupyter Notebooks

We recommend installing Jupyter using Anaconda as it will also install other regularly used packages for scientific computing and data science. Some pointers to setup Jupyter notebooks on your system:

- Video link - *https://www.youtube.com/watch?v=MvN7Wdh0Juk*
- Medium Link - *https://medium.com/@neuralnets/beginners-quick-guide-for-handling-issues-launching-jupyter-notebook-for-python-using-anaconda-8be3d57a209b*
- Tutorials link - *https://www.dataquest.io/blog/jupyter-notebook-tutorial/*, *https://www.youtube.com/watch?v=3C9E2yPBw7s*

Before you Begin

- Visually explore the data sets in the experiments below, and consider the following:

  types of attributes

  class distribution

  which attributes appear to be good predictors, if any

  possible correlation between attributes

  any special structure that you might observe

  Note: The discussion of this exploration is not required in the report, but this step will help you get ready to answer the questions that follow

Report and Submission

- Write a report addressing the experiment questions. Your project will be evaluated based only on what you write on the report. Submit the report as a PDF file on Canvas.

- Collect output from all your experiments and submit your Jupyter notebooks (cells displaying output) electronically as a separate zipped file on Canvas. We will look at your outputs if something is ambiguous in your report. Copy and paste the output from the Jupyter notebook into your report only to the limited extent needed to support your answers.

1 Problem 1 [17 points]

The files for this problem are under Experiment 1 folder. Datasets to be used for experimentation: **2d_data, chameleon, elliptical, and vertebrate**. Jupyter notebook: **cluster_analysis.ipynb**. In this experiment, you will use different clustering techniques provided by the scikit-learn library package to answer the following questions:

1. (4 points points) On the movie ratings dataset, k-means clustering assign users to two clusters: cluster 0 has users with more affiinity for horror movies, and cluster 1 has users with more affinity for action movies. Given the cluster centroids, assign the following users to their respective cluster assignment:

User	Exorcist	Omen	Star Wars	Jaws
Paul	4	5	2	4
Adel	1	2	3	4
Kevin	2	3	5	5
Jessi	1	1	3	2

2. (2 points points) To determine the optimal value of K in K-means, a common approach is to use the Elbow Method, where the idea is to find a K value that shows the sharpest change in slope of the SSE curve. For the movie rating dataset, what value of K would you arrive at by applying the Elbow Method visually? Briefly explain your reasoning.

3. (4 points points) On the Vertebrate dataset, we illustrate the results of using three hierarchical clustering algorithms (1) single link (MIN), (2) complete link (MAX), and (3) group average. Given the class label in the original dataset, compute the cophenetic correlation coefficient of the clustering produced by each algorithm. Which clustering algorithm shows the best match with the class labels?

4. (5 points points) On the chameleon dataset, how many clusters are produced by DBSCAN when the minimum number of points (min_samples) is set to 1, 2, 3, 4, and 5, respectively, while neighborhood radius (eps) is set to a constant value of 15.5. For each instance, copy and paste the plot of the clusters.

5. (2 points points) For elliptical and 2D data, we applied k-means with k = 2. What happens if we use k = 10 for both these datasets? Copy and paste the clusters formed.

2 Problem 2 [4 points]

The files for this problem are under Experiment 2 folder. Datasets to be used for experimentation are : **samsung_test_labels, samsung_train_labels, samsung_train, samsung_test**. Jupyter notebook: **pca_and_clustering.ipynb**. The data comes from the accelerometers and gyros of Samsung Galaxy S3 mobile phones
(*https://archive.ics.uci.edu/ml/datasets/Human+Activity+Recognition+Using+Smartphones*).

In this data, the type of activity a person was performing with a phone in their pocket is also known - whether they were walking, standing, lying down, sitting, walking up or walking down the stairs. Answer the following questions:

1. (2 points points) Let us look at the correspondence between the cluster labels and the original activity class labels. We see that each cluster has points coming from multiple classes, and is thus impure. Let's look at the maximum percentage of points in a cluster that are coming from a single class, which we can call as the 'purity' metric. For example, if a cluster consisting of 300 points has the following distribution of class labels:

   - class 1 - 200
   - class 3 - 50
   - class 6 - 50

   then the purity metric for this cluster will be 200/300, which is approximately 0.67. A higher value of this metric for a cluster signifies higher purity of the cluster. Compute this metric for all of the 6 clusters produced by running Kmeans with K = 6 on the given dataset. What is the maximum purity metric across all 6 clusters?

2. (2 points points) What is the maximum purity metric for any cluster if we run Kmeans with K = 10 on the same dataset? Explain the rise/fall in purity as we increase K from 6 to 10.

3 Problem 3 [18 points]

The files for this problem are under Experiment 3 folder. Jupyter notebook: **covid-19-research-challenge.ipynb**. In this experiment, we will be looking at the problem of clustering real-world research articles related to COVID-19. **Dataset Download URL**: `https://drive.google.com/file/d/1ICOs9QoBLWFN9tRI-z2QbJJWgngfAm8w/view?usp=sharing` (Filename: CORD-19-research-challenge.zip, File size: 1.58 GB). Please download and unzip this file in the Experiment 3 folder before running the Python notebook for this problem.
Dataset Description: In response to the COVID-19 pandemic, the White House and a coalition of leading research groups have prepared the COVID-19 Open Research Dataset (CORD-19). CORD-19 is a resource of over 29,000 scholarly articles, including over 13,000 with full text, about COVID-19, SARS-CoV-2, and related coronaviruses. This freely available dataset is provided to the global research community to apply recent advances in natural language processing and other AI techniques to generate new insights in support of the ongoing fight against this infectious disease. There is a growing urgency for these approaches because of the rapid acceleration in modern coronavirus literature, making it difficult for the medical research community to keep up.
Answer the following questions.

1. (2 points points) After handling duplicates, what is the count, mean, standard deviation minimum, and maximum values for the abstract word count and body word count?

2. (2 points points) Given the following word list: ['the', '2019', 'novel', 'coronavirus', 'sarscov2', 'identified', 'as', 'the', 'cause'], what is its corresponding list of 2-grams ?

3. (4 points points) When we applied k-means clustering with K = 10 on the data created using HashingVectorizer features from 2-grams, we could see that some clusters still had some overlap in the t-SNE plot. Can you improve this by changing the number of clusters? What value of K visually leads to good separation among the clusters in the t-SNE plot? Copy and paste the corresponding t-SNE plot.

4. (4 points points) By using tf-idf vectorizer and plain text features instead of 2-grams, we could see that the clusters obtained from K-means clustering (with K = 10) are more separable in the t-SNE plot. What happens when we apply the tf-idf vectorizer on the 2-gram representation of documents instead of plain text, and then apply K-means clustering with K = 10? Copy and paste the corresponding t-SNE plot.

5. (6 points points) In the interactive t-SNE with 20 clusters, can you do a manual analysis of different clusters to see what articles are clustered together? Choose any 5 clusters and write 4-5 keywords that describe it. Hover your mouse over the cluster point and you can see the article that it refers. You can use the box zoom feature to choose to display points of only one cluster in the plot, to simplify your analysis. Also, name the clusters that include articles involving social and economic impacts of the coronavirus?