CSE 5243
Instructor: Thomas Bihari
Homework 4: Clustering

Objective:

In this lab, you will perform clustering on three datasets. Your will choose suitable clustering algorithms, evaluate them on the datasets, and compare their performance.

The objectives of this assignment are:

1. Understand how to select and evaluate suitable off-the-shelf clustering algorithms based on the characteristics of a dataset and the outcomes you need.
2. Understand how to tune and evaluate a clustering algorithm to achieve good performance.

The Datasets:

- The file **small_Xydf.csv** is a two-dimensional dataset with 100 records. It contains columns X0, X1, and y. The y column is the actual cluster number that was produced by the dataset generation algorithm. Do not use it for the clustering algorithm. It will be used to evaluate your clustering algorithm below.

- The file **large1_Xydf.csv** is a two-dimensional dataset with 2000 records. It contains columns X0, X1, and y. The y column is the actual cluster number that was produced by the dataset generation algorithm. Do not use it for the clustering algorithm. It will be used to evaluate your clustering algorithm below.

- The file **large2_Xydf.csv** is another two-dimensional dataset with 2000 records, and characteristics different from the "large1" dataset. It contains columns X0, X1, and y. The y column is the actual cluster number that was produced by the dataset generation algorithm. Do not use it for the clustering algorithm. It will be used to evaluate your clustering algorithm below.

Approach:

This homework will make use of the Clustering Algorithms Offered by the SciKitLearn Library. Study the information at https://scikit-learn.org/stable/modules/clustering.html .

1. **Evaluate the K-Means Algorithm on the Small Dataset**
   - **A.** Given that you know the **true** clusters (from column y in the original data), compute the **true** within-cluster WSS, the between-cluster BSS, and the overall SSE.
   - **B.** Configure and run the SciKitLearn K-Means algorithm on the Small dataset. Explain all configuration parameter values you chose, and why you chose them.
   - **C.** Run your algorithm for K=2, 3, 4. For each run, compute the within-cluster WSS, the between-cluster BSS, and the overall SSE, and compute the running time (see Python Time reference below – see %%time, time.process_time(), etc.).
   - **D.** For the K=3 case above:
     1. Create a scatterplot, overlaying the true cluster with the cluster produced by your algorithm.
     2. Create a cross tabulation matrix (i.e., confusion matrix) comparing the true and assigned clusters, and the basic measures (precision, recall, F1, accuracy, etc. – see classification_report()).
   - **E.** What do you observe or conclude from these experiments? Which is your "preferred" clustering (K value in particular), and why? Support this with statistics and/or graphs.

2. **Evaluate the K-Means Algorithm on the Large1 Dataset**
   - **A.** Given that you know the **true** clusters (from column y in the original data), compute the **true** within-cluster WSS, the between-cluster BSS, and the overall SSE.
   - **B.** Configure and run the SciKitLearn K-Means algorithm on the Large1 dataset. Explain all configuration parameter values you chose, and why you chose them.
   - **C.** Run your algorithm for K=6, 8, 10. For each run, compute the within-cluster WSS, the between-cluster BSS, and the overall SSE, and compute the running time.
   - **D.** For the K=8 case above:
     1. Create a scatterplot, overlaying the true cluster with the cluster produced by your algorithm.
     2. Create a cross tabulation matrix (i.e., confusion matrix) comparing the true and assigned clusters, and the basic measures (precision, recall, F1, accuracy, etc. – see classification_report()).
   - **E.** What do you observe or conclude from these experiments? Which is your "preferred" clustering (K value in particular), and why? Support this with statistics and/or graphs.

3. **Evaluate the K-Means Algorithm on the Large2 Dataset**
   - **A.** Given that you know the **true** clusters (from column y in the original data), compute the **true** within-cluster WSS, the between-cluster BSS, and the overall SSE.

B. Configure and run the SciKitLearn K-Means algorithm on the Large2 dataset. Explain all configuration parameter values you chose, and why you chose them.

C. Run your algorithm for K=2, 3, 4. For each run, compute the within-cluster WSS, the between-cluster BSS, and the overall SSE, and compute the running time.

D. For the K=2 case above:

1. Create a scatterplot, overlaying the true cluster with the cluster produced by your algorithm.
2. Create a cross tabulation matrix (i.e., confusion matrix) comparing the true and assigned clusters, and the basic measures (precision, recall, F1, accuracy, etc. – see classification_report()).

E. What do you observe or conclude from these experiments? Which is your "preferred" clustering (K value in particular), and why? Support this with statistics and/or graphs.

4. Evaluate a Second Clustering Algorithm on the Large2 Dataset

A. Choose a second clustering algorithm from the SciKitLearn library. Explain why you chose it.

B. Configure and run your algorithm for (at least) **two** variations of the configuration settings (if any). Explain all configuration parameter values you chose, and why you chose them. For each run:

1. Compute the within-cluster WSS, the between-cluster BSS, and the overall SSE.
2. Compute the running time.
3. Create a scatterplot, overlaying the true cluster with the cluster produced by your algorithm.
4. Create a cross tabulation matrix (i.e., confusion matrix) comparing the true and assigned clusters, and the basic measures (precision, recall, F1, accuracy, etc. – see classification_report()).

C. Which is your "preferred" clustering (configuration settings, if any), and why? Support this with statistics and/or graphs. What do you observe or conclude from these experiments?

5. Evaluate a Third Clustering Algorithm on the Large2 Dataset

A. Choose a third clustering algorithm from the SciKitLearn library. Explain why you chose it.

B. Configure and run your algorithm for (at least) **two** variations of the configuration settings (if any). Explain all configuration parameter values you chose, and why you chose them. For each run:

1. Compute the within-cluster WSS, the between-cluster BSS, and the overall SSE.
2. Compute the running time.
3. Create a scatterplot, overlaying the true cluster with the cluster produced by

your algorithm.

4. Create a cross tabulation matrix (i.e., confusion matrix) comparing the true and assigned clusters, and the basic measures (precision, recall, F1, accuracy, etc. – see classification_report()).

C. Which is your "preferred" clustering (configuration settings, if any), and why? Support this with statistics and/or graphs. What do you observe or conclude from these experiments?

6. Comparison of the Three Clustering Algorithms on the Large2 Dataset

Compare the results of your experiments on the Large2 dataset.

A. What was their relative performance (quality and timing), and their performance versus the true clustering? What characteristics of the data might impact the relative performance?

B. Choose one of the clustering algorithms as best, and explain why.

7. Conclusions

Write a paragraph on what you discovered or learned from this homework.

Collaboration:

For this assignment, you should work as an individual. You may informally discuss ideas with classmates, to get advice on general Python usage, etc., but your work should be your own. **Please make use of Piazza!**

What you need to turn in:

1) **Code**

A. For this homework, the code **is** the Jupyter Notebook. Use the provided Jupyter Notebook template and fill in the appropriate information.

B. You may use common Python libraries for I/O, data manipulation, data visualization, etc. (e.g., NumPy, Pandas, MatPlotLib,… See reference below.)

C. You may ***not*** use library operations that perform, in effect, the "core" computations for this homework (e.g., If the assignment is to **implement** a K-Means algorithm, you may not use a library operation that, in effect, does the core work needed to implement a K-Means algorithm.). When in doubt, ask the grader or instructor.

D. The code must be written by you, and any significant code snips you found on the Internet and used to understand how to do your coding for the core functionality must be attributed. (You do not need to attribute basic functionality – matrix operations, IO, etc.)

E. The code must be commented sufficiently to allow a reader to understand the algorithm without reading the actual Python, step by step.
F. When in doubt, ask the grader or instructor.

2) Written Report

A. For this homework, the report **is** the Jupyter Notebook. The report should be well-written. Please proof-read and remove spelling and grammar errors and typos.
B. The report should discuss your analysis and observations. Key points and findings **must** be written in a style suitable for consumption by non-experts. Present charts and graphs to support your observations. If you performed any data processing, cleaning, etc., please discuss it within the report.

Grading Criteria:

1. **Overall readability and organization of your report (10%)** - Is it well organized and does the presentation flow in a logical manner; are there many grammar and spelling mistakes; do the charts/graphs relate to the text, etc.
2. **Evaluation of the KNN Clustering Algorithmon the Small Dataset (15%)** – Is your configuration sound? Have you made an effort to tune the algorithm for good performance? Are your analyses and evaluations sound, and supported by suitable statistics and/or visualizations?
3. **Evaluation of the KNN Clustering Algorithm on the Large1 Dataset (15%)** – Is your configuration sound? Have you made an effort to tune the algorithm for good performance? Are your analyses and evaluations sound, and supported by suitable statistics and/or visualizations?
4. **Evaluation of the KNN Clustering Algorithm on the Large2 Dataset (15%)** – Is your configuration sound? Have you made an effort to tune the algorithm for good performance? Are your analyses and evaluations sound, and supported by suitable statistics and/or visualizations?
5. **Evaluation of the Second Clustering Algorithm on the Large2 Dataset (15%)** – Is your choice of algorithm and your configuration sound? Have you made an effort to tune the algorithm for good performance? Are your analyses and evaluations sound, and supported by suitable statistics and/or visualizations?
6. **Evaluation of the Third Clustering Algorithm on the Large2 Dataset (15%)** – Is your choice of algorithm and your configuration sound? Have you made an effort to tune the algorithm for good performance? Are your analyses and evaluations sound, and supported by suitable statistics and/or visualizations?
7. **Comparison of the Three Clustering Algorithms (10%)** - Is the comparison sound? Did

you choose a specific clustering algorithm as best and explain why?

8. **Takeaways (5%)** – Did you document your overall insights?

How to turn in your work on Carmen:

Please follow these instructions exactly - it helps the grading process. If you have questions, please ask. Submit to Carmen any code that you used to process and analyze this data. You do not need to include the input data. All the related files (code and/or report) except for the data should be archived in a *.zip file (with no folder trees inside) and submitted via Carmen. The submitted file should be less than 5MB. Use this naming convention: **Homework2_*Surname_DotNumber*.zip**

References and Acknowledgements:

1. SciKit-Learn:
    a. https://scikit-learn.org/stable/modules/classes.html#module-sklearn.cluster
    b. https://scikit-learn.org/stable/modules/clustering.html
    c. https://docs.python.org/3/library/time.html