Northeastern University
DS5230 – Unsupervised Machine Learning
Spring 2025 Semester, Jonathan Mwaura

Group Problem Set #1

This group problem set #1 requires the presentation of results at **6.00 PM on 26th, February 2025** . The presentation will include all research and experiments carried out over the next two weeks. The code repository and the slides must be submitted by **11.30 PM on 27th, Feb 2025** and will be submitted via Canvas (note that there shall be no late deadline).

The experimentation, presentation of the results and the report will carry 100 points. Thorough work will be required.

Note that you need to carry out this piece of work with the group that you shall be working with on your end-of-semester project.

The written report is due at **11.30 PM on 11th, March 2025**

Although a solution must be absolutely perfect to receive full marks, I will be generous in awarding partial marks for incomplete solutions that demonstrate progress.

So that there is no ambiguity, there are two non-negotiable rules. A violation of either rule constitutes plagiarism and will result in you receiving an F for this course.

(a) This is a group assignment; I expect the work to be consensus work drawn from each member of the group. The process of finding a solution might take 3 - 5 iterations or even more, BUT you learn from all these attempts, and your confidence grows with each iteration.

(b) These problems might seem difficult at first glance. They are designed to be such. We learn by attempting problems, struggling through them, and coming on top. I encourage you to make this learning exercise worth it. What do I mean? Open the problem sets as early as you get them, then do not look at hints or answers any where (including on the Internet and consulting other students for direct answers), give it the best shot you can. If you get stuck, come to Professor or TA's office hour and we shall be glad to listen to your rationale and work with you till you are able to tackle the problem sets.

(c) Note that the use of LLM is allowed for research and streamlining of your thought but make sure that any ideas coming from LLM are properly cited and referenced. Do not pass LLM work as yours because then you will not have learned how to carry out the exercise and you will not have built the data science skill targeted by this assignment.

The goal of this assignment is to explore and compare five different clustering techniques in five categories of algorithms. Each group will implement one clustering technique for each category and evaluate their performance on two real-world datasets.

Group	Partitional	AHC with:	Density	Grid	Model & Graph
Group 1	Spectral Clustering	BIRCH	DBSCAN	CLIQUE	Gaussian Mixture Model
Group 2	MiniBatch K-Means	Complete Linkage	OPTICS	STING	Spectral Co-Clustering
Group 3	Bisecting K-Means	Average Linkage	HDBSCAN	WaveCluster	Gaussian Mixture Model
Group 4	AffinityPropagation	Centroid	DENCLUE	CLIQUE	Bayesian Clustering
Group 5	K-Medoids	Ward Linkage	MiniShift	STING	Spectral Bi-Clustering

Table 1: Clustering Algorithm Assignments for Each Group, note BIRCH (under hierarchica) is a clustering algorithm not a linkage

(i) **Dataset Selection (10)**

- Select **two real-world datasets** from sources such as Kaggle, UCI Machine Learning Repository, or OpenML.
- Justify the dataset selection based on clustering relevance.
- Preprocess the data (handling missing values, encoding categorical data, normalization/scaling).

(ii) **Dimensionality Reduction (10 points)**

- Apply **at least one** dimensionality reduction technique (PCA, t-SNE, UMAP).
- Justify the choice and visualize the reduced data.

(iii) **Clustering Implementation (30 points)**

- Implement **all five clustering techniques** assigned to your group. You can leverage existing libraries and frameworks
- Tune hyperparameters for each algorithm (e.g., K for K-Means, epsilon for DBSCAN, bandwidth for DENCLUE).
- For dissimilarity metrics, you case Euclidean distance
- Visualize clustering results using scatterplots, dendrograms, or density plots.
- Compare clustering results across the five techniques.
- For each clustering algorithm, a slide is required that discusses the given clustering algorithms. Note some of these may not have been covered in class. The class will rely on you to educate us.

(iv) Cluster Validation (15 points)

- Evaluate clustering results using **at least three** validation techniques:
  - Silhouette Score
  - Davies-Bouldin Index
  - Gap Statistic
  - Calinski-Harabasz Index
  - Elbow Method (if applicable)
- Interpret the results and determine the most effective clustering method for the datasets.

(v) Comparative Analysis & Discussion (20 points)

- Compare clustering techniques **within the same dataset**.
- Discuss **strengths, weaknesses, and best use cases** for each algorithm.
- Compare performance on **both datasets**.
- Analyze the impact of dimensionality reduction on clustering performance.
- Summarize key takeaways and make recommendations based on findings.

(vi) Report & Presentation (15 points)

- Prepare a 15-minute presentation for class discussion (Must be presented on Feb 26th).
- Submit a structured report (4-6 pages) summarizing:
  - Introduction to clustering
  - Dataset details
  - Methodology
  - Results & Discussion
  - Conclusion
- The presentation (how you engage with participates and depth of knowledge will be awarded 5 points, the report (due a week later) will be awarded 10 points.

Deliverables

- **Jupyter Notebook / Python Script** with:
  - Data preprocessing
  - Dimensionality reduction
  - Clustering implementations
  - Visualization
  - Cluster validation
- **Jupyter Notebook / Python Script - These are due on Feb 27th**
- **Presentation Slides on or before 27th Feb, 2025**
- **Written Report (Due on 11th March, 2025)**

Additional Notes

- Use **scikit-learn, SciPy, HDBSCAN, pyclustering, or other relevant libraries**.
- **Matplotlib, Seaborn, or Plotly** should be used for visualization.
- Provide **inline code documentation** in Jupyter Notebooks.
- Clearly explain **hyperparameter choices** for each algorithm.