**Project 2: Clustering Algorithms**

Group #: 1 (you own group number, same as in Proj1)

Tianle Ma, 50096690, [tianlema@buffalo.edu](mailto:tianlema@buffalo.edu)

Other members

1. **KMeans**
2. How it works: workflow of the algorithm and your implementation. (3 points)
3. When it works and when it doesn’t. State pros and cons. (4 points)
4. For each dataset, change the number of clusters – K (you only need to try a few representative, informative numbers which help us see the trend). Calculate an external index (Rand Index or Jaccard Coefficient) and an internal index (Silhouette or Correlation) for each K. Show the results in a table(s) and a plot(s). (8 points)
5. Use PCA to visualize one of the “best” clustering results. (Note: there is no single criteria to measuring clustering results, you can decide on your own based on the above calculated metrics and visualization results.) (2 points)
6. **Hierarchical clustering**

(1) How it works: workflow of the algorithm and your implementation (3 points)

(2) When it works and when it doesn’t. State pros and cons. (4 points)

(3) For each dataset, change the number of clusters – K (you only need to try a few representative, informative numbers which help us see the trend). Calculate an external index (Rand Index or Jaccard Coefficient) and an internal index (Silhouette or Correlation) for each K. Show the results in a table(s) and a plot(s). (8 points)

(4) Use PCA to visualize one of the “best” clustering results. (Note: there is no single criteria to measuring clustering results, you can decide on your own based on the above calculated metrics and visualization results.) (2 points)

1. **One from density-based, mixture model, and spectral clustering**
2. How it works: workflow of the algorithm and/or your implementation (3 points)
3. When it works and when it doesn’t. State pros and cons. (4 points)
4. For DBSCAN, change epsilon and MinPts; for mixture models and spectral clustering, change the number of clusters. Calculate an external index (Rand Index or Jaccard Coefficient) and an internal index (Silhouette or Correlation) for each parameter. Show the results in a table(s) and a plot(s). (8 points)
5. Use PCA to visualize one of the “best” clustering results. (Note: there is no single criteria to measuring clustering results, you can decide on your own based on the above calculated metrics and visualization results.) (2 points)
6. **Comparative analysis** (9 points)

For each dataset, compare the performance among all three (or more) algorithms. Which works best? What are the possible reasons? Do you have any other findings?

1. **Other considerations** (10 points)

If you have done extra good work, depending on what it is, we may or may not consider to give you extra score. Please state it in your report **as well as during the demo**.