COSI 126A: Homework 1

Due by Feb.17

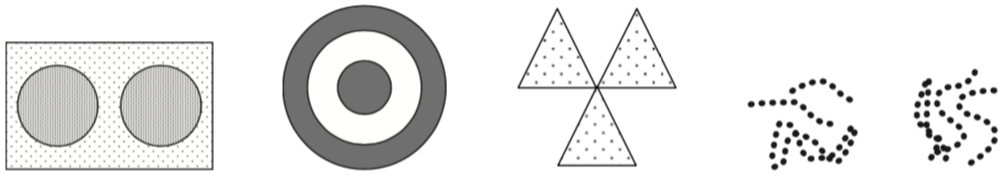
Section I: Clustering Problems

Problem 1 (5 points)

Many partitional clustering algorithms that automatically determine the number of clusters claim that this is an advantage. List two situations in which this is not the case.

Problem 2 (5 points)

Identify the clusters below using the center-, contiguity-, and density- based definitions. Also indicate the number of clusters for each case and give a brief indication of your reasoning. Note that darkness or the number of dots indicates density. If it helps, assume center-based means K-means, contiguity-based means single link, and density-based means DBSCAN.



(a) (b) (c) (d)

Problem 3 (10 points)

Hierarchical clustering algorithms require $O(m^2 \log(m))$ time, and consequently, are impractical to use directly on larger data sets. One possible technique for reducing the time required

is to sample the data set. For example, if K clusters are desired and $\sqrt{m}$ points are sampled from the m points, then a hierarchical clustering algorithm will produce a hierarchical clustering in roughly $O(m)$ time. K clusters can be extracted from this hierarchical clustering by taking the clusters on the K^{th} level of the dendrogram. The remaining points can then be assigned to a cluster in linear time, by using various strategies. To give a specific example, the centroids of the K clusters can be computed, and then each of the $m - \sqrt{m}$ remaining points can be assigned to the cluster associated with the closest centroid.

For each of the following types of data or clusters, discuss briefly if (1) sampling will cause problems for this approach and (2) what those problems are. Assume that the sampling technique randomly chooses points from the total set of m points and that any unmentioned characteristics of the data or clusters are as optimal as possible. In other words, focus only on problems caused by the particular characteristic mentioned. Finally, assume that K is very much less than m.

(a) Data with very different sized clusters.

(b) High-dimensional data.

(c) Data with outliers, i.e., atypical points.

(d) Data with highly irregular regions.

(e) Data with globular clusters.

(f) Data with widely different densities.

(g) Data with a small percentage of noise points.

(h) Non-Euclidean data.

(i) Euclidean data.

(j) Data with many and mixed attribute types.

Problem 4 (8 points)

(a) Compute the entropy and purity for each cluster in the confusion matrix below.

(b) Compute the total entropy and total purity.

(c) Compute the following F-measure: F(#3,Water)

Cluster	Normal	Water	Grass	Fire	Electric	Ground	Flying	Ghost
#1	8	22	0	0	767	4	45	22
#2	654	34	89	123	12	76	13	2
#3	6	301	2	3	98	23	31	1001
#4	4	21	34	2	3	543	112	0

Problem 5 (7 points)

(a) Given the set of cluster labels and similarity matrix shown below, compute the correlation between the similarity matrix and the ideal similarity matrix, i.e., the matrix whose ij^{th} entry is 1 if two objects belong to the same cluster, and 0 otherwise.

(b) Compute the silhouette coefficient for each point, each of the three clusters, and the overall clustering.

Point	Cluster Label
P1	1
P2	1
P3	2
P4	2
P5	3

	P1	P2	P3	P4	P5
P1	1	0.92	0.33	0.61	0.82
P2	0.92	1	0.43	0.01	0.22
P3	0.33	0.43	1	0.75	0.11
P4	0.61	0.01	0.75	1	0.17
P5	0.82	0.22	0.11	0.17	1

Section II: Programming

Please submit a Python package implementing the features specified below. A skeleton file will be provided.

Part I: Cluster Validity (15 points)

1. Implement the following external measurements:
   (a) Accuracy
   (b) Normalized Mutual Information
   (c) Normalized Rand Index
2. Implement the following internal measurements:
   (a) Silhouette Index
   (b) Clustering Validation Index based on Nearest Neighbors (CVNN)

References:

- Liu, Y., Li, Z., Xiong, H., Gao, X., Wu, J., & Wu, S. (2013). "Understanding and enhancement of internal clustering validation measures." *IEEE transactions on cybernetics*, 43(3), 982-994.
- Wu, J., Xiong, H., & Chen, J. (2009, June). "Adapting the right measures for k-means clustering." *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 877-886). ACM.

Part II: K-Means Clustering (20 points)

1. Implement Lloyd's K-means algorithm. Ensure that it works for the basic dataset *three_globs.csv*. Provide the following three initialization methods:
   (a) Random
   (b) K-means++
   (c) Global K-means
2. Implement Hartigan's K-means algorithm.
3. Use the internal measurements in Part 1 to find the proper cluster number for the *image_segmentation.csv* dataset.
4. Given the true cluster number, run your Lloyd's K-means algorithm on the *image_segmentation.csv* dataset, and evaluate the results in terms of the external measurements completed in Part I.

5. Run the algorithm ten times, and record the average, standard deviation, and execution time.

References:

- Arthur, D., & Vassilvitskii, S. (2007, January). "k-means++: The advantages of careful seeding." *Proceedings of the eighteenth annual ACM-SIAM Symposium on Discrete Algorithms* (pp. 1027-1035). Society for Industrial and Applied Mathematics.
- Likas, A., Vlassis, N., & Verbeek, J. J. (2003). "The global k-means clustering algorithm." *Pattern Recognition*, 36(2), 451-461.

Part III: DBSCAN(15 points)

1. Implement your own DBSCAN. 5-point bonus for an implementation within 40 lines.
2. Run your algorithm on the *anthill.csv* dataset. You can use the recommended parameters in the textbook, or you can choose your own. Evaluate the performance in terms of the external measurements in Part 1.

Part IV: Spectral Clustering and Kernel K-means (15 points)

1. Implement your own spectral clustering algorithm.
2. Implement your own Kernel K-means algorithm.
3. Run both algorithms on the *eye.csv* dataset. Demonstrate empirically whether they are the same or not.

References:

- Dhillon, I. S., Guan, Y., & Kulis, B. (2004, August). "Kernel k-means: spectral clustering and normalized cuts." *Proceedings of the tenth ACM SIGKDD international conference on knowledge discovery and data mining* (pp. 551-556). ACM.

Final Notes

- Do not hard-code your work. Expect that it will be run on datasets other than the ones provided.
- Supplementary libraries such as pandas and numpy may be used. However, do not use any features directly related to the algorithms in question.
- Submissions will be checked for plagiarism using MOSS. Plagiarized work will result in a 0 for the assignment and possible disciplinary action.