CSE 4820 / CSE 5819 Fall 2025-Section 001
Instructor: Jinbo Bi
Introduction to Machine Learning

Assignment 7 (100 pts)
Due: Nov 2nd (Sunday) mid-night
Cluster Analysis, Cluster Validity

Name: _____ NetID: _____

Part 1: (ChatGPT Self-Learning) 20pts

This semester we are trying to use some generative AI to help learn ML concepts. Keep in mind that generative AI does not always search from authoritative websites and does not always answer the questions correctly. It is your responsibility to judge and learn from multiple sources based on our in-class discussion. You can start from the following prompts and create subsequent questions that attempt to understand the ML concepts. For each question you have asked chatGPT, just copy the first three lines of its answer into your answer sheet (so to save the space but show that you have studied). Please include this part in the end of your HW after you answer HW part 2.

1. What is cluster analysis?
2. How many kinds of clustering methods are there?
3. What are the pros and cons of k-means?
4. What are the pros and cons of hierarchical clustering?
5. What are the pros and cons of DBSCAN?
6. If data is sparse, is there a faster algorithm to run k-means?
7. What are the best validity indices to measure the validity of clusters?
8. How to compare different clustering results?

Part 2: Answer the Following Problems (80pts).

This part will be graded based on correctness, accuracy and clarify. Please prepare your answers in a pdf file to submit through HuskyCT assignment portal and clearly label your file with part 2 of assignment number. If you decide to use handwriting, make sure your handwriting is readable; or otherwise TAs have all rights to give 0 pt for answers that they cannot read. Please provide your calculation process, based on which you may get partial scores even if your answer is not correct.
Please try not to use ChatGPT (or other generative AI) to answer this part. If you rely too much on generative AI for this part, you may not learn enough and be well prepared for exams.

[**Cluster Analysis – K-means]** [25 pts]
Consider the following set of one-dimensional data points: {0.1, 0.25, 0.45, 0.55, 0.8, 0.9}. All the points are located in the range between [0, 1].

(a) [15 pts] Suppose we apply K-Means clustering to obtain three clusters, A, B, and C. If the initial centroids are located at {0, 0.4, 1}, respectively, show the cluster assignments and locations of the updated centroids after the first three iterations by filling out the following table. (hint: although you can directly calculate all distances, if you draw a real axis, and put down the points, it might be easier to see the cluster assignment.)

Iter	Cluster assignment of data points						Centroid Location		
	0.10	0.25	0.45	0.55	0.80	0.90	A	B	C
1									
2									
3									

(b) [5 pts] Find the sum-of-squared errors (SSE) of the clustering after the third iteration.

(c) [5 pts] For the dataset given in part (1), is it possible to obtain empty clusters? Why?

[**Cluster Analysis – Hierarchical Clustering]** [25 pts]

You are given the following data (5 points with 2 features).

5	60
3	50
1	80
9	50
8	70

(a) [10 pts] Run hierarchical clustering with the MIN and MAX similarity metrics respectively, and draw dendrograms. (You can also plot the points into a figure and circle points for clusters)

(b) [15 pts] Normalize both features with standard normalization. Run again hierarchical clustering with the MIN and MAX similarity metrics respectively, and draw dendrograms. (You can also plot the points into a figure and circle points for clusters)

Part 3 [**Programming**] (30 pts) In earlier HWs, you have used Seaborn Iris dataset to create supervised learning models.

Now let us do more experiments on this data but with cluster analysis. For all of the following clustering results, computer an internal index (e.g., Silhouette index) and an external index (e.g., entropy or purity) for the clusters you obtained.

(1) [10 pts] Use scikit-learn **sklearn.cluster.KMeans** to run K-means with K=3. (you can play with parameters but choose one to report).

(2) [10 pts] Run hierarchical clustering **sklearn.cluster.AgglomerativeClustering** to obtain a dendrogram up to 5 clusters. Then cut down at the level when there are three clusters. Report the indices you choose on this clustering solution.

(3) [10 pts] Run DBSCAN **sklearn.cluster.DBSCAN** on this dataset and report the clusters you obtained and the indices.