DATA 240-21, Fall 2024

Assignment #2

Release on Sept 28th, 2024
Due 11:59pm on Oct 15th, 2024

Notes

This assignment should be submitted in Canvas as a format of ipython notebook (assignment1.ipynb).

No late assignments will be accepted. Do not accept any other format. Minimum penalty is 2pts with acceptable excuse. You may collaborate on homework but must **independently** write code/solutions. Copying and other forms of cheating will not be tolerated and will result in a **zero score** for the homework (minimal penalty) or a failing grade for the course. Your work will be graded in terms of correctness, completeness, and clarity, not just the answer. Thus, correct answers with no or poorly written supporting steps may receive very little credit.

NOTE: Please do not use any package/library including scikit-learn library except NumPy, Pandas, and Matplotlib.

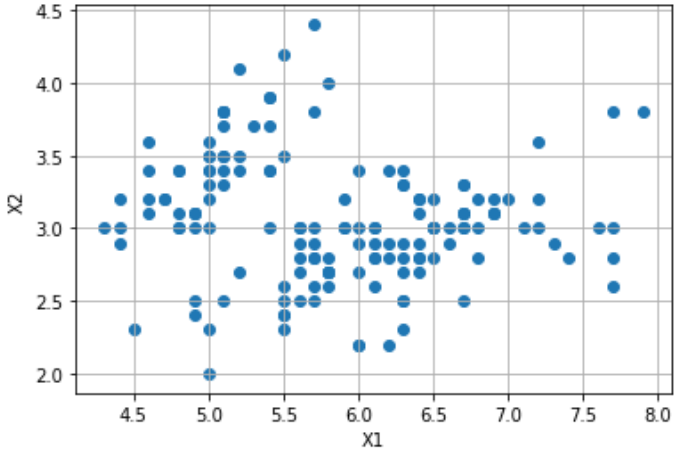
1. (3 pts) Implanting K-means clustering algorithm

Please download cluster_data1.csv.
The sample data is shown below.

X1	X2	X3	X4
6.7	3	5	1.7
6.3	2.9	5.6	1.8
5.6	3	4.5	1.5
7.6	3	6.6	2.1
6	3.4	4.5	1.6
6.4	3.2	5.3	2.3
7.7	2.8	6.7	2
4.8	3	1.4	0.3
5	3	1.6	0.2
5	3.4	1.6	0.4

K-means algorithm is a method to automatically cluster similar data examples together. K-means is an iterative procedure that starts by guessing the initial centroids, and then refines this guess by repeatedly assigning examples to their closest centroids and then recomputing the centroids based on the assignments until they converge.

Let's assume K=3. Please implement K-means clustering algorithm from scratch. Put random seed as '123'. Please plot the location of k centroids and their assignment for each cluster in 2D with different colors to distinguish each cluster and its centroid for the first 5 steps (the initial setting of the centroids, Then, after iteration 1/2/3/4). In your plot, set x-axis as 'X1' and y-axis as 'X2' as the below figure.



2. (3 pts) Implanting K-means++ clustering algorithm and finding K

The converged solution may not always be ideal and depends on the initial setting of the centroids.
To address this issue, K-means++ was introduced. You should implement K-means++ algorithm as described during the class.

Please implement K-means++ algorithm from scratch. Put random seed as '123'.
Please plot the location of k centroids and their assignment for each cluster in 2D with different colors to distinguish each cluster and its centroid for the first 5 steps (the initial setting of the centroids, Then, after iteration 1/2/3/4). In your plot, set x-axis as 'X1' and y-axis as 'X2' as the above figure.

3. (4 pts) Implanting KNN classification

Please download cluster_data2.csv.
The sample data is shown below.

X1	X2	X3	X4
5.21	3.65	1.42	0.25
5.07	3.41	1.43	0.19
5.85	2.65	4.14	1.27
5.64	2.73	4.03	1.23
6.55	2.9	5.54	2.05

Please implement KNN classification algorithm from scratch.
Please assign class for each data point based on the result of problem #2 using K-NN method.