ENGR-UH 4560
Selected Topics in Information and Computational Systems

Mini Project - K-means Clustering Algorithm

Due Date: Refer to NYU Class

Introduction

K-means is one of the widely used unsupervised learning algorithms that solve the well-known clustering problem. The procedure follows a simple and easy way to classify a given data set through a certain number of clusters (assume k clusters). The main idea is to define k centers, one for each cluster. These centers should be placed in a cunning way because of different location causes different result.

Dataset

In this assignment, you will get familiar with generating dataset by yourself using the third-party library.

Requirements

1. Use sklearn library to generate the synthetic data for k-means clustering.
    a. We set the total number of instances to be 300
    b. The number of centers is 4 with the standard deviation 0.6

2. Plot the generated data with labels by using matplotlib.

3. Implement the K-means function return the labels and centers.

4. Apply your implemented K-means on the dataset and plot the figure with seed = 0

5. Apply your implemented K-means on the dataset and plot the figure with seed=2

6. Compare the results from 4 and 5. Is there any differences? If yes, why?

7. Implement the K-means++ function return the labels and centers.

8. Apply your implemented K-means++ on the dataset and plot the figure with seed = 0

9. Apply your implemented K-means++ on the dataset and plot the figure with seed=2

10. Compare the results from 8 and 9. Is there any differences? If yes, why?

11. Compare the results from 4,5,8 and 9. State your observations.

Deliverables

A .ipynb file containing the following:

1. Source code
2. Detailed description of the project if needed

3. Answers to the project questions.

Before submitting your project, please make sure to test your program on the given dataset.

Notes

You may discuss the general concepts in this project with other students, but you must implement the program on your own. ***No sharing of code or report is allowed.*** *Violation of this policy can result in a grade penalty.*

Late submission is acceptable with the following penalty policy:
10 points deduction for every day after the deadline