

# UH-3332 - Applied Machine Learning

## Clustering

Due Date: Refer to Brightspace

# K-means clustering

## 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 into 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 different locations will result in different results.

## Dataset

Test your algorithm on two separate datasets (1) Use `make_blobs` function to generate synthetic data set from `sklearn` library. (2) Use an RGB image to cluster the R,G,B data into  $K$  clusters to demonstrate image compression. Display images before and after

## Requirements

1. Use `sklearn` library to generate the synthetic data for k-means clustering.
  - $\alpha$ . We set the total number of instances to be 300
  - $\beta$ . 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. Fit the model on the dataset (default seed) and plot the figure
5. Fit the model on the dataset (seed=2) and plot the figure
6. Implement the K-means++ function return the labels and centers
7. Fit the model on the dataset (default seed) and plot the figure
8. Fit the model on the dataset (seed=2) and plot the figure
9. Compare the results from 4,5,7 and 8. State your observations

# Hierarchical clustering

## Introduction

Hierarchical clustering involves creating clusters that have a predetermined ordering. For example, all files and folders on the hard disk are organized in a hierarchy.

## Dataset

In this project you will work on the Mall Customer dataset (Mall\_Customers.csv)

## Requirements

Implement a hierarchical clustering model using Ward distance and plot the dendrogram.

## Deliverables

A .ipynb file containing the following:

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

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**