C S 487/519 Applied Machine Learning
Compare clustering methods

1 Objective

In this *individual* homework, you are required to understand and compare several clustering algorithms.

2 Requirements

2.1 Tasks

(1) (45 points) Write code to conduct clustering by
   (a) (15 points) using the K-means algorithm offered by scikit-learn library,
   (b) (15 points) using a hierarchical approach offered by SciPy library, and
   (c) (15 points) using a hierarchical approach offered by scikit-learn library,.

(2) (10 points) Use elbow approach to decide a reasonable K for K-means algorithm.

(3) (20 points) Each cluster algorithm needs to be tested using two datasets: (a) The Iris dataset (`iris.data`) with description (`iris.names.txt`), which can be downloaded from this page, and (b) a subset of the MNIST dataset. You need to think how to utilize such datasets to conduct clustering because these datasets are generally used for classification.

(4) (20 points) Properly analyze the clustering algorithms' behavior by applying the knowledge that we discussed in class. Such analysis should include running time. You can include Sum Squared Error (SSE) analysis. You can also use class labels as ground truth to examine the clustered results.

(5) (5 points) Write a readme file `readme.txt` with detailed instructions to run your program.

2.2 Other requirements

- Your Python code should be written for **Python version 3.5.2 or higher**.
- Please write proper **comments** in your code to help the instructor and teaching assistants to understand it.
- Please properly organize your Python code (e.g., create proper classes, modules).
- You can put your code to Jupyter Notebook or a `.py` file.

3 Submission instructions

Put all your files (Python code, readme file, report, etc.) to a zip file named `hw.zip` and upload it to Canvas.

4 Grading criteria

(1) **ZERO point will be given if your code does not work. Please do not submit code that you did not test and make sure it works.**

(2) The score allocation has been put beside the questions.

(3) FIVE points will be deducted if files are not submitted in the required format.

(4) If the total points are more than 100. Your grades will be scaled to the range of [0,100].

(5) Please make sure that you test your code thoroughly by considering all possible test cases. For this homework, your code will NOT be tested using more datasets. Thus, it does not need to be flexible to accept different datasets as input.