C S 487/519 Applied Machine Learning I
Fall 2018
Project 5: Compare clustering methods

1 Objective

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

2 Requirements

- (45 points) Write code to conduct clustering by
  - (10 points) using the K-means algorithm offered by scikit-learn library,
  - (15 points) using a hierarchical approach offered by SciPy library,
  - (10 points) using a hierarchical approach offered by scikit-learn library, and
  - (10 points) using the DBSCAN density based method offered by scikit-learn library.
- (5 points) Use elbow approach to decide a reasonable K for K-means algorithm.
- (10 points) Write code to decide a reasonable *MinPts* and *eps* for the DBSCAN method.
- (20 points) Each cluster algorithm needs to be tested using two datasets: (1) the `iris` dataset, which is on Canvas, and (2) the Faulty Steel Plates dataset at `https://www.kaggle.com/uciml/faulty-steel-plates`. You need to think how to utilize such datasets to conduct clustering because these datasets are generally used for classification.
- (15 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 points) Write a readme file `readme.txt` with the commands to run your code.
- Your Python code should be written for Python version 3.5.2 or higher.
- Please properly organize your Python code.

3 Submission instructions

- Compress your python code to a zip file named `proj5.zip` and upload it to Canvas.

4 Grading criteria

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

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

(3) At least 5 points will be deducted if submitted files (including files types, file names, etc.) do not follow the instructions.