**C S 487/519 Applied Machine Learning I**

# Compare clustering methods

1. **Objective**

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

# 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 hierarchi cal approach offered by scikit-learn library, and
    - (10 points) using the DBSCAN density based method offered by scikit-learn library. (I may remove this task later)
  + (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) part of the MNIST dataset or the Faulty Steel Plates dataset at [here](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 applyin g 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 code needs to run in command line, accepting as input parameters the algorithm name, the dataset filename, and any required parameter. For example, “python main.py ridge dataset.csv –k 3”
  + Your Python code should be written for Python version 3.5.2 or higher.
  + Please properly organize your Python code. Each required task had better be implemented in a separate python file and imported into the main script. For example, to use the K-means method you can create the script **mykmeans.py**, then use “import mykmeans” in the main.py file to test your implementation.

# Submission instructions

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

# 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.