CS 550 -- Machine Learning
Homework #2
Due: 17:30, December 13, 2019

In this homework, you will implement five clustering algorithms and compare their results on three datasets. All datasets contain instances with two input features. They are provided as three text files (*dataset1*, *dataset2*, and *dataset3*). An individual line of each file corresponds to an instance where the first number is the first feature of that instance and the next one is its second feature.

First implement the following clustering algorithms. For each algorithm, use the Euclidean distance as a (dis)similarity measure. Note that in the remaining part of this document, the names given in bold are used to refer the clustering algorithms. Please also use these names in your report when you refer them.

- **K-means:** K-means clustering algorithm
- **Single-linkage**: Agglomerative hierarchical clustering algorithm with single linkage
- **Complete-linkage**: Agglomerative hierarchical clustering algorithm with complete linkage
- **Group-average**: Agglomerative hierarchical clustering algorithm with group average similarity
- **DBSCAN**: DBSCAN algorithm

Then run your clustering algorithms as specified below.

1. Run all your algorithms on *dataset1*. Select k = 7 for all algorithms except **DBSCAN**. For **DBSCAN**, run the algorithm with different parameters, select the one that you "favor" the most, and use the clusters obtained by this selection (of course, this selection should follow common sense). Note that your selection may yield a number of clusters different than 7. Then, for each algorithm
   - Plot the data points using a different color for each cluster,
   - Calculate the sum of the squared errors $E = \sum_{i=0}^{k} \sum_{x \in D_i} |x - \mu_i|^2$, where μ_i is the mean of the data points belonging to the i^{th} cluster, and
   - Measure the computational time.
2. Run all your algorithms on *dataset2*. Select k = 3 for all algorithms except **DBSCAN**. Then, repeat all steps and get your results similar to the first part.
3. Run all your algorithms on *dataset3*. Select k = 2 for all algorithms except **DBSCAN**. Then, repeat all steps and get your results similar to the first part.

At the end prepare your report. Your report should include

- Three sections, each of which is prepared for the run taken on a particular dataset. Each section should include
  - Five plots, one for each clustering algorithm,
  - A table containing the sum of the squared errors for the five clustering algorithms,
  - A table containing the computational times for the five clustering algorithms (use the same computer to get the runs of all algorithms),
  - The parameters you select for the **DBSCAN** algorithm, and
  - The number of iterations for the **K-means** algorithm.
- A fourth section that includes the discussion for the comparison of the five clustering algorithms based on your findings in your experiments. This section should be no longer than a half page.

This homework asks you to implement five clustering algorithms by **writing your own codes**. Thus, you are not allowed using any machine learning package. In your implementation, you may use any programming language you would like.

Similar to the first assignment, prepare your report neatly and properly. Your report should be a maximum of 4 pages, where the fourth section should be no longer than a half page.

Please submit the hardcopy of your report before the deadline. But this time, submit **a color printout** of your report since it uses different colors to plot the data points of different clusters. DO NOT submit the printout of your source code. However, you need to email the source code of your implementations before the deadline. The subject line of your email should CS 550: HW2.