

Graded Lab Assignment 2: Clustering

Implement the k-means clustering algorithm discussed in Unit 3. Please do go through the lecture material in detail before you start on your implementation to ensure that you fully understand the algorithm. Please run your program with different values of k and with progressively larger datasets and choose the k initial clusters in different ways to get a better perspective on how the k-means algorithm works.

Test the algorithm using any spatial dataset that is freely available in the public domain. Lots of spatial datasets are downloadable from websites such as:

- <http://freegisdata.rtwilson.com/>
- <http://water.usgs.gov/lookup/getgislist>
- <http://www.diva-gis.org/Data>
- [The above are just examples of sites where you can download spatial datasets. Please feel free to download any spatial data that is freely available in the public domain and is legal to download i.e., no copyright restrictions etc.]

Deliverable: Your k-means clustering program should be able to run with at least 10,000 spatial points. You are required to provide a demo of your program on or before the deadline specified below in this document.

Please note the following points:

1. This lab assignment will contribute to **5% of your grades** for the course.
2. Please write the Names and Roll Numbers of your group members in the Comments section at the beginning of your program file. All programs should be submitted on Blackboard on or before the stated deadline. Please note that your programs may be subjected to plagiarism checks.
3. The **deadline** for assignment submission is **Feb 16, 2017, 11.59 pm IST.**
4. This is a **HARD deadline** and no points will be awarded for the assignment if you submit after the deadline, unless there are extenuating circumstances.
5. On or before the deadline, you need to show the demo of your program running with at least 10000 points in 2-dimensions. Of course, you should try to run your program with a much larger number of points, but 10000 points is a bare minimum criterion for the completion of this assignment. If your program does not run with at least 10000 points, you will receive a grade of 0 for this assignment.
6. The grading criteria for this assignment will be based on effort, adherence to learning points from your previous non-graded assignment, code quality, visualization, results and scalability.
7. ***Any act of plagiarism will result in a zero for the entire assignment. Hence, please avoid any form of plagiarism.***
8. This is a group assignment, hence please do NOT collaborate with your fellow students in other groups towards the completion of this assignment. However, you are obviously required to collaborate effectively with members of your own group to ensure that you are able to function as a team player.

This assignment carries 5 points.