Data Mining September 22, 2018
IIIT Delhi CSE 506, Monsoon 2018
Sumit Bhatia Assignment

Assignment 2
Due Sunday, 11:59PM, October 7, 2018

Policy

- This assignment is to be done individually. You will be required to demo your code to the TAs.
- Any help taken from any outside source *must* be reported in your report.
- Institute's policy on plagiarism will apply. *We will use plagiarism detection tool to check for plagiarism in code.* Violation will result in 0 points for this assignment.
- Last date for doubt clearance from instructor: Friday, September 28, 4 PM.
- Last date for doubt clearance from TAs: Friday, October 5, 4 PM.
- Late submissions penalized by 10% every three hours.
- Programming can be done either in Java or Python.

1. (20 points) Write a skeleton program to implement the generalized representative based clustering algorithm discussed in class. Note that the program should be written in a modular style so that you can easily use it for a variety of distance functions and representative definitions. You should have different methods/functions/subroutines for computing distances and determining representatives.
2. (25 points) Implement *k-means* clustering using the above skeleton program.
3. (25 points) Implement *k-medians* clustering using the skeleton program developed in Question 1.
4. Consider the following 1-dimensional dataset with three natural clusters:
   1. The first cluster contains points $\{1, 2, 3, 4, 5\}$.
   2. The second cluster contains points $\{8, 9, 10, 11, 12\}$.
   3. The third cluster contains points $\{24, 28, 32, 36, 40\}$.

   (a) (10 points) Use the k-means implementation done in Question 2 above to cluster the data with initial seed clusters as $\{1, 11, 28\}$. Is the algorithm able to find the right clusters?

   (b) (10 points) Use the k-means implementation done in Question 2 above to cluster the data with initial seed clusters as $\{1, 2, 3\}$. Is the algorithm able to find the right clusters?

   (c) (10 points) What do you interpret from the results of above two parts?