

In this assignment, you are asked to program a simple k -means (as discussed in class) clustering algorithm, *kmeans*, using the Euclidean distance for 2-dimensional numerical data.

You have the flexibility to choose a programming language. Your program should be executed as follows:

kmeans k input.txt

where input parameter $k > 1$ is an integer, specifying the number of clusters. *input.txt* is an input file containing many 2-dimensional data points in the following format,

274	119
317	144
267	164
233	137
272	99
297	116
268	142
522	286
468	308
441	263

Your program should output a txt file called output.txt, in the following format:

274	119	1
317	144	1
267	164	1
233	137	1
272	99	1
297	116	1
268	142	1
522	286	2
468	308	2
441	263	2

In output.txt, 1 and 2 are cluster labels. Each data point should be labeled using one of the labels from 1 to k . In the above example, there are 10 data points and $k = 2$.

For your convenience, a Windows data generator, gen.exe, is posted on the course webpage. You can use it to generate and visualize 2-dimensional data as well as clustering results.