CSCE 474/874: Introduction to Data Mining
Spring 2018

Assignment No. 3 February 18, 2018

Assignment

Implement *k*-means algorithm to perform clustering and compare your results with the results from Weka.

- Assume that all the attributes are continuous variables.
- Your program must allow the number of clusters (*k*) to be specified as input.
- Your program must allow the epsilon (change in the sum of the distances from the cluster centers) to be specified as input.
- Your program must allow the number of iterations to be specified as input.

Your program should stop if either the number of iterations is reached or if the change in the total sum of the squares of the distances (SSD) falls below epsilon.

Plot the runtime of the algorithm as a function of number of clusters, number of dimensions and size of the dataset (number of transactions).

Plot the goodness of clustering as a function of the number of clusters and determine the optimal number of clusters.

Compare the performance of your algorithm with that of Weka and summarize your results.

For this assignment you will work in teams. Use the dataset from the domain you will be working on for the project. If the data is not suitable, you may use one from the Weka dataset.

All code must be written by the members of your team. You may NOT use any code from ANY OTHER source, including other students and the Internet.

Due Date

The assignment is due on March 5 is worth 75 points.

Handin

Hand in a report along with the listing of your program, the output generated from the run of the test file on Canvas. Make sure that you have uploaded a signed copy of the Contributions form. Prepare and submit two files as follows:

- Your report named as “Lastname1_Lastname2.pdf” in pdf format. The signed contributions form should be used as the cover page of your report.
- A zip file named “Lastname1_Lastname2.zip” that includes everything else (your program, the output generated from the run of the test file, etc.). You must include a README file that describes the usage of your program. Make sure your implementation can successfully execute on the CSE server.

Grading Guidelines

Implement the *k*-means algorithm to perform clustering in a dataset. (40 points)

- Your implementation will be tested on cse.unl.edu server using the command you provided in the README file. (30 points)
- In the report, you should write a paragraph about your program design (10 points)

Plot the runtime of the algorithm as a function of number of clusters, number of dimensions and size of the dataset (number of transactions). (10 points)

- In the report, you should write a paragraph to summarize the observation and elaborate on it.

Plot the goodness of clustering as a function of the number of clusters and determine the optimal number of clusters. (15 points)

- In the report, you should write a paragraph to summarize the observation and elaborate on it.

Compare the performance of your algorithm with that of Weka and summarize your results. (10 points)

- Summarize the differences (if there is any) and elaborate on it (why/how).