

CS 445/545: Machine Learning
Programming Assignment #3

Dataset: Cluster dataset (txt format)

Assignment1: K-means -

The selectCentroids function initializes the centroids by shuffling the dataset and then selecting random k points for centroids without replacement. We calculate the Euclidean distance for random points k for the centroids without replacement. We take the number of clusters and interactions from the user as input. We calculate the sum of squared distances between the points and centroids. If the sum of the squared distance value reaches epsilon, we notice it as an early stop condition. The plots with the squared error are displayed when we execute the code.

All squareErrors sums:

[987.0437204371035, 1072.163038292803, 707.5834889380822, 671.9901768410426, 664.8112702114249, 661.9837724757779, 660.7516104525172, 659.9283176937054, 658.4663216674323, 657.4033977570986, 657.0339299401193, 656.8888419284038, 656.9027547302545]

Best squareError sum:

656.8888419284038

Assignment 2: Fuzzy C-means -

The difference between K-means and C-means is that the data point can belong to one or more clusters. We take the number of clusters and interactions from the user as input.

The membership weights are assigned randomly for every data point. We calculate the centroid for each cluster and calculate the membership grade with $m = 2$ (the fuzzifier parameter).

Membership grade is the one that indicates the degree to which a data point can belong to a cluster.

Epsilon reached, stopping early

New best square error: 2 All

squareErrors sums:

[1186.0318322123521, 1185.9885494668977]

Best squareError sum:

1185.9885494668977