Project #1 Report

Starting Centroids:

The starting centroids were generated randomly by using the Random function in java to choosing a number between 0-200 and using that random number as the row index of the data array holding all data points.

In step 6 the starting centroids were chosen based on the numeric value of the data points y values. Visually, centroids one, two and three were chosen based on being the highest, median and lowest y values respectfully.

Details on implementation:

To implement that k means algorithm, 3 data points were chosen at random to be assigned as centroids. These centroids were then used to assign other data points in the data set to 1 of either cluster based on its distance from the closest centroid. This way each data point is forced to be in one of three clusters.

Table for observing results:

|  |  |  |  |
| --- | --- | --- | --- |
| Starting Centroids | Intercluster Variability (IV) | Extracluster Variability (EV) | IV/EV |
| (44.3301,35.4358)  (41.7999,41.7424)  (47.9783,40.3569) | 8062.71564727776 | 1517.0840483118307 | 5.314 |
| (39.0, 23.0) (28.0,40.0) (41.0, 40.0) | 8062.71564727776 | 1512.6865920247606 | 5.330 |
| (39.0, 24.0)  (26.0, 40.0)  (41.0, 40.0) | 8062.71564727776 | 1499.8488397990843 | 5.375 |
| (39.0, 24.0)  (25.0, 41.0)  (41.0, 41.0) | 8062.71564727776 | 1485.401248190066 | 5.4279715 |
| (39.0, 24.0)  (24.0, 41.0)  (41.0, 41.0) | 8062.71564727776 | 1485.401248190066 | 5.4279715 |
| (28.4938,51.5998)  (37.0062,36.1636)  (31.5179,15.5518) | 8062.71564727776 | 1485.401248190066 | 5.4279715 |

Results discussion:

Based on the results it shows that every time the values of the centroids are changed the IV/EV value goes up. This is because each time centroids are moved the mean distance from the centroid is reduced. This implementation does the proper partitioning but based on the observation it cannot be said to globally optimal. It also lacks consistency because each time the centroids are shifted it will lead to a different result in EV.