**KMeans Worksheet**

Consider the following dataset that is being analyzed with a clustering algorithm:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| (17, 2) | (2, 16) | (16, 3) | (4, 16) | (1, 15) | (15, 2) |
| (2, 18) | (3, 17) | (18, 1) | (18, 3) | (1, 18) | (17, 4) |

1. Neatly plot these points below, using a dot to represent each point (a ruler might help).

20

20

1. As an intelligent human being, visually inspect the plot.  
     
   How many clusters do you see in the dataset?  
   Don’t overthink it, the answer is obvious.
2. There are obviously two clusters, each one will be identified by a centroid. Assume that one centroid, #0, is at (2, 17) and the other centroid, #1, is at (5, 15). Use Euclidean Distance to calculate the distance from every sample to both centroids. Assign the closest centroid as the label. Hint: create a program or spreadsheet to do the math.

|  |  |  |  |
| --- | --- | --- | --- |
| Sample | Dist to C#0  (2, 17) | Dist to C#1  (5, 15) | Closest Centroid Label (0 or 1) |
| (17, 2) | 21.21 | 17.69 | 1 |
| (2, 16) |  |  |  |
| (16, 3) |  |  |  |
| (4, 16) |  |  |  |
| (1, 15) |  |  |  |
| (15, 2) |  |  |  |
| (2, 18) |  |  |  |
| (3, 17) |  |  |  |
| (18, 1) |  |  |  |
| (18, 3) |  |  |  |
| (1, 18) |  |  |  |
| (17, 4) |  |  |  |

Explain the calculation that you performed in this step.

1. Now that you know which cluster each sample belongs to (although one of the samples may seem out of place, but at this point the math dictates otherwise), recalculate the centroids. Each centroid is the mean of all samples in the same cluster.

|  |  |
| --- | --- |
| Samples in Cluster #0 | Samples in Cluster #1 |
| ( , ) | ( , ) |
| ( , ) | ( , ) |
| ( , ) | ( , ) |
| ( , ) | ( , ) |
| ( , ) | ( , ) |
|  | ( , ) |
|  | ( , ) |

Mean point for Cluster #0: ( , )  
  
Mean point for Cluster #1: ( , )  
  
The “k” mean points are your new centroids (k=2). Lightly draw two really-small x's on your original plot to represent these centroids. If you compare to the original centroids, you’ll notice that one has barely moved and the other has changed significantly.

1. Visually inspect the plot and see if any of the samples have changed and are now closer to a different centroid.  
     
   Which of the samples changed labels? ( , )
2. Now that a sample has changed labels, visually estimate the new centroids and place two larger x's on your original plot to represent these two new centroids. If the new centroids make the clusters are stable (e.g., no sample labels would change in the future), then the algorithm is finished. The final set of labels are the clusters and the final mean points are the centroids.  
     
   Will the clusters be stable?
3. I gave you values for the initial centroids, which made it a bit easier. A more realistic situation is that you would only have the dataset and you would need to choose your own set of initial centroids. How might you do this?
4. Now shift to programmer mode, thinking about the overall design and organization of a class that would implement the KMeans algorithm. Based on your previous calculations, (A) identify the main functions that you would use and (B) lay out some pseudo code that shows how they work together to find the clusters.