**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. Plot these points below using a dot to represent each point.  
     
     
   
2. Assume that there are two clusters in this dataset. Use common sense to draw circles around the two clusters on your plot. This will be our final answer, but it might take us a few steps to arrive at these clusters.
3. Assume that there are two centroids. The first, Centroid #0, is at (2, 17) and the second, Centroid #1, is at (5, 15). Use the formula for Euclidean Distance to calculate the distance from every sample to both centroids. Assign the closest centroid as the label.

| Sample | Dist to C#0 | Dist to C#1 | Label (0 or 1) |
| --- | --- | --- | --- |
| (17, 2) | 21.21 | 16.97 | 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 have assigned each sample to a cluster, it's time to recalculate the centroids. Each centroid is the mean point for 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 mean points are your new centroids. Draw two really-small x's on your original plot to represent these two new centroids.

1. Visually inspect the plot and see if any of the samples have changed and are now closer to a different centroid. If the clusters are stable, then the algorithm is finished. The final set of labels are the clusters and the final mean points are the centroids.  
     
   Note: One of the samples should have 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.  
     
   At this point, the clusters should 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.