CX4010 HW2

Thomas Hervieu and IsaK Thomas

KMEANS:

If the amount of centroids is to large, some of the clusters will become empty.

If the amount of centroids is to small then it wouldn't be an accurate representation of the data.

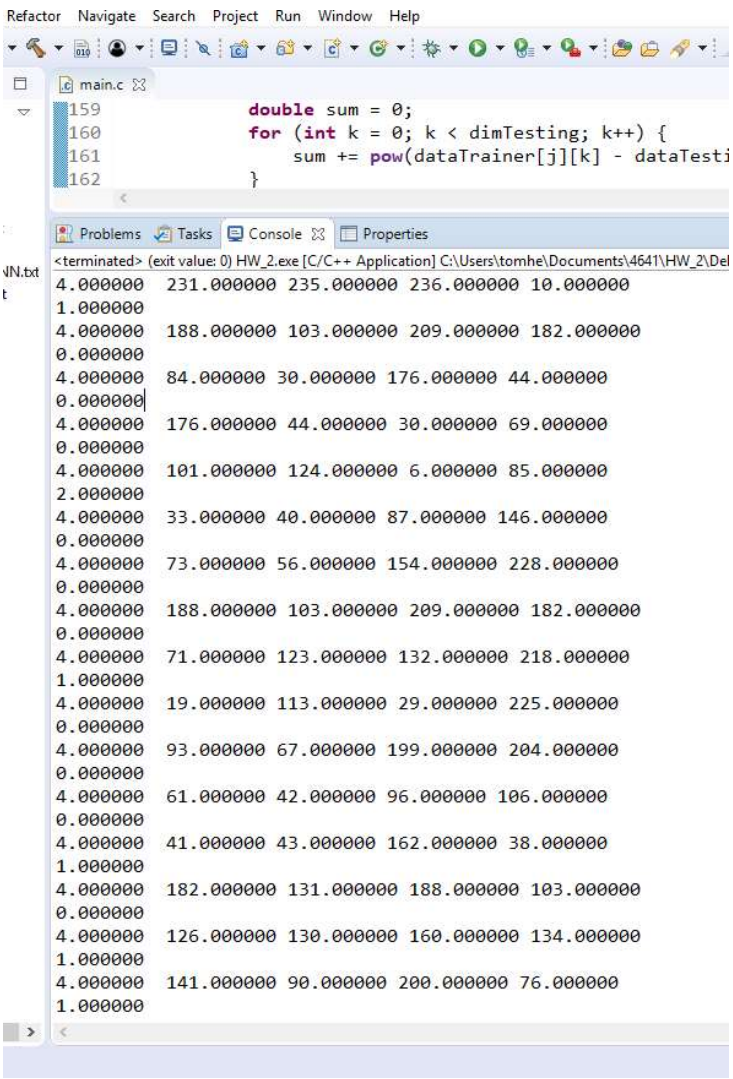
The way we determined that the optimal amount of centroids is three is that there was a fair distribution among the centroids. One always had significantly more, but the distribution of the other two were fairly close.

Exit full screen

```
numItems: 250 numAttr: 2
Number of points assigned to centroid 0 = 133
The coordinates of centroid 0 = 6.541293 133.278195
Number of points assigned to centroid 1 = 96
The coordinates of centroid 1 = 9.062417 184.645833
Number of points assigned to centroid 2 = 21
The coordinates of centroid 2 = 41.428190 844.095238
The distance from datapoint 0 to assigned centroid is 0.000000
The distance from datapoint 1 to assigned centroid is 1.474072
The distance from datapoint 2 to assigned centroid is 1.474072
The distance from datapoint 3 to assigned centroid is 7.025667
The distance from datapoint 4 to assigned centroid is 7.025667
The distance from datapoint 5 to assigned centroid is 0.000000
The distance from datapoint 6 to assigned centroid is 0.000000
The distance from datapoint 7 to assigned centroid is 0.000000
The distance from datapoint 8 to assigned centroid is 4.107370
The distance from datapoint 9 to assigned centroid is 4.107370
The distance from datapoint 10 to assigned centroid is 1.450000
etc...
```

KNN:

Code successfully ran for a synthetic version of the Kmeans output file. If K is less than the number of rows of the training set, the code will run correctly. I realized that a k of around 10 will be ideal for labeling the testing data points. A k that is too low tends to label the points prematurely because there may be many more data points of another class near the testing point that are in closest proximity to the testing point apart from only a couple points from another class. Also, if k is too large (k = 100) the bigger clusters (which contain the most labels) will dominate almost every testing point, which will result in all the testing points being classified to one class.



The 2 input files and output file have been included under the KNN folder to show that the calculations are correct. To the left is a sample execution with K=4 nearest neighbors.