**Project 3 Questions**

* 1. What are the assumptions that the k-means algorithm makes of the input data?
     1. *Clusters are spherical/ spread out uniformly from centroids*
     2. *Clusters are separate/ have clear boundaries between them*
     3. *Clusters are similarly sized and have a similar density*
     4. *Clusters separated correctly by Voronoi partitioning*
  2. Each input file has the original generating clusters in the filename before the “.dat” extension. For example, bullseye2.dat was generated with 2 clusters. For each input file provided, run *k*-means 3 times with the k value indicated by the filename.

Record the WCSS of each run in the Columns B-D of the given p3-results.xlsx Excel file. Be sure to visualize the results for each run. For which files does *k*-means clustering appear to succeed always/sometimes/never in your sample of 3 runs?

* + 1. *Always: easygaussian2, easygaussian3, hardgaussian2, hardgaussian3*
    2. *Sometimes: easygaussian4, easygaussian5, easygaussian6, easygaussian7, easygaussian8, easygaussian9, hardgaussian4, hardgaussian5, hardgaussian6,*
    3. *Never: bullseye2, diffdensity2, diffstddev2, easygaussian1, hardgaussian1, hardgaussian7, hardgaussian8, hardgaussian9, stretched2*
  1. For those problems where *k*-means appears to always fail, which assumptions (if any) of the *k*-means algorithm are violated?
     1. *Bullseye2: 1 and 4*
     2. *Diffdensity2: 3*
     3. *Diffstddev2: 3*
     4. *Easygaussian1: only one cluster so 2, 3, and 4 are not applicable*
     5. *Hardgaussian1: only one cluster so 2, 3, and 4 are not applicable*
     6. *Hardgaussian7: 3*
     7. *Hardgaussian8: 3*
     8. *Hardgaussian9: 3*
     9. *Stretched2: 1 and 4*
  2. Is it possible for *k*-means to fail if no assumptions are violated? Why or why not?
     1. *Yes:* 
        1. *If centroids converge on a local minimum. This can happen due to the random initialization (Forgy initialization) choosing bad centroids, or k-means not going through enough iterations.*
        2. *If K does not match the number of clusters. If K is too high, groups can get split. If K is too low, groups can be merged.*
  3. Why should iterated use of the *k*-means algorithm help in some cases with the quality of output clusters?
  4. For each input file provided, run iterated *k*-means one time each. Enter the resulting WCSS in Column H of the Excel sheet. For which files does *k*-means clustering appear to succeed in your iterated sample run?
  5. For which types of clustering problems does/doesn’t this iterated approach help?
  6. For each input file provided, run your program 8 times with *kMin*=2 and *kMax*=10 and *iter*=10. From each run, record the number of clusters *k* that yielded the maximum gap statistic in Columns J-Q of the spreadsheet.
  7. For which data set(s) does this technique consistently succeed in discerning the correct number of clusters?
  8. For which data set(s) does the discerned number of clusters vary from run to run?
  9. For which data set(s) does this technique consistently return the incorrectnumber of clusters? What do you observe about the nature of the data in these case(s)?