Week4

* single-linkage (MIN) hierarchical clustering with two clusters?
* Identify the number of clusters using the center-based, contiguity-based, and density-based definitions for below figure
* What would be the clusters if we use complete-linkage (MAX) hierarchical clustering with three clusters?
* What would be the clusters if we use DBSCAN clustering with following parameters? How many core samples?
  + Eps: 3
  + Minimum samples (MinPts): 2
  + Distance measurement: Euclidean distance
    - SQRT((X1-x2)^2 + (y1-y2)^2)
* compute average Silhouette coefficient for all samples of this clustering.

Hierarchical clustering

* basic algorithm
  + compute proximity matrix
  + let each data point be a cluster
  + repeat
    - merge 2 closest clusters
    - update the proximity matrix
      * for example if P3 and P6 are the coordinates of the closet point in the prox matrix, we will merge those. To update the matrix we consolidate those points and the new value depends on if we use MAX or MIN. if max it’s the greater of the 2 values.
  + until a single cluster remains
* Min has good performance when size of clusters is different. Very susceptible to noise.
  + Closest points in 2 different clusters
* Max tries to break down clusters which is a problem
  + Farthest points in 2 different clusters

**Live recording 4/4:**

* What is the final values of x and y for centroids after the first iteration?
  + Get the distance from each point to the noted centroid
  + The coordinate point belongs to the centroid that its closest to
  + The value of the centroid is the average of the x and y from the coordinates belonging to the centroid
* What is the final cluster assignments?
  + Process is the same, the only thing that changes is the cluster centroids (values derived in step 3 from above)
  + If theres no change we can stop
* Cluster validation
  + Supervised techniques
    - We don’t use class names for clustering, just for validating
  + Unsupervised
    - We don’t have class names, but we still want to measure the accuracy of the clustering. So we use purity which evalues the accuracy of the clustering.
    - Purity of cluster, or total of clustering
    - Purity of a cluster = (1/ total data in the cluster) \* (class total that appears most)
    - Purity for TOTAL clusters = (1/ total data) \* (class total that appears most in each cluster aka doesn’t have to be the same class for each cluster)

**Practice checks:**

1. Calc centroids is practice check 1
2. complete linkage (MAX) hierarchical clustering, in which faces will the nose, eyes, and mouth be well clustered based on the patterns
3. compute the final clusters
   1. compute the SSE for this clustering
   2. compute the sum of Silhouette coefficient for all samples of this clustering