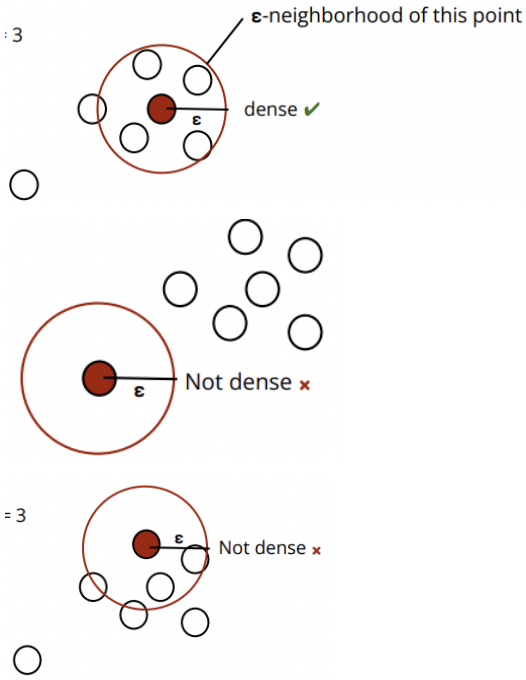
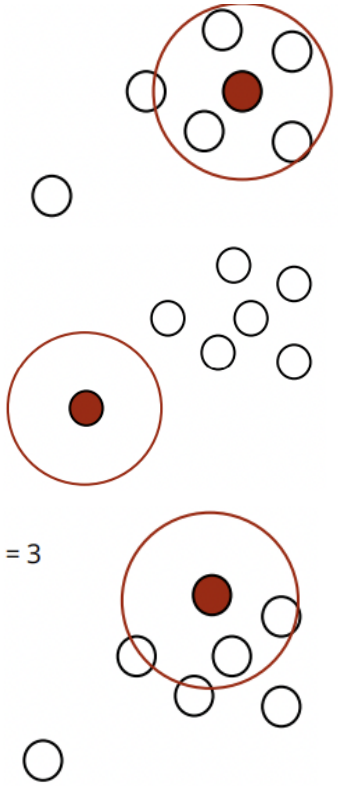
CAS CS 506
Lec 07

Density-Based Clustering

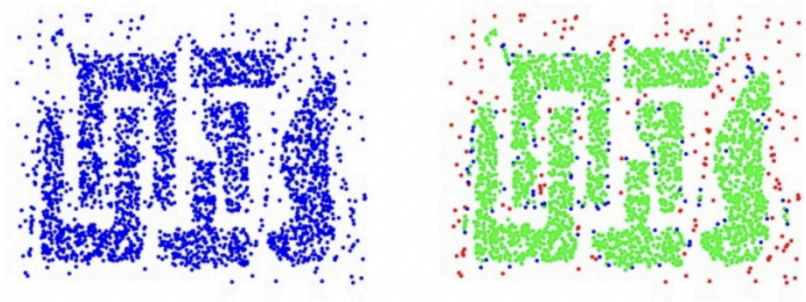
1. Density-Based Clustering
    a. Goal: cluster together points that are densely packed together
    b. How should we define density?
        i. Given a fixed radius **ε** around a point, if there are at least min_pts number of points in that area, then this area is dense
2. Example
    a. Min_pts = 3

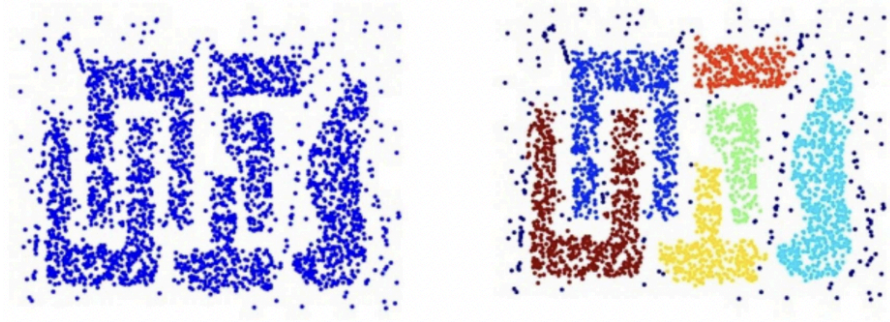


    b. But… That point was part of a dense section earlier…
3. Density-Based Clustering
    a. Need to distinguish between points at the core of a dense region and points at the border of a dense region
    b. Define
        i. Core point: if its **ε**-neighborhood contains at least min_pts
        ii. Border point: if it is in the **ε**-neighborhood of a core point
        iii. Noise point: if it is neither a core nor border point
4. Example
    a. Min_pts = 3



5. Density-Based Clustering

a. 

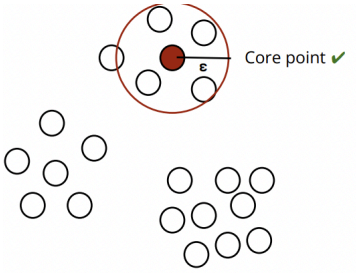
b. 

   i. Create clusters by connecting core points

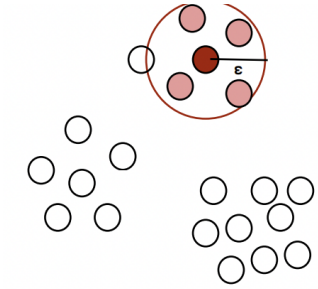
6. DBScan Algorithm
   a. **ε** and min_pts given
      i. Find the **ε**-neighborhood of each point

ii. Label the point as core if it contains at least min_pts
iii. For each core point, assign to the same cluster all core points in its neighborhood (crux of the algorithm)
iv. Label points in its neighborhood that are not core as border
v. Label points as noise if they are neither core nor border
vi. Assign border points to nearby clusters

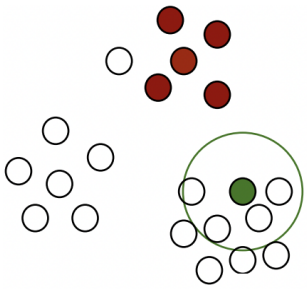
7. DBScan Visualized
    a. Iterate through the dataset



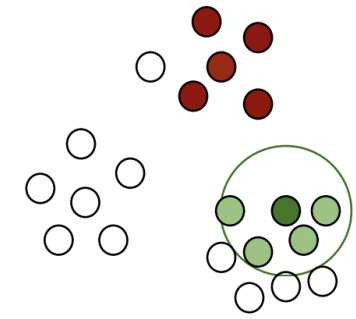
    b. If core point, iterate through its neighborhood to fine more core points that should also be part of this cluster



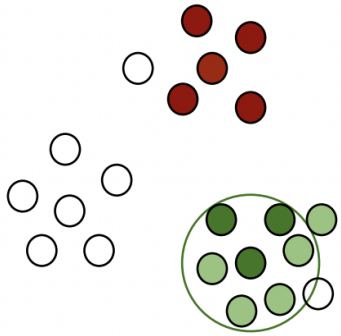
    c. Go to next data point in the dataset



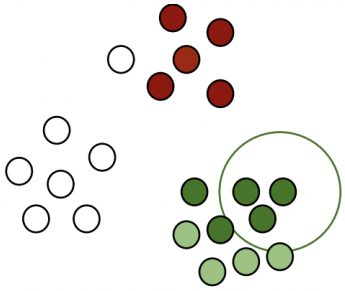
    d. Iterate over its neighborhood since it's a core point



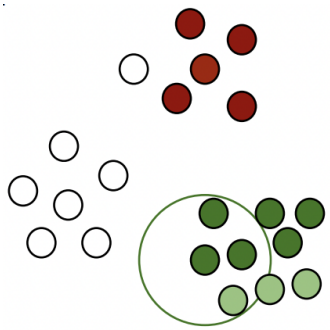
e. Found more core points so need to iterate over its neighborhood too



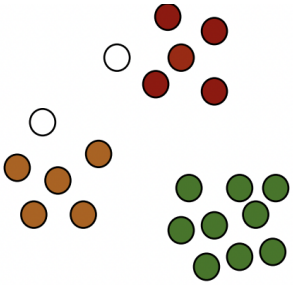
f. Border point but let's assign it to the cluster now



g. Core point but all its neighborhood is already tracked



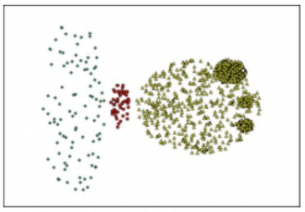
h. Final result



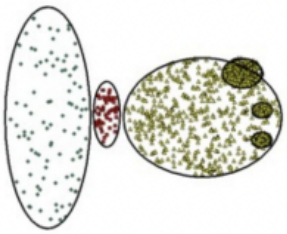
8. DBScan - Benefits
    a. Can identify clusters of different shapes and sizes
    b. Resistant to noise
9. DBScan - Limitations
    a. Can fail to identify clusters of varying densities
    b. Tends to create clusters of the same density

c. Notion of density is problematic in high-dimensional spaces

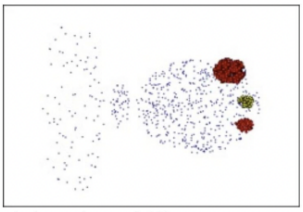
d.



(MinPts=4, Eps=9.75).



→



(MinPts=4, Eps=9.92)