Hannah Wolfe
Data-mining course

1. What is unsupervised learning?
    a. Computer given inputs
    b. Computer figures out structure from inputs
2. What are the pros of unsupervised learning
    a. Algorithm finds groups without knowledge of groups
3. Clustering
    a. Most common type of unsupervised learning algorithm
    b. How to take a dataset and split it up into meaningful groups
4. Questions to ask of clustering algorithms
    a. How do you define similarity?
    b. How many clusters?
        i. Do you choose or the algorithm?
        ii. What variables control this?
    c. When do you stop?
        i. What variables control this?
    d. Are there outliers?
    e. Are there “hard” or “soft” groups?
    f. How flexible is it?
    g. Are groups variable sizes?
    h. How efficient is it?
        i. How big is your dataset?
5. Types of clustering
    a. Centroid-Based Clustering
        i. Defined number of clusters
        ii. Place centroids, measure distance, create new centroids
    b. Connectivity-Based Clustering
        i. Find similarities, cluster them together
    c. Distribution-Based Clustering
        i. Soft-clustering, defines the probability item is in a cluster
    d. Density-Based Clustering
        i. Each cluster is a high density region in the space
        ii. Each point is Core, border or noise (ignores outliers)
    e. Self-organizing maps (SOM)
        i. Used for visualizing low density views of high dimensional data
6. Schedule
    a. This class: centroid and connectivity-based clustering
    b. Next class: distribution and density-based clustering
    c. Neural Net class: SOM
7. Centroid-Based Clustering
    a. K-means clustering

    i. Place k centroids in random locations around the space
    ii. Repeat until convergence (no clusters change)
      1. For each point x
        a. Find nearest centroid c (compute distance between point and all centroids)
        b. Assign the point x to centroid cluster
      2. For each cluster
        a. For each centroid recompute it’s position
          i. Average vectors in cluster

8. Connectivity-Based Clustering
  a. Hierarchical Clustering
    i. Cannot be used for big datasets that can’t fit in memory
    ii. Top down (Divisive)
      1. Start with one cluster and recursively split it
    iii. Bottom up (Agglomerative)
      1. Initially each point is a cluster
      2. Repeatedly combine two nearest clusters into one
      3. Questions
        a. How do you represent a cluster of more than one point
          i. Euclidean space
            1. Centroid (average of points)
          ii. Non-Euclidean space
            1. Clusteroid (pick closest point to others)
            2. Smallest maximum distance
            3. Smallest average distance
            4. Smallest sum of squares distance
        b. How do you determine the “nearness of clusters”
          i. Measure distance from centroid/clustroid
        c. When do we stop combining clusters?
          i. Pick number k up front
          ii. Stop when next merge would create a cluster with low “cohesion”
            1. Diameter of the merged cluster =maximum distance between points
            2. Radius = maximum distance of a point from a centroid
            3. Density based approach=divide number of points per unit volume
      4. Create a dendrogram
        a. Like family tree of how evolved

Questions	Centroid	Connectivity	Distribution	Density
Similarity?	distance	distance		
How many clusters?	User defined	Program or user defined		
When do you stop?	Convergence (no clusters change)	When chosen number of groups or the next addition would cause low cohesion		
Are there outliers?	no	Yes, when cluster number is program defined, no when cluster number is user defined		
How efficient is it?	Better than hierarchical	Works with small datasets		
Are groups variable sizes	No, for area Yes, for group	yes		