Data Mining, CSCI 347, Fall 2019
Clustering, Oct. 30

Clustering - Divide instances into "natural" groups.
Unsupervised learning

Groups can be:

- Disjoint or overlapping
- Deterministic or probabilistic
- Flat or hierarchical

k-means clustering algorithm:

1. Input k to indicate the number of clusters to create
2. k points are randomly chosen to serve as cluster centers (these don't need to represent an instance). These are called "seeds"
3. Loop while cluster centers are changing
    a. Assign each instance to its closest cluster center using the Euclidean distance
    b. Calculate the mean point of each cluster, the centroid, to serve as the new cluster center

k-means clustering algorithm:

- Typically don't know how many natural clusters
- It is important to select good seeds
- Can get trapped in a local minimum
- Better way to select seeds, called k-means++
    - Chose the initial seed at random from the entire space, with a uniform probability distribution
    - Choose the second seed with a probability that is proportional to the square of the distance from the fist
    - Proceed, at each stage choosing the next seed with a probability proportional to the square of the distance from the closest seed that has already been chosen.

Can create a hierarchical clustering by applying the algorithm with k=2 and then repeating, recursively, within each cluster

It is difficult to know the best k to use.

- Try different values
- Use the one that creates the best (tightest) clusters