Clustering 2 Lecture

Introduce Research Assignment 3
Pull up assignment description on Gauchospace

1. Changed it from being an individual assignment to a group assignment.
2. It is a 2-3 page report.
3. 5+ slides. Should summarize in bullet points the report.
4. Choose 1-2 datasets.
    a. Prioritize what datasets you need.
    b. Collect and clean data by Friday at the latest.
5. Focus on one overarching question that can be answered by the chosen dataset

##

Clustering Review

1. 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. Density-Based Clustering
        i. Each cluster is a high density region in the space
        ii. Each point is Core, border or noise (ignores outliers)
    d. Distribution-Based Clustering
        i. Soft-clustering, defines the probability item is in a cluster

Density Based Clustering

1. Assumption is that the data is caused by a probability density function that we will threshold to find an area. Anything in that area is part of our dataset. This is hard to do, so we have find algorithms that simulate it.
2. https://www.naftaliharris.com/blog/visualizing-dbscan-clustering/
3. DBSCAN-- Density Based Spatial Clustering of Applications with Noise
4. Categorize points as:
    a. core (has a greater number of neighbors than the minPoints within the radius)
    b. border (connected to one core point)
    c. noise (near no points)
5. DBSCAN goal is to put each core or border point in a cluster.
6. Algorithm
    a. choose a core point, see what's connected to it and put it in the cluster
    b. recursively for each added point, put it's neighbors in the cluster
    recursive: function refers to itself
    c. repeat til all border and core points are added to a cluster
7. Positive:

    a. find any shape
    b. detect and ignore outliers
8. Negative:
    a. Have to fine tune neighborhood and minPoints parameters
        i. sensitive to the neighborhood parameter

##

Distribution based clustering:

1. Soft-clustering, defines the probability item is in a cluster
2. Therefore we need to know the probability of each item in relationship to each cluster
3. Probability distributions
    a. Mean --average! μ = mu
    b. Variance --measures the spread or variability of the distribution
        i. Variance, σ^2
        ii. Standard Deviation σ = sigma
        iii. $\sigma_X^2 = \sum(x_i - \mu_x)^2 p_i$
        iv. p_i = probability that x happens or 1/n with n = count(x)
4. Bayes Rule

    Bayes' theorem is stated mathematically as the following equation:[2]

    $$P(A \mid B) = \frac{P(B \mid A)\, P(A)}{P(B)},$$

    where A and B are events and $P(B) \neq 0$.

    - $P(A)$ and $P(B)$ are the probabilities of observing A and B without regard to each other.
    - $P(A \mid B)$, a conditional probability, is the probability of observing event A given that B is true.
    - $P(B \mid A)$ is the probability of observing event B given that A is true.
5. EXAMPLE 1:
    a. 1D
    b. We know which are in which group
    c. Calculate mu and variance
        i. mean(x)
        ii. take average of (x-mean(x))^2
    d. That was easy!
6. What if we don't know which is in which group?
    a. If we know the parameters we can guess which one is more likely in which gaussian, using Bayes rule
    b. Chicken and egg problem
        i. we need the parameters to guess the sources
        ii. we need to know the source to estimate the parameters

7. EXAMPLE 2:
    a. 1D
    b. EM Algorithm (expectation-maximization algorithm)
        i. start with two randomly placed Gaussians
        ii. Expectation:
            1. For each point calculate the probabilities it is in a group
            (I will not be doing the derivation of this)

$$P(x_i|b) = \frac{1}{\sqrt{2\pi\sigma_b^2}} exp\left(-\frac{(x_i-\mu_b)^2}{2\sigma_b^2}\right)$$

            2. Take the Bayesian Posterior

$$b_i = P(b|x_i) = \frac{P(x_i|b)P(b)}{P(x_i|b)P(b)+P(x_i|a)P(a)}$$

            $a_i = P(a|x_i) = 1 - b$
        iii. Maximization: adjust the means and variances to fit the points assigned to them
            1. Take all the ones that have a higher probability to be in one group and calculate its mean and variance
            Mean:

$$\mu_b = \frac{b_1x_1 + b_2x_2 + ... + b_nx_n}{b_1 + b_2 + ... + b_n}$$

            Variance: (mu of b)

$$\sigma_b^2 = \frac{b_1(x_1-\mu_1)^2 + ... + b_n(x_n-\mu_n)^2}{b_1 + b_2 + ... + b_n}$$

            2. Take all the ones that have a higher probability to be in one group and calculate its mean and variance
        iv. Iterate until it converges

##

Questions	**Centroid**	**Connectivity**	**Density**	**Distribution**

Similarity?	distance	distance	distance	distance
How many clusters?	User defined	Program or user defined	Program defined	User defined
When do you stop?	Convergence (no clusters change)	When chosen number of groups or the next addition would cause low cohesion	When all core and border points are in a group	Convergence (no change to mean and variance when recalculated)
Are there outliers?	no	Yes, when cluster number is program defined, no when cluster number is user defined	Yes	No (sort of)
How efficient is it? (complexity)	Better than hierarchical K-Means $O(n^{dk+1})$ k = # of clusters d = dimension	Works with small datasets AGGLOMERATE: $O(n^2 \log(n))$	DBSCAN: $O(n \log(n))$	Dependent on model, number of steps and time it takes to calculate
Are groups variable sizes	No, for area Yes, for group	yes	yes	yes
Are there hard or soft groups?	hard	hard	hard	soft
How flexible are the shapes?	Not flexible	Not flexible	flexible	Not flexible