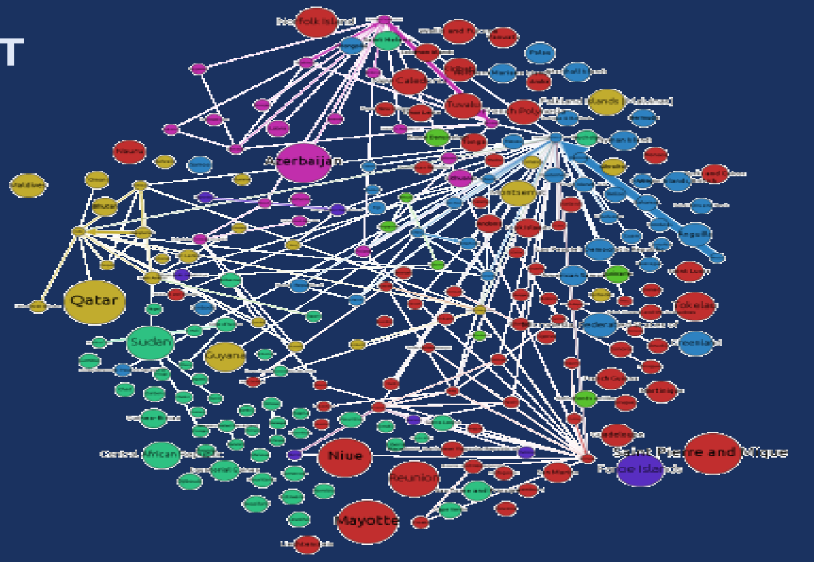
IS422P - DATA MINING
CLUSTERING-PART2

AMIRA REZK

INFORMATION SYSTEM DEPARTMENT



AGENDA

The Basics

- What is Cluster Analysis?
- Requirements for Cluster Analysis
- Overview of methods

Partitioning Methods

- K-Means

Hierarchical Methods

- Agglomerative vs. Divisive
- Distance Measures

Density-Based Methods

- DBSCAN

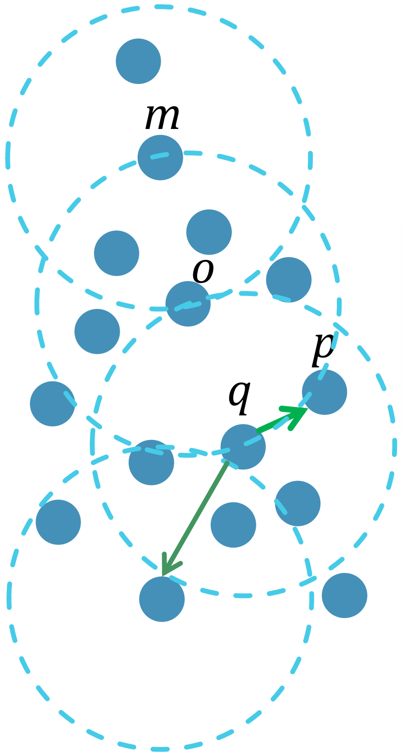
Evaluation of Clustering

- Assessing Clustering Tendency
- Measuring Clustering Quality

DENSITY-BASED METHODS
DBSCAN: DENSITY-BASED SPATIAL CLUSTERING OF APPLICATIONS WITH NOISE

- Find *core objects* (with dense neighborhoods)
- Connect core objects to form dense clusters
- User provides:
  - **ϵ-neighborhood** of object $\boldsymbol{o}$ → space within a radius ϵ centered at $\boldsymbol{o}$
    - **Neighborhood density** → # objects in that neighborhood
  - ***MinPts*** → **density threshold** for a neighborhood
- **Core object** → object whose ϵ-neighborhood contains at least *MinPts* objects

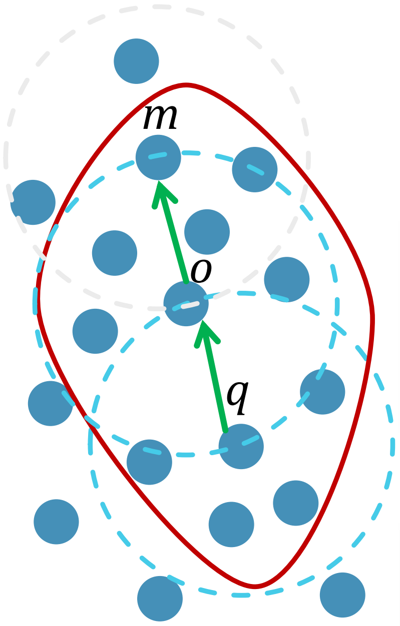
DENSITY-BASED METHODS
DBSCAN



Given $\epsilon = 4$ and $MinPts = 5$

an object $\boldsymbol{p}$ is directly **density-reachable** from another object $\boldsymbol{q}$ if and only if $\boldsymbol{q}$ is a core object and $\boldsymbol{p}$ is in the ϵ-neighborhood of $\boldsymbol{q}$

DENSITY-BASED METHODS
DBSCAN



Given $\epsilon = 4$ and $MinPts = 5$

an object ***p*** is directly **density-reachable** from another object ***q*** if and only if ***q*** is a core object and ***p*** is in the -neighborhood of ***q***

objects ***q*** & ***m*** are **density-connected** if there is an object ***o*** such that ***q*** & ***m*** are both density-reachable from ***o***

DENSITY-BASED METHODS DBSCAN

Algorithm: DBSCAN: a density-based clustering algorithm.

Input:

- D: a data set containing n objects,
- ϵ: the radius parameter, and
- $MinPts$: the neighborhood density threshold.

Output: A set of density-based clusters.

Method:

(1) mark all objects as unvisited;
(2) **do**
(3)   randomly select an unvisited object p;
(4)   mark p as visited;
(5)   **if** the ϵ-neighborhood of p has at least $MinPts$ objects
(6)   create a new cluster C, and add p to C;
(7)   let N be the set of objects in the ϵ-neighborhood of p;
(8)   **for** each point p' in N
(9)   **if** p' is unvisited
(10)   mark p' as visited;
(11)   **if** the ϵ-neighborhood of p' has at least $MinPts$ points,
                      add those points to N;
(12)   **if** p' is not yet a member of any cluster, add p' to C;
(13)   **end for**
(14)   output C;
(15)   **else** mark p as noise;
(16) **until** no object is unvisited;

AGENDA

The Basics
- What is Cluster Analysis?
- Requirements for Cluster Analysis
- Overview of methods

Partitioning Methods
- K-Means

Hierarchical Methods
- Agglomerative vs. Divisive
- Distance Measures

Density-Based Methods
- DBSCAN

Evaluation of Clustering
- Assessing Clustering Tendency
- Measuring Clustering Quality

EVALUATION OF CLUSTERING
ASSESSING CLUSTERING TENDENCY

- Determines whether a given data set has a non-random structure
- **Hopkins Statistic** → Statistical tests for spatial randomness
  - Sample ***n*** points, $p_1, \ldots, p_n$ uniformly from ***D***
  - For each point, $\boldsymbol{p_i}$ find its nearest neighbor in ***D***

    $$distance\ x_i = \min_{v \in D}\{dist(p_i, v)\}$$

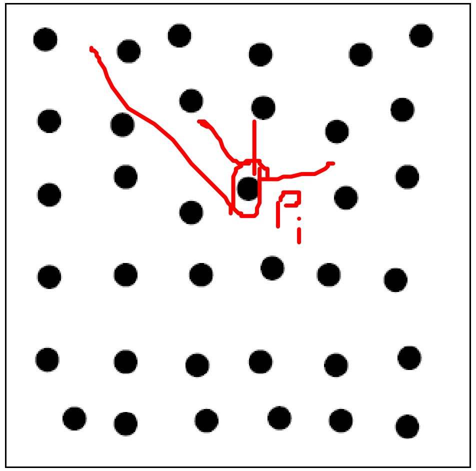
  - Sample ***n*** points, $q_1, \ldots, q_n$, uniformly from ***D***
  - For each q_i find its nearest neighbor of q_i in D-$\{q_i\}$

    $$distance\ y_i = \min_{v \in D, v \neq q_i}\{dist(q_i, v)\}$$

  - Calculate the <u>*Hopkins Statistic*</u> ***H***:

    $$H = \frac{\sum_{i=1}^{n} y_i}{\sum_{i=1}^{n} x_i + \sum_{i=1}^{n} y_i}$$

  - If ***D*** is uniformly distributed, $\sum_{i=1}^{n} x_i$ *and* $\sum_{i=1}^{n} y_i$ are roughly equal, and ***H*** ≈ 0.5



EVALUATION OF CLUSTERING
MEASURING CLUSTERING QUALITY – EXTRINSIC METHODS

Extrinsic methods → compare clustering against **ground truth** (*supervision*)

- Assign a score $Q(C, C_g)$ to capture:
  - **Cluster homogeneity** → the purer the better – clusters represent separate class labels
  - **Cluster completeness** → an object with a class label belongs to the cluster representing that class label
  - **Rag bag** → objects that can't be merged into clusters belong to a *rag bag* – penalize a *misc. object* when put in a *pure cluster* more than in a rag bag
  - **Small cluster preservation** → splitting a small category is *more harmful* than splitting a large category
- Ex. **BCubed** *precision* and *recall* of every object in dataset:
  - Precision → how many objects in the same cluster ∈ the same category as the object
  - Recall → how many objects of the same category are assigned to the same cluster

EVALUATION OF CLUSTERING
MEASURING CLUSTERING QUALITY – INTRINSIC METHODS

- Intrinsic methods → measure how well the clusters are separated
- Ex. **The silhouette coefficient** → difference between:
  - average distance between object o and all other objects in the cluster to which o belongs (captures cluster correctness) – smaller is better (more compact)
  - minimum average distance from o to all clusters to which o does not belong (captures degree of separation from other clusters) – larger is better
- Compute average silhouette coefficient for all objects in a cluster or over all of the dataset
  - +ve → clustering is good
  - -ve → clustering is bad

EVALUATION OF CLUSTERING
MEASURING CLUSTERING QUALITY – INTRINSIC METHODS

•Compute **the silhouette coefficient** for object **x1.**
What is the meaning of the computed value?

C1= {x1,x4,x8}={(2,10), (5,8), (4,9)} Mean of C1= $(2\frac{2}{3},9)$

C2= {x3,x5,x6}={(8,4), (7,5), (6,4)} Mean of C2 = $(7, 4\frac{1}{3})$

C3= {x2,x7}={(2,5), (1,2)} Mean of C3 = $(1\frac{1}{2}, 3\frac{1}{2})$

- $a(o) = \frac{\sum_{o' \in c_i} dis(o,o')}{|c_i|-1} = \frac{5+3}{2} = 4$
- $b(o) = \min\left\{\frac{\sum_{o' \in c_j} dis(o,o')}{|c_j|}\right\} = \min\left\{\frac{12+10+10}{3}, \frac{5+9}{2}\right\} = 7$
- $S(o) = \frac{b(o)-a(o)}{\max\{a(o),b(o)\}} = \frac{7-4}{7} \rightarrow +ve$
- This mean the cluster containing o is compact and o is far from other cluster

QUESTION?

NEXT …….