School of Computing
UNIVERSITY OF GEORGIA

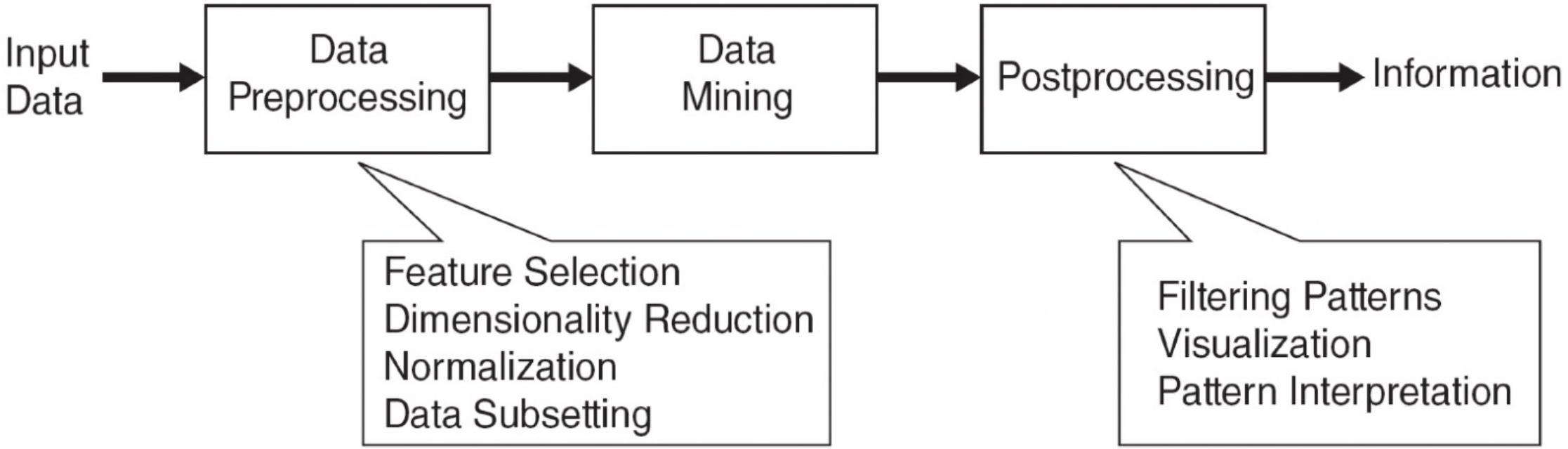
CSCI 4380/6380 DATA MINING

Fei Dou

Assistant Professor
School of Computing
University of Georgia

November 02, 2023

Recap: Data Mining Process



Density based Clustering

Density Based Clustering

- Clusters are regions of high density that are separated from one another by regions on low density.

Density Based Clustering

- Basic idea
  - A cluster is defined as a maximal set of density-connected points
  - Discovers clusters of arbitrary shape

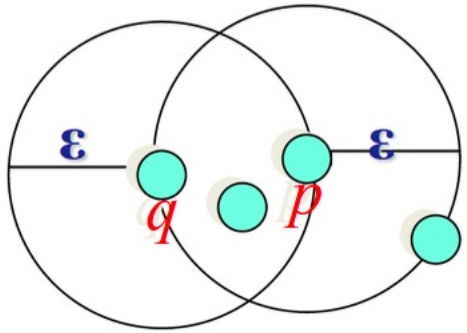
- Method
  - DBSCAN: Density-based spatial clustering of applications with noise

Density Based Clustering

- Density Definition
  - ε-Neighborhood, samples within a radius of ε from a sample.

$$N_\varepsilon(\mathbf{p}) : \{d(\mathbf{p}, \mathbf{q}) \leq \varepsilon\}$$

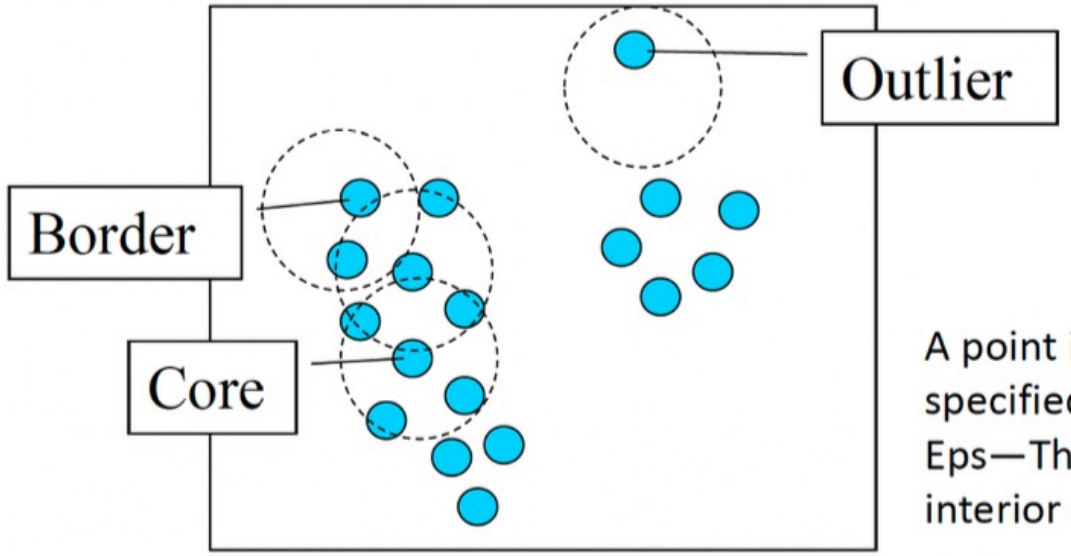
- High density: ε-Neighborhood of a sample contains at least MinPts of samples.



DBSCAN

- DBSCAN is a density-based algorithm.
  - Density = number of points within a specified radius (ε or Eps)
  - A point is a core point if it has at least a specified number of points (MinPts) within Eps
    - These are points that are at the interior of a cluster
    - Counts the point itself
  - A border point is not a core point, but is in the neighborhood of a core point
  - A noise point is any point that is not a core point or a border point

DBSCAN: Core, Border, and Outlier



$\varepsilon = 1\text{unit}, \text{MinPts} = 5$

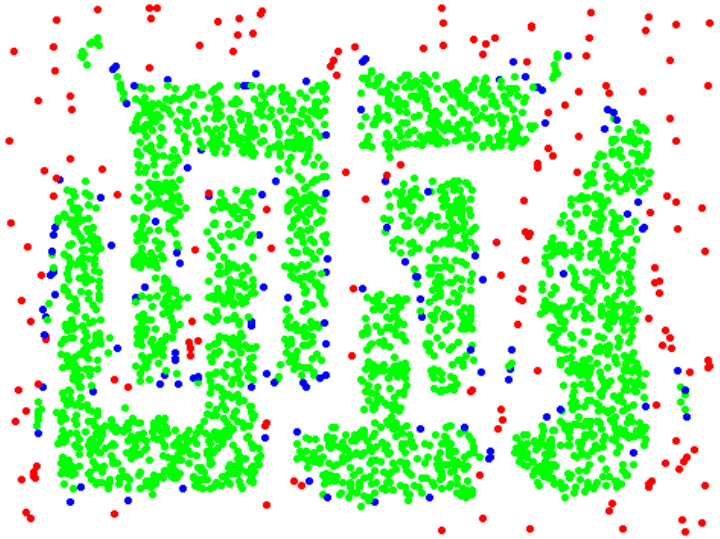
A point is a core point if it has more than a specified number of points (MinPts) within Eps—These are points that are at the interior of a cluster.

A border point has fewer than MinPts within Eps, but is in the neighborhood of a core point.

A noise point is any point that is not a core point nor a border point.

DBSCAN: Core, Border and Noise Points

Original Points

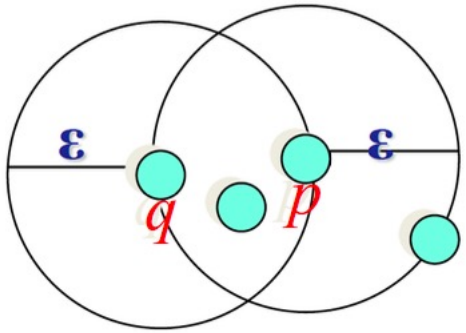


Point types: core, border and noise

Eps = 10, MinPts = 4

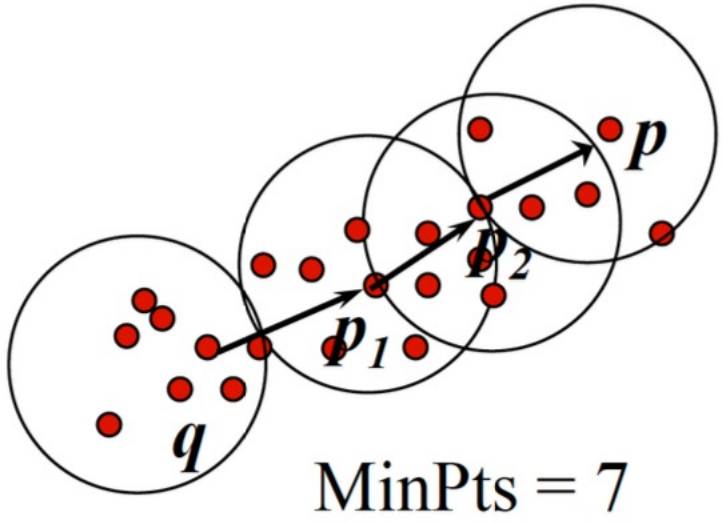
DBCAN: Density-reachability

- **Directly density-reachable**: A sample **q** is directly density-reachable from sample **p** if **p** is a core sample and **q** is in **p**'s ε-neighborhood
  - **q** is directly density-reachable from **p**
  - **p** is not directly density-reachable from **q**
  - Density-reachability is asymmetric



DBCAN: Density-reachability

- Density-Reachable (directly and indirectly):
  - A point **p** is directly density-reachable from $\mathbf{p}_2$
  - $\mathbf{p}_2$ is directly density-reachable from $\mathbf{p}_1$
  - $\mathbf{p}_1$ is directly density-reachable from **q**
  - $\mathbf{p} \leftarrow\!- \mathbf{p}_1 \leftarrow\!- \mathbf{p}_2 \leftarrow\!- \mathbf{q}$ form a chain



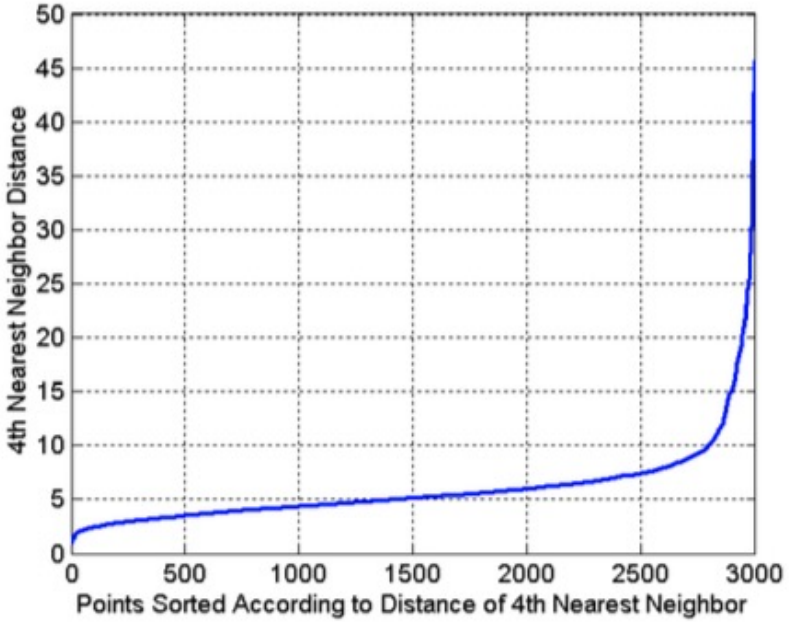
DBSCAN Algorithm

- Form clusters using core points, and assign border points to one of its neighboring clusters
  1: Label all points as core, border, or noise points.
  2: Eliminate noise points.
  3: Put an edge between all core points within a distance *Eps* of each other.
  4: Make each group of connected core points into a separate cluster.
  5: Assign each border point to one of the clusters of its associated core points

ε = 2, and MinPts=3

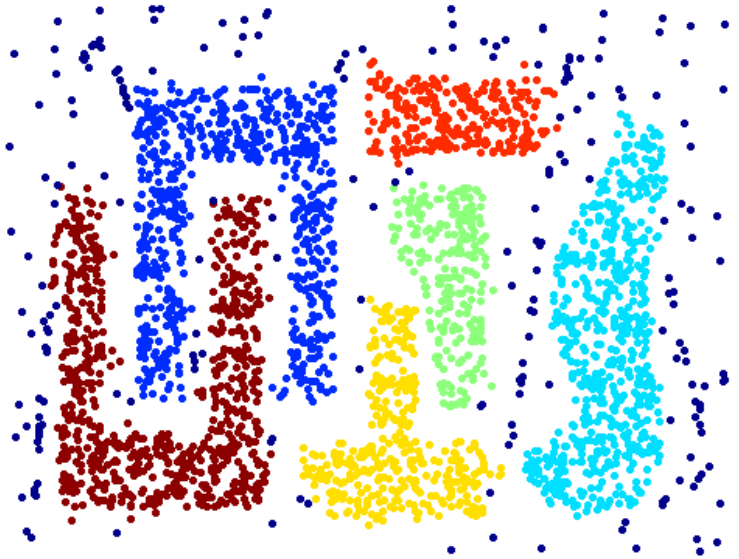
DBCAN: Determining ε and MinPts

- Idea is that for samples in a cluster, their k-th nearest neighbors are at roughly the same distance
- Noise samples have the k-th nearest neighbor at farther distance
- So, plot sorted distance of every sample to its k-th nearest neighbor



When DBSCAN Works Well

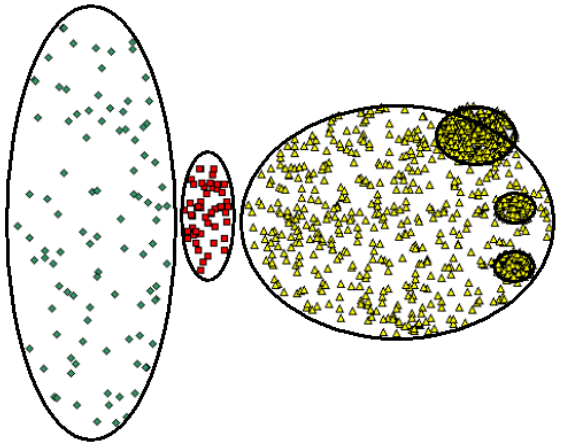
Original Points



Clusters (dark blue points indicate noise)

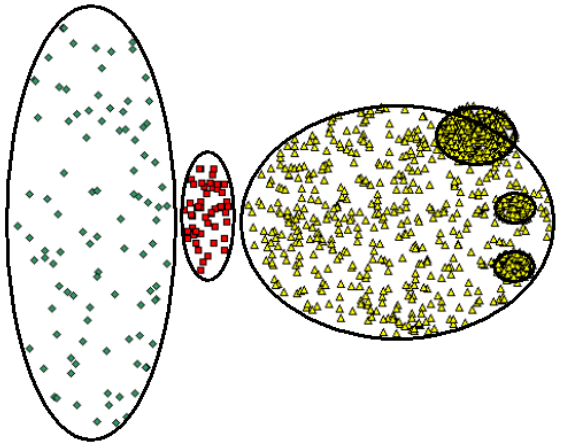
- Can handle clusters of different shapes and sizes
- Resistant to noise

When DBSCAN Does NOT Work Well



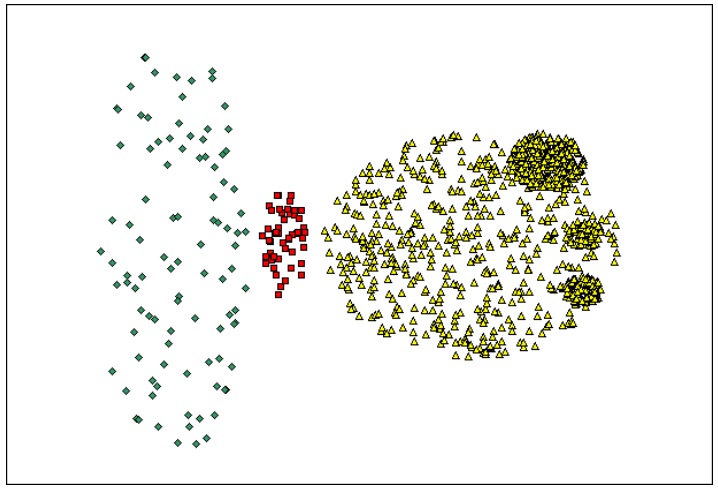
Original Points

When DBSCAN Does NOT Work Well

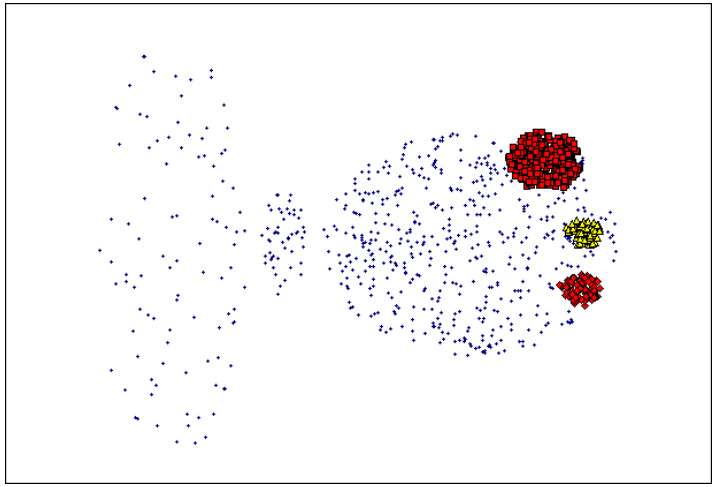


Original Points

- **Varying densities**
- **High-dimensional data**



(MinPts=4, Eps=9.92).



(MinPts=4, Eps=9.75)

Density based Clustering

- **Pros**
  - Resistant to Noise
  - Can handle clusters of different shapes and sizes
- **Cons**
  - Cannot handle varying densities
  - Sensitive to parameters—hard to determine the correct set of parameters