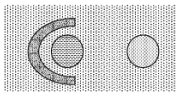
CS522 Advanced Database Systems
Clustering: Density-Based Methods

Chengyu Sun
California State University, Los Angeles

Density-based Clusters



- A cluster is a dense region of objects surrounded by a region of low density

DBSCAN

- Density-Based Spatial Clustering of Applications with Noise

Classification of Points

- Given a radius ε and the minimum number of points `MinPts` within a radius of ε (ε-neighborhood)
  - Core point
    - Points in its ε-neighborhood ≥ `MinPts`
  - Border points
    - Within the ε-neighborhood of a core point
  - Noise points

Point Examples

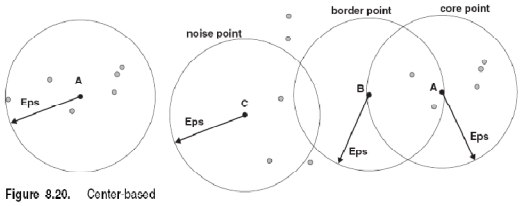


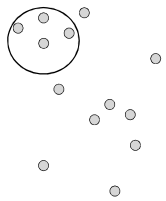
Figure 8.20. Center-based density.

Figure 8.21. Core, border, and noise points.

The DBSCAN Algorithm

- Label all points as core, border, or noise
- Remove all noise points
- Put an edge between all core points that are within ε of each other
- Make each connected group of core points a cluster
- Assign border points to *one of the clusters* of their associated core points

DBSCAN Example



Select DBSCAN Parameters

- k-dist: distance to the kth nearest neighbor
- k=4 is usually reasonable for most 2-D datasets

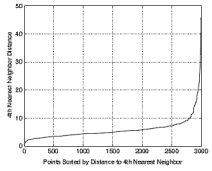
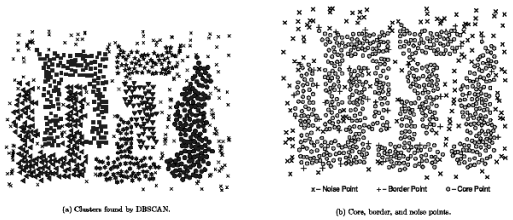


Figure 8.23. K-dist plot for sample data.

More DBSCAN Examples



(a) Clusters found by DBSCAN.

(b) Core, border, and noise points.

About DBSCAN

- Handle clusters with arbitrary shapes and sizes
- Limitations
  - Clusters with varying densities
  - High dimensional data
- Could be expensive because of nearest neighbor computation
  - Use a spatial index structure like R tree or k-d tree

Readings

- Textbook 7.6.1