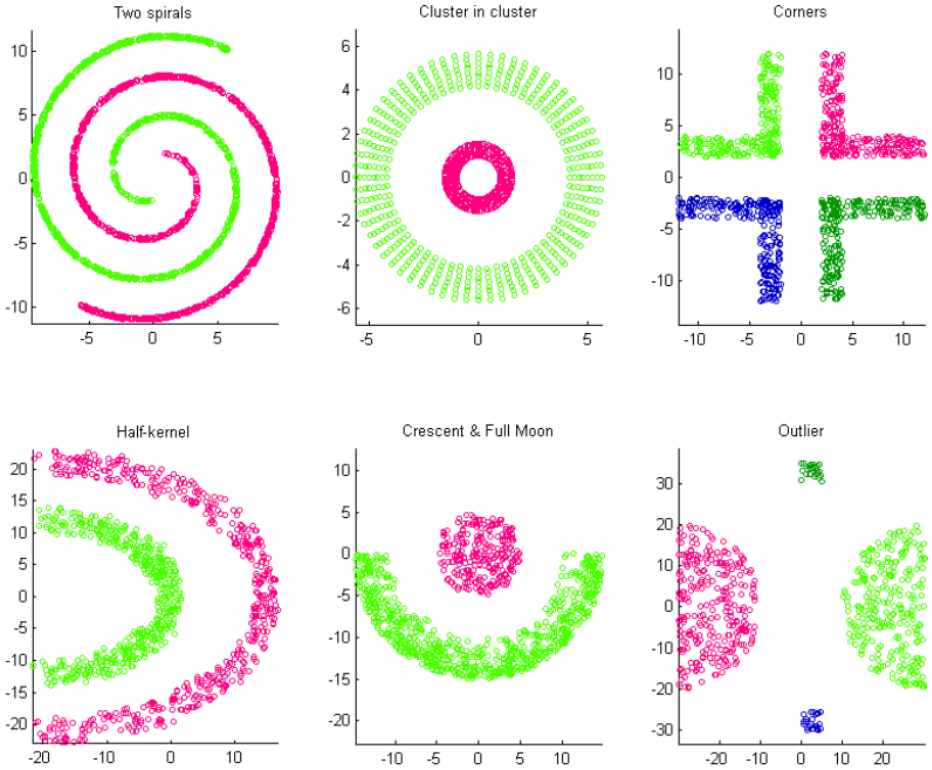
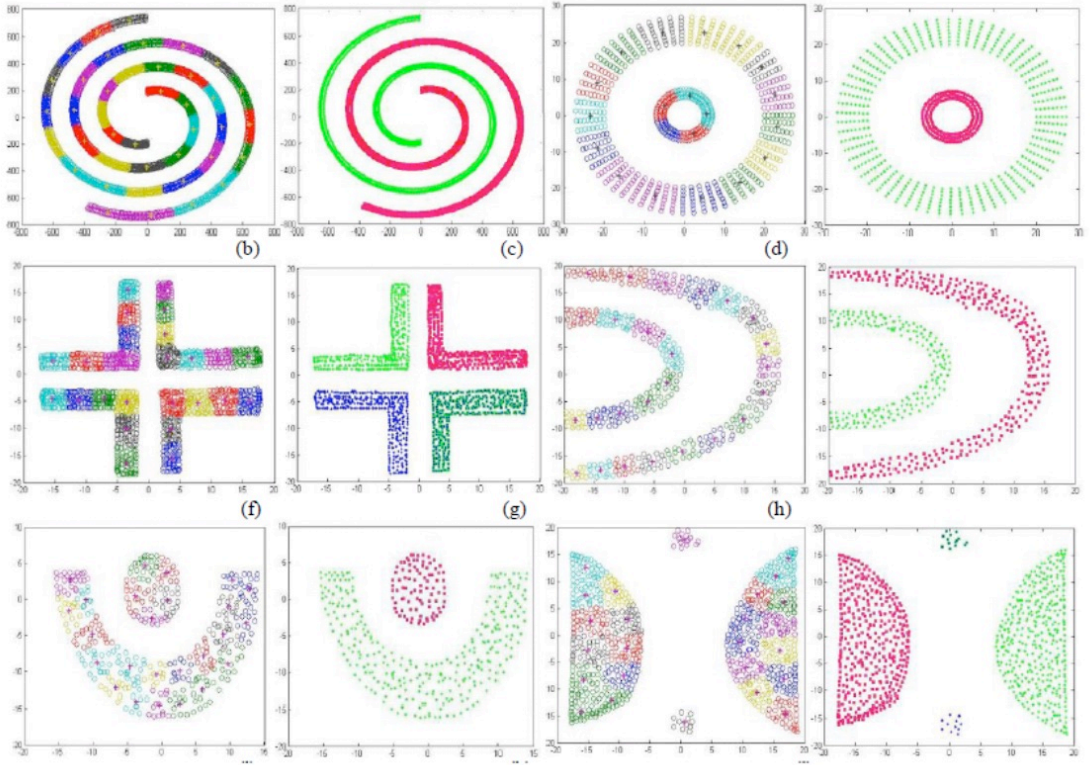
Clustering

Part IV

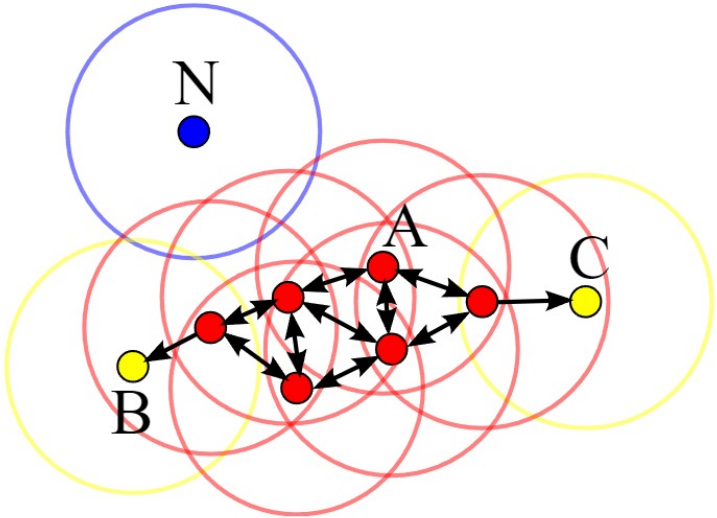
Challenging Examples



Challenging Examples, Continued



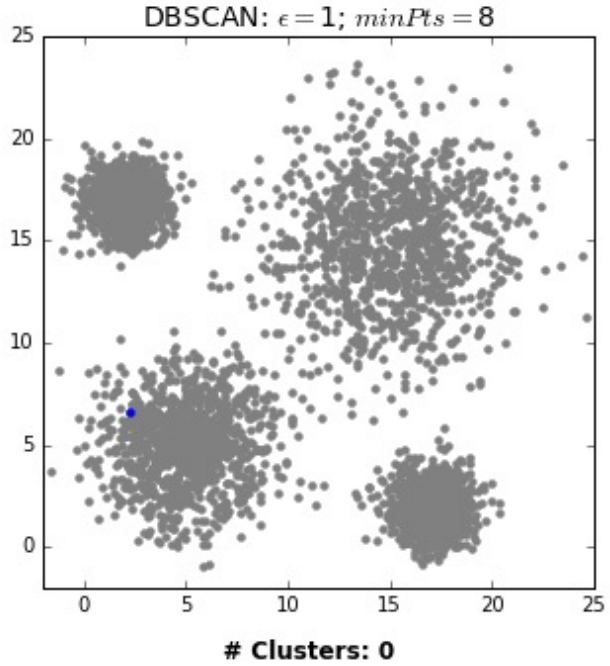
DBSCAN (Density-Based Spatial Clustering of Applications With Noise)



- eps: neighborhood radius
- min_samples: 4

- A: Core
- B, C: not core
- N: noise

DBSCAN Illustration



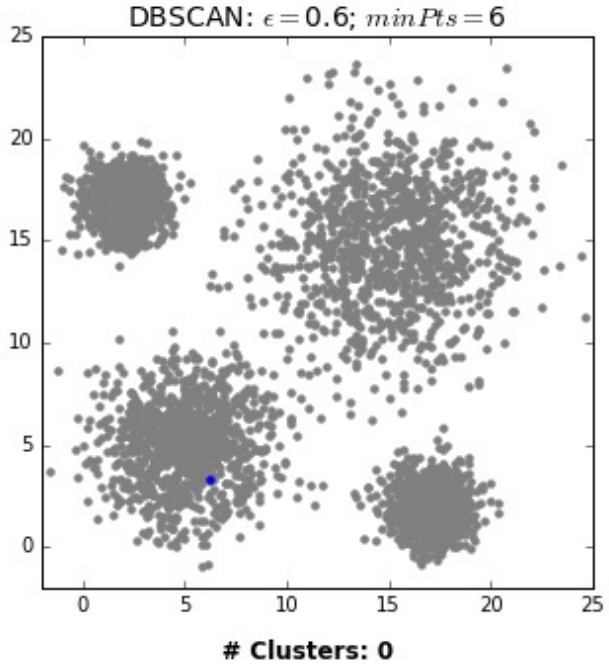
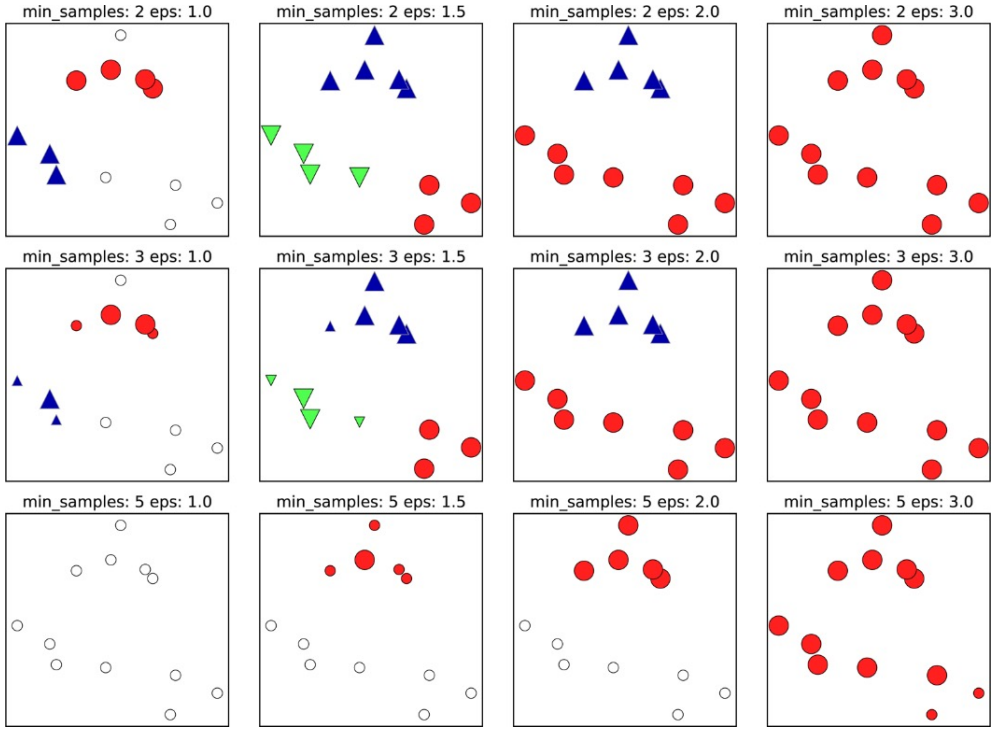
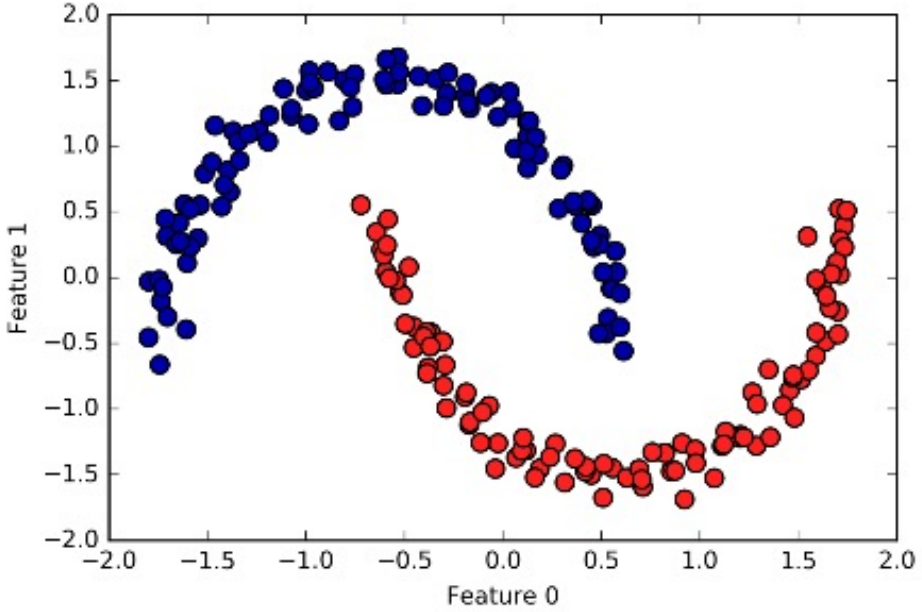


Illustration of Parameters



Pros and Cons

Pro: Can learn arbitrary cluster shapes
Pro: Can detect outliers
Con: Needs two (non-obvious?) parameters to adjust



Silhouette Coefficient

- Input:
  - Data
  - Cluster assignment
- Output:
  - A score from -1 to 1 where:
    - 0 is random clustering
    - 1 is perfect clustering
- For each point i, compute:
  - a_i = mean distance between i and all other points in the same cluster
  - b_i = mean distance between i and the points in the next nearest cluster
  - $s_i = \frac{b_i - a_i}{\max(a_i, b_i)}$
- The overall silhouette coefficient is the mean value of $s_i, i = 1 \ldots n$