Artificial Intelligence

——Partitional and Density-based Clustering

Yanghui Rao
Assistant Prof., Ph.D
School of Data and Computer Science,
Sun Yat-sen University
raoyangh@mail.sysu.edu.cn

Clustering

设计 程度

- **Task :** Evolve measures of similarity to cluster a collection of documents/terms into groups within which similarity within a cluster is larger than across clusters.

注释

- **Cluster Hypothesis:** Given a 'suitable' clustering of a collection, if the user is interested in document/term *d/t*, he is likely to be interested in other members of the cluster to which *d/t* belongs.
- **Similarity measures**
  - Represent documents by vectors
    - Distance between document vectors
    - Cosine of angle between document vectors
- **Issues**

  规模
  - Large number of noisy dimensions
  - Notion of noise is application dependent

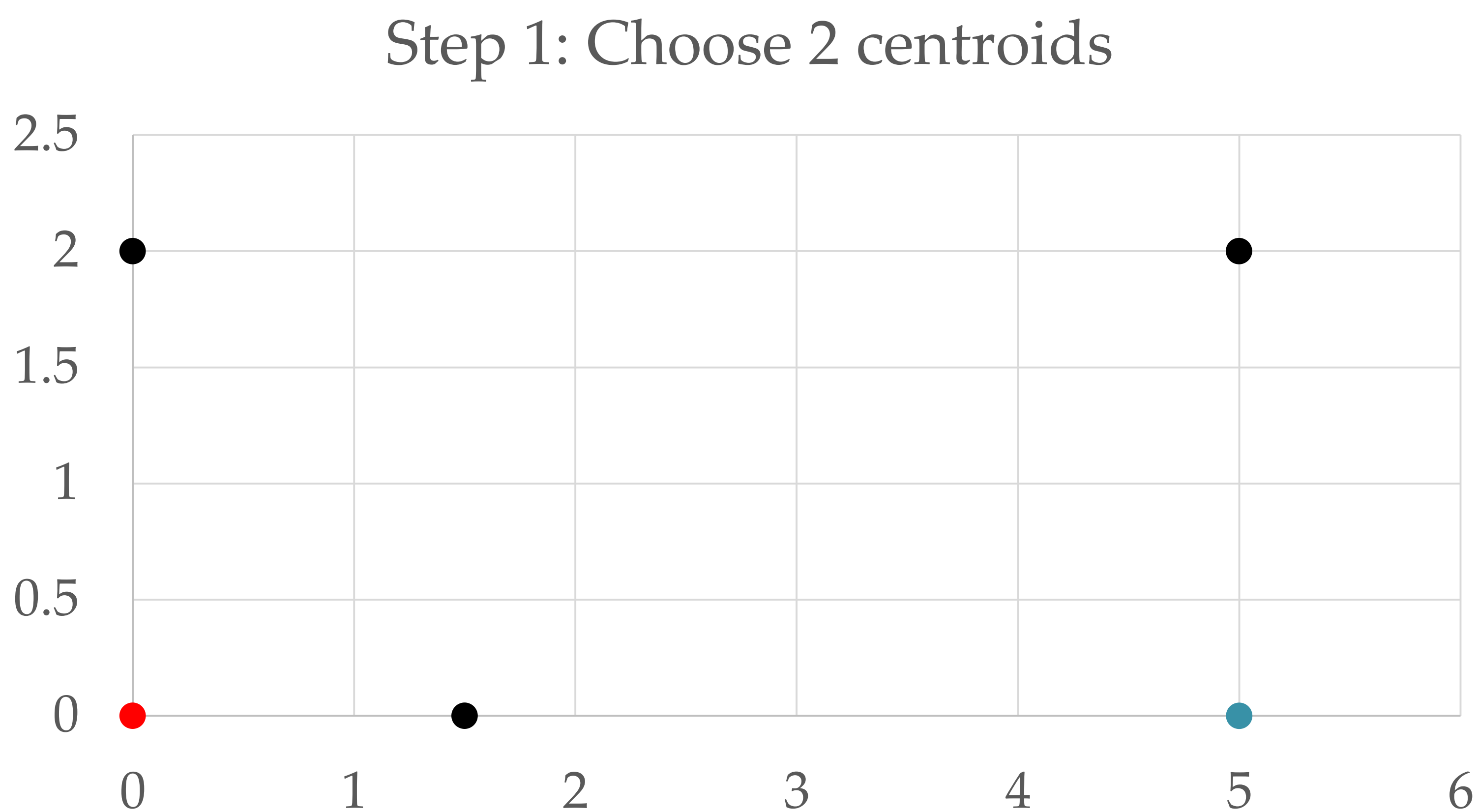
  概念

Partitional Clustering

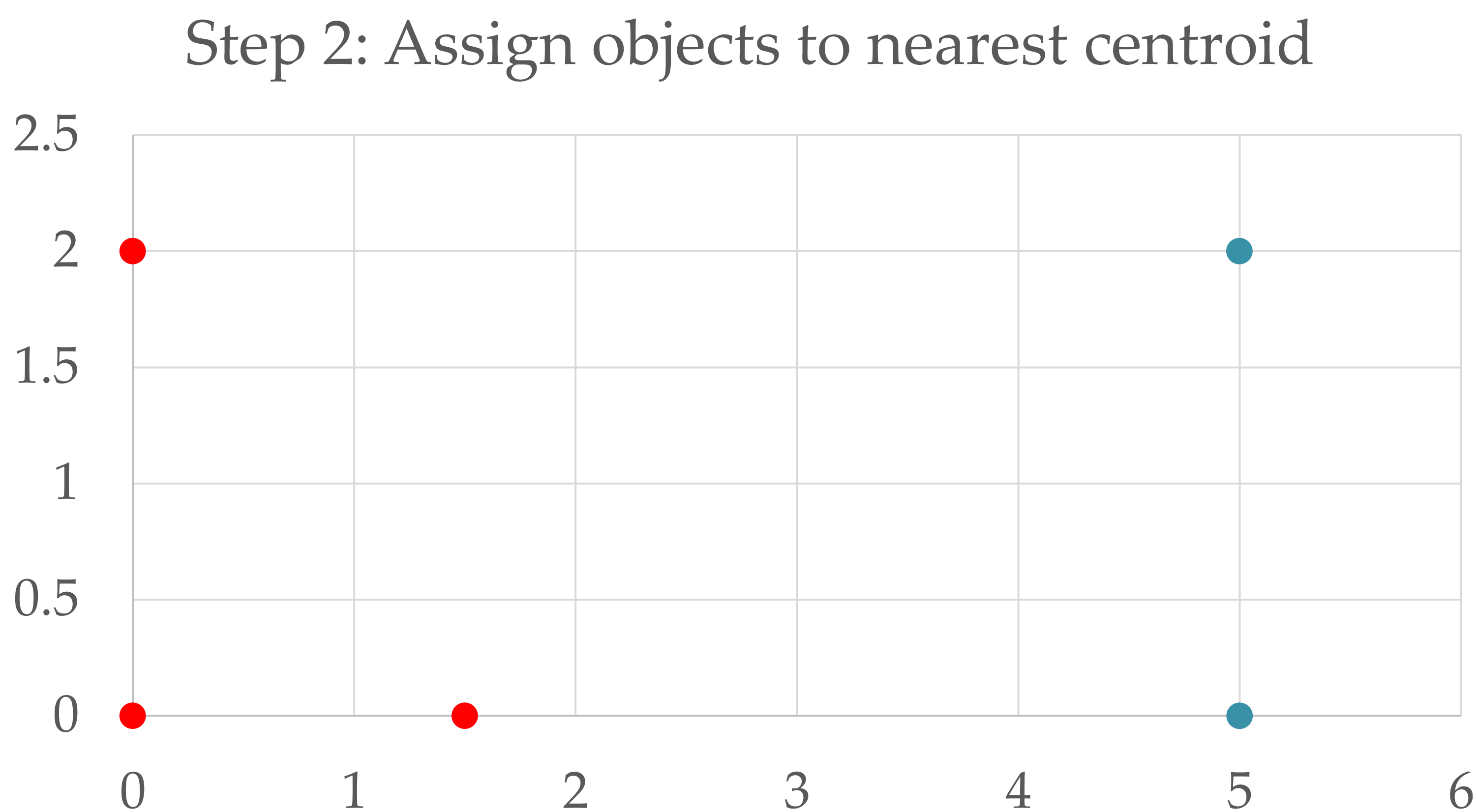
- *k*-Means: Repeat…
  形心，矩心
  - Choose *k* arbitrary 'centroids'
  - Assign each document to nearest centroid
  - Re-compute centroids

- Example of *k*-Means (划分法)
  - x1 = (0, 2), x2 = (0, 0), x3 = (1.5, 0), x4 = (5, 0), x5 = (5, 2)
  - $k = 2$

k-Means

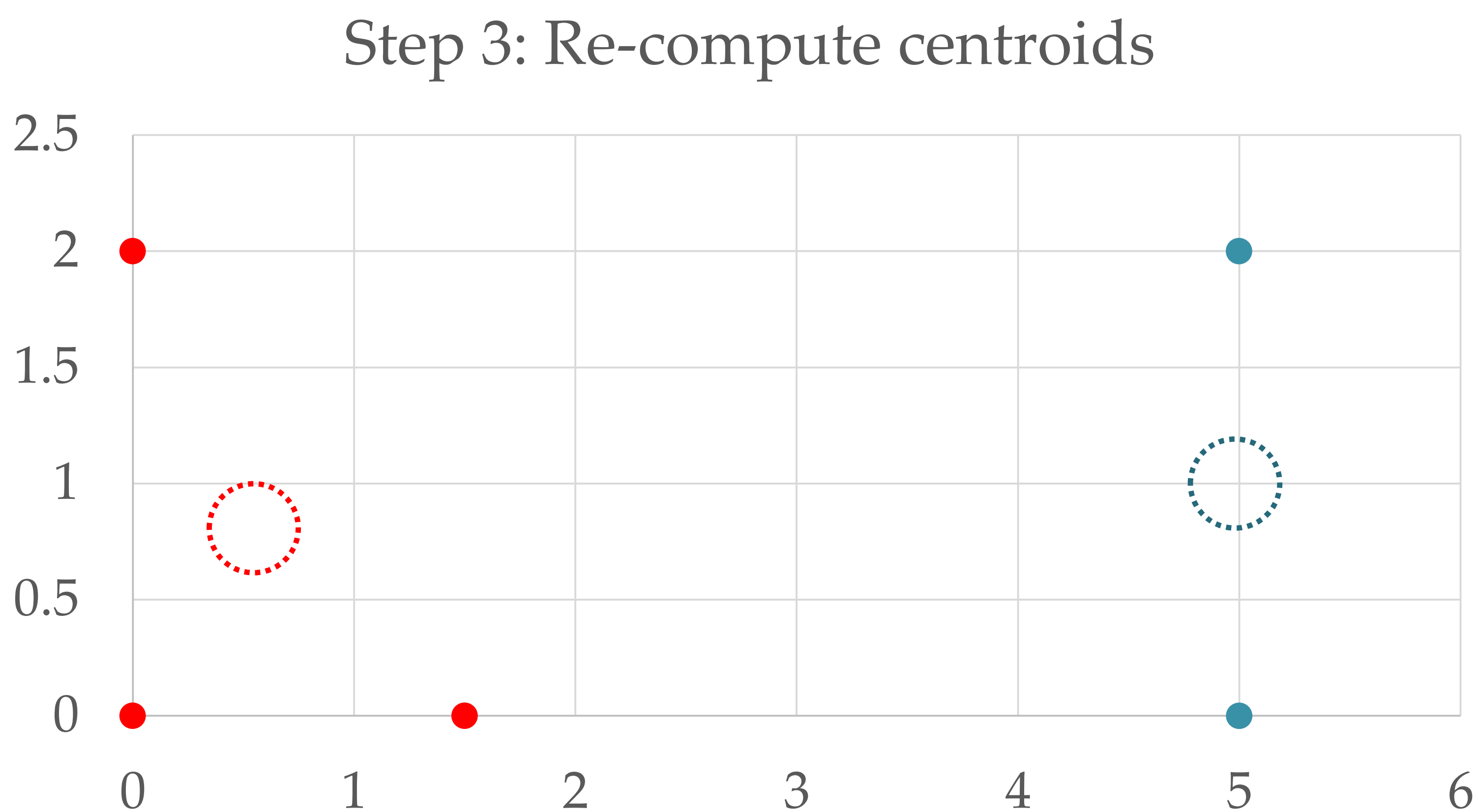


k-Means

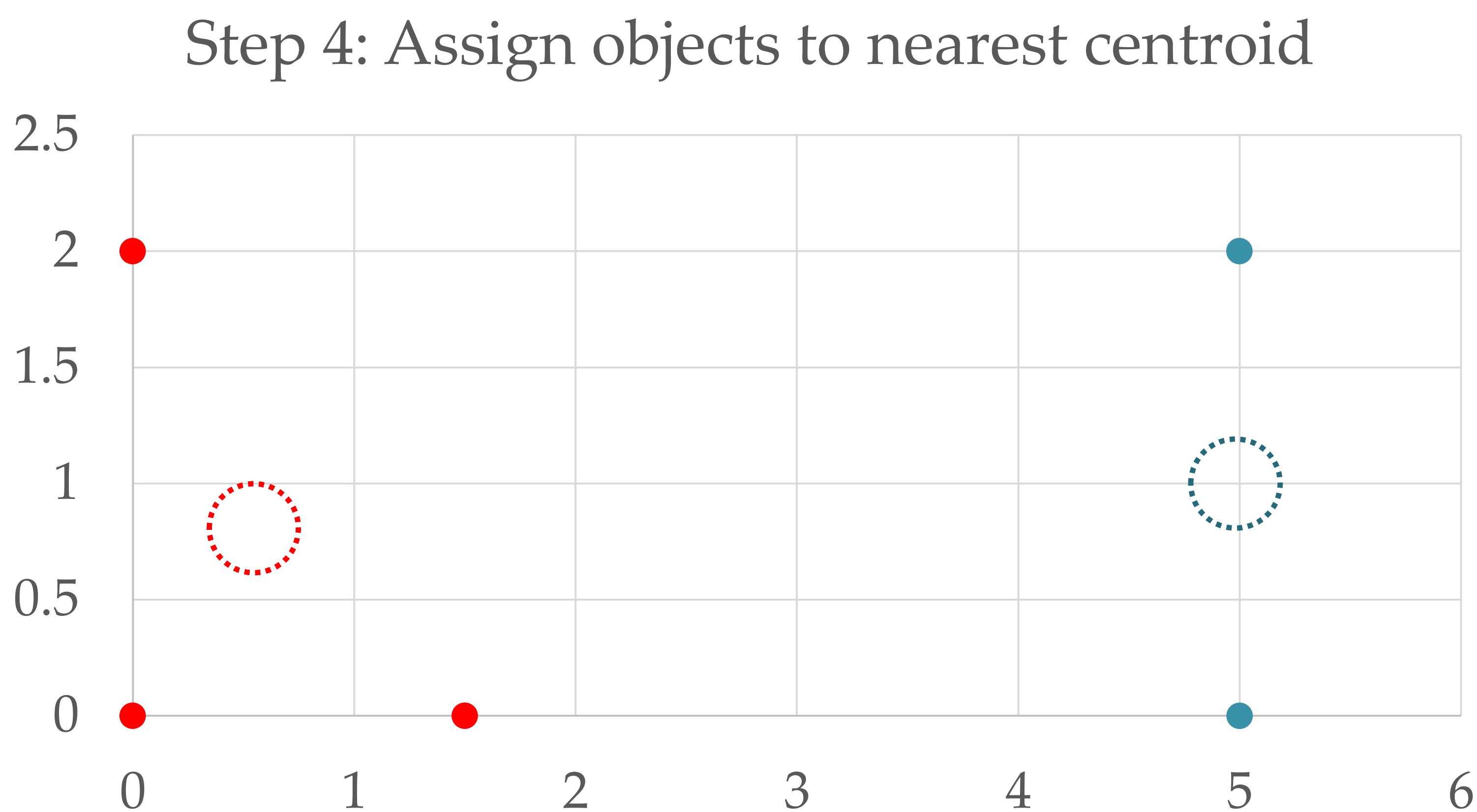


Step 2: Assign objects to nearest centroid
2.5
2
1.5
1
0.5
0
0
1
2
3
4
5
6

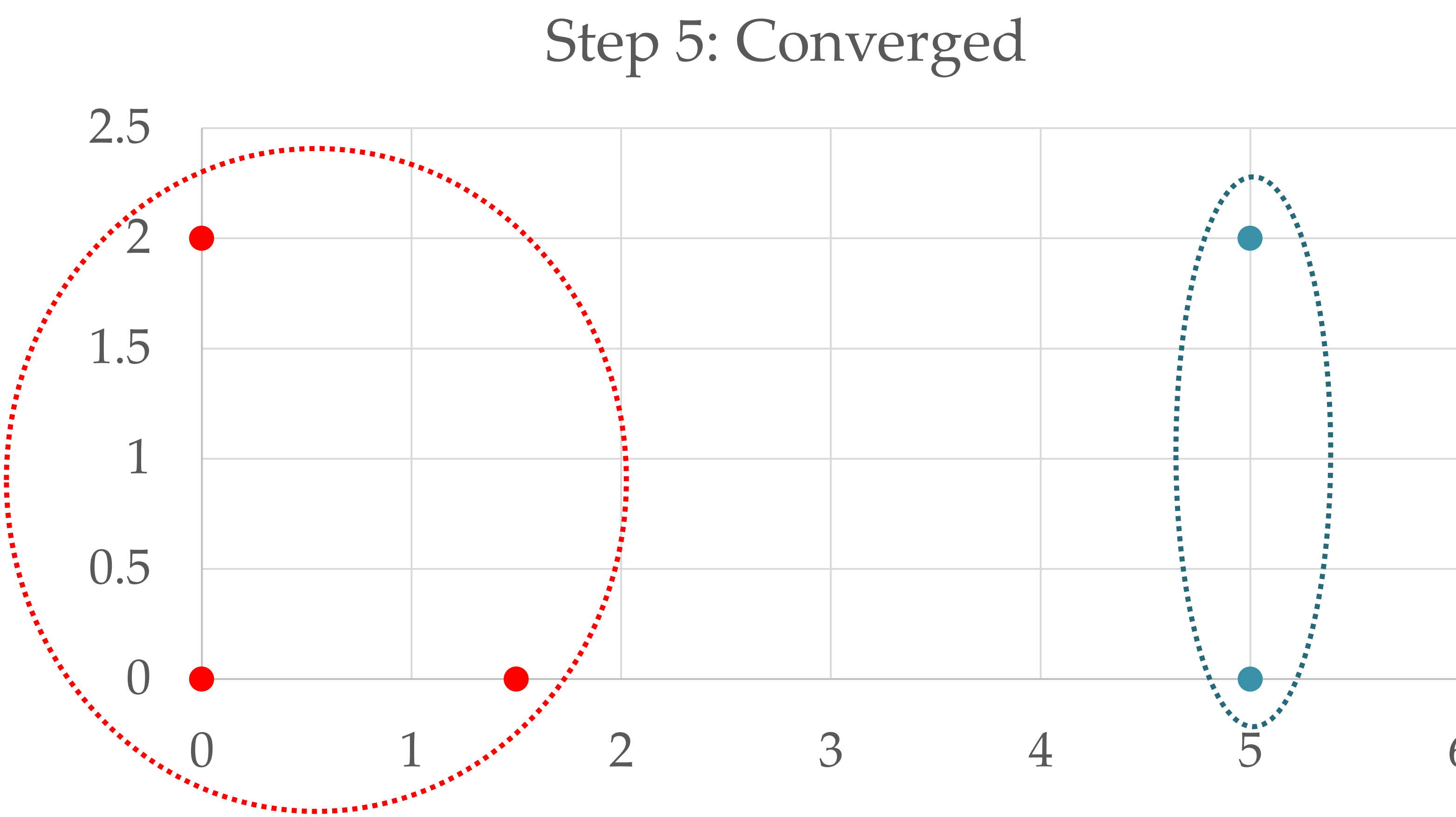
k-Means



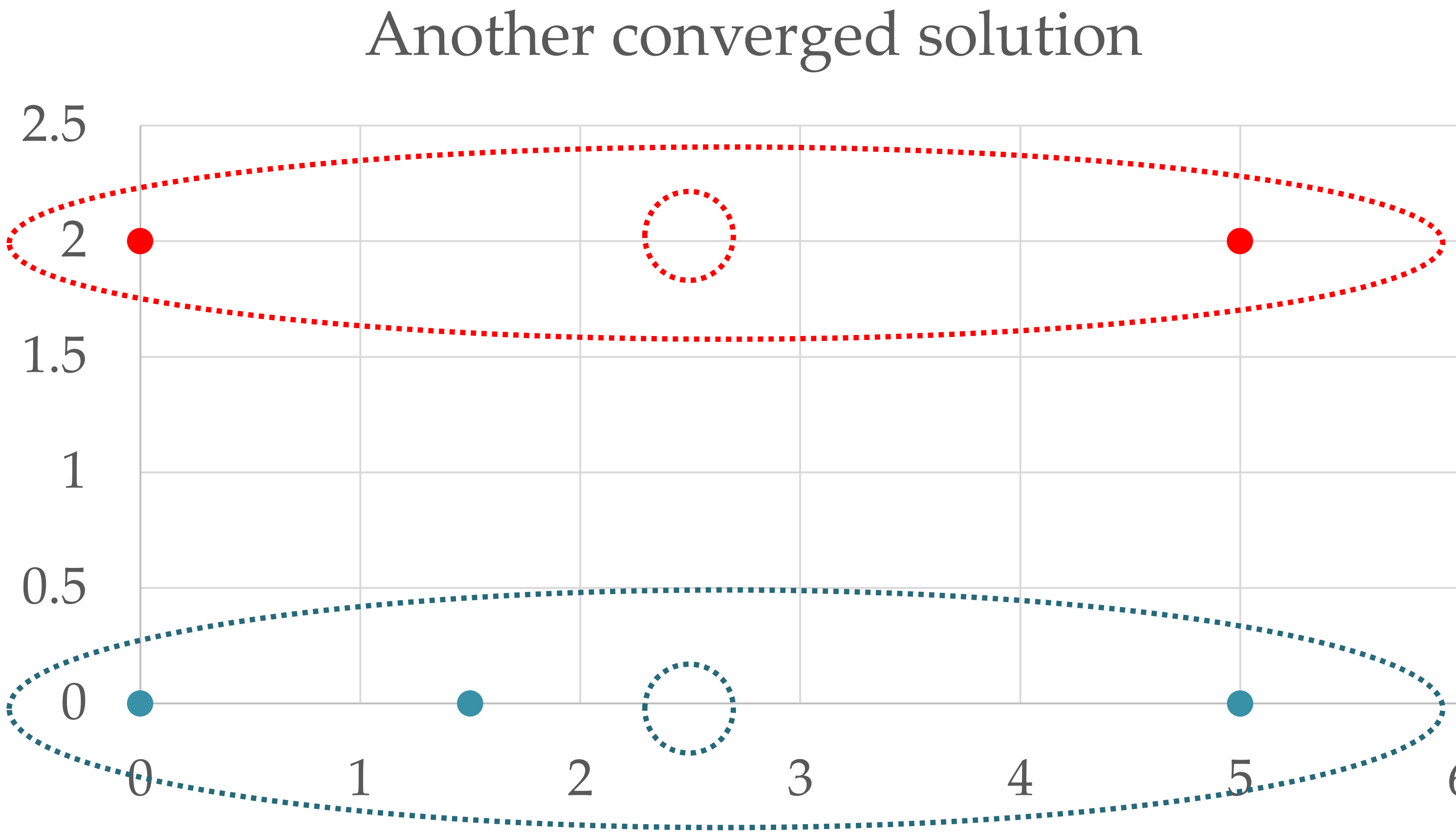
k-Means


Step 4: Assign objects to nearest centroid
2.5
2
1.5
1
0.5
0
0
1
2
3
4
5
6

k-Means



k-Means



k-Means: Choosing *k*

- Mostly problem driven
- Could be 'data driven' only when either
  - Data is not sparse 稀疏的
  - Measurement dimensions are not too noisy

Density-based Clustering

- Density-based clustering locates regions of high density that are separated from one another by regions of low density.

空间的

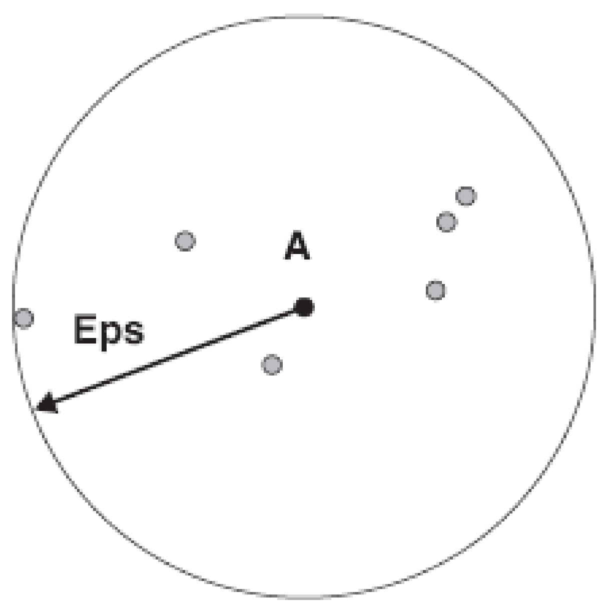
- DBSCAN (Density-Based Spatial Clustering of Applications with Noise) is a simple and effective density-based clustering (基于密度的聚类) algorithm.

DBSCAN

- For DBSCAN, we need to estimate the density (密度) for a particular point in the data set.
- This is performed by counting the number of points within or at a specified radius, Eps, of that point.
- The count includes the point itself.

DBSCAN

- This technique is illustrated in the following figure.
- The number of points within or at a radius of Eps of point A is 7, including A itself.



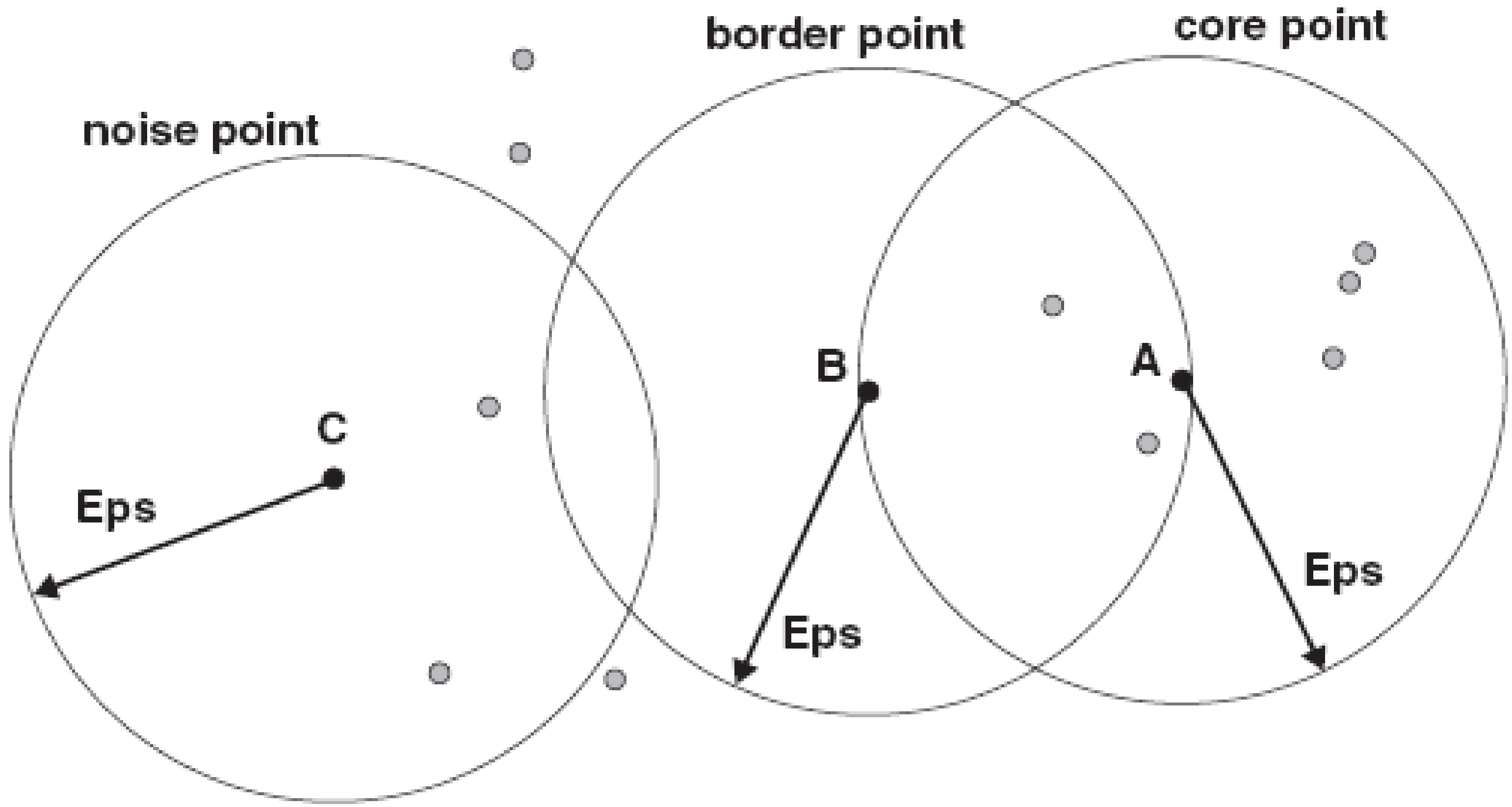
DBSCAN

- The density of any point will depend on the specified radius.
- Suppose the number of points in the data set is m.
- If the radius is large enough, then all points will have a density of m.
- If the radius is too small, then all points will have a density of 1.

DBSCAN

- We need to classify a point as being
  内部
  - In the interior of a dense region (a core point, 核心点).
  - At the edge of a dense region (a border point, 边界点)
    稀疏地
  - In a sparsely occupied region (a noise or background point, 噪音点).
- The concepts of core, border and noise points are illustrated as follows.

DBSCAN



DBSCAN

- Core points are in the interior of a density-based cluster.
- A point is a core point if the number of points within or at the boundary of a given neighborhood of the point is greater than or equal to a certain threshold MinPts.
- The size of the neighborhood is determined by the distance function and a user-specified distance parameter, Eps.
- The threshold MinPts is also a user-specified parameter.
- In the above figure, A is a core point for the indicated radius (Eps) if MinPts=7.

DBSCAN

- A border point is not a core point, but falls within or at the boundary of the neighborhood of a core point.
- In the above figure, B is a border point.
- A border point can fall within the neighborhoods of several core points.

- A noise point is any point that is neither a core point nor a border point.
- In the above figure, C is a noise point.

DBSCAN

- The DBSCAN can be summarized as follows:

- If all points have been processed, stop.
- For a particular point which has not been previously processed, check whether it is a core point or not.
- If it is not a core point
  - Label it as a noise point (This label may change later).
- If it is a core point, label the point and
  - Form a new cluster C_{new} using this point and include all points within or at the boundary of its Eps-neighborhood in the cluster.
  - Insert all these neighboring points into a queue.
  - While the queue is not empty,
    - Remove the first point from the queue
    - If this point is not a core point, label it as a border point.
    - If this point is a core point, label it and check every point in its neighborhood which was not previously assigned to a cluster. For each of these unassigned neighboring points,
      - Assign the point to the current cluster C_{new}.
      - Insert the point into the queue.

DBSCAN

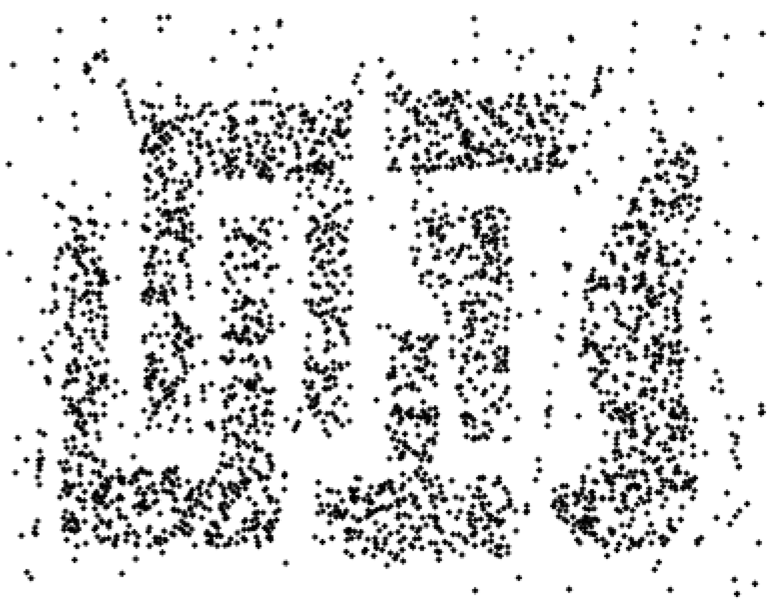
幻灯片

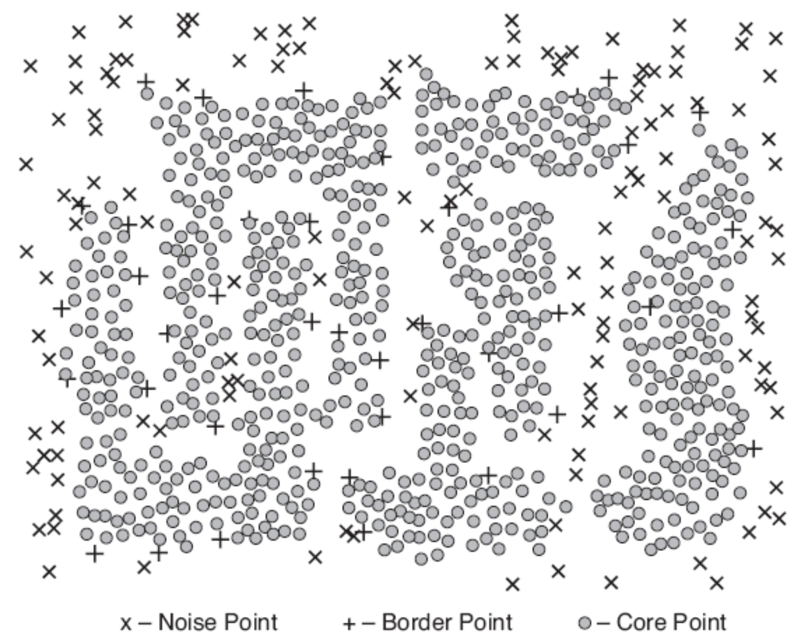
- The left figure on the next slide shows a sample data set with 3000 2-D points.

- The right figure shows the resulting clusters found by DBSCAN.

- The core points, border points and noise points are also displayed.

DBSCAN





DBSCAN

有抵抗力的

- DBSCAN is relatively resistant to noise and can handle clusters of arbitrary shapes and sizes.

- As a result, it can find many clusters that cannot be found using *k*-Means.

无监督学习参考

Unsupervised Learning Reference

- S.J. Rizvi and J.R. Haritsa. Maintaining data privacy in association rule mining. *Proceedings of the 28th VLDB Conference*, 34(6):682-693, 2002.
- A.K. Jain, M.N. Murty, and P.J. Flynn. Data clustering: a review. *ACM Computing Surveys*, 31(2):264-323, 1999.
- A. Rodriguez and A. Laio. Clustering by fast search and find of density peaks. *Science*, 344(6191):1492-1496, 2014.