机器学习引论

彭玺

pengxi@scu.edu.cn

www.pengxi.me

提纲

一．Clustering

二．Performance Metric

三．Hierarchical Clustering

四．Density based Clustering

提纲

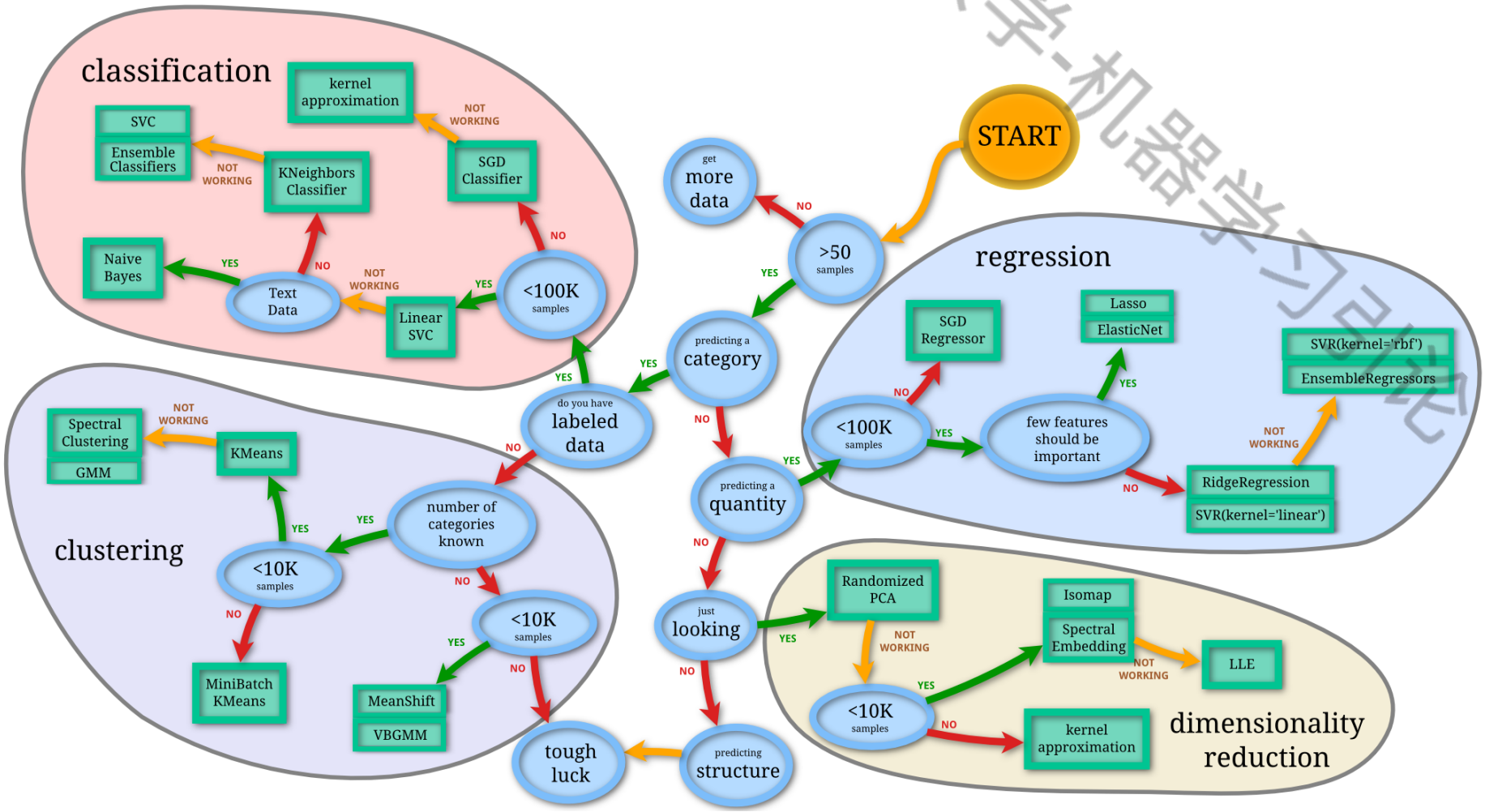
一．Clustering

二．Performance Metric

三．Hierarchical Clustering

四．Density based Clustering

一、Clustering



一、Clustering

Classification

- Given training instances (x, y)
- Learn a model/mapping f(.)
- Such that f(x) = y
- Use f(.) to predict new x

- y denotes the label

一、Clustering

Classification

- Given training instances (x, y)
- Learn a model/mapping f(.)
- Such that f(x) = y
- Use f(.) to predict new x

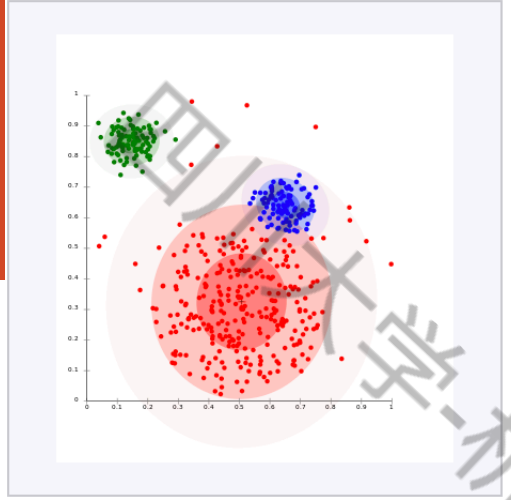
- y denotes the label

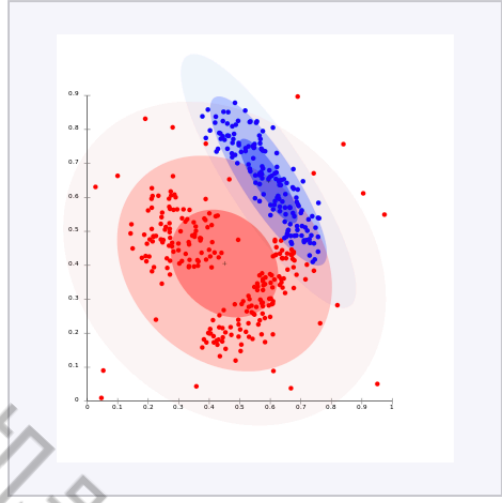
Dimension Reduction

- Given training instances (x, y)
- Learn a model/mapping f(.)
- Such that g(f(x,y))=x
- Use f(.) to predict new x

- y could be label or prior or assumption

一、Clustering

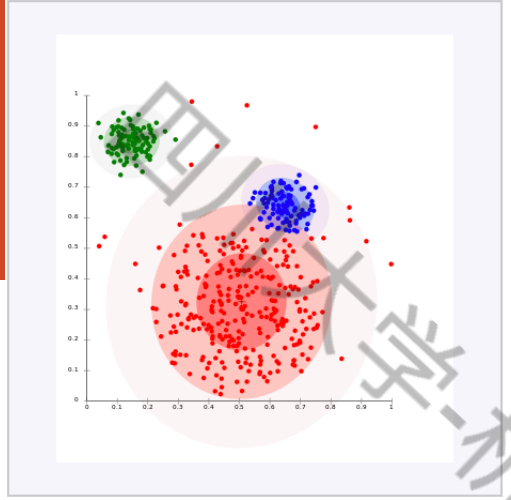


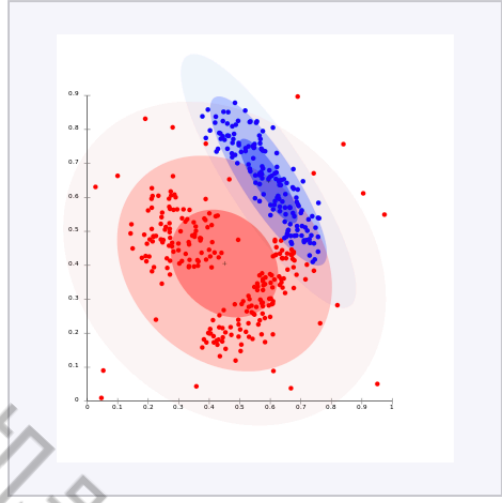


Problem Statement :

- Given a set of data points, group them into multiple clusters so that:
  - points within each cluster are similar to each other
  - points from different clusters are dissimilar

一、Clustering

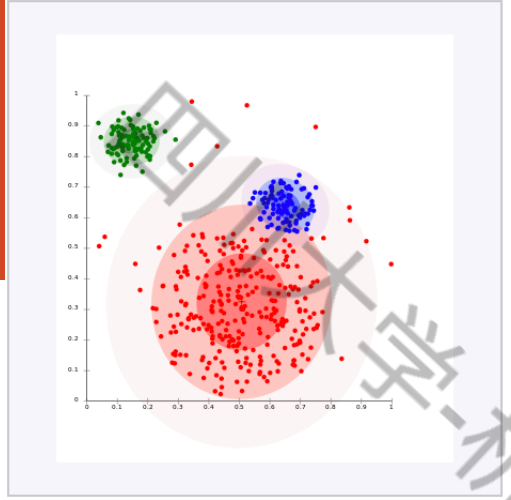


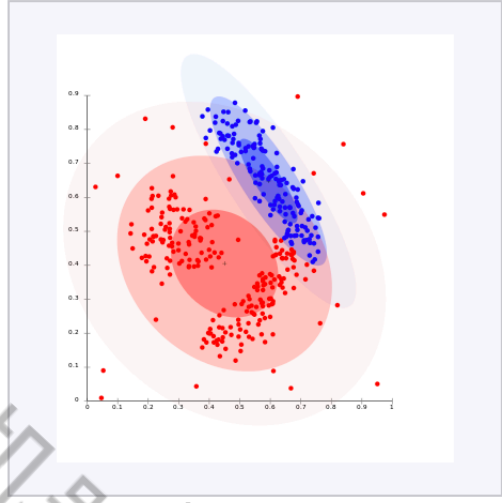


Problem Statement :

- Given a set of data points, group them into multiple clusters so that:
  - points within each cluster are similar to each other $\min \sum_{j} \sum_{\mathbf{x}_i \in \mathrm{C}_j} \|\mathbf{x}_i - \mathbf{u}_j\|_2^2$
  - points from different clusters are dissimilar $\max \sum_{i} \sum_{j} \|\mathbf{u}_i - \mathbf{u}_j\|_2^2$

一、Clustering



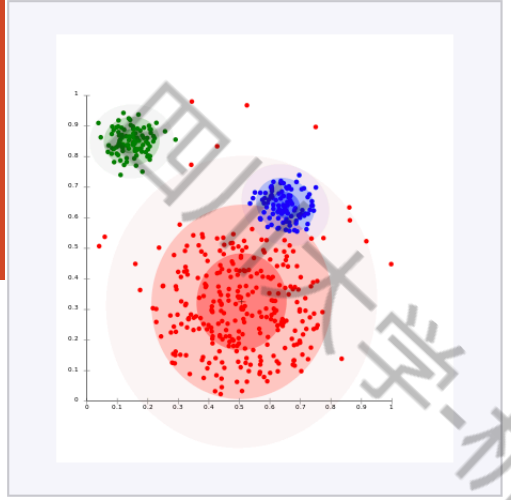


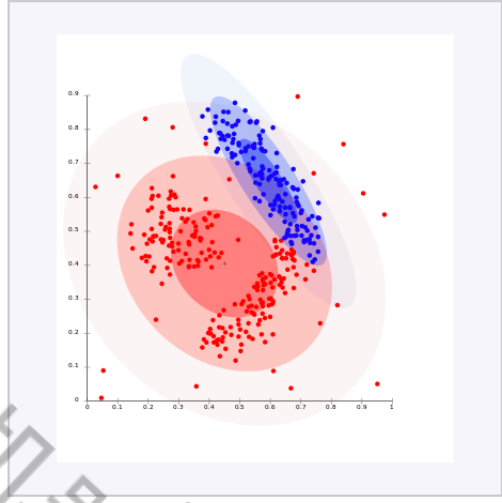
Problem Statement：

- Given a set of data points, group them into multiple clusters so that:
  - points within each cluster are similar to each other $\min \sum_{j} \sum_{\mathbf{x}_i \in \mathrm{C}_j} \|\mathbf{x}_i - \mathbf{u}_j\|_2^2$
  - points from different clusters are dissimilar $\max \sum_{i} \sum_{j} \|\mathbf{u}_i - \mathbf{u}_j\|_2^2$

Challenges 1 (key problem of clustering analysis): The major difficulty is that the label is unknown so that the within-/between- class scatter is unavailable.

一、Clustering





Problem Statement：

- Given a set of data points, group them into multiple clusters so that:
  - points within each cluster are similar to each other $\min \sum_j \sum_{\mathbf{x}_i \in \mathrm{C}_j} \|\mathbf{x}_i - \mathbf{u}_j\|_2^2$
  - points from different clusters are dissimilar $\max \sum_i \sum_j \|\mathbf{u}_i - \mathbf{u}_j\|_2^2$

Challenges 1 (key problem of clustering analysis): The major difficulty is that the label is unknown so that the within-/between- class scatter is unavailable.

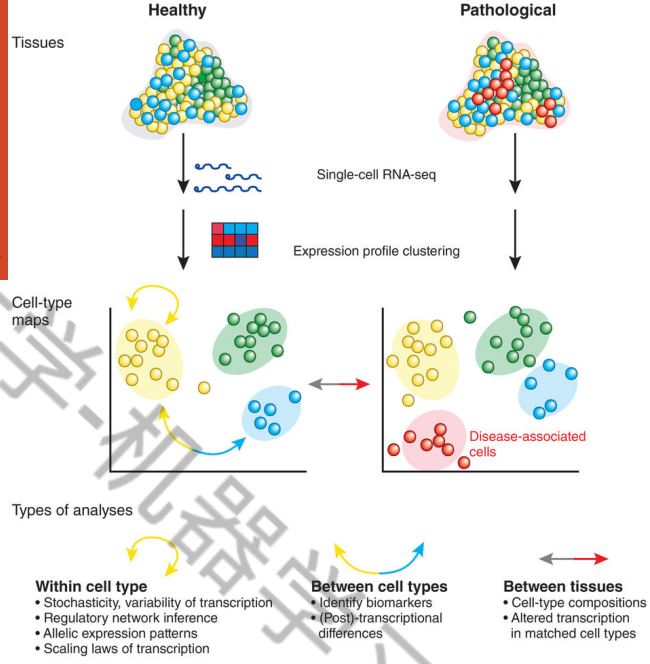
Challenges 2 (high-dimensional clustering analysis): Usually, points are in a high-dimensional space, and similarity is defined using a distance measure

- Euclidean, Cosine, Jaccard, edit distance, ...

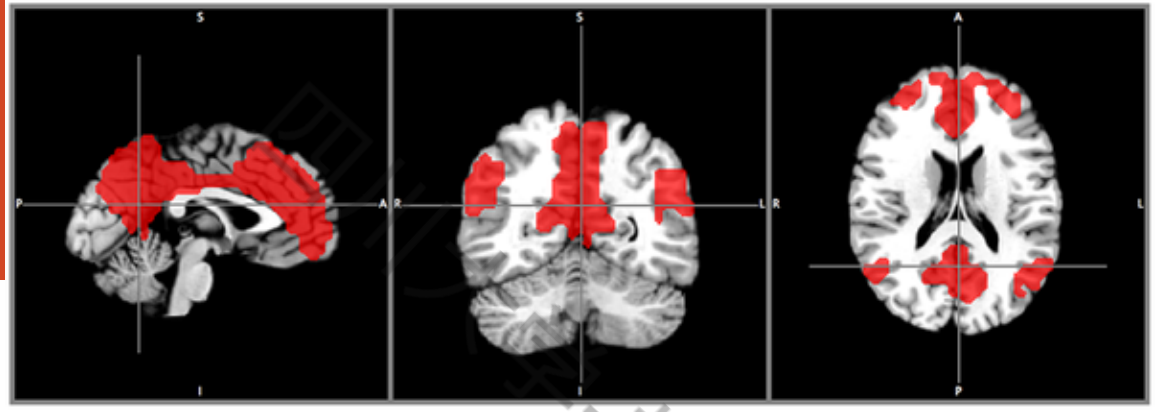
一、Clustering

Applications in biology, computational biology and bioinformatics

- Plant and animal ecology: describe and to make spatial and temporal comparisons of communities of organisms in heterogeneous environments
- Transcriptomics(转录组学): build groups of genes with related expression patterns
- Sequence analysis: group homologous sequences (同源序列) into gene families, gene duplication
- …

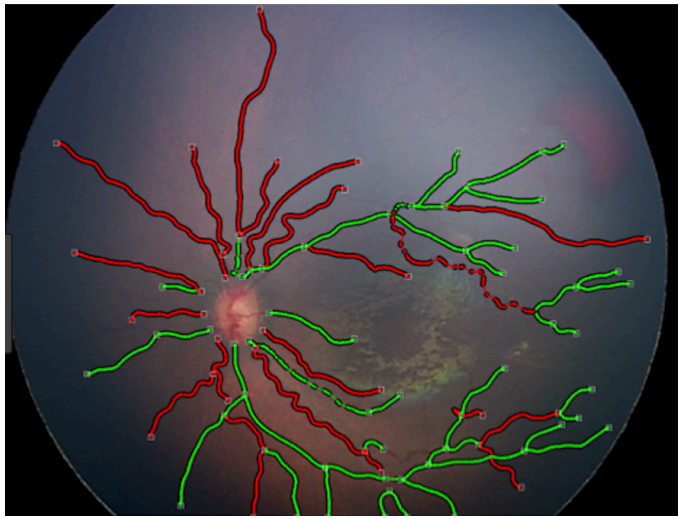


一、Clustering



Applications in Medicine

- Medical imaging: differentiate between different types of tissue in a three-dimensional image
- Image segmentation: divide a fluence map into distinct regions



一、Clustering

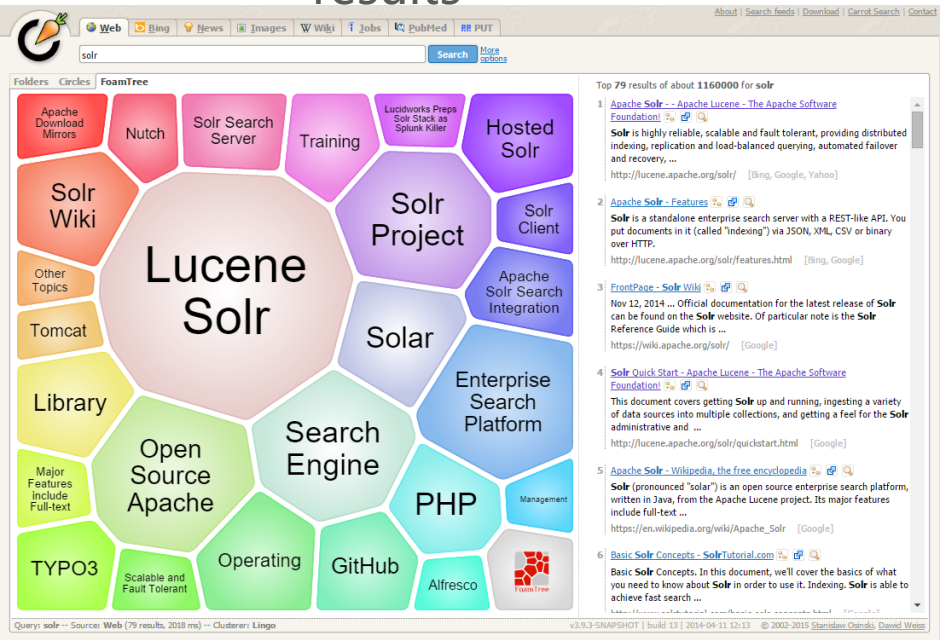
Applications in Business and marketing

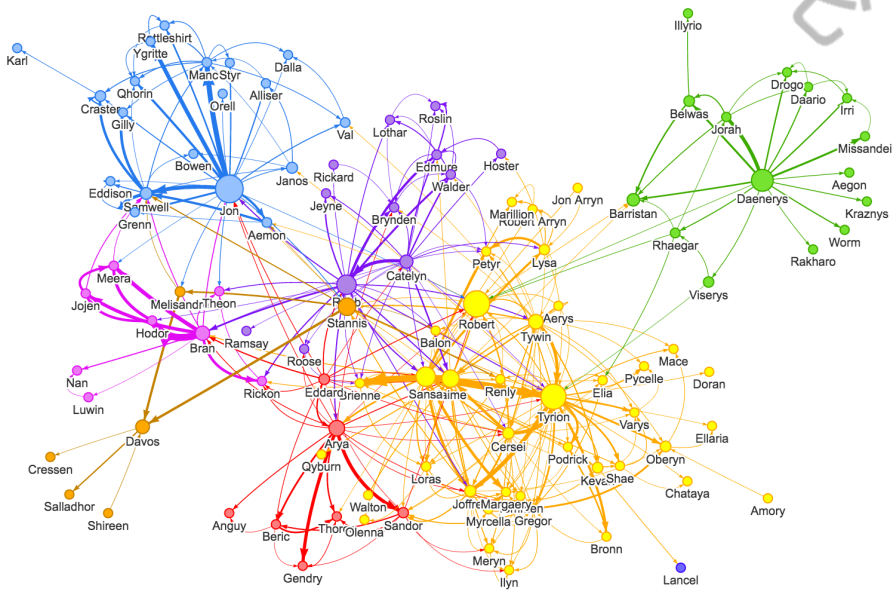
- Market research: partition the general population of consumers into market segments and to better understand the relationships between different groups of consumers/potential customers, and for use in market segmentation, Product positioning, New product development and Selecting test markets.
- Grouping of shopping items: group all the shopping items available on the web into a set of unique products

一、Clustering

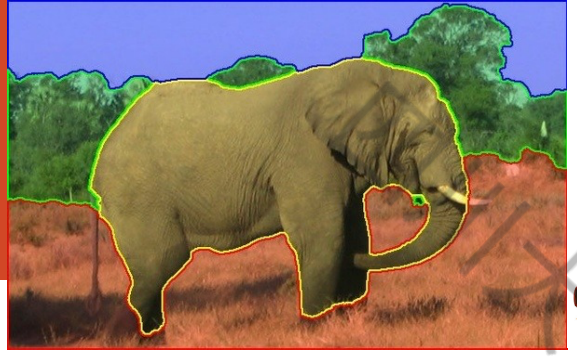
Applications in World wide web

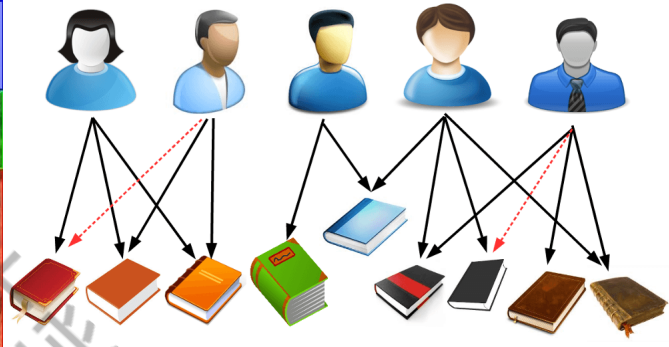
- Social network analysis: recognize communities within large groups of people
- Search result grouping: create a more relevant set of search results





一、Clustering

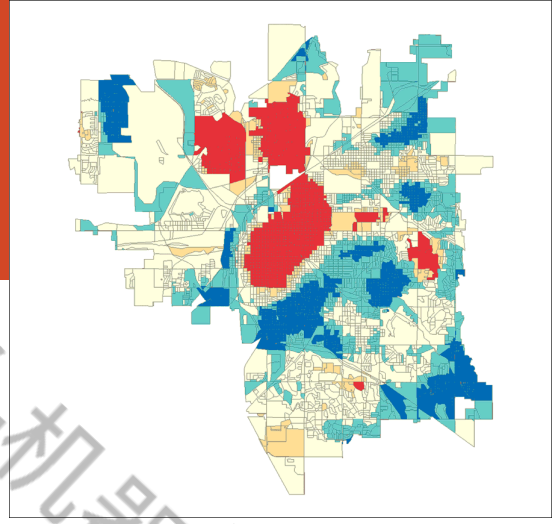




Applications in Computer science

- Image segmentation: divide a digital image into distinct regions for border detection or object recognition
- Evolutionary algorithms: identify different niches within the population of an evolutionary algorithm
- Recommender systems: predict a user's preferences based on the preferences of other users in the user's cluster.
- Anomaly/outlier detection

一、Clustering



Applications in Social science:

- Crime analysis: identify areas where there are greater incidences of particular types of crime. By identifying these distinct areas or "hot spots" where a similar crime has happened over a period of time, it is possible to manage law enforcement resources more effectively.
- Educational data mining: identify groups of schools or students with similar properties.

提纲

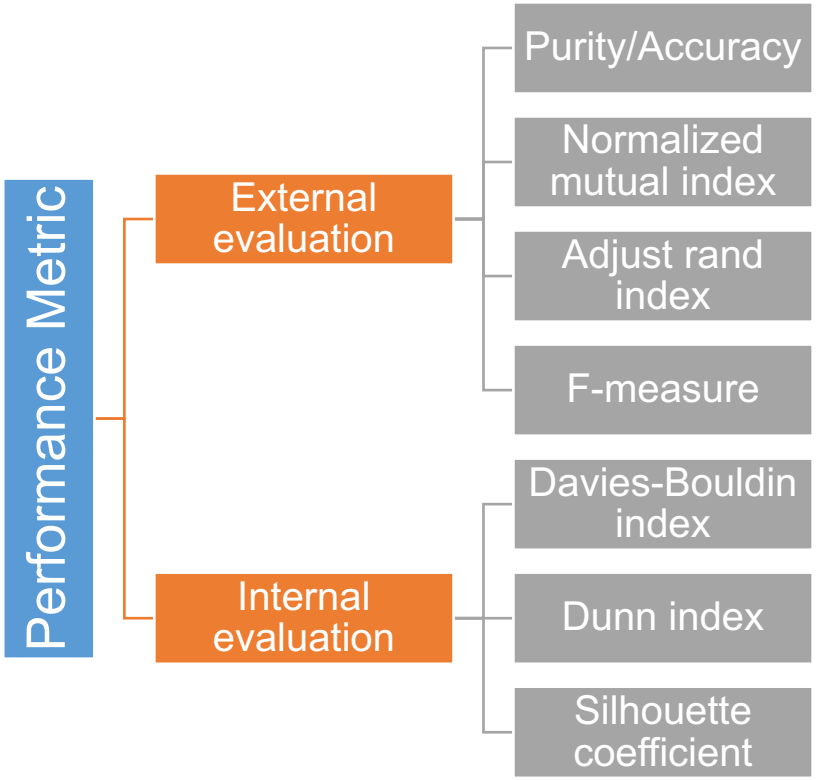
一．Clustering

二．Performance Metric

三．Hierarchical Clustering

四．Density based Clustering

二、Performance Metric



提纲

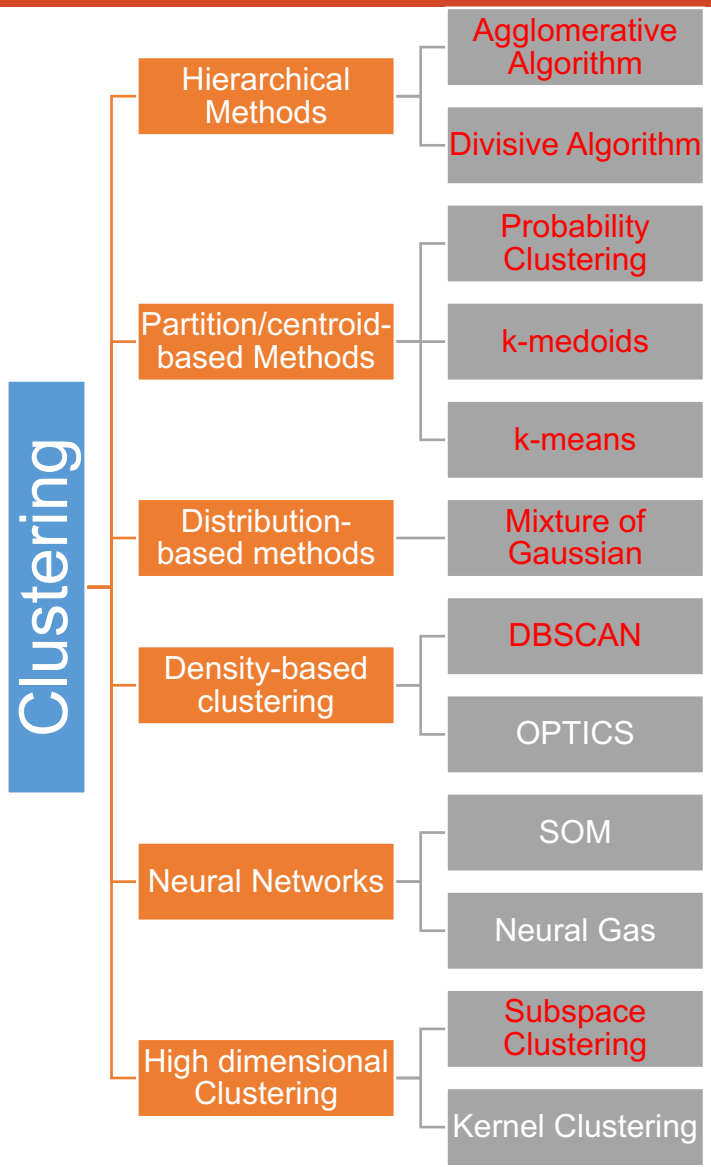
一．Clustering

二．Performance Metric

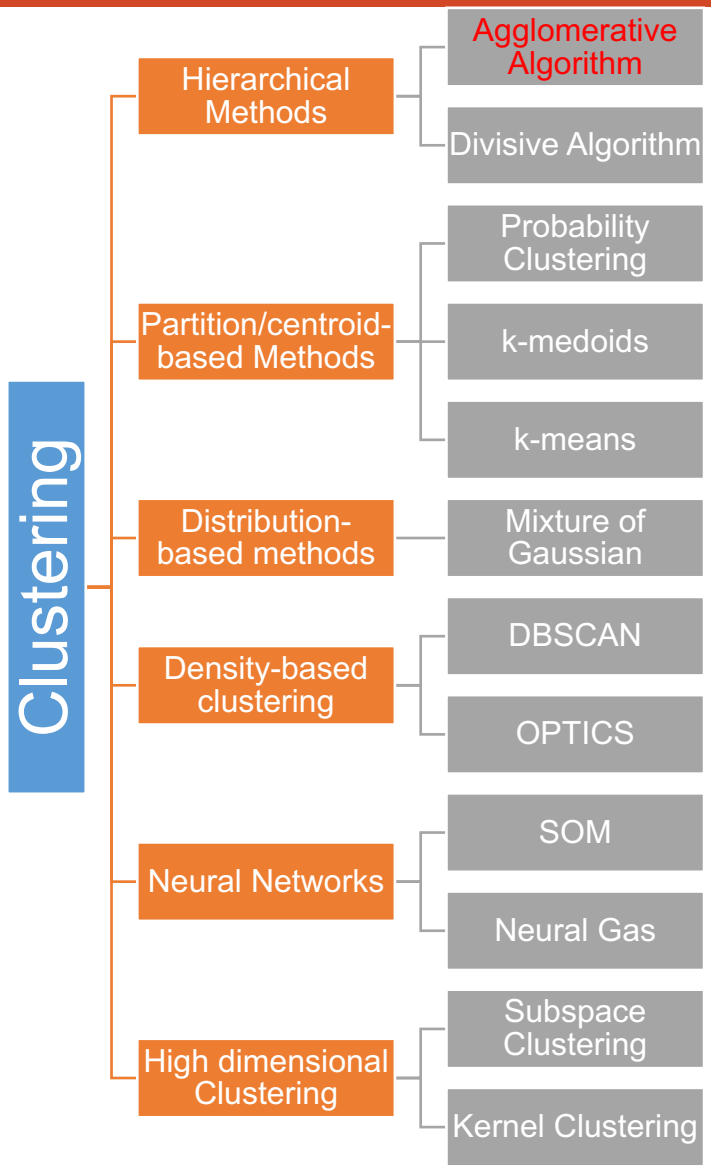
三．Hierarchical Clustering

四．Density based Clustering

三、Hierarchical Clustering



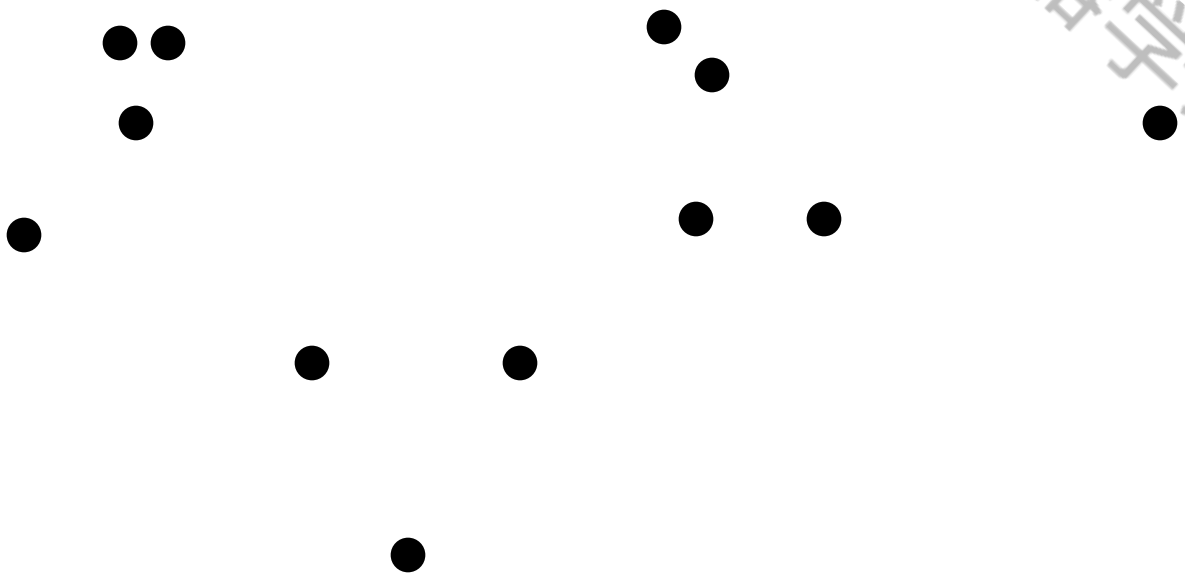
三、Hierarchical Clustering



- Agglomerative (Bottom-up)
  - Compute all pair-wise pattern-pattern similarity coefficients
  - Place each of *n* patterns into a class of its own
  - Merge the two most similar clusters into one
    - Replace the two clusters into the new cluster
    - Re-compute inter-cluster similarity scores w.r.t. the new cluster
  - Repeat the above step until there are *k* clusters left (*k* can be 1)

三、Hierarchical Clustering

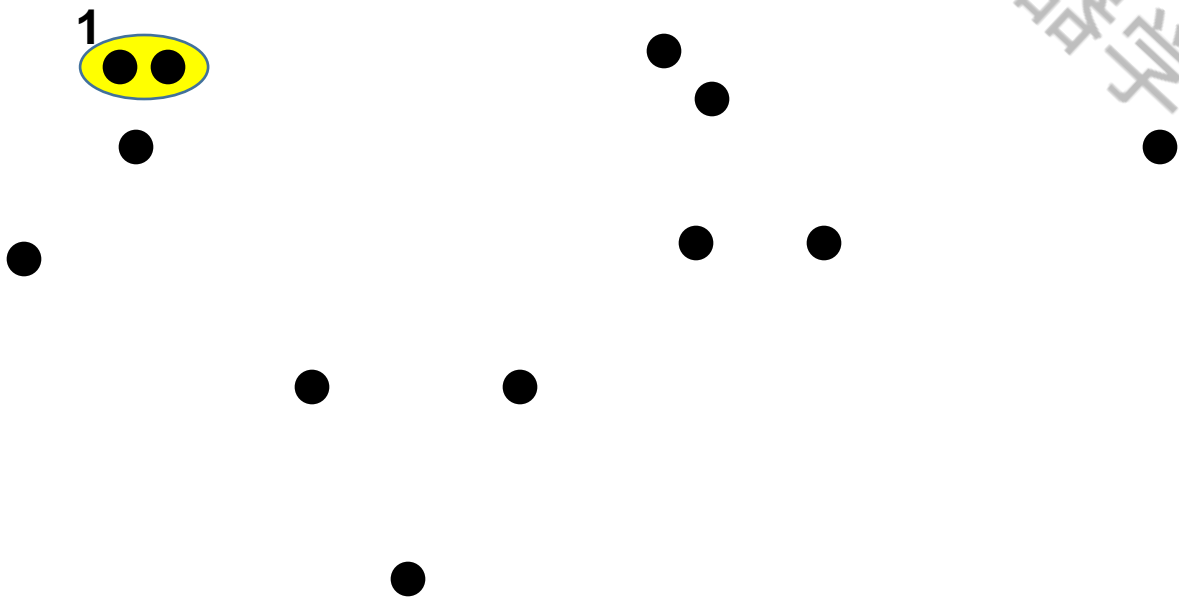
Agglomerative (Bottom up)



三、Hierarchical Clustering

Agglomerative (Bottom up)

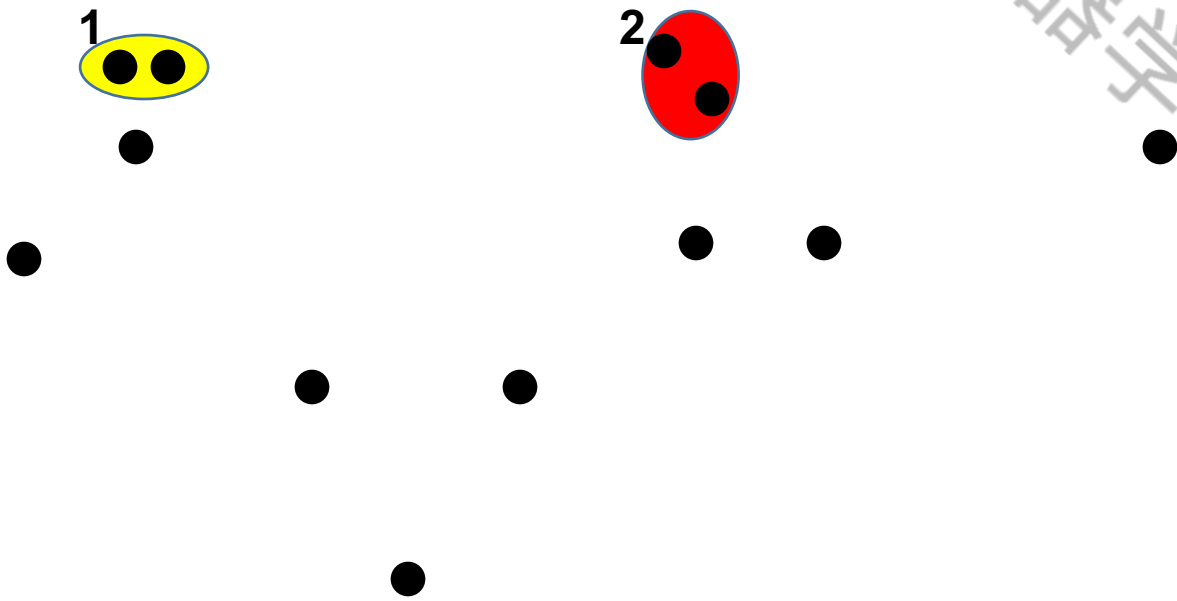
1st iteration



三、Hierarchical Clustering

Agglomerative (Bottom up)

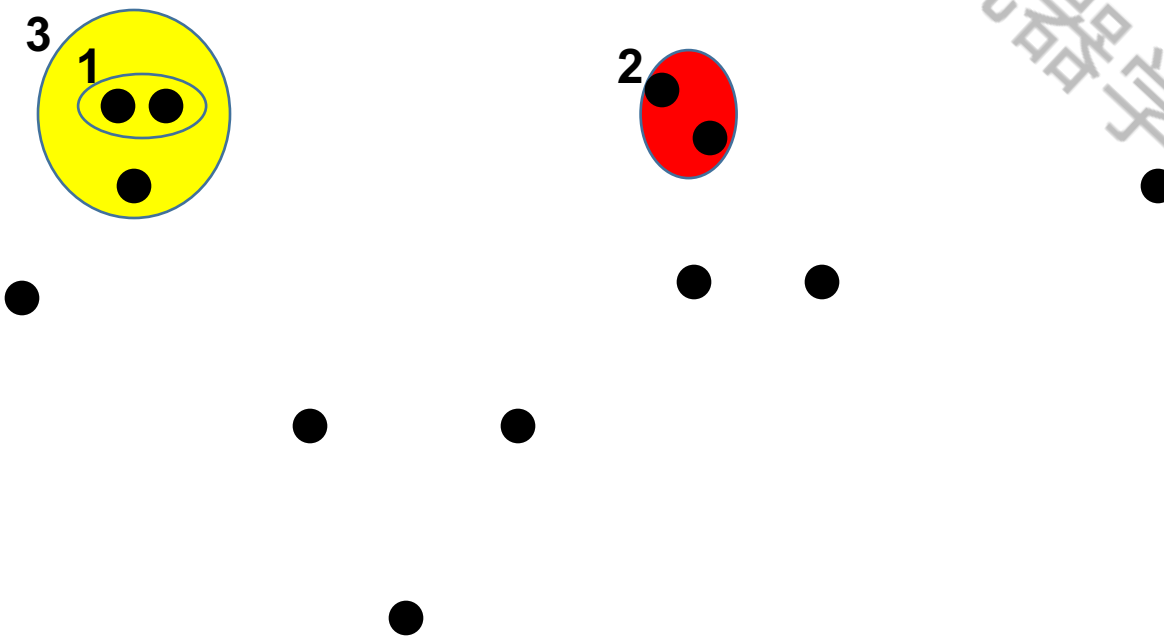
2nd iteration



三、Hierarchical Clustering

Agglomerative (Bottom up)

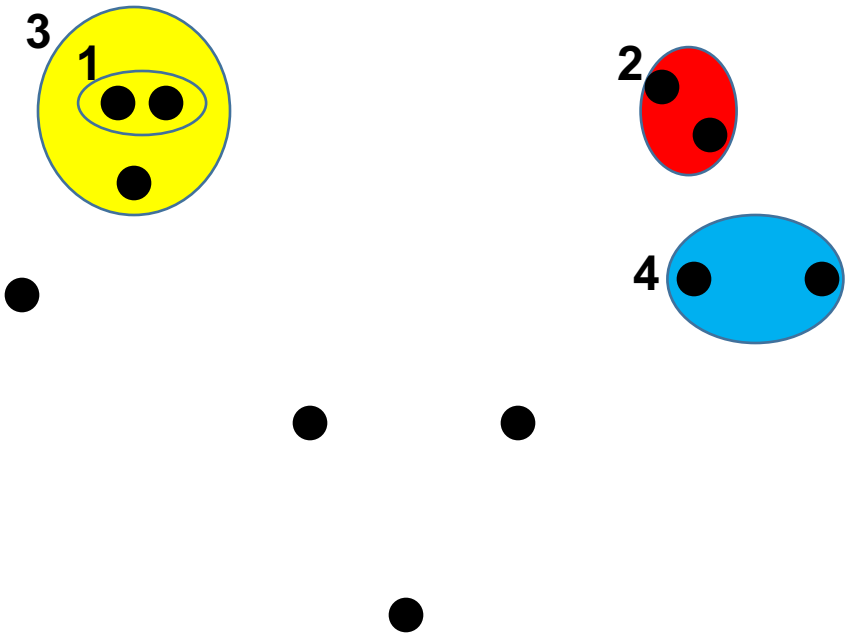
3rd iteration



三、Hierarchical Clustering

Agglomerative (Bottom up)

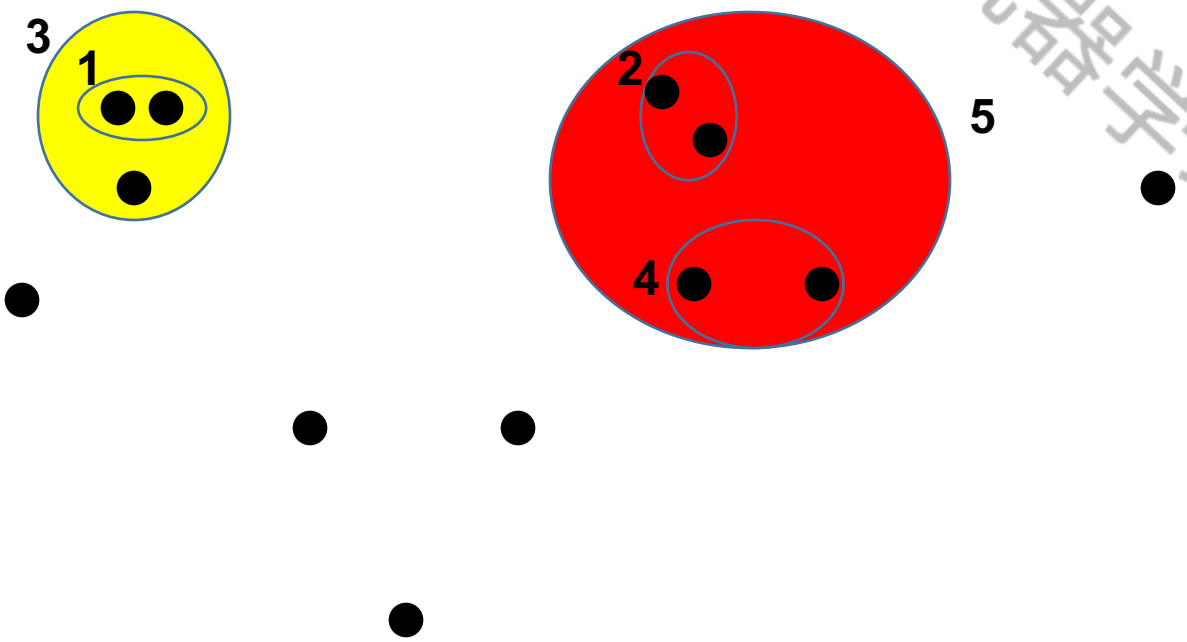
4th iteration



三、Hierarchical Clustering

Agglomerative (Bottom up)

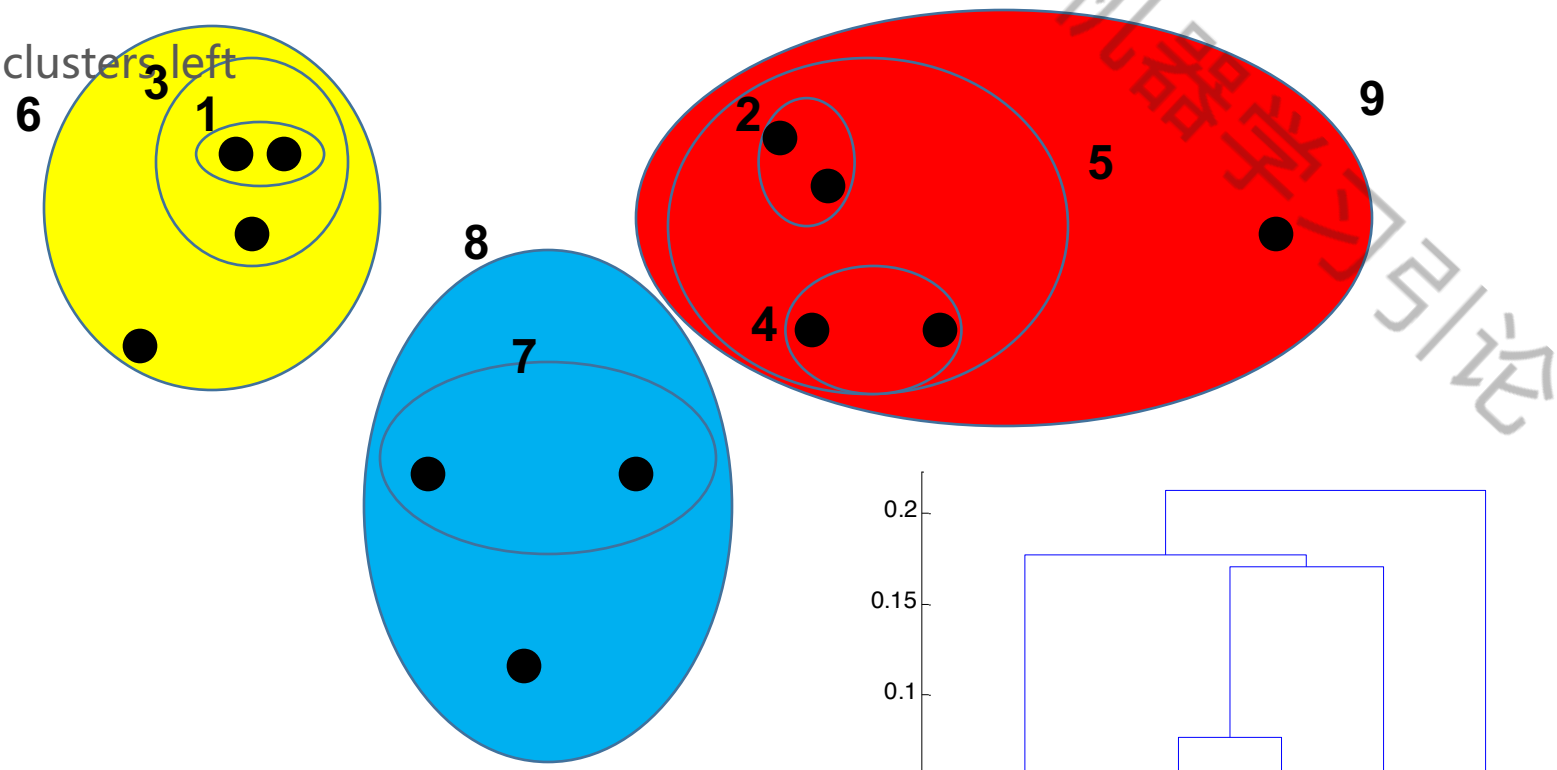
5th iteration



三、Hierarchical Clustering

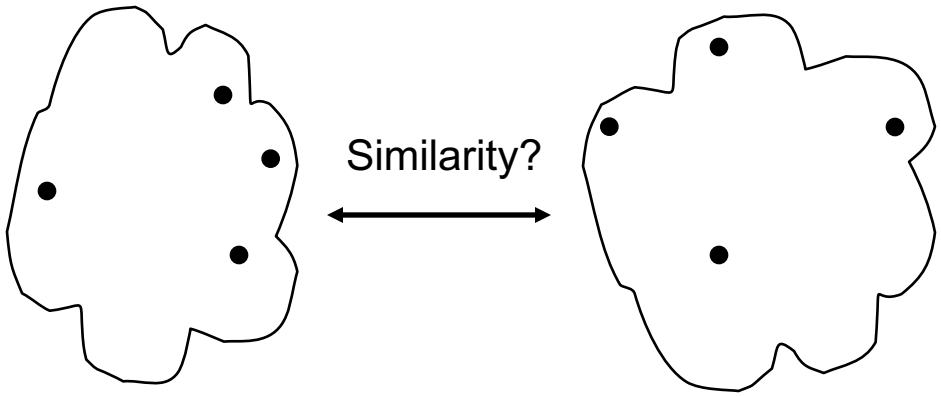
Agglomerative (Bottom up)

Finally k clusters left



dendrogram

Tip: How to Define Inter-Cluster Similarity

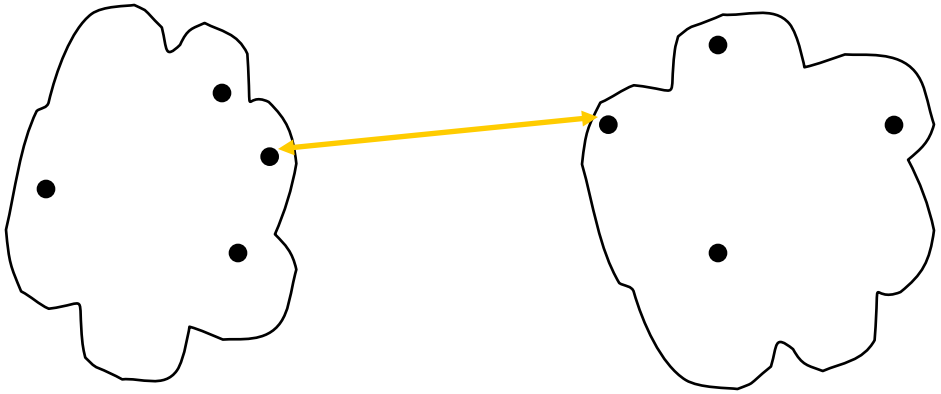


	p1	p2	p3	p4	p5	. . .
p1						
p2						
p3						
p4						
p5						
.						
.						
.						

Proximity Matrix

- MIN
- MAX
- Group Average
- Distance Between Centroids
- Other methods driven by an objective function
  - Ward's Method uses squared error

Tip: How to Define Inter-Cluster Similarity

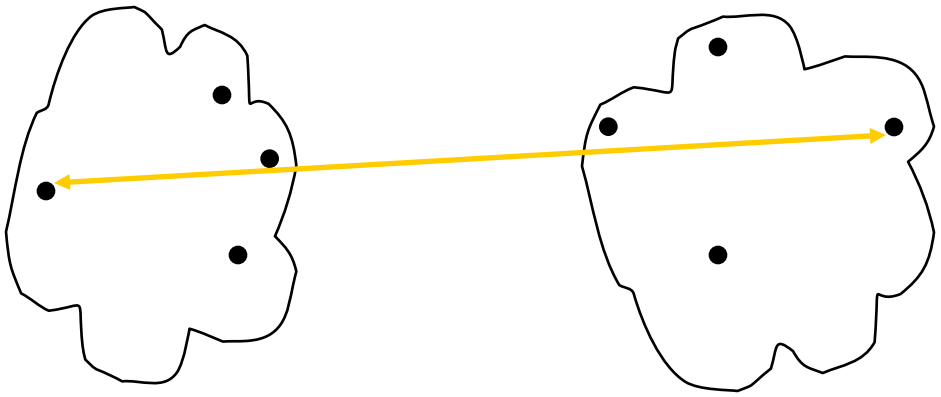


	p1	p2	p3	p4	p5	. . .
p1						
p2						
p3						
p4						
p5						
.						
.						
.						

Proximity Matrix

- MIN
- MAX
- Group Average
- Distance Between Centroids
- Other methods driven by an objective function
  - Ward's Method uses squared error

Tip: How to Define Inter-Cluster Similarity

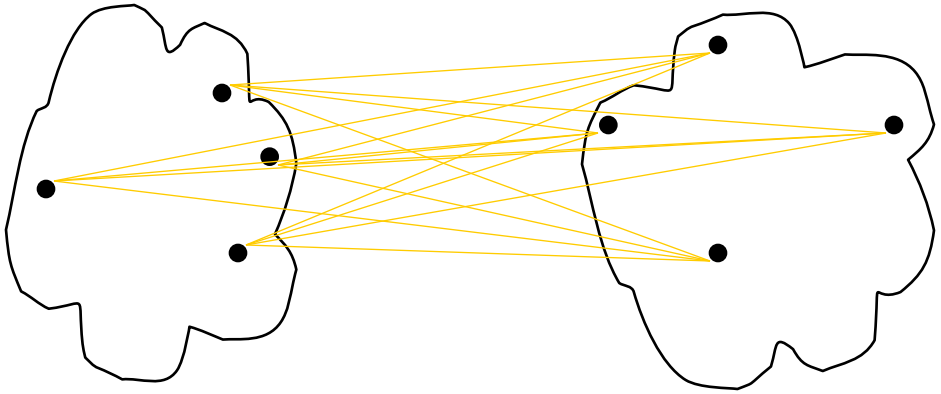


	p1	p2	p3	p4	p5	. . .
p1						
p2						
p3						
p4						
p5						
.						
.						
.						

Proximity Matrix

- MIN
- MAX
- Group Average
- Distance Between Centroids
- Other methods driven by an objective function
  - Ward's Method uses squared error

Tip: How to Define Inter-Cluster Similarity

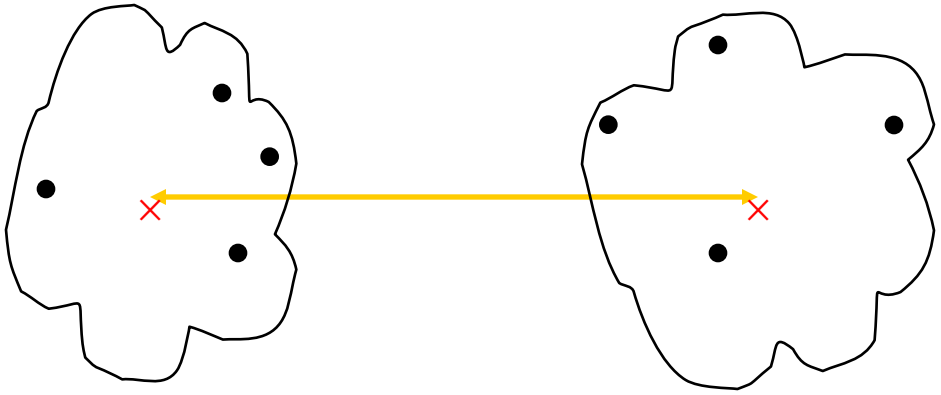


	p1	p2	p3	p4	p5	. . .
p1						
p2						
p3						
p4						
p5						
.						
.						
.						

Proximity Matrix

- MIN
- MAX
- Group Average
- Distance Between Centroids
- Other methods driven by an objective function
  - Ward's Method uses squared error

Tip: How to Define Inter-Cluster Similarity

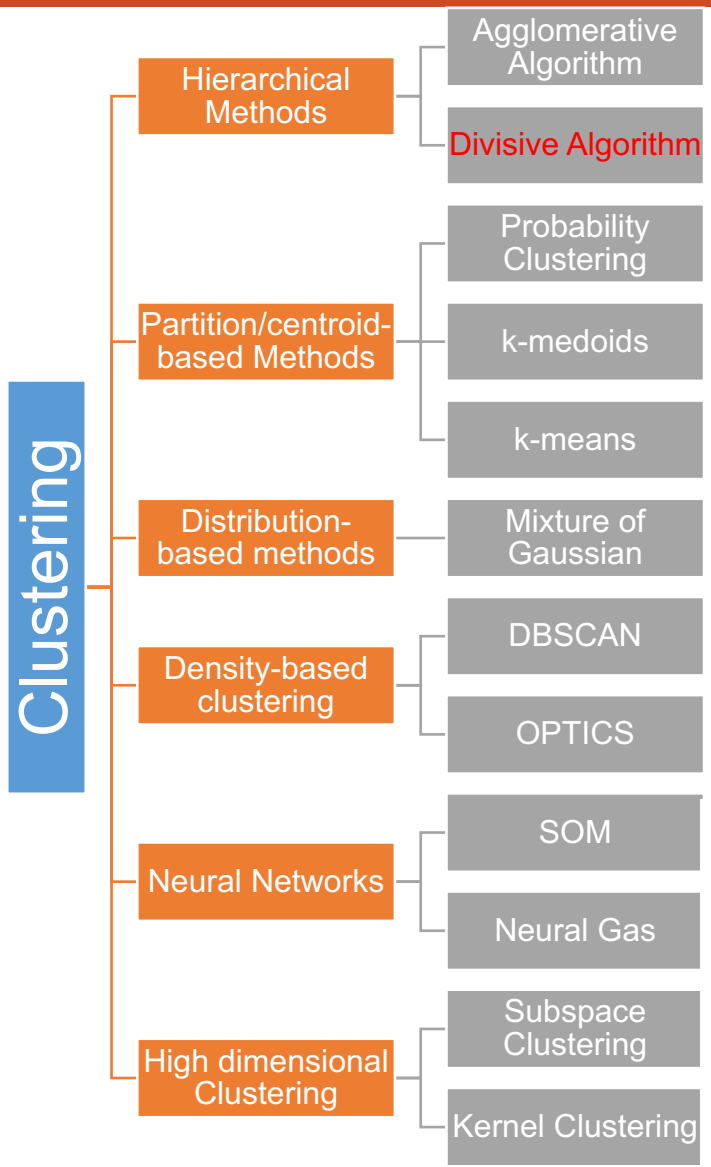


	p1	p2	p3	p4	p5	. . .
p1						
p2						
p3						
p4						
p5						
.						
.						
.						

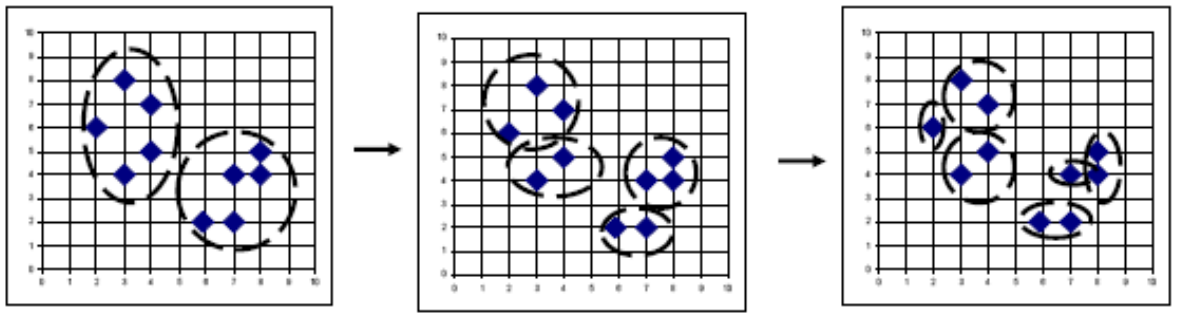
Proximity Matrix

- MIN
- MAX
- Group Average
- Distance Between Centroids
- Other methods driven by an objective function
  - Ward's Method uses squared error

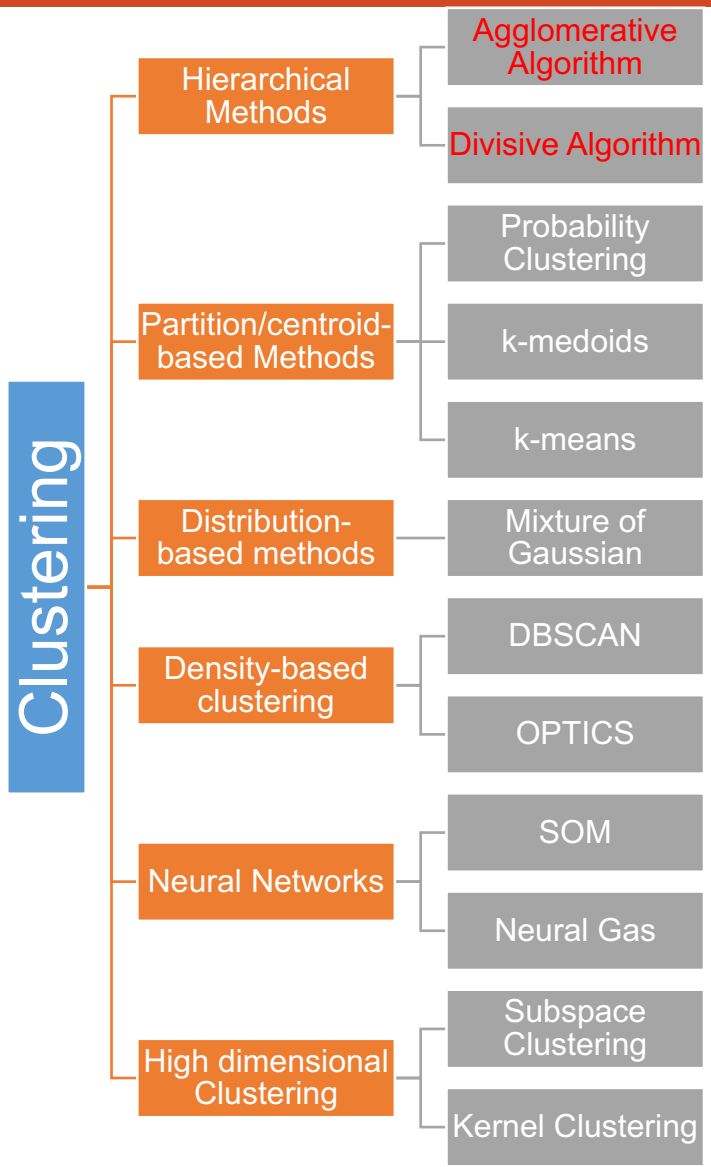
三、Hierarchical Clustering



- Divisive (Top-down)
  - Start at the top with all patterns in one cluster
  - The cluster is split using a flat clustering algorithm
  - This procedure is applied recursively until each pattern is in its own singleton cluster



三、Hierarchical Clustering



- Computational complexity in time and space
- Once a decision is made to combine two clusters, it cannot be undone
- No objective function is directly minimized
- Different schemes have problems with one or more of the following:
  - Sensitivity to noise and outliers
  - Difficulty handling different sized clusters and convex shapes
  - Breaking large clusters

提纲

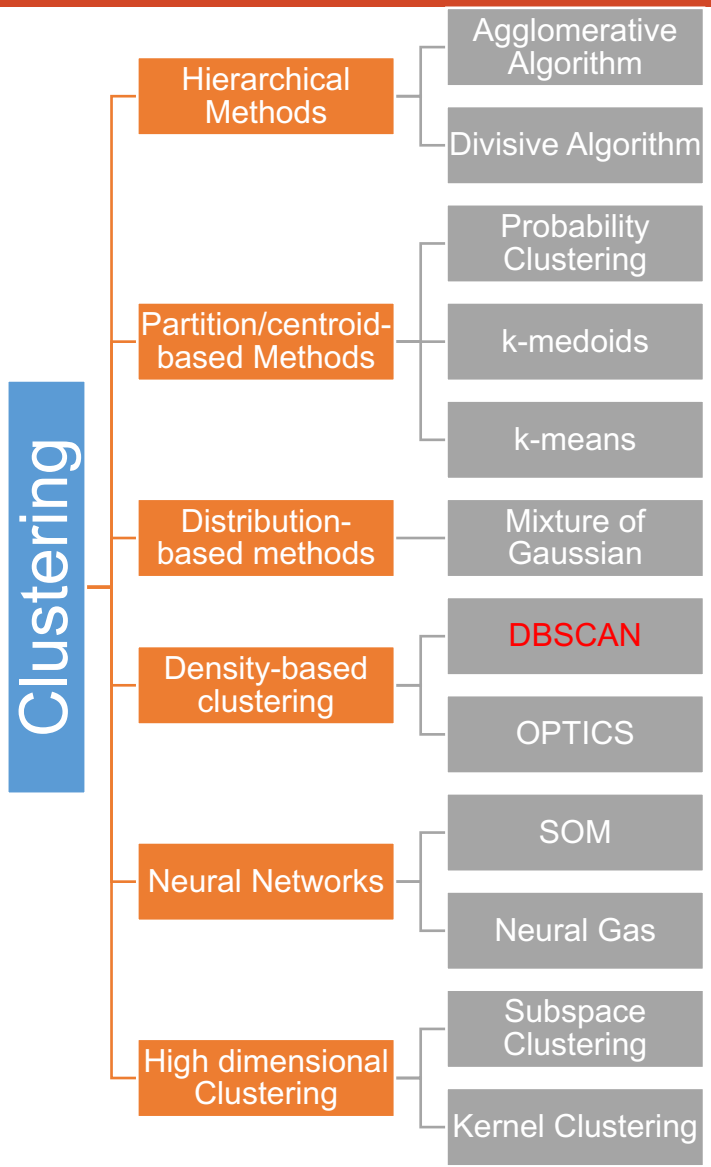
一．Clustering

二．Performance Metric

三．Hierarchical Clustering

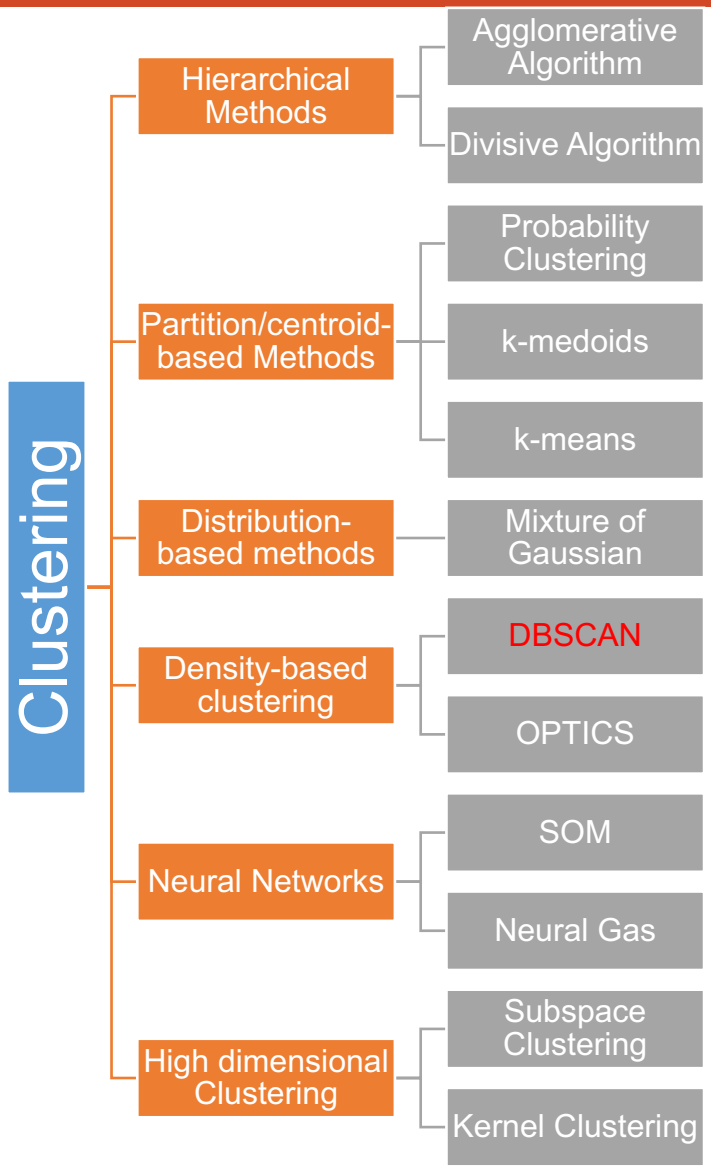
四．Density based Clustering

四、Density based Clustering



- Important Questions:
  - How do we measure density?
  - What is a dense region?

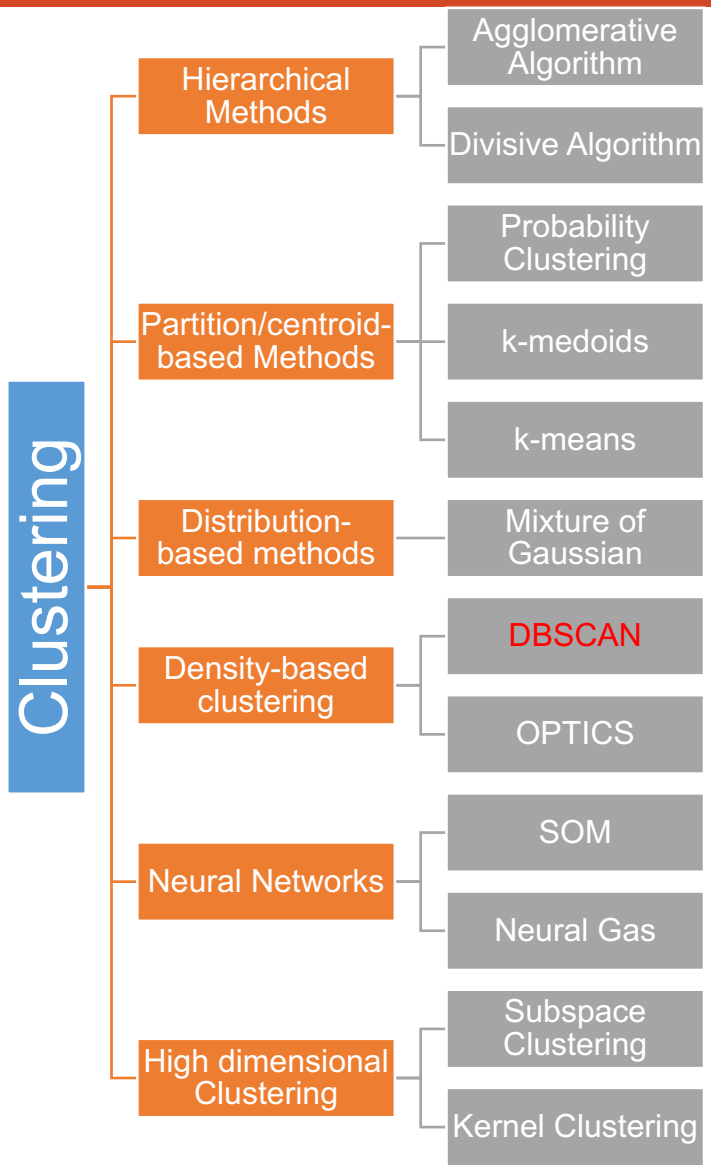
四、Density based Clustering



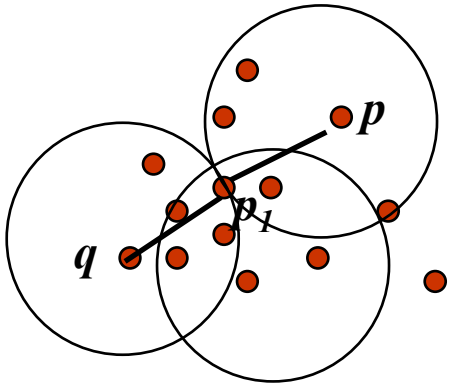
- Important Questions:
  - How do we measure density?
  - What is a dense region?

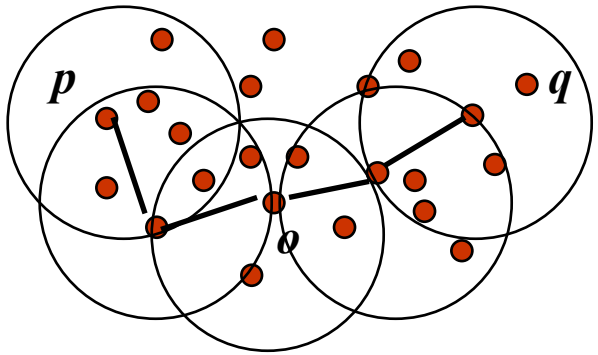
- Density at point p: number of points within a circle of radius Eps
- Dense Region: A circle of radius Eps that contains at least MinPts points

四、Density based Clustering

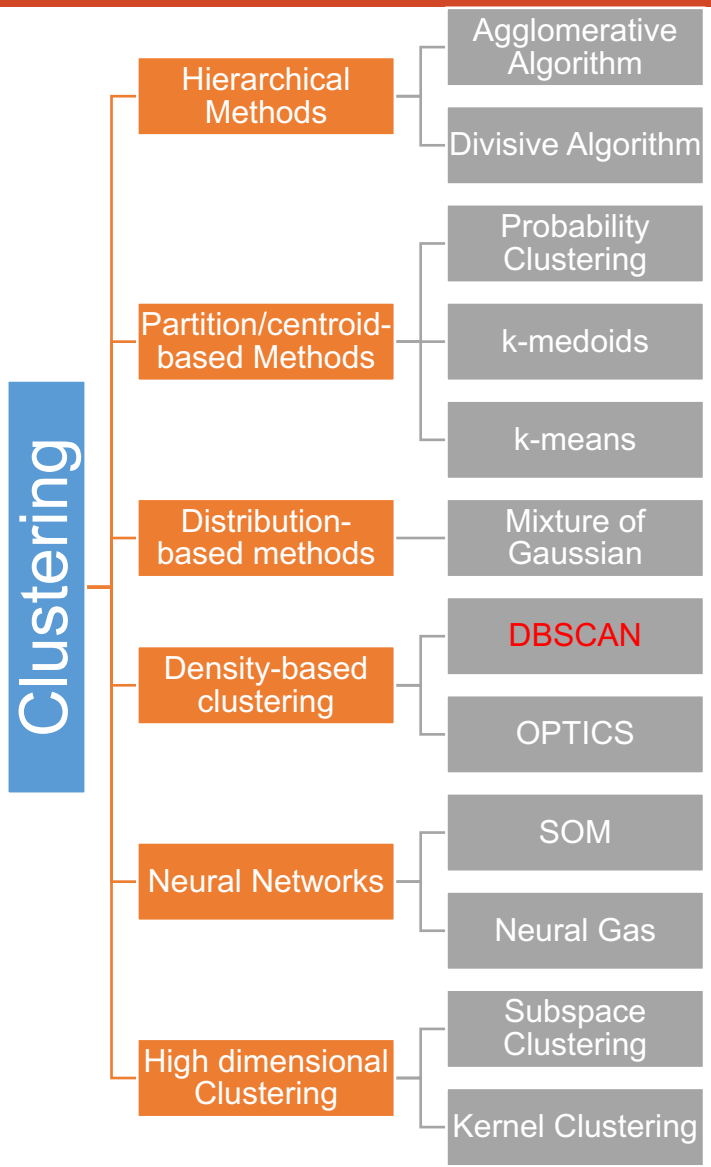


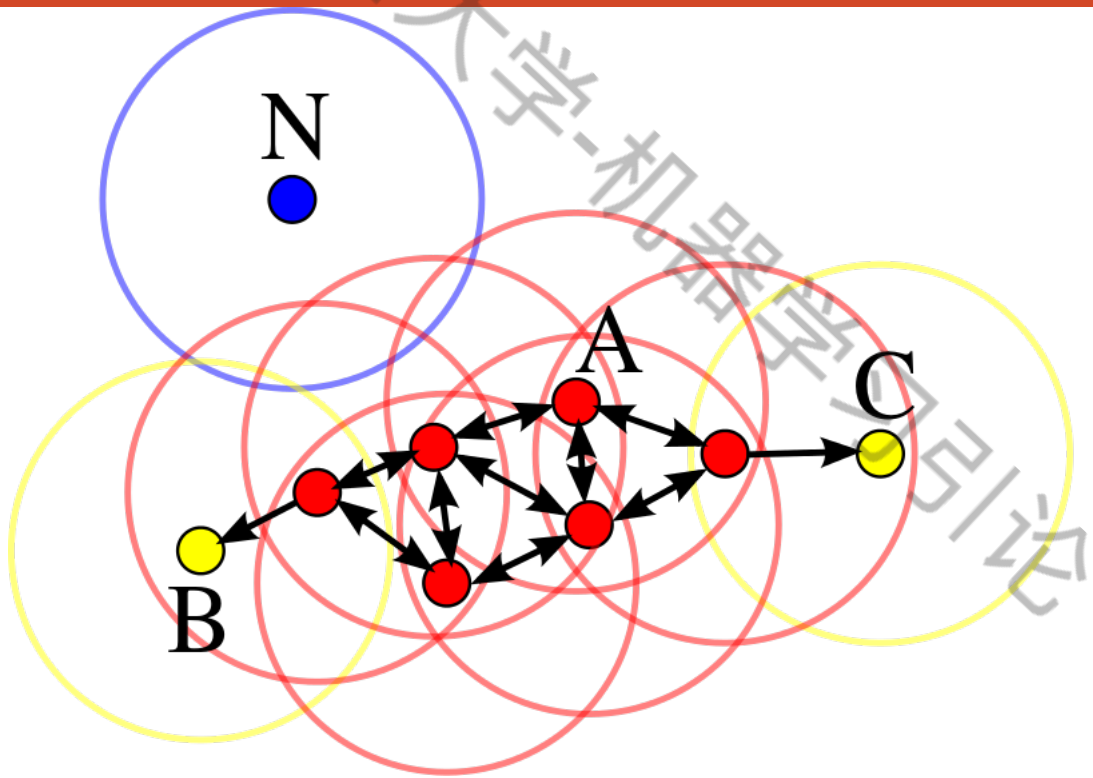
- Density edge
  - We place an edge between two core points q and p if they are within distance Eps.
- Density-connected
  - A point p is density-connected to a point q if there is a path of edges from p to q





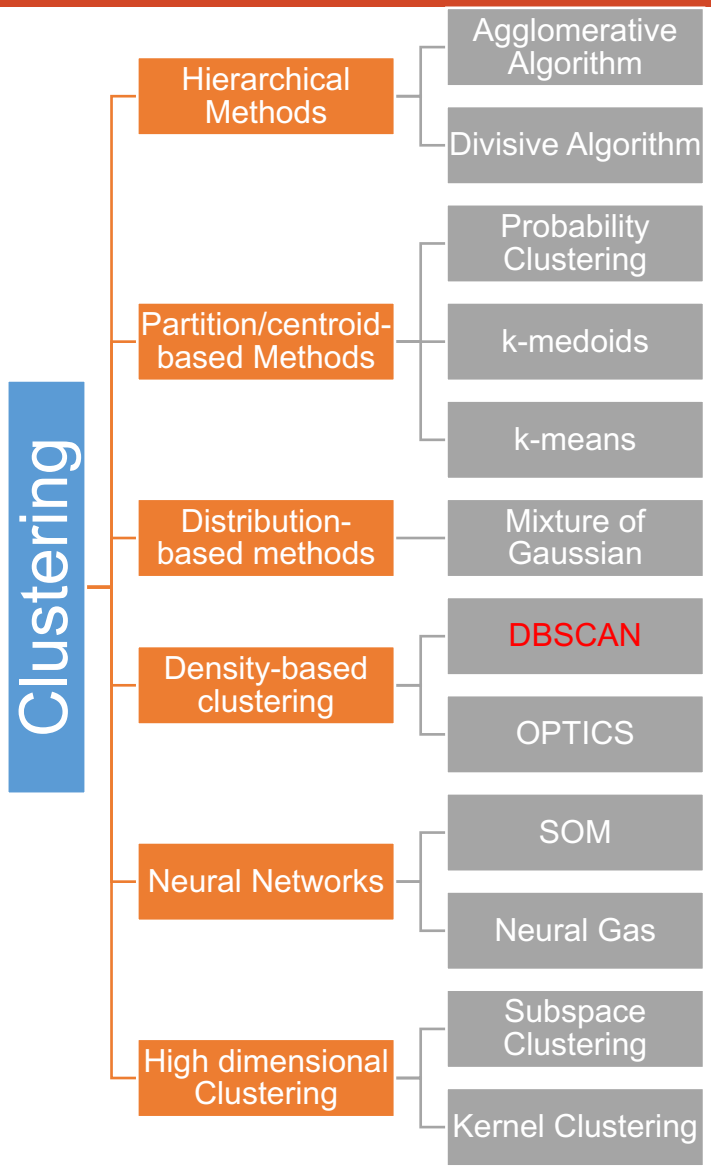
四、Density based Clustering





In this diagram, minPts = 4. Point A and the other red points are core points, because the area surrounding these points in an ε radius contain at least 4 points (including the point itself). Because they are all reachable from one another, they form a single cluster. Points B and C are not core points, but are reachable from A (via other core points) and thus belong to the cluster as well. Point N is a noise point that is neither a core point nor directly-reachable.

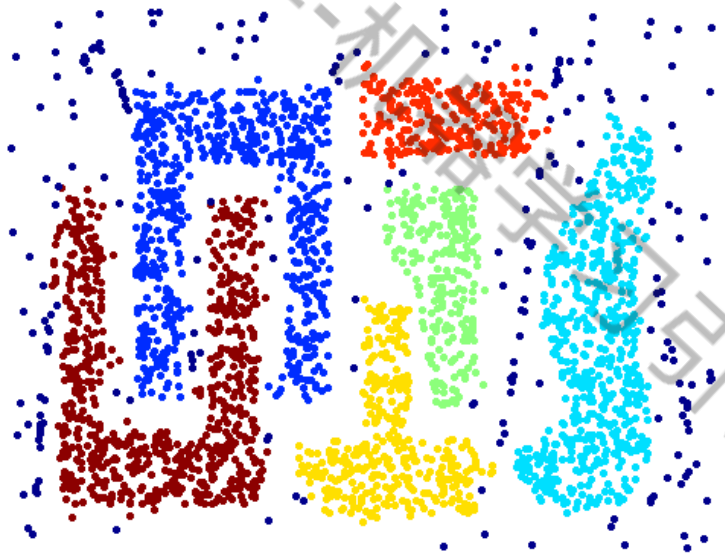
四、Density based Clustering



- Label points as core, border and noise
- Eliminate noise points
- For every core point p that has not been assigned to a cluster
  - Create a new cluster with the point p and all the points that are density-connected to p.
- Assign border points to the cluster of the closest core point.

四、Density based Clustering

Original Points



Clusters

- Resistant to Noise/outlier
- Can handle clusters of different shapes and sizes
- Do not need to input the cluster number

Test Questions

- What difference between classification and clustering?
- What is the key (most fundamental) of clustering method? why?
- What the similarity and dissimilarity between Hierarchical Clustering and Density based clustering?
- What advantages of Hierarchical Clustering and Density based clustering?
- What disadvantages of Hierarchical Clustering and Density based clustering?

Others~

Further reading:

Alex Rodriguez, Alessandro Laio, Clustering by fast search and find of density peaks, Science, 2014.

Q&A

THANKS!