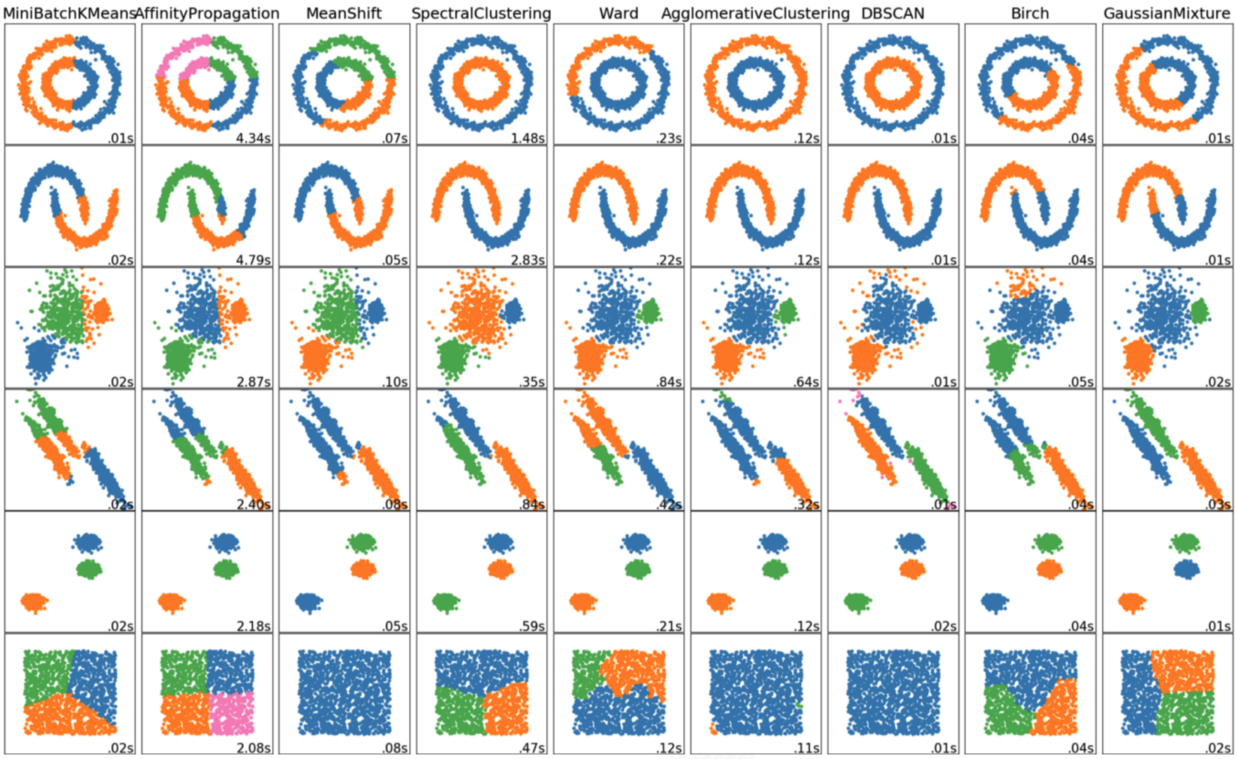
Clustering

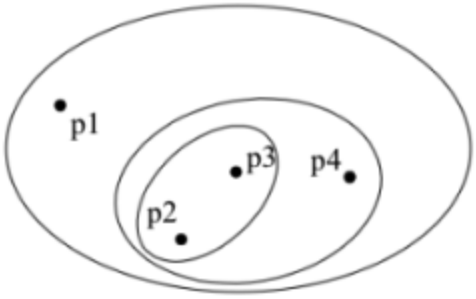
主要分Hierarchical（nested）和Partitional（overlapping）



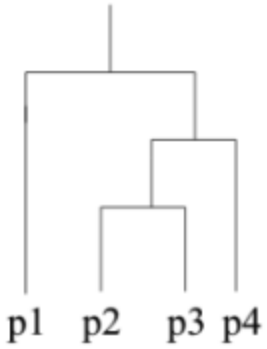
接下来主要focus在第二种

Hierarchical的就是：

Hierarchical Clustering



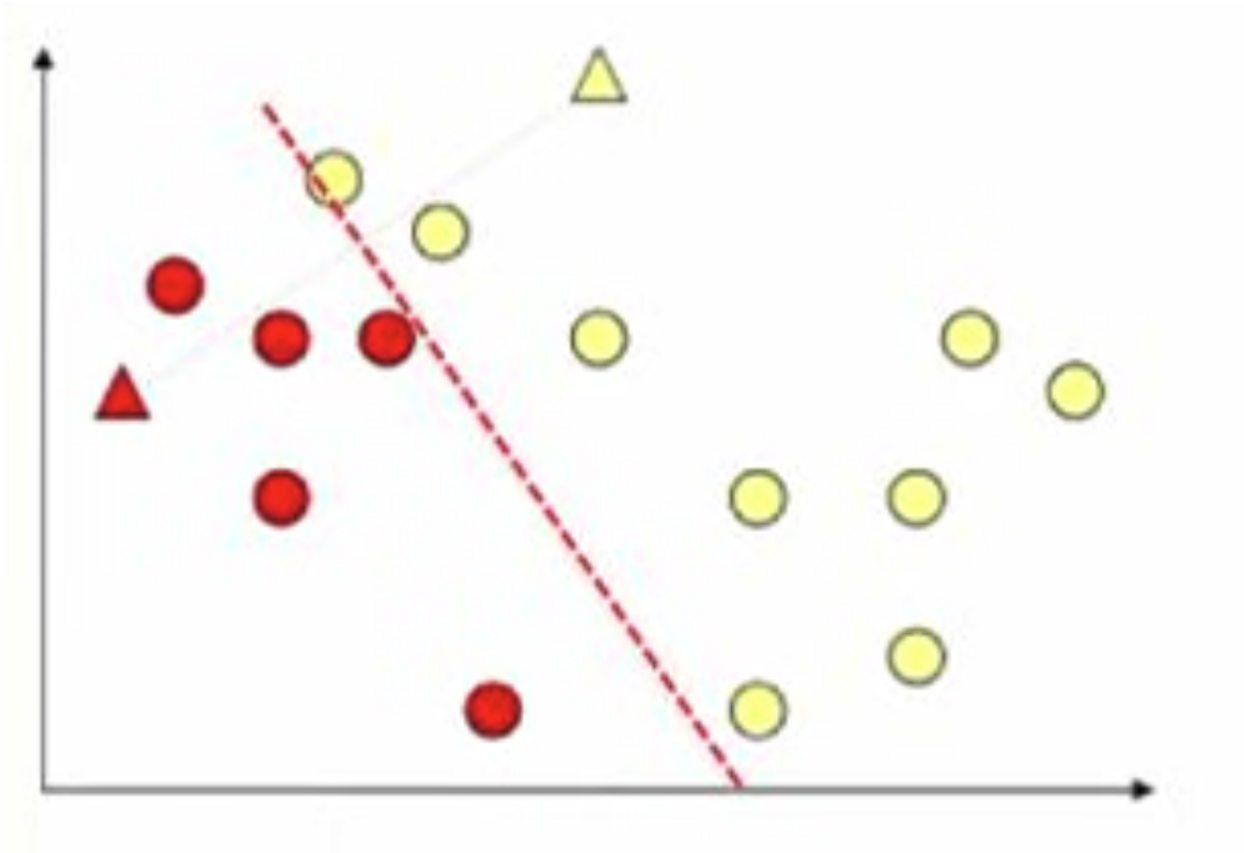
Hierarchical Clustering

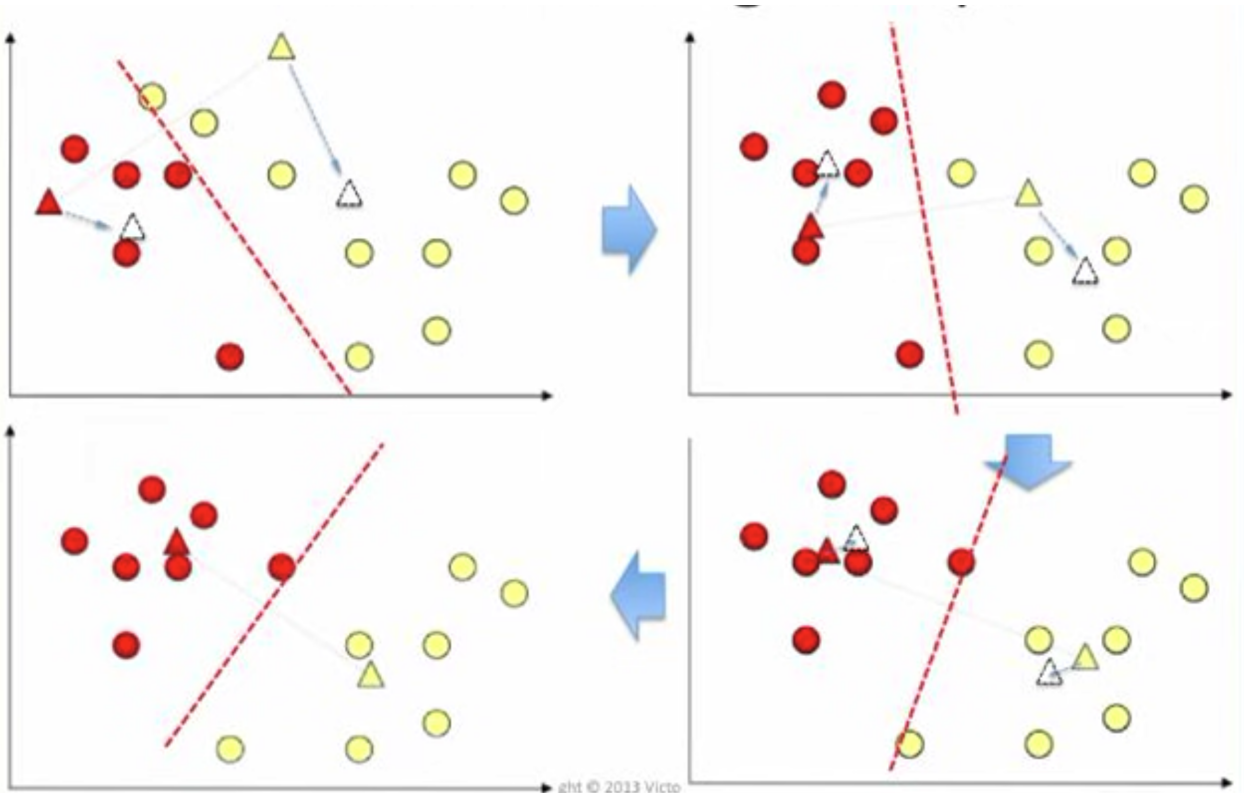


Dendrogram

K-Means

1. 定义K （group 数）
2. random center points, 然后垂直线劈开数据，归组
3. 所有组员的mean设定成新的center point，重新归，这个时候会有data换组
4. 直到没有data变换分组

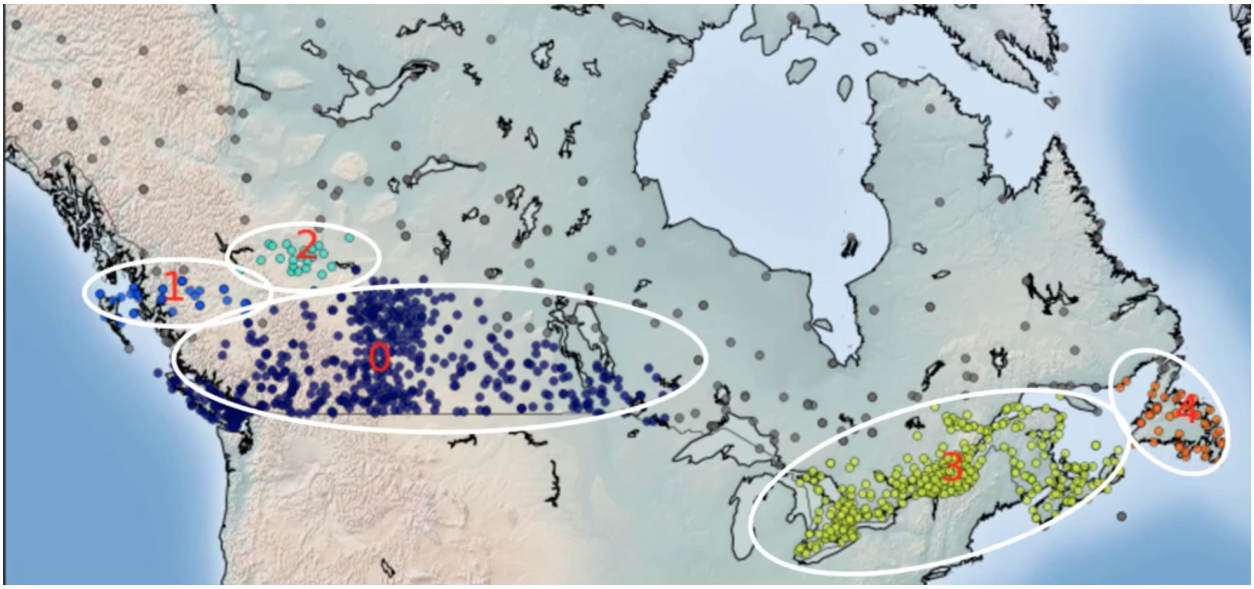



ght © 2013 Victo

- 优点:
  - 速度快
- 缺点：
  - 需要定义group的数量
  - 只能numerical ，不能categorical
  - 不能很好的deal noises，outliers，differing densities

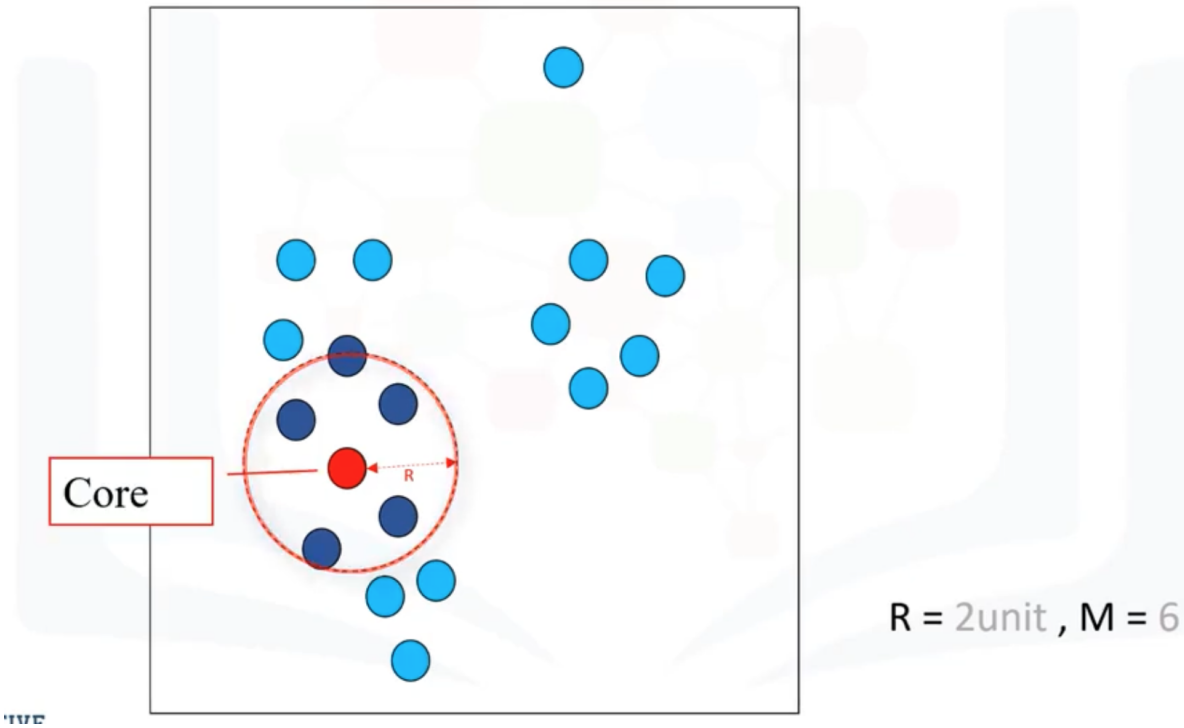
Density-Based Spatial Clustering of Applications with Noise (DBSCAN)

很适合搞map类的



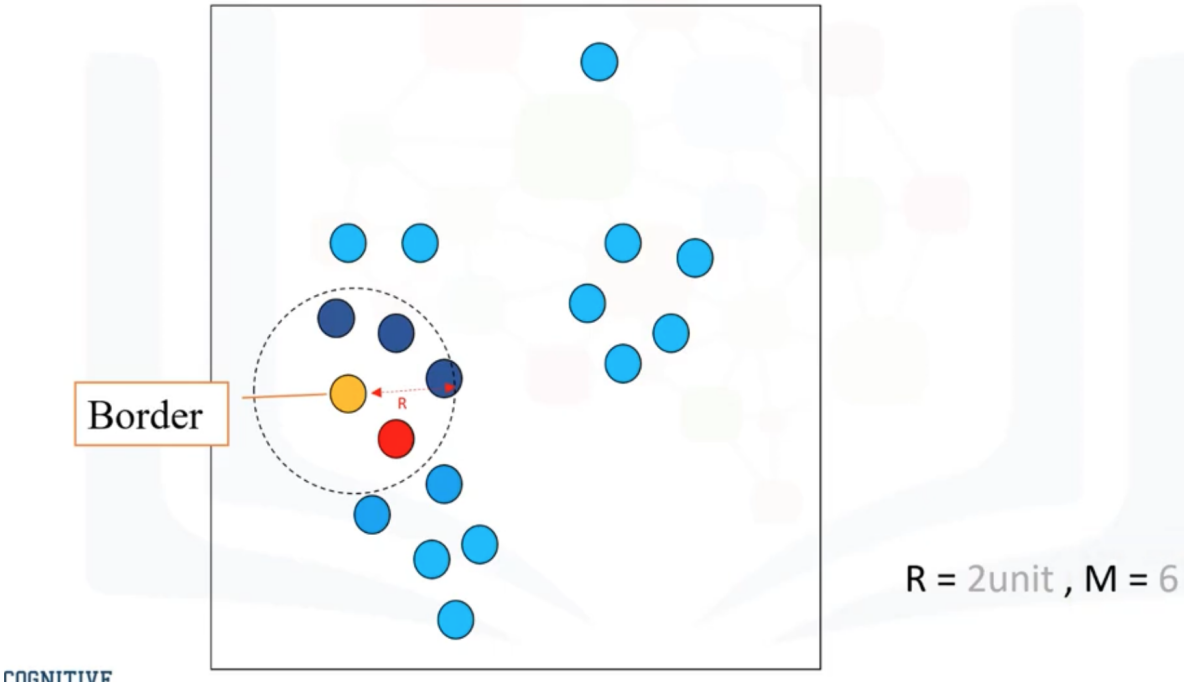
一共两个参数，R（半径）和M（最少point数）

Core point



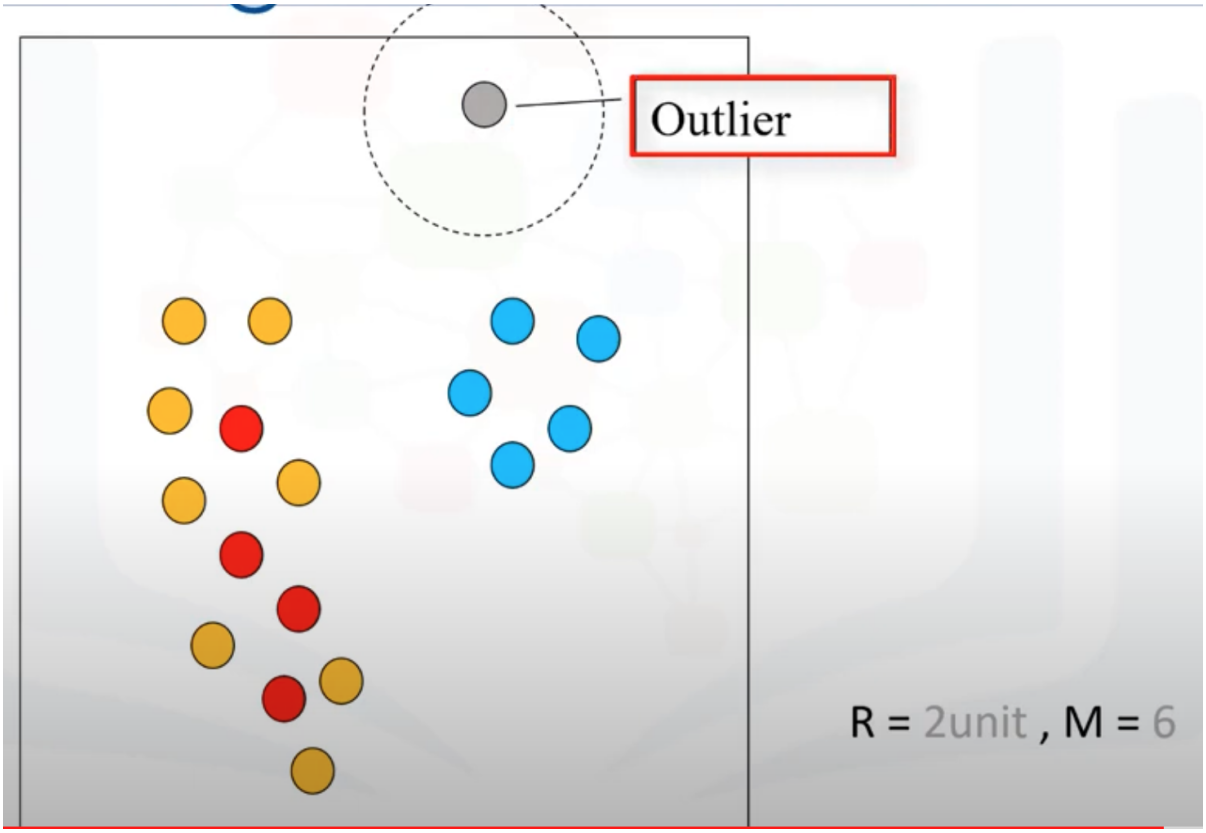
在R的范围内至少有M个点

Border point

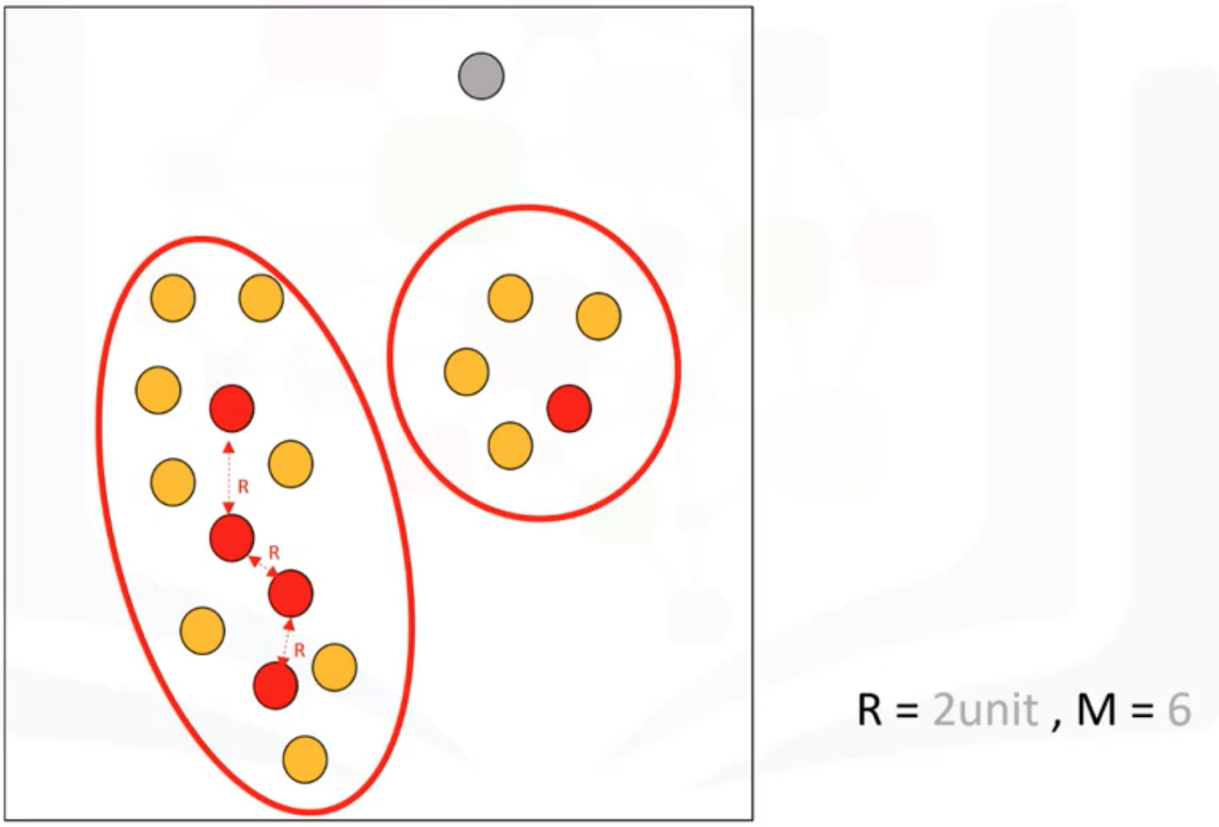


在R内不够M个点，但是能包含一个core point

outlier

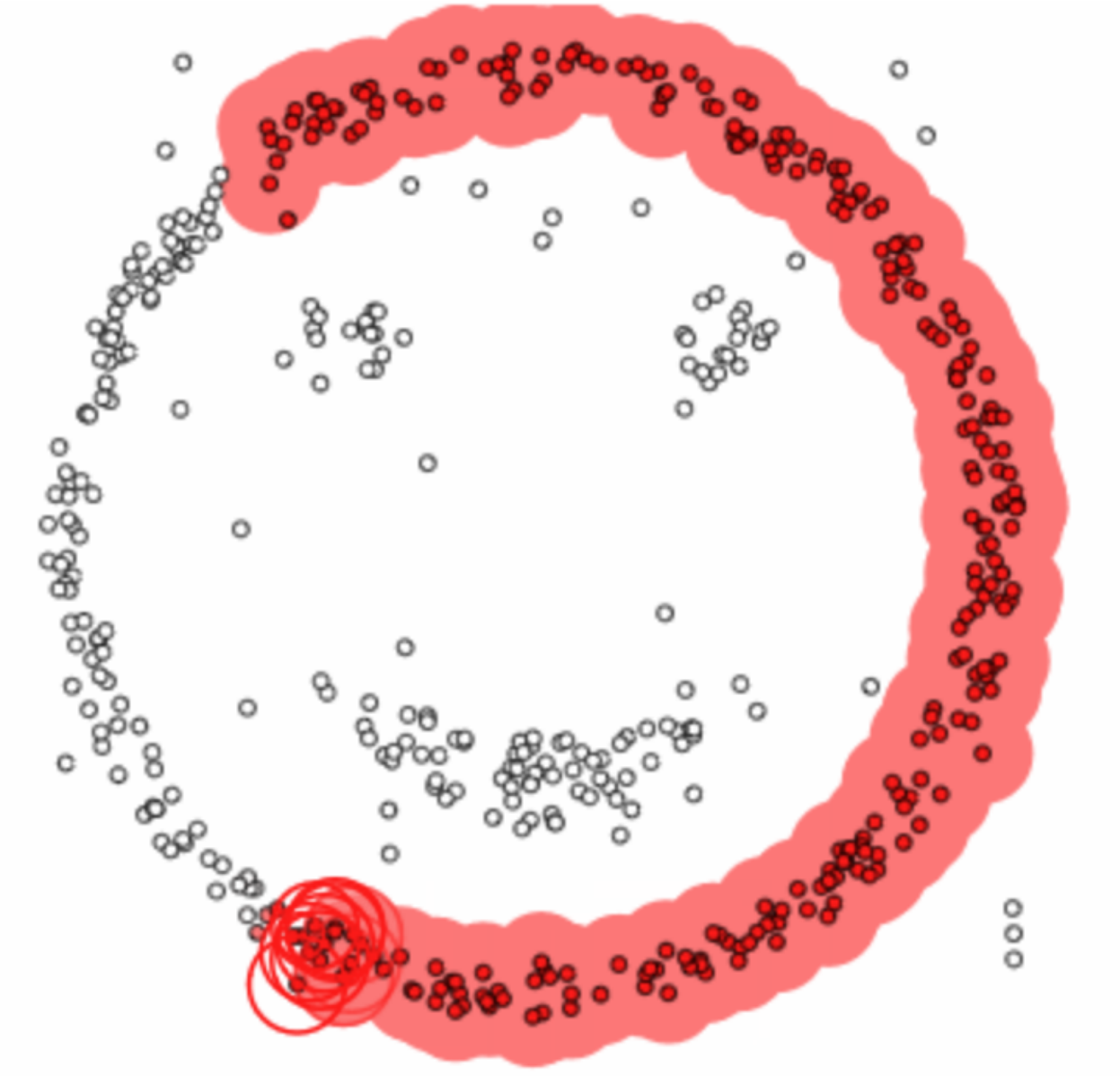


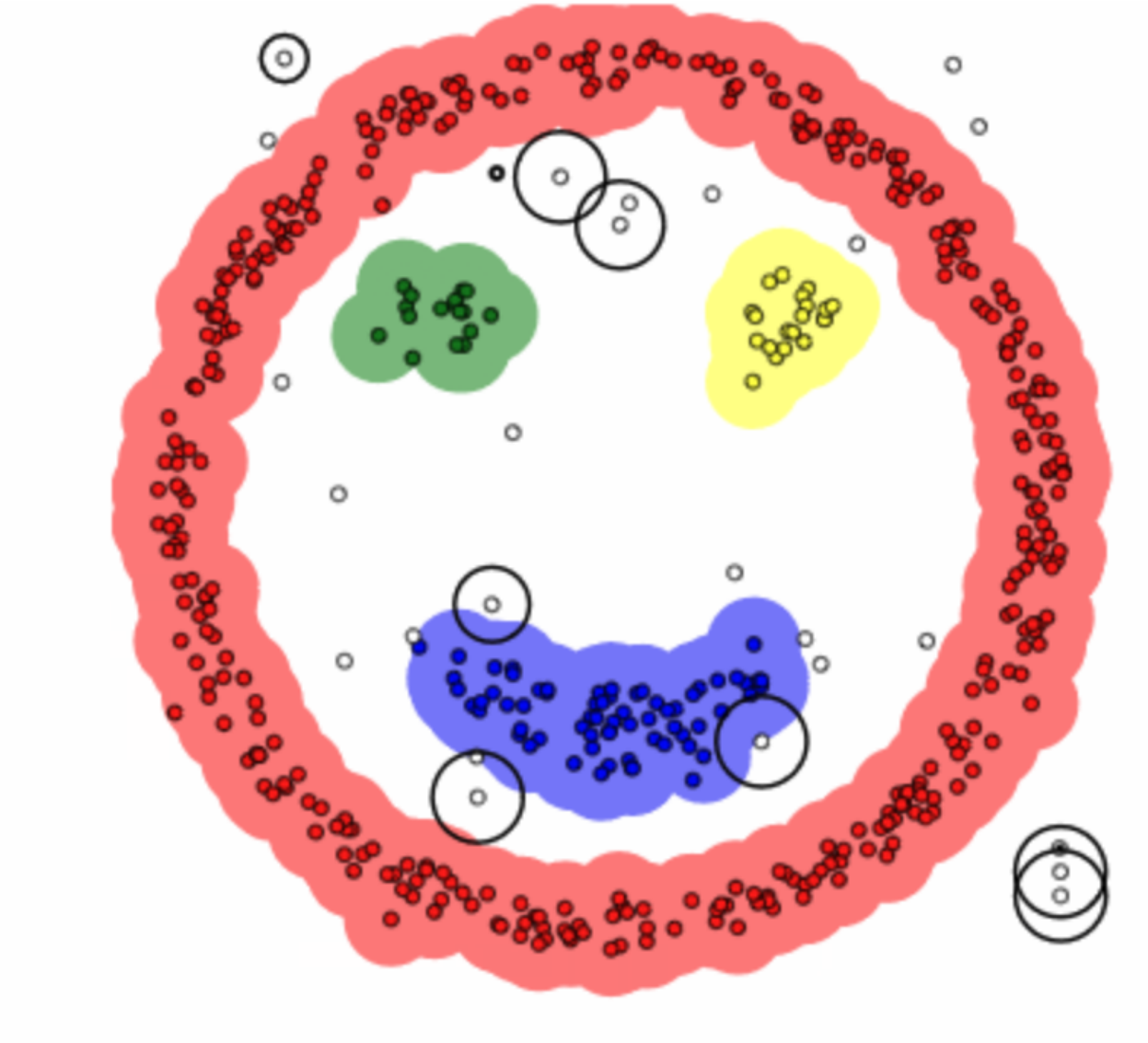
在R内不够M个点，也没有core point



分组按照1个 core point + reachable core + 他们各自的boarder

1. DBSCAN begins with an arbitrary starting data point that has not been visited. The neighborhood of this point is extracted using a distance epsilon ε (All points which are within the ε distance are neighborhood points).
2. If there are a sufficient number of points (according to minPoints) within this neighborhood then the clustering process starts and the current data point becomes the first point in the new cluster. Otherwise, the point will be labeled as noise (later this noisy point might become the part of the cluster). In both cases that point is marked as “visited”.
3. For this first point in the new cluster, the points within its ε distance neighborhood also become part of the same cluster. This procedure of making all points in the ε neighborhood belong to the same cluster is then repeated for all of the new points that have been just added to the cluster group.

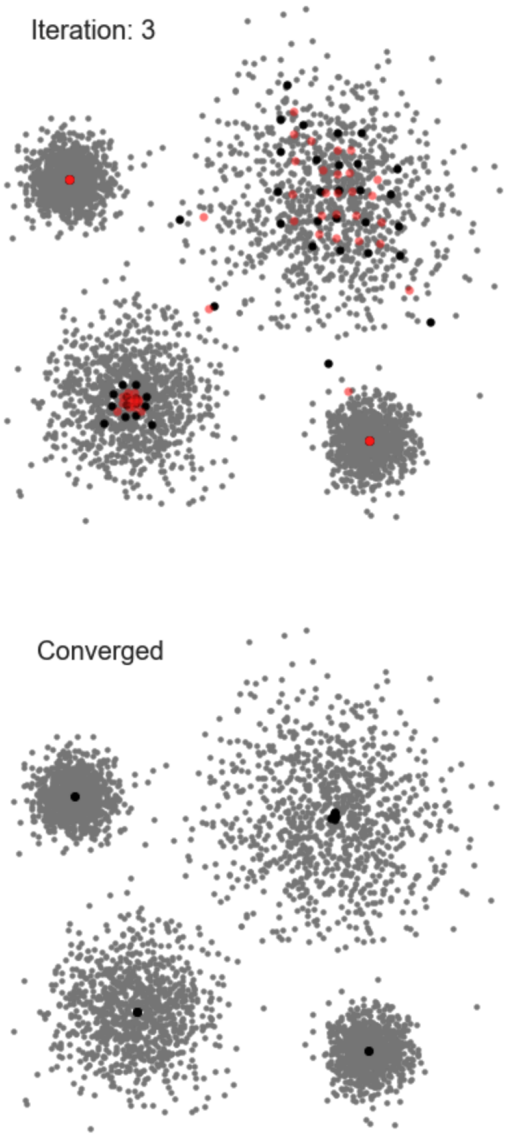


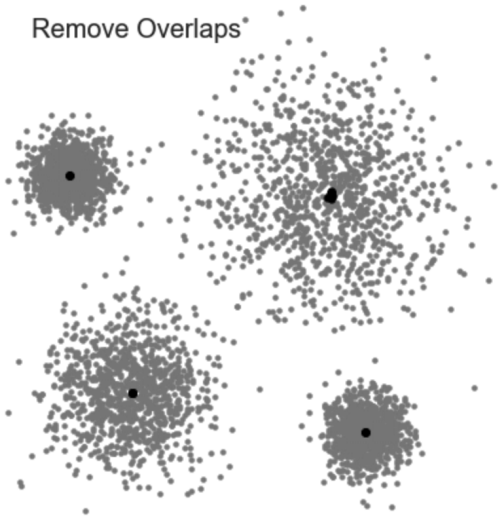


- 优点
  - 不需要pre set cluster #
  - identifies outliers as noises （mean-shift which simply throws them into a cluster even if the data point is very different.）
  - it can find arbitrarily sized and arbitrarily shaped clusters quite well.
- 缺点
  - it doesn’t perform as well as others when the clusters are of varying density不同密度.

Mean-Shift Clustering

1. Mean shift clustering is a sliding-window-based algorithm that attempts to find dense areas of data points.
2. the goal is to locate the center points of each group/class, which works by updating candidates for center points to be the mean of the points within the sliding-window.

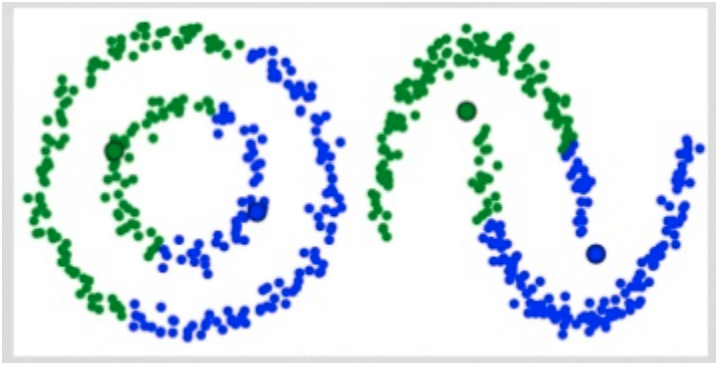




- 优点：
  - 不需要预设group数量，mean-shift automatically discovers this
- 缺点：
  - the selection of the window size/radius “r” can be non-trivial.

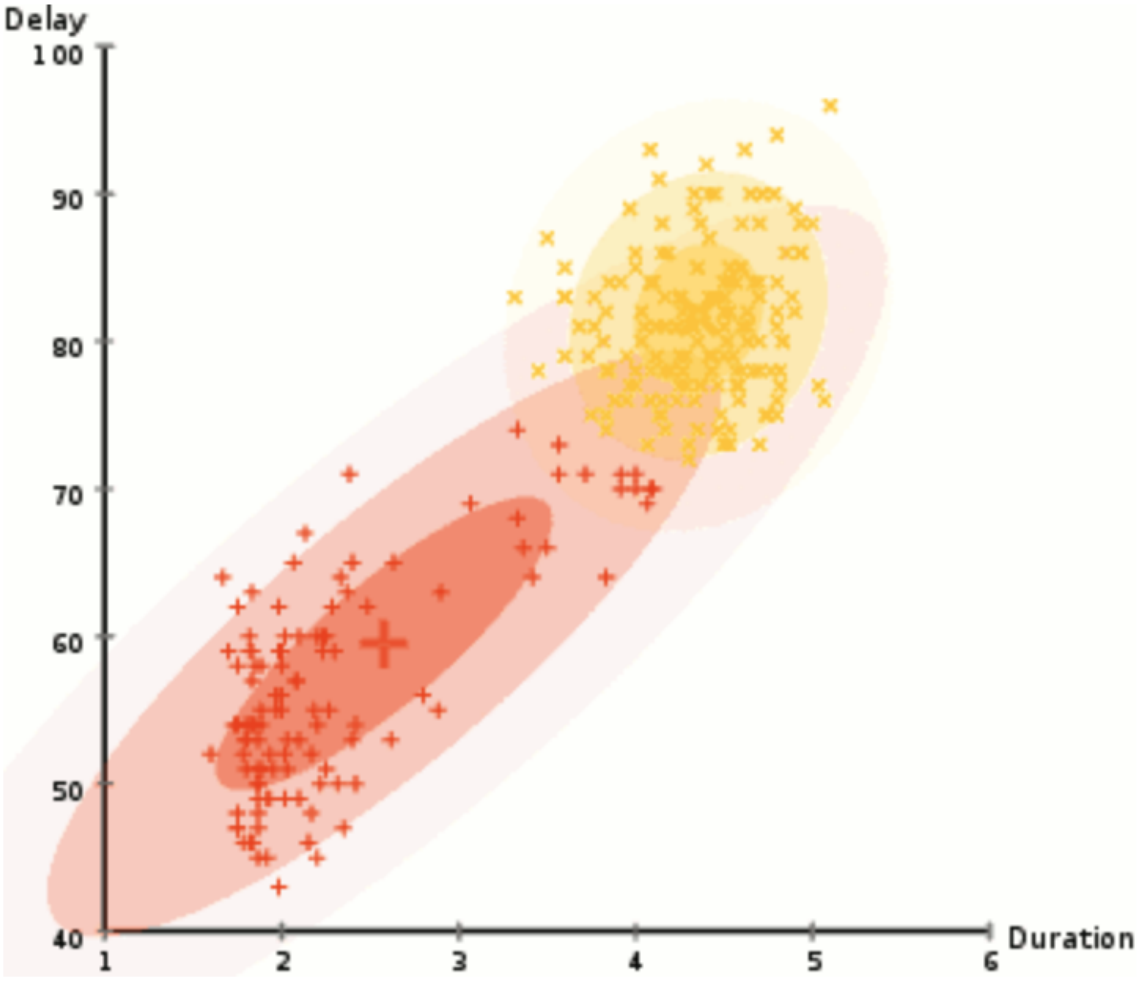
Gaussian Mixture Models

K-mean的升级版？ K-mean 一个大问题就是只单纯的用mean去分组，当数据有线性规律的时候就完蛋，比如

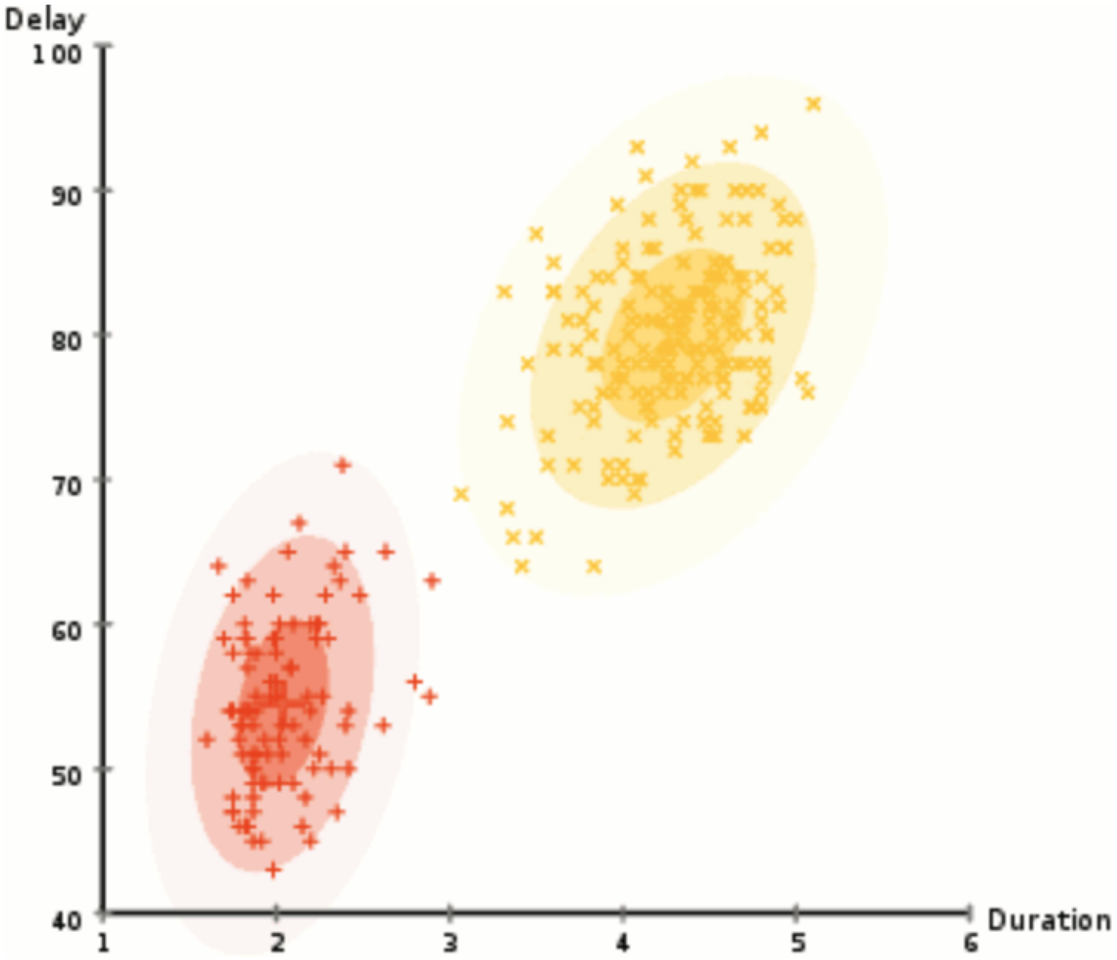


Two failure cases for K-Means

GMM用的是**mean** 和 **standard deviation** 作为parameters



EM Clustering using GMMs



- 优点
  - more **flexible** in terms of **cluster covariance** than K-means
  - Since GMMs use probabilities, they can have multiple clusters per data point. **GMM支持多重membership （可以有overlap！）**
- 缺点
  - 需要设定number of clusters