6222005 Học máy

Bài giảng: Ví dụ về bài toán phân cụm

Chương 3: Phân cụm

VTPham@IUH, 2024

Ôn lại bài học trước

- Bạn có nhớ ? % ?

Nội dung chính

- 3.1 Khái niệm về phân cụm
- 3.2 Mô tả bài toán phân cụm
- 3.3 Hàm mục tiêu
- 3.4 K-Means
- 3.5 Ví dụ về bài toán phân cụm

Một số thuật toán phân cụm khác

- DBSCAN
- Gaussian mixture model
- Hierarchical clustering

DBSCAN

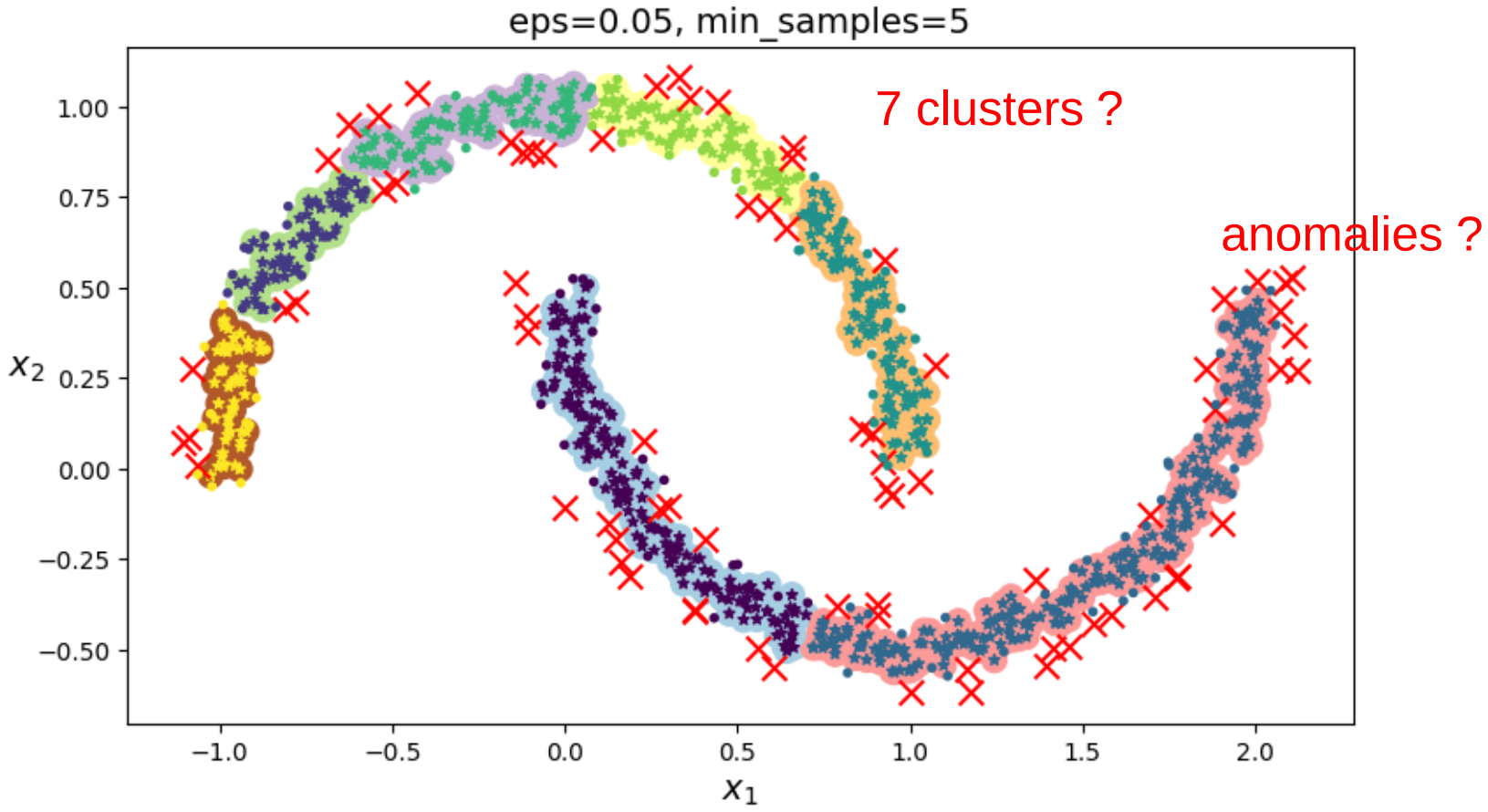
- DBSCAN (Density-based spatial clustering of applications with noise)
  - Thuật toán này định nghĩa các cụm như là các vùng liên tục với mật độ cao
  - Thuật toán này hoạt động tốt nếu tất cả các cụm đủ dày đặc và nếu chúng được phân tách tốt bởi các vùng có mật độ thấp

DBSCAN

- Các bước thực hiện
  - Đối với mỗi mẫu, đếm số lượng mẫu nằm trong khoảng cách nhỏ ε từ mẫu đó. Vùng này được gọi là vùng lân cận ε của mẫu
  - Nếu một mẫu có ít nhất min_samples mẫu trong vùng lân cận ε của nó (bao gồm cả chính nó), thì mẫu đó được coi là một mẫu lõi (core instance). Nói cách khác, các mẫu lõi là những mẫu nằm trong các vùng “dày đặc“.
  - Tất cả các mẫu trong vùng lân cận của một mẫu lõi đều thuộc cùng một cụm. Vùng lân cận này có thể bao gồm các mẫu lõi khác. Do đó, một chuỗi dài các mẫu lõi lân cận tạo thành một cụm duy nhất.
  - Bất kỳ mẫu nào không phải là mẫu lõi và không có mẫu nào trong vùng lân cận của nó đều được coi là một bất thường.

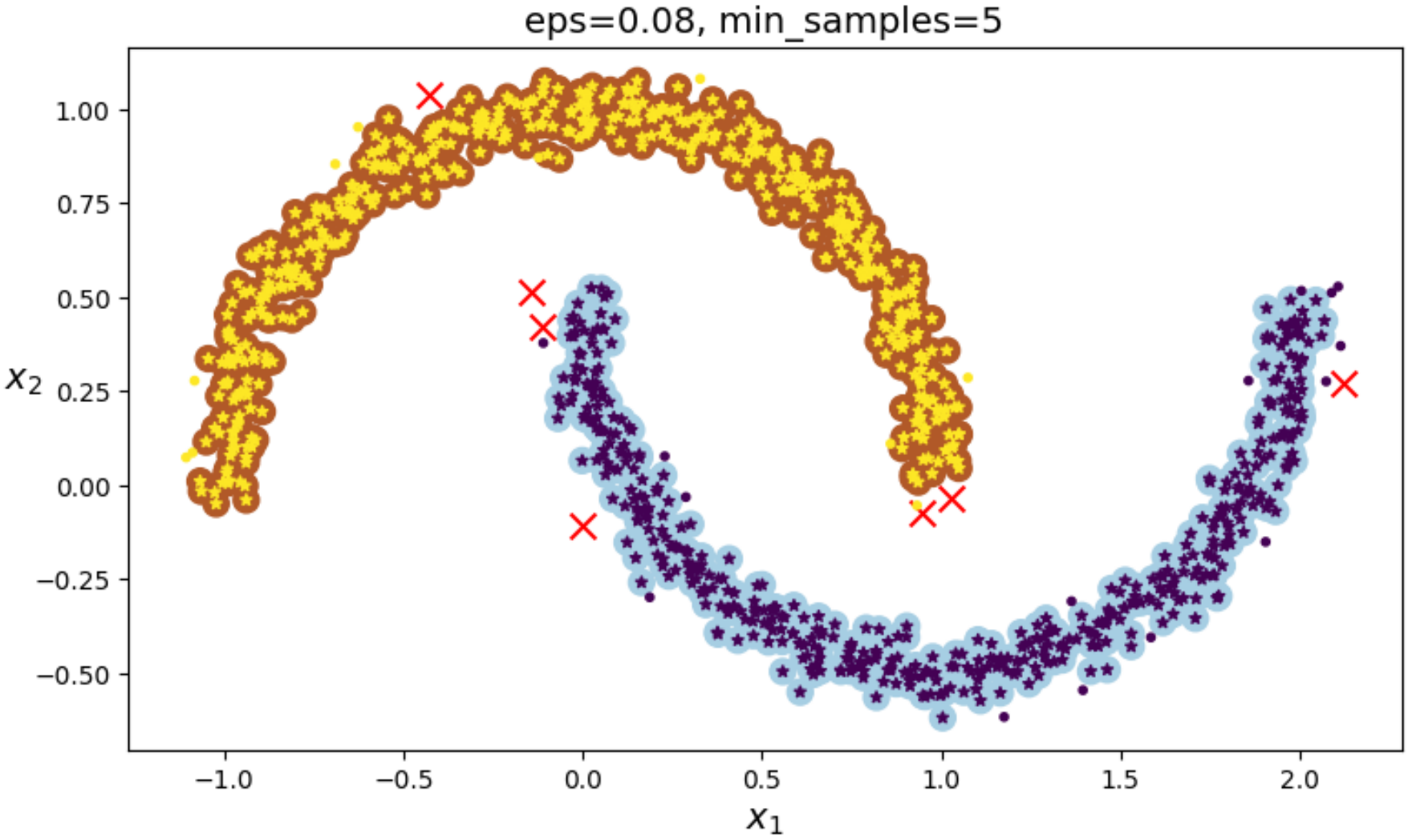
DBSCAN

- Trong sklearn
  - DBSCAN
  - Chỉ có 2 hệ số là: eps và min_samples



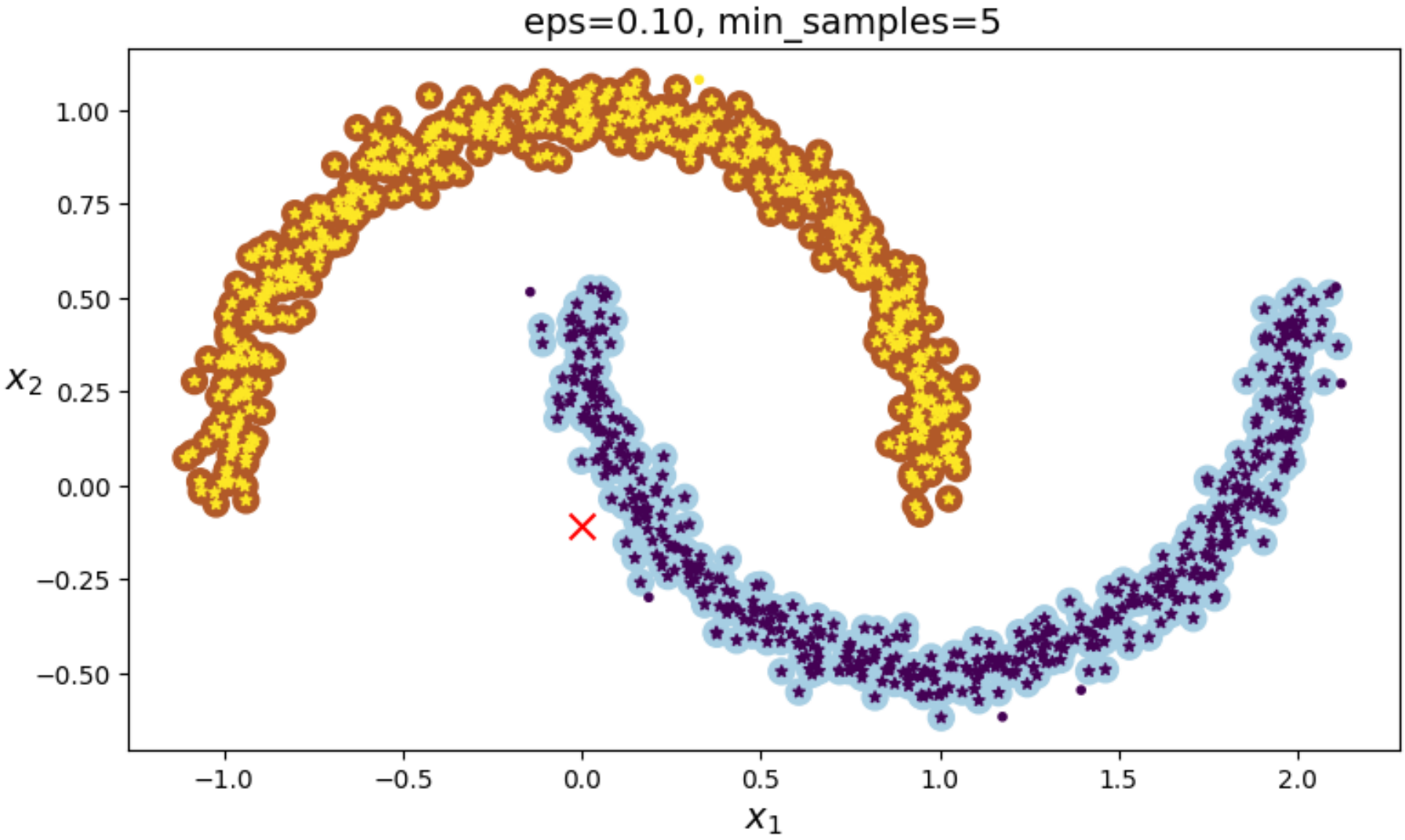
DBSCAN

- eps=0.08



DBSCAN

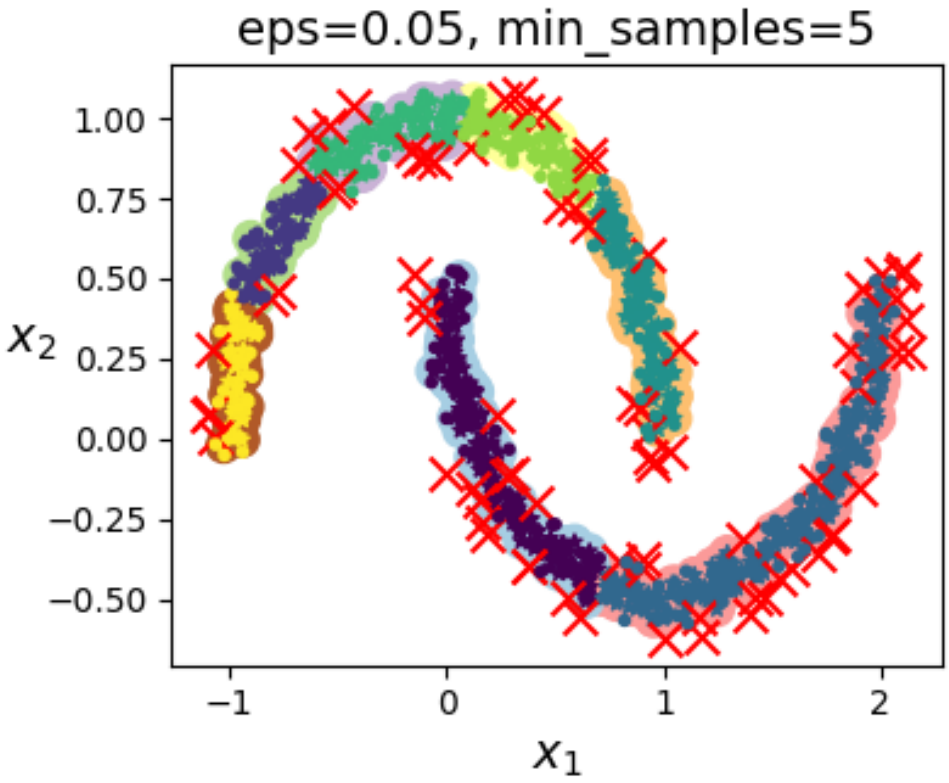
- eps=0.10

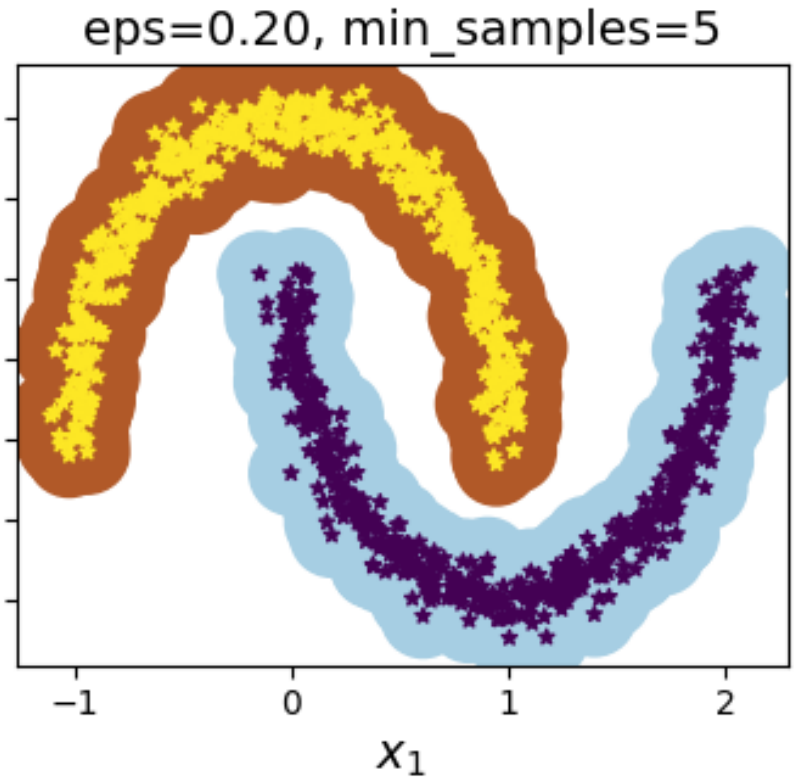


DBSCAN

- Trong sklearn
  - DBSCAN
  - Chỉ có 2 hệ số là: eps và min_samples

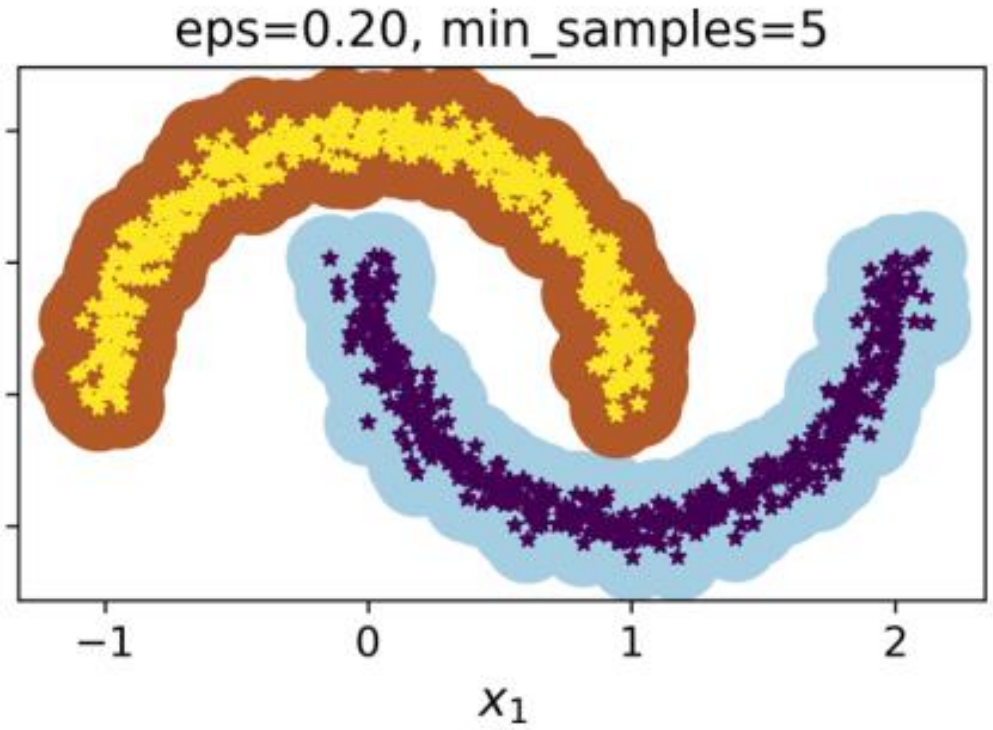
2 clusters 0 anomaly





DBSCAN

- Thuật toán đơn giản nhưng mạnh mẽ, cho phép phân cụm cho bất kỳ hình dạng nào



Gaussian mixture

- GMM (Gaussian mixture model)

  - Là mô hình xác suất với giả sử rằng các mẫu được tạo ra từ sự pha trộn của một số phân phối Gaussian (có các tham số ta không biết)

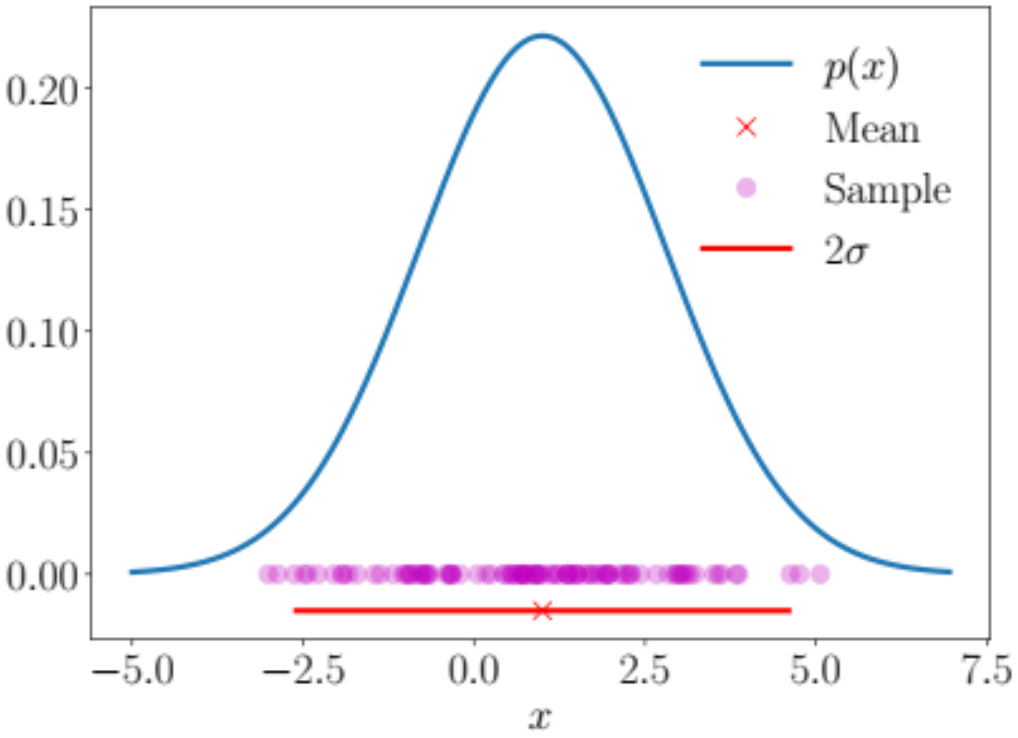
  - Tất cả các mẫu tạo bởi cùng 1 phân phối Gaussian đơn hợp thành 1 cụm, thông thường có dạng elip

Gaussian distribution

- Univariate (one-dimensional) Gaussian distribution

$$\mathcal{N}(x|\mu,\sigma^2) = \frac{1}{(2\pi\sigma^2)^{1/2}} \exp\left\{-\frac{1}{2\sigma^2}(x-\mu)^2\right\}$$

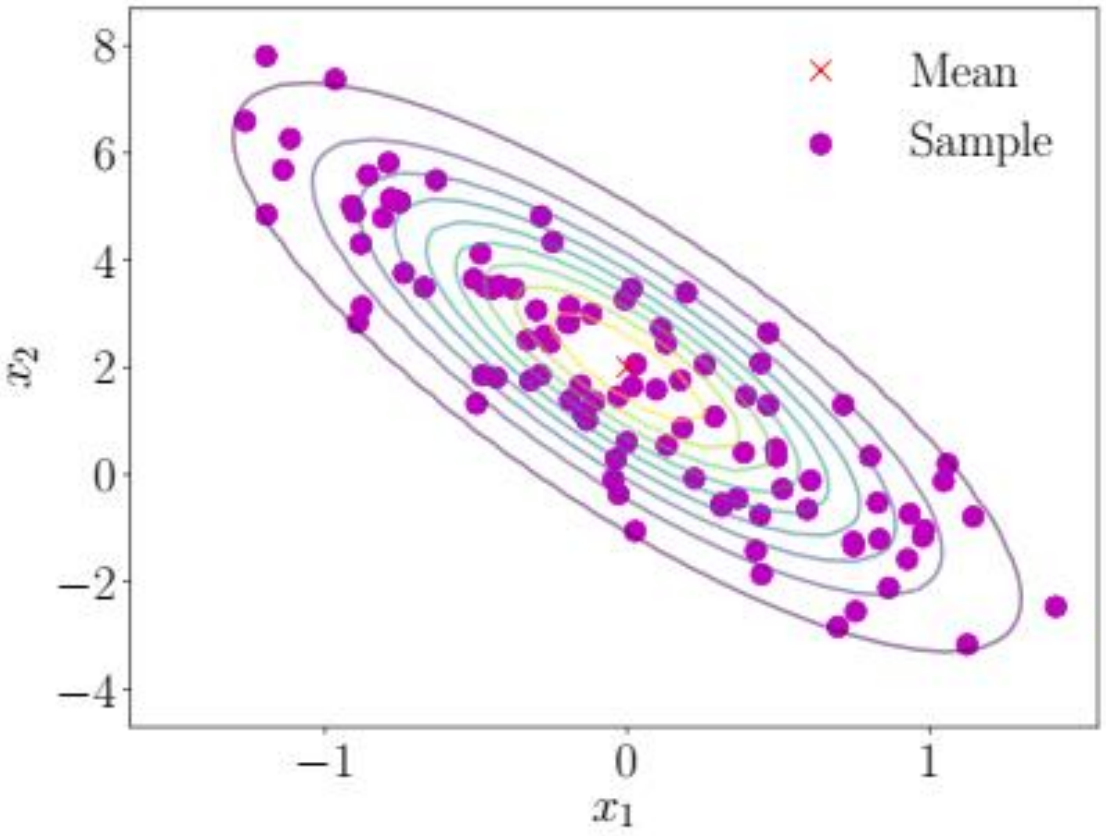
μ : kỳ vọng (mean) σ^2 : phương sai (variance)



Gaussian distribution

- Multivariate Gaussian distribution

$$\mathcal{N}(\mathbf{x}|\boldsymbol{\mu},\boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{D/2}}\frac{1}{|\boldsymbol{\Sigma}|^{1/2}}\exp\left\{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right\}$$



$\boldsymbol{\mu}$: vector kỳ vọng (mean vector) $\boldsymbol{\mu} \in \mathbb{R}^D$

$\boldsymbol{\Sigma}$: ma trận hiệp phương sai (covariance matrix)

Biến ngẫu nhiên *D* chiều

Mixtures of Gaussians

- Gaussian mixture distribution

$$p(\mathbf{x}) = \sum_{k=1}^{K} \pi_k \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$$

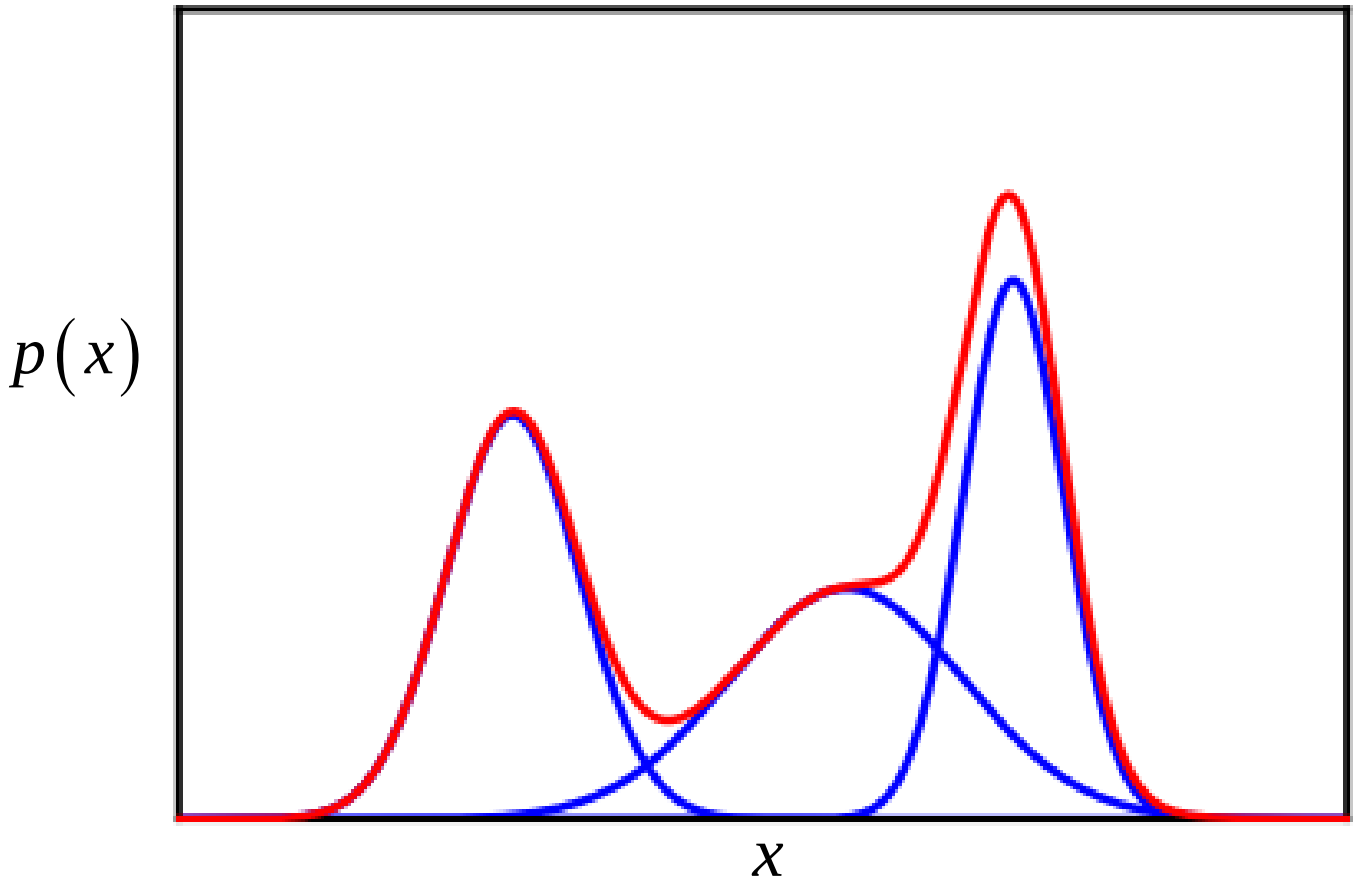
π_k : mixing coefficients

$$0 \leqslant \pi_k \leqslant 1 \qquad \sum_{k=1}^{K} \pi_k = 1$$

$\mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$: components

Mixtures of Gaussians

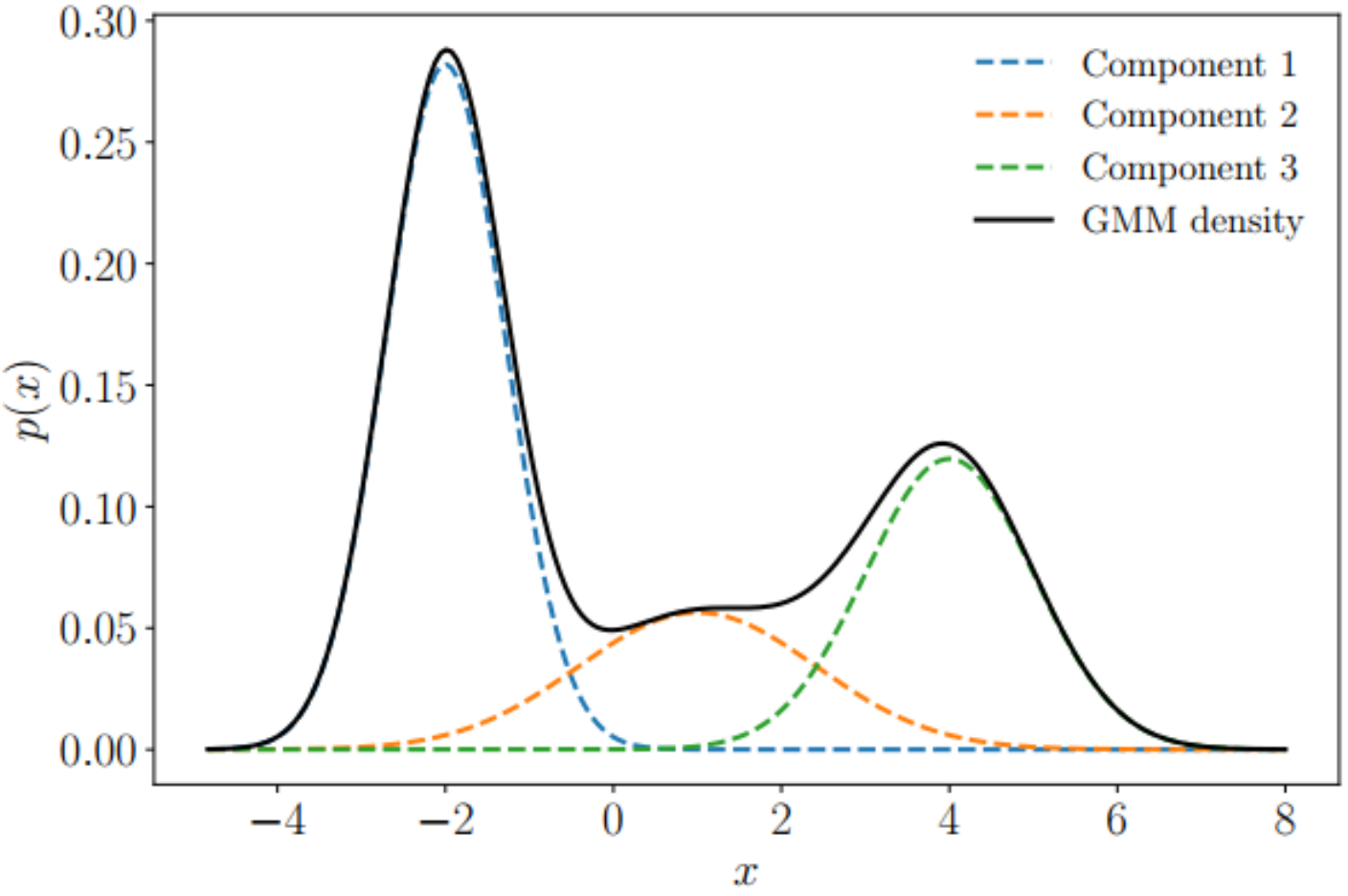
- Ví dụ: về phân phối hỗn hợp Gaussian (màu đỏ) hiển thị ba phân phối Gaussian một chiều (màu xanh nước biển)



Mixtures of Gaussians

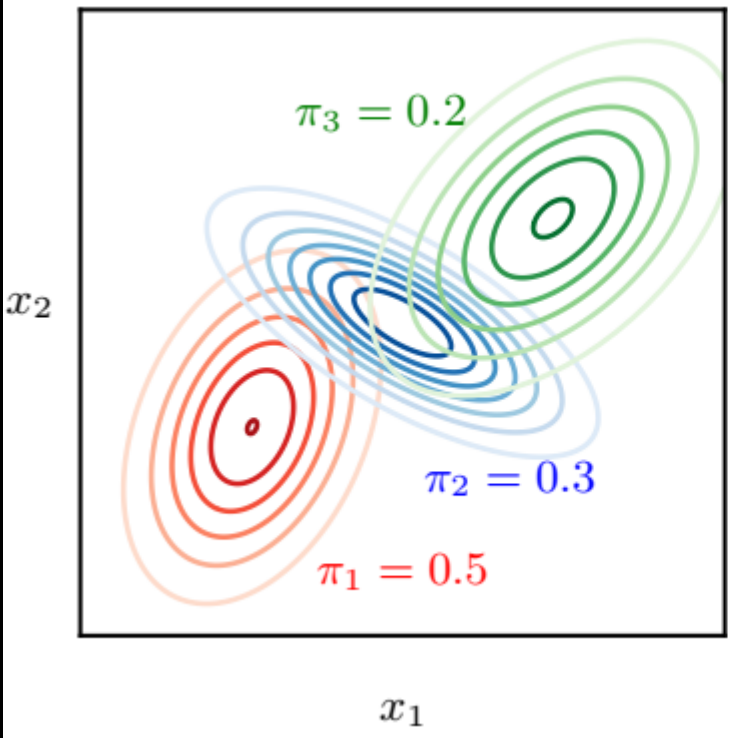
- Ví dụ: về phân phối hỗn hợp Gaussian

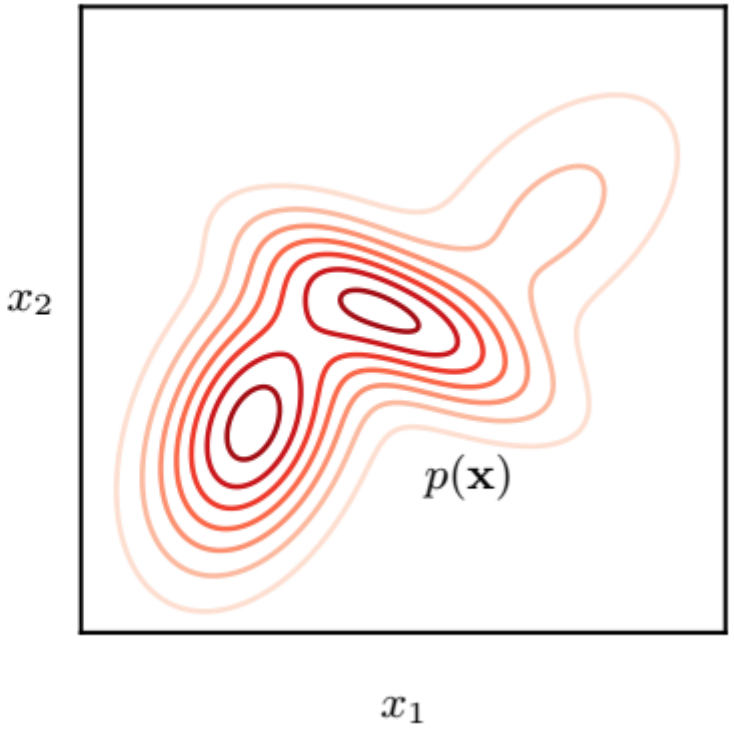
$$0.5\mathcal{N}\left(x \mid -2, \tfrac{1}{2}\right) + 0.2\mathcal{N}\left(x \mid 1, 2\right) + 0.3\mathcal{N}\left(x \mid 4, 1\right)$$

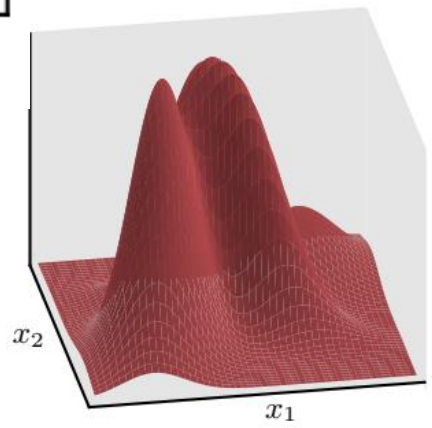


Mixtures of Gaussians

- Ví dụ: 3 Gaussian trong không gian hai chiều

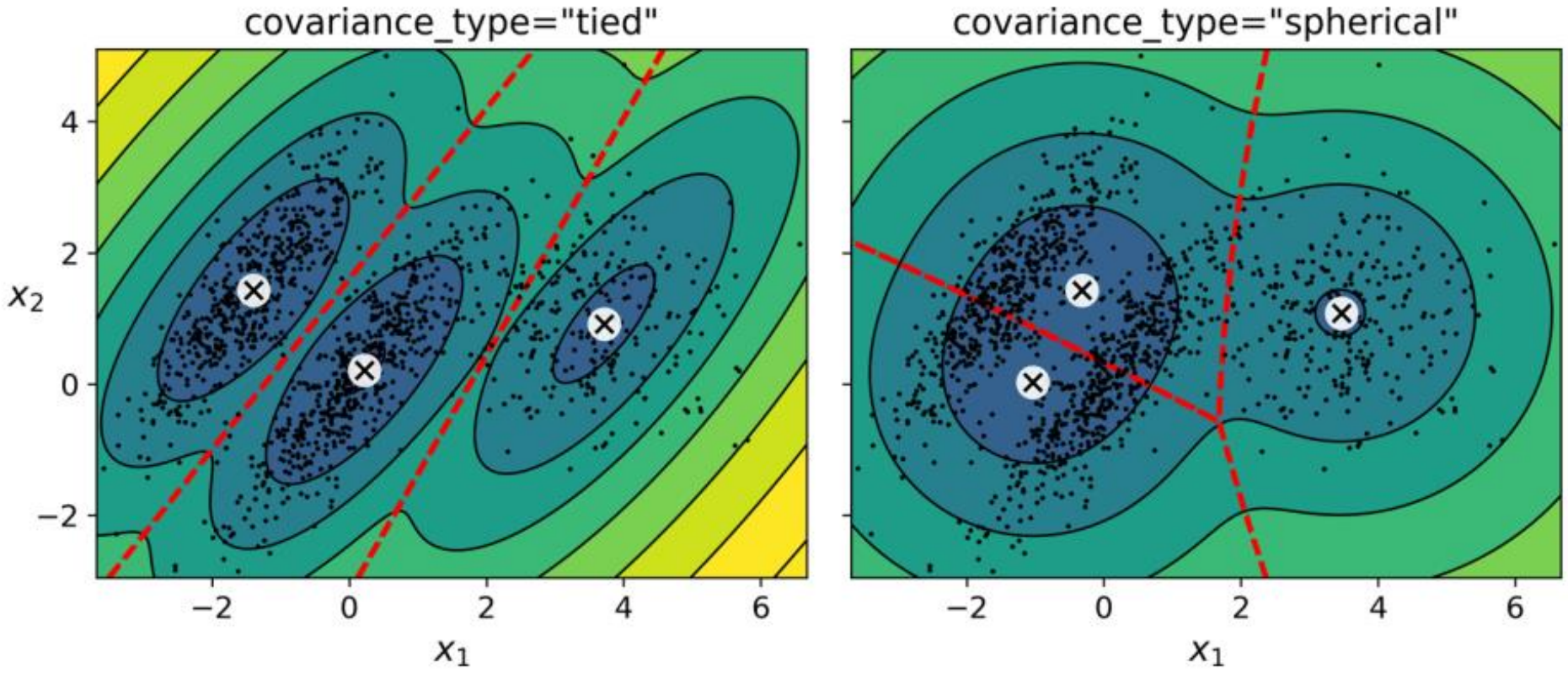






Gaussian mixture

- sklearn: GaussianMixture
  - Chú ý lựa chọn cho hệ số covariance_type: spherical, diag, tied



Hierarchical Clustering [B11TLTK1]

- Phân cụm phân cấp
  - Không yêu cầu phải lựa chọn số lượng cụm *k* cụ thể
  - Tạo ra một biểu diễn quan sát dạng cây trực quan, được gọi là dendrogram hay sơ đồ phân cấp/nhánh

Bottom-up or agglomerative clustering

- Phương pháp phân cụm phân cấp tổng hợp
  - Bắt đầu từ dưới cùng của biểu đồ phân nhánh
  - Mỗi mẫu (trong số N mẫu của tập dữ liệu) được coi là cụm
  - Hai cụm giống nhau nhất sau đó được hợp nhất để có N−1 cụm
  - Tiếp theo, hai cụm giống nhau nhất lại được hợp nhất lần nữa để có N−2 cụm
  - Thuật toán tiến hành lặp lại theo cách này cho đến khi tất cả các mẫu đều thuộc về một cụm duy nhất và khi đó biểu đồ phân nhánh hoàn tất

Bottom-up or agglomerative clustering

- Phương pháp phân cụm phân cấp tổng hợp

1. Begin with n observations and a measure (such as Euclidean distance) of all the $\binom{n}{2} = n(n-1)/2$ pairwise dissimilarities. Treat each observation as its own cluster.

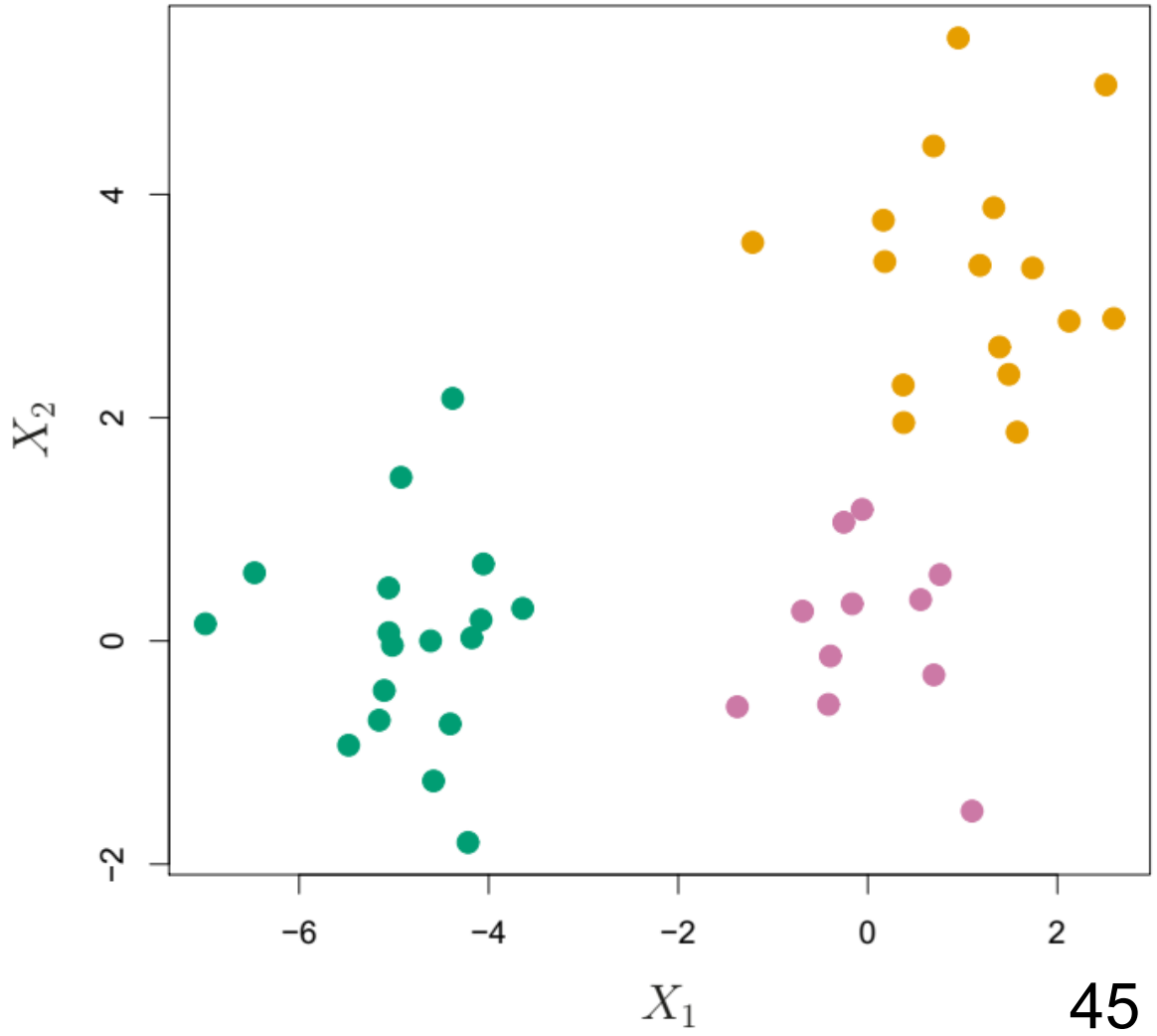
2. For $i = n, n-1, \ldots, 2$:

   (a) Examine all pairwise inter-cluster dissimilarities among the i clusters and identify the pair of clusters that are least dissimilar (that is, most similar). Fuse these two clusters. The dissimilarity between these two clusters indicates the height in the dendrogram at which the fusion should be placed.

   (b) Compute the new pairwise inter-cluster dissimilarities among the $i-1$ remaining clusters.

Agglomerative clustering

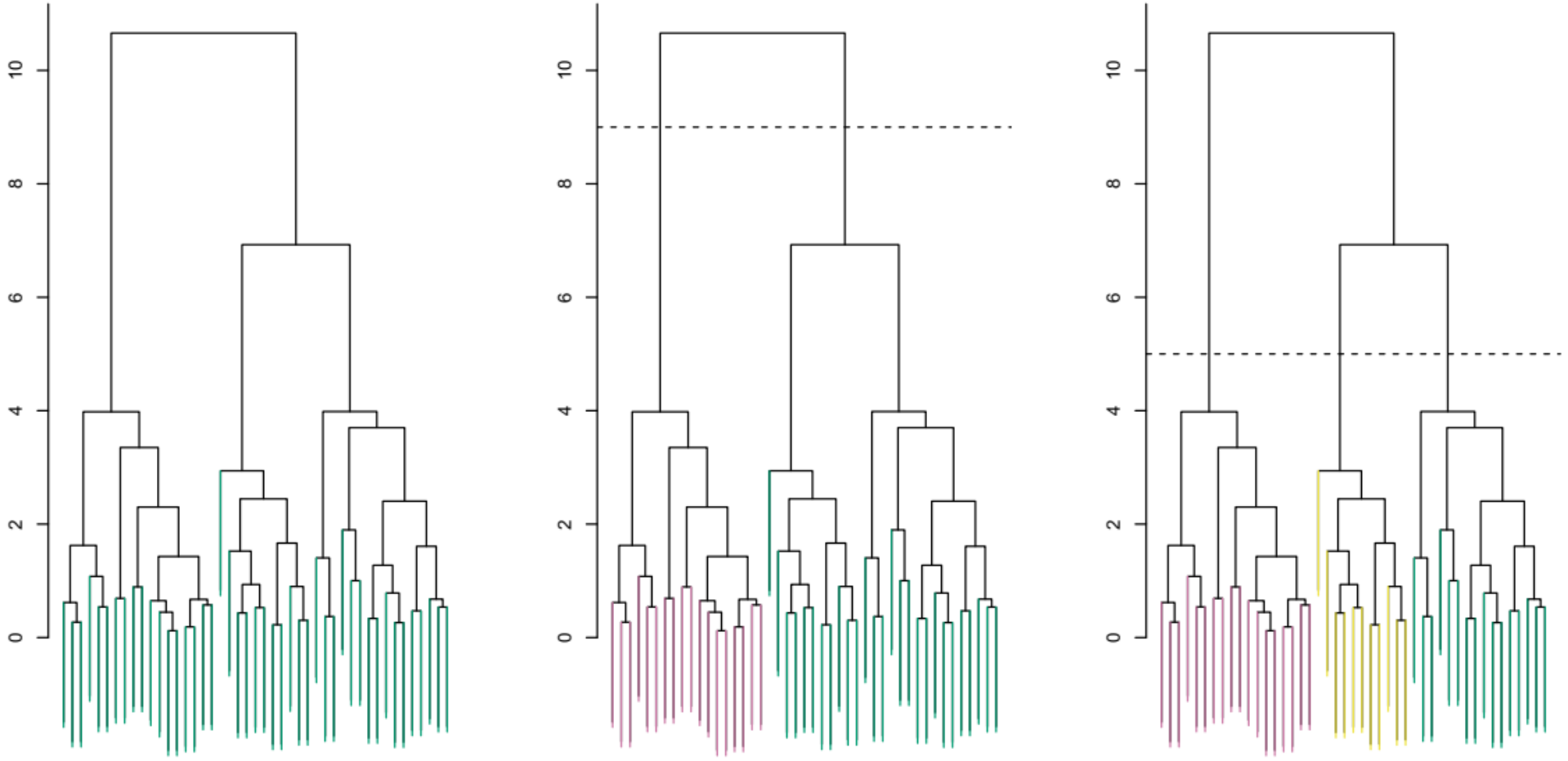
- Ví dụ: xét bộ dữ liệu gồm 45 mẫu



45 mẫu (3 cụm)

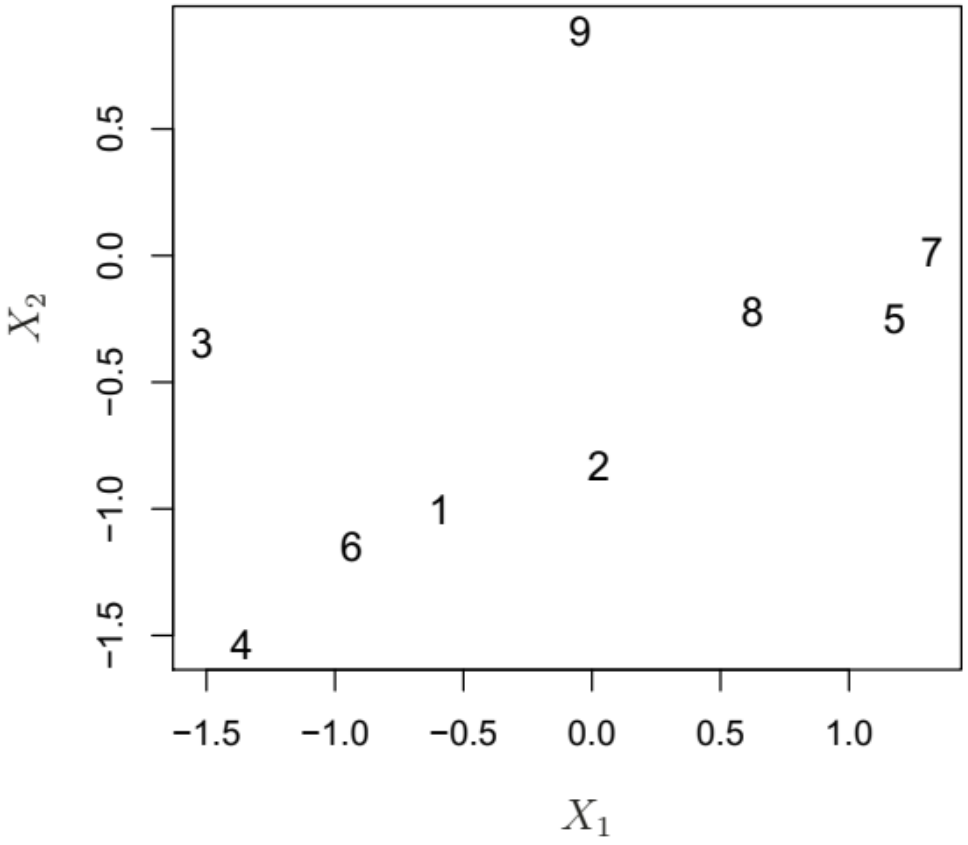
Dendrogram

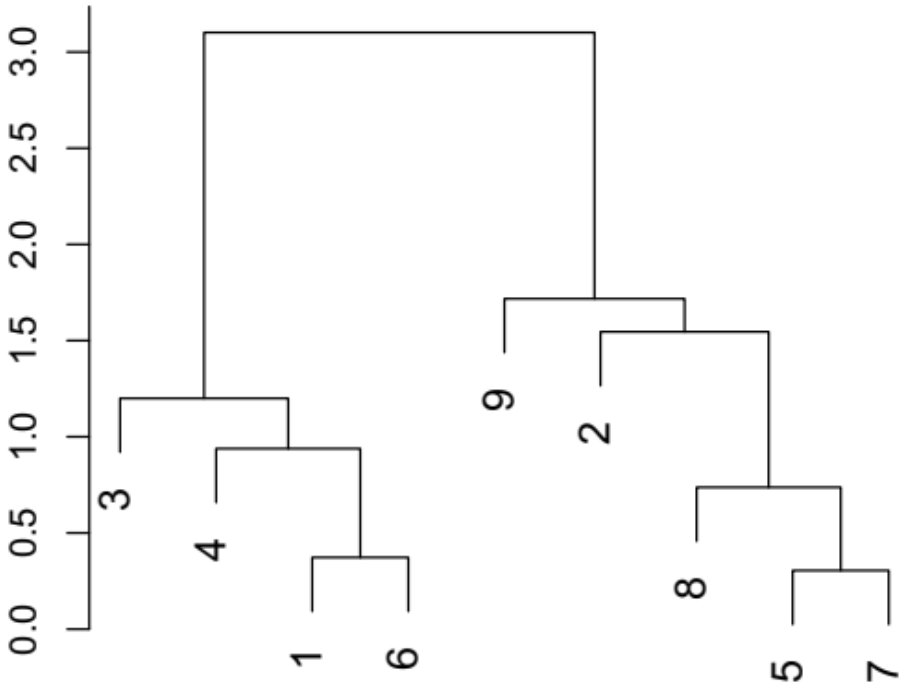
- Dendrogram using complete linkage and Euclidean distance.



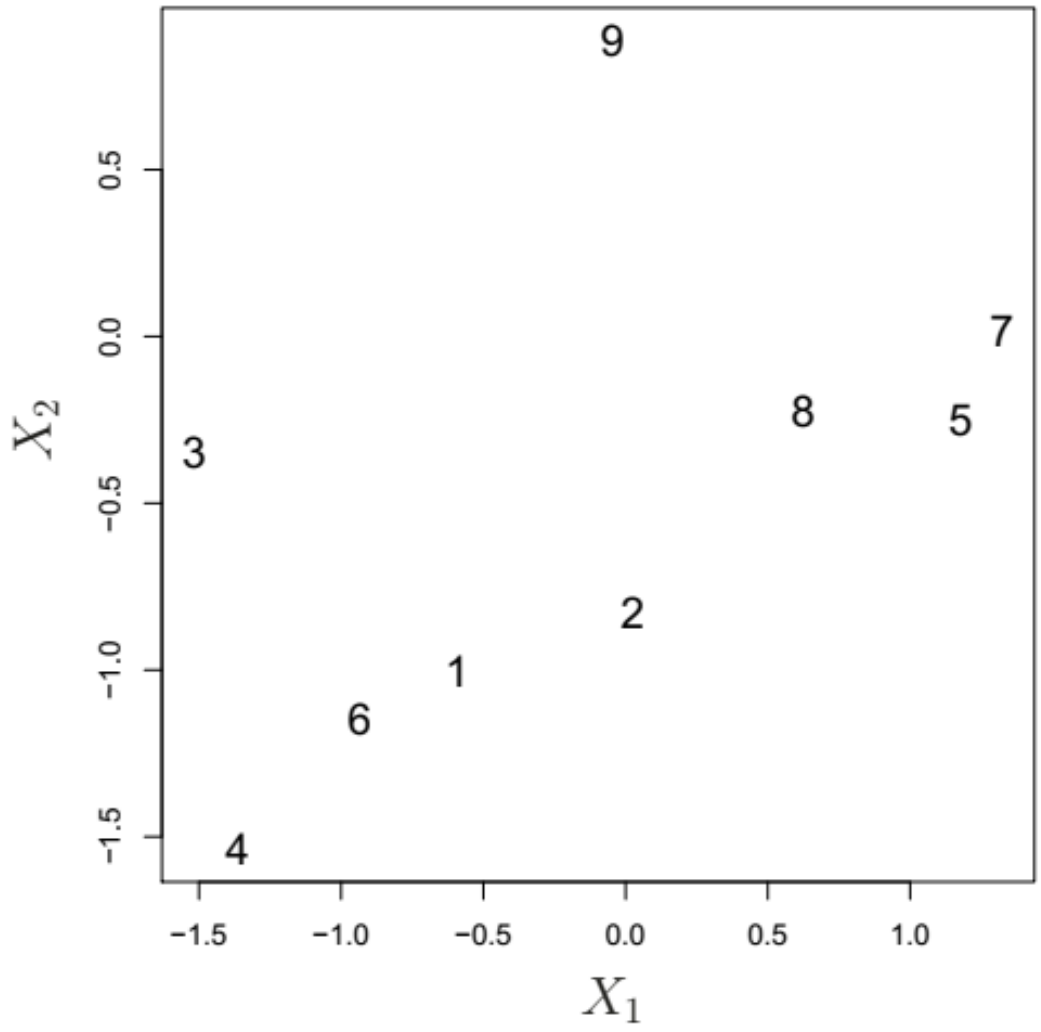
Agglomerative clustering

- Ví dụ

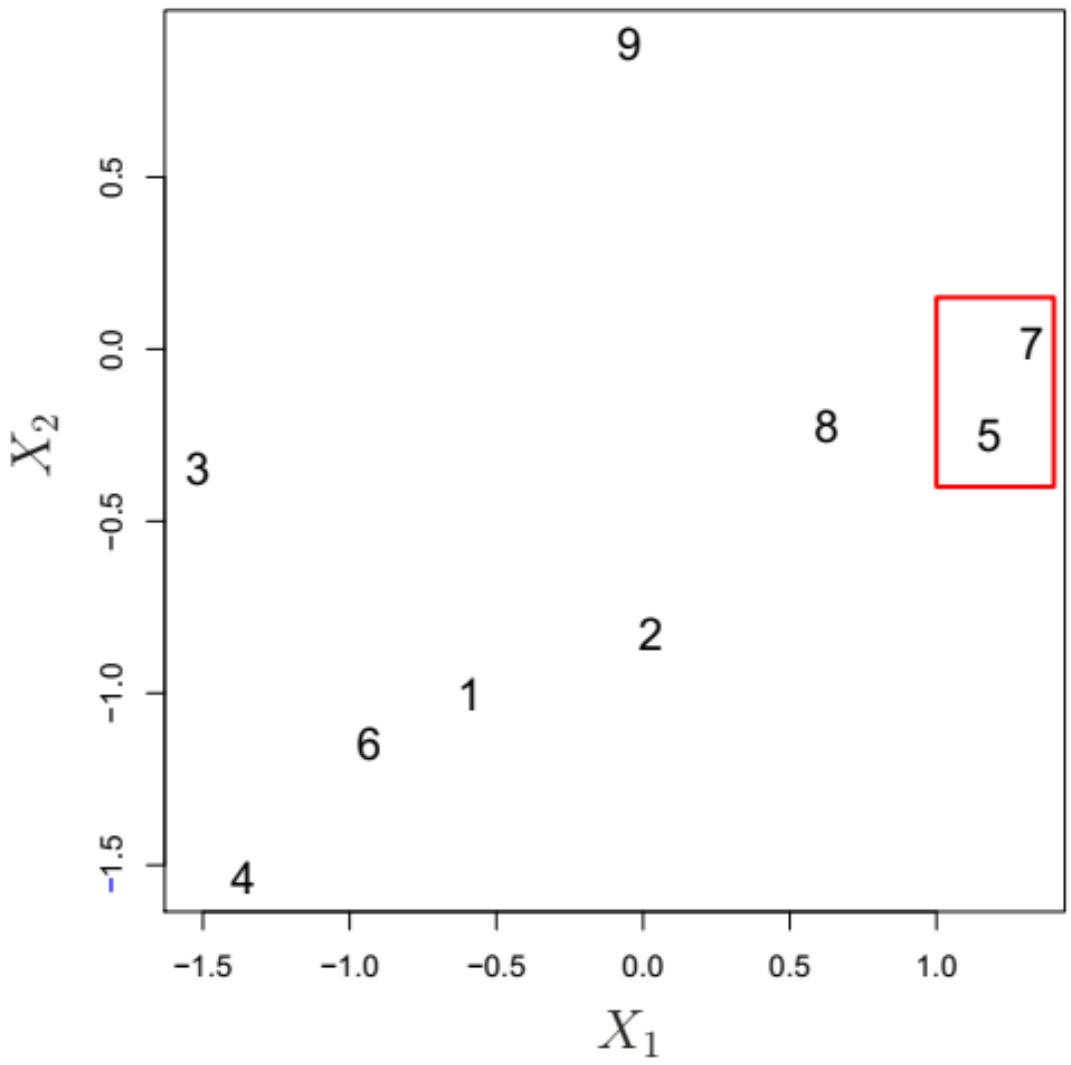




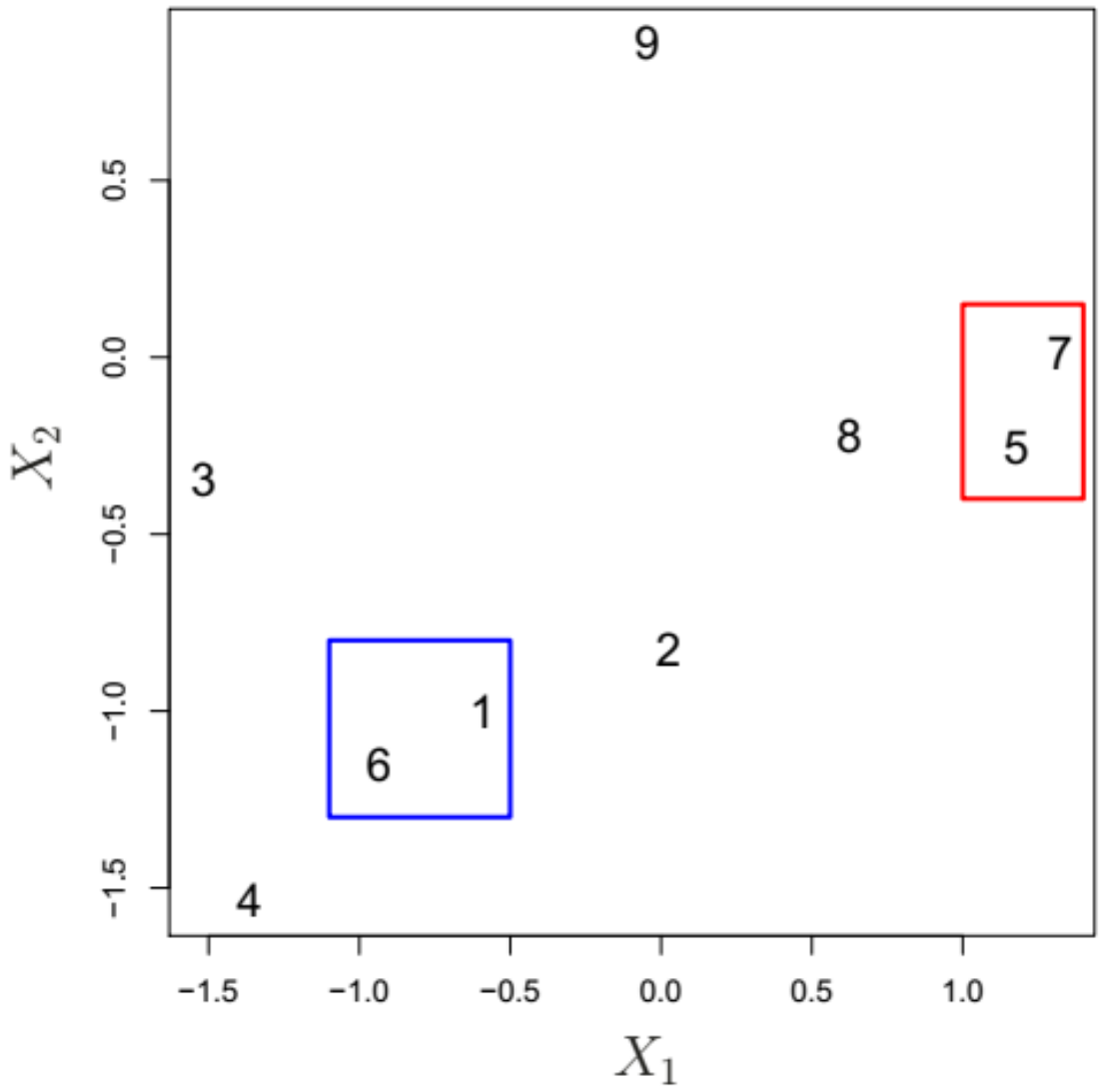
Agglomerative clustering



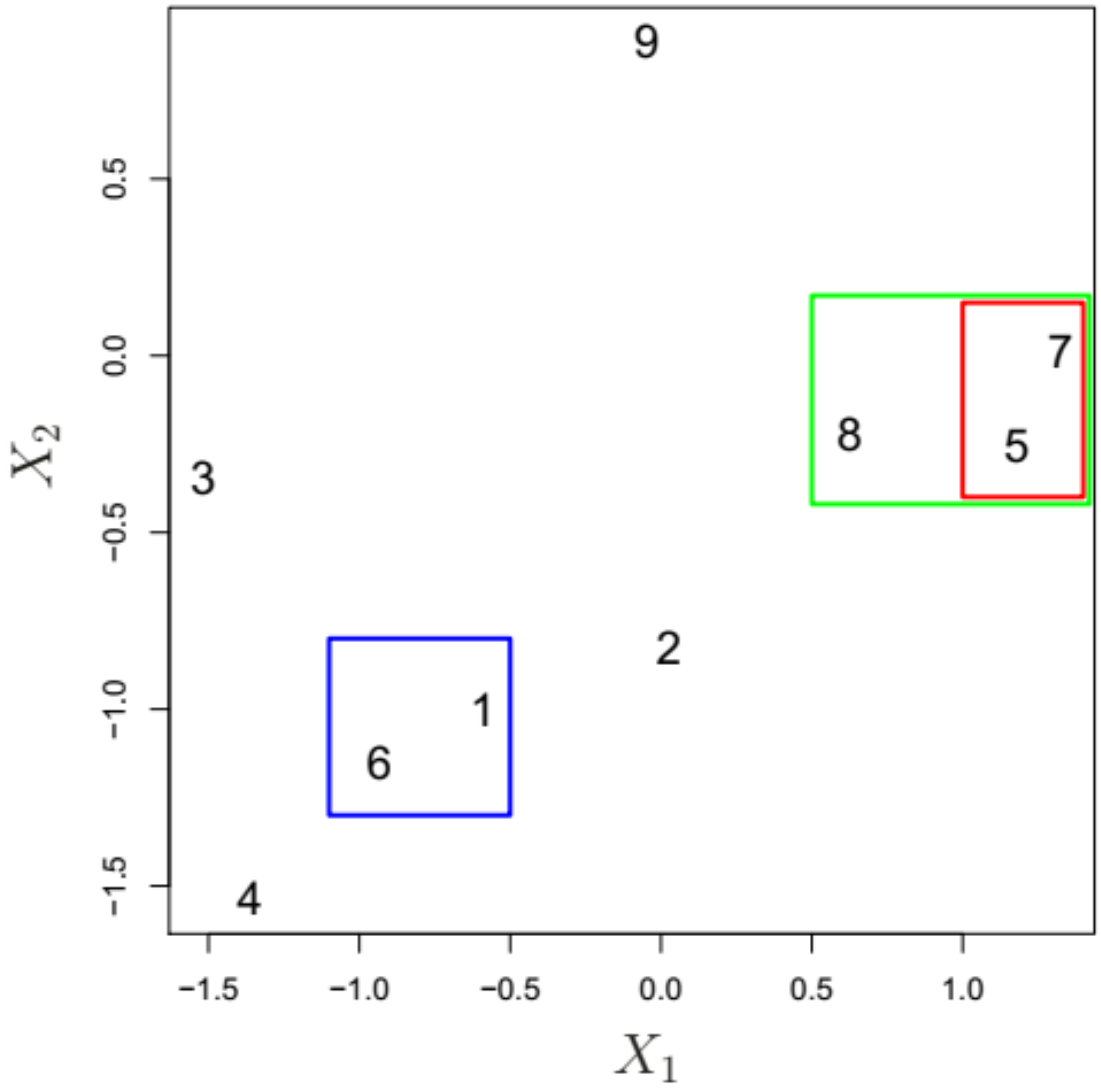
Agglomerative clustering



Agglomerative clustering



Agglomerative clustering



Linkage

- ~ biểu diễn sự khác biệt giữa các cụm
- Sử dụng cho các chiến lược hợp nhất cụm
- Ba kiểu thông dụng nhất là: complete, average, single
  - Liên kết trung bình và hoàn chỉnh thường được ưa chuộng hơn liên kết đơn vì chúng có xu hướng tạo ra các sơ đồ phân nhánh cân bằng hơn

Linkage

- Complete linkage
  - Độ khác biệt giữa các cụm tối đa
    - Tính toán tất cả các độ khác biệt nhau theo cặp giữa các mẫu trong cụm A và các mẫu trong cụm B
    - Lấy độ khác biệt lớn nhất trong số các độ khác biệt đã tính

  => “khoảng cách“ cụm ~ “khoảng cách“ giữa hai điểm “xa nhất“ trong hai cụm

Linkage

- Single linkage
  - Độ khác biệt giữa các cụm tối thiểu
    - Tính toán tất cả các độ khác biệt nhau theo cặp giữa các mẫu trong cụm A và các mẫu trong cụm B
    - Lấy độ khác biệt nhỏ nhất trong số các độ khác biệt đã tính

  => “khoảng cách“ cụm ~ “khoảng cách“ giữa hai điểm “gần nhất“ trong hai cụm

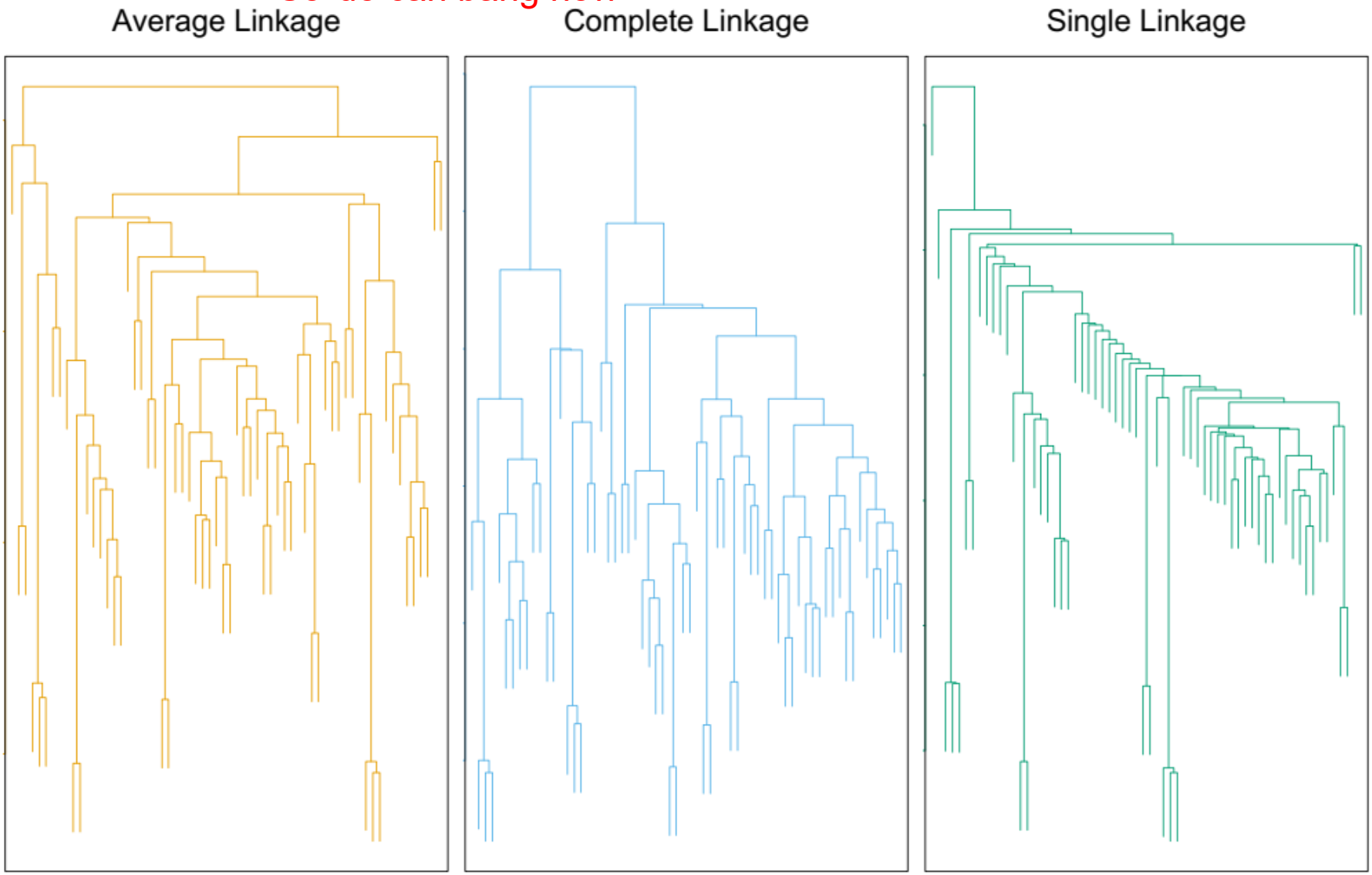
Linkage

- Average linkage
  - Độ khác biệt giữa các cụm trung bình
    - Tính toán tất cả các độ khác biệt nhau theo cặp giữa các mẫu trong cụm A và các mẫu trong cụm B
    - Lấy độ khác biệt trung bình của các độ khác biệt đã tính

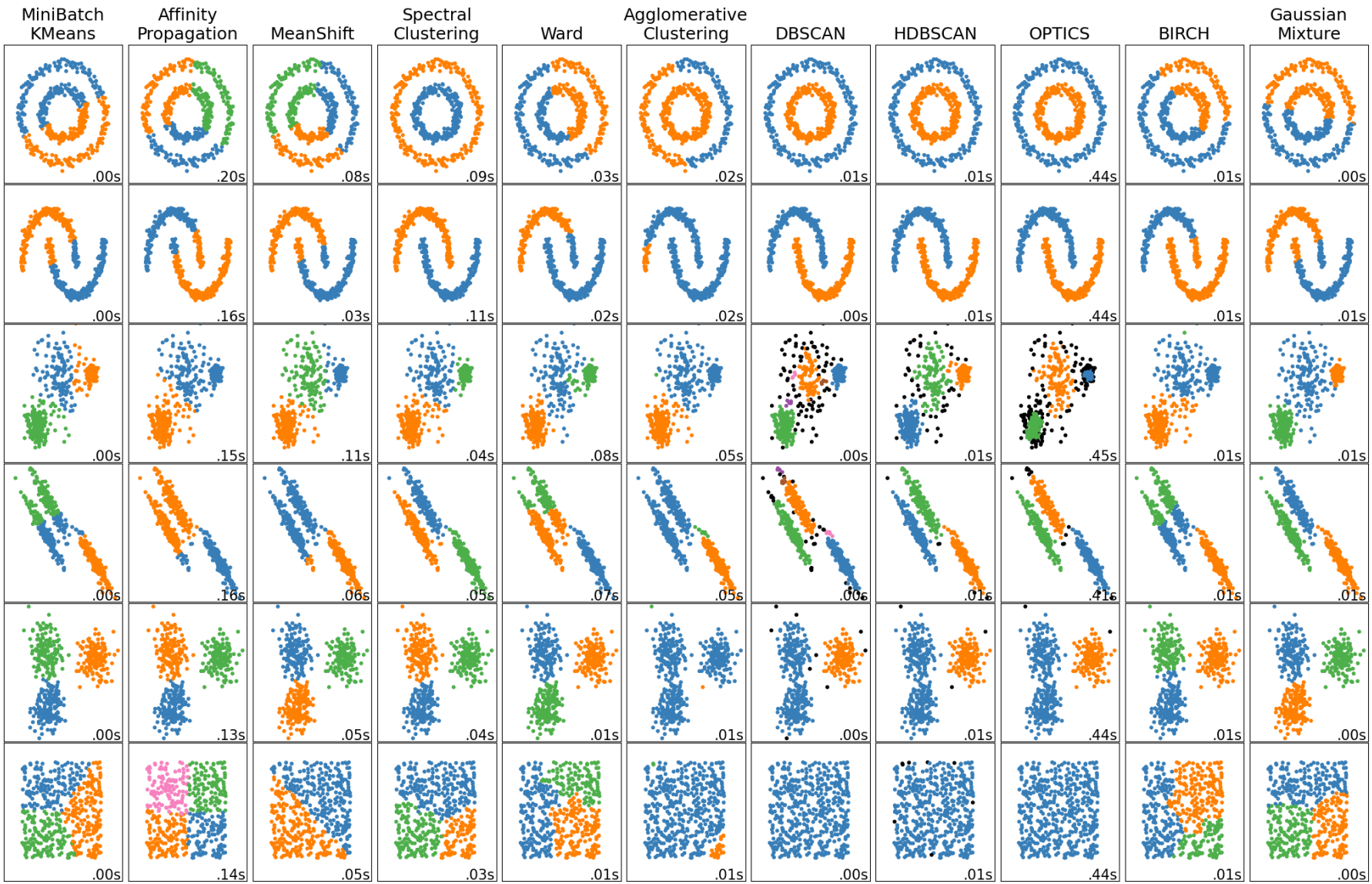
  => “khoảng cách“ cụm ~ “khoảng cách“ trung bình của các “khoảng cách“ của mỗi cặp trong hai cụm

Linkage

Sơ đồ cân bằng hơn



Các thuật toán phân cụm khác



https://scikit-learn.org/stable/auto_examples/cluster/plot_cluster_comparison.html

3.5 Ví dụ về bài toán phân cụm

- Ví dụ 1: Phân đoạn màu sắc

https://picryl.com/media/lady-bug-bugs-animals-ffd157

Ví dụ 1

Mô hình ML

k cụm trong bức ảnh

Sử dụng lớp Kmeans trong scikit-learn

Các pixel có cùng màu sắc được phân vào cùng một cụm

Ví dụ 1

Original image

10 colors

8 colors

6 colors

4 colors

2 colors

Ví dụ 2

- Sử dụng phân cụm để tiền xử lý
  - Phân cụm có thể là một phương pháp hiệu quả để giảm chiều dữ liệu ~ như một bước tiền xử lý trước thuật toán học có giám sát
  - Áp dụng cho tập dữ liệu đơn giản giống MNIST chứa 1797 hình ảnh theo thang độ xám 8×8, biểu diễn các số từ 0 đến 9.

Ví dụ 2

- Sử dụng Pipeline

```
log_reg = LogisticRegression()
log_reg.fit(X_train, y_train)
```

```
>>> log_reg.score(X_test, y_test)
0.9688888888888889
```

```
from sklearn.pipeline import Pipeline

pipeline = Pipeline([
    ("kmeans", KMeans(n_clusters=50)),
    ("log_reg", LogisticRegression()),
])
pipeline.fit(X_train, y_train)
```

```
>>> pipeline.score(X_test, y_test)
0.9777777777777777
```

Ví dụ 3

- Sử dụng phân cụm cho học bán giám sát

```
n_labeled = 50
log_reg = LogisticRegression()
log_reg.fit(X_train[:n_labeled], y_train[:n_labeled])
```

```
>>> log_reg.score(X_test, y_test)
0.8333333333333334
```

```
k = 50
kmeans = KMeans(n_clusters=k)
X_digits_dist = kmeans.fit_transform(X_train)
representative_digit_idx = np.argmin(X_digits_dist, axis=0)
X_representative_digits = X_train[representative_digit_idx]
y_representative_digits = np.array([4, 8, 0, 6, 8, 3, ..., 7, 6, 2, 3, 1, 1])
```

```
>>> log_reg = LogisticRegression()
>>> log_reg.fit(X_representative_digits, y_representative_digits)
>>> log_reg.score(X_test, y_test)
0.9222222222222223
```

Ví dụ 3

- Sử dụng phân cụm cho học bán giám sát

  label propagation

```
y_train_propagated = np.empty(len(X_train), dtype=np.int32)
for i in range(k):
    y_train_propagated[kmeans.labels_==i] = y_representative_digits[i]


>>> log_reg = LogisticRegression()
>>> log_reg.fit(X_train, y_train_propagated)
>>> log_reg.score(X_test, y_test)
0.9333333333333333
```

Tổng kết

- Sinh viên nắm được một số thuật toán phân cụm thông dụng bên cạnh k-means
- Sinh viên biết các ứng dụng sklearn để thực hiện phân cụm trên các bộ dữ liệu khác nhau

Hoạt động sau buổi học

- Sinh viên sử dụng sklearn để thực hiện phân cụm, sử dụng các bộ dữ liệu khác nhau

Chuẩn bị cho buổi học tiếp theo

- Sinh viên tìm hiểu về học tăng cường (Reinforcement Learning) và ứng dụng

Tài liệu tham khảo

- [B11TLTK1] G. James, D. Witten, T. Hastie, R. Tibshirani, J. Taylor, An Introduction to Statistical Learning with Applications in Python, Springer, 2023
  - Chapter 12 Unsupervised Learning