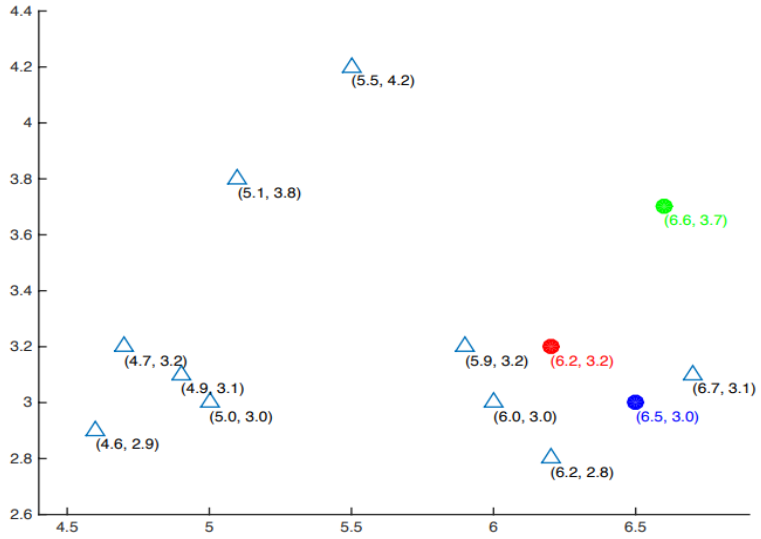
Phân cụm dữ liệu

4.1. K-means

Cho tập dữ liệu có các điểm sau: X = [5.9, 3.2], [4.6, 2.9], [6.2, 2.8], [4.7, 3.2],[5.5, 4.2], [5.0, 3.0]

[4.9 3.1], [6.7, 3.1], [5.1, 3.8], [6.0 3.0].

Phân cụm k-means với k = 3 và độ đo khoảng cách giữa các điểm $(x_i - y_i)$ là khoảng cách Euclid như sau: $d = \sqrt{\sum_{i=1}^{p}(x_i - y_i)^2}$. Tất cả các điểm của X được minh họa trong hình 1. Giả sử rằng 3 tâm cụm khởi tạo là 3 điểm μ_1 = (6.2, 3.2) (màu đỏ), μ_2 = (6.6, 3.7) (màu xanh lá), μ3 = (6.5, 3.0) (màu xanh da trời).



Hình 1. Biểu đồ phân tán của tập dữ liệu và 3 tâm khởi tạo của 3 cụm

A. Thực hiện tính toán thủ công các yêu cầu sau:

1. Cho biết tọa độ của tâm cụm đầu tiên (màu đỏ) sau một lần lặp, tương tự với 2 cụm còn lại

2. Cho biết tọa độ của tâm cụm thứ hai (màu xanh lá) sau lần lặp thứ 2

3. Cho biết tọa độ của tâm thứ ba (màu xanh da trời) khi quá trình phân cụm kết thúc

4. Cho biết số lần lặp khi quá trình phân cụm kết thúc

B. Sử dụng thuật toán k-means trong thư viện scikit-learn phân cụm tập dữ liệu DATA

Data = { 'x': [25, 34, 22, 27, 33, 33, 31, 22, 35, 34, 67, 54, 57, 43, 50, 57, 59, 52, 65, 47, 49, 48, 35, 33, 44, 45, 38, 43, 51, 46],
'y': [79, 51, 53, 78, 59, 74, 73, 57, 69, 75, 51, 32, 40, 47, 53, 36, 35, 58, 59, 50, 25, 20, 14, 12, 20, 5, 29, 27, 8, 7]

Tham khảo ví dụ và liên kết sau đây:

https://scikit-learn.org/stable/modules/generated/sklearn.cluster.KMeans.html#sklearn.cluster.KMeans

```
from sklearn.cluster import KMeans
import numpy as np
X = np.array([[1, 2], [1, 4], [1, 0],
              [10, 2], [10, 4], [10, 0]])
kmeans = KMeans(n_clusters=2, random_state=0).fit(X)
print(kmeans.labels_)
print(kmeans.predict([[0, 0], [12, 3]]))
print("tam cum la:", kmeans.cluster_centers_)
```

C. Viết báo cáo thuật toán k-means trong scikit-learn với các ví dụ minh họa về Thuộc tính, về Phương thức nạp vào LMS

4.2. Phân cụm phân cấp (thứ bậc)

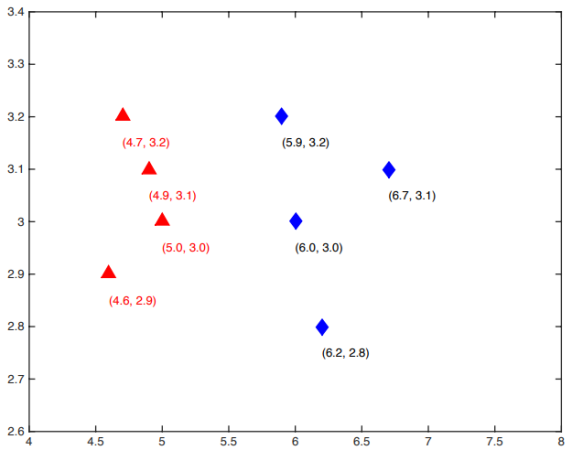
A. Hình 2 là 2 cụm A (đỏ) và B (xanh da trời), mỗi cụm có 4 điểm và tọa độ tương ứng của các điểm. Tính khoảng cách Euclid giữa 2 cụm A và B bằng các phương pháp sau:

1. Single linkage

2. Complete linkage

3. Trung bình nhóm

4. Ba phương pháp trên phương pháp nào có khả năng đáp ứng với nhiễu và dữ liệu ngoại lai tốt nhất?



Hình 2. Phân bố dữ liệu của 2 cụm A và B

B. Sử dụng thuật toán phân cụm thứ bậc Agglomerative trong thư viện scikit-learn phân cụm tập dữ liệu DATA.

```
DATA =[5,3],[10,15],[15,12],[24,10],[30,30],[85,70],[71,80],
[60,78],[70,55],[80,91]
```

. Tham khảo ví dụ và liên kết sau đây:

https://scikit-learn.org/stable/modules/generated/sklearn.cluster.AgglomerativeClustering.html

C. Viết báo cáo thuật toán phân cụm Agglomerative trong scikit-learn với các ví dụ minh họa về Thuộc tính, về Phương thức nạp vào LMS