**Hoàng Đắc Hùng**

**18103200013**

1. Thuật toán phân cụm K-means

**Thuật toán K-MEAN**

Cho số K, mỗi nhóm được biểu diễn bằng giá trị trung bình của dữ liệu trong nhóm

**Bước 1**: Chọn ngẫu nhiên k đối tượng như là những điểm trung tâm của các nhóm

**Bước 2**: Gán từng đối tượng còn lại vào nhóm có trung tâm nhóm gần nó nhất (dựa trên độ

đo khoảng cách khoảng cách **Euclide**)

D (p, q) = =

Ví dụ: p(1, 2) q(3, 4)

D (p, q) =

**Bước 3**: Tính lại giá trị trung tâm của từng nhóm

**Bước 3.1**. Di chuyển trung tâm nhóm về = giá trị trung bình mới của nhóm

**Bước 3.2**. Cho nhóm Ki = , giá trị trung bình của nhóm là:

mi =

**Bước 4**: Nếu các trung tâm của nhóm không thay đổi thì dừng, ngược lại quay lại **Bước 2**

**Bài tập**

Có 2 dạng bài: Cho trước “tâm” để phân cụm và tự chọn “tâm” để phân cụm

|  |  |  |
| --- | --- | --- |
| Đối tượng | Đặc Tính X | Đặc Tính Y |
| A | 1 | 1 |
| B | 2 | 1 |
| C | 4 | 3 |
| D | 5 | 4 |

*Trong bài này chúng ta sẽ tự chọn tâm cho 2 nhóm*

* Khởi tạo tâm cho 2 nhóm
  + **A** là tâm nhóm thứ nhất **C1(1, 1)**
  + **D** là tâm nhóm thứ hai **C2(5, 4)**
* Tính khoảng cách từ các đối tượng đến tâm nhóm

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | **A** | **B** | **C** | **D** |  |
| **D0** | **0** | **1** | **3.61** | **5** | **C1 (1, 1)** |
|  | **5** | **4.21** | **1.41** | **0** | **C2 (5, 4)** |

Từ bảng trên ta xét từng cột:

* Giá trị ô nào nhỏ nhất thì ta điền 0 (gọi 0 là điểm không thuộc nhóm đó)
* Giá trị lớn ta điền 1 (gọi 1 là điểm thuộc nhóm)
* Từ đó ta suy ra bảng sau

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | **A** | **B** | **C** | **D** |  |
| **G0** | **1** | **1** | **0** | **0** | **C1 (1, 1)** |
|  | **0** | **0** | **1** | **1** | **C2 (5, 4)** |

Như vậy ta có:

* Nhóm C1 gồm có 2 điểm A (1, 1) ; B(2, 1)
* Nhóm C2 gồm có 2 điểm C (4, 3) ; D (5, 4)

Tính lại điểm trung tâm của các nhóm bằng cách tính trung bình của hoành độ và tung độ của tất cả các điểm trong mỗi nhóm:

Từ kết quả trước ta có:

* + m1 là tâm nhóm thứ nhất C1( 1.5 , 1 )
  + m2 là tâm của nhóm thứ C2(4.5 , 3.5)

Tính lại khoảng cách từ các đối tượng đến tâm của nhóm

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | **A** | **B** | **C** | **D** |  |
| **D1** | **0.5** | **0.5** | **3.2** | **3.6** | **C1 (1, 1)** |
|  | **4.3** | **3.53** | **0.7** | **0.7** | **C2 (5, 4)** |

Ta lại tiếp tục xét trên từng cột :

* Giá trị nhỏ nhất ta điền 0
* Giá trị lớn nhất ta điền 1

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | **A** | **B** | **C** | **D** |  |
| **G1** | **1** | **1** | **0** | **0** | **C1 (1, 1)** |
|  | **0** | **0** | **1** | **1** | **C2 (5, 4)** |

Ta nhận thấy G0 và G1 có các điểm như nhau không thay đổi. Ta không tiếp tục phân nhóm.

Kết luận:

* Nhóm C1 có các điểm ( A, B), tâm m1 (1.5, 1)
* Nhóm C2 có các điểm (C, D), tâm m2 (4.5, 3.5)

**Cài đặt chương trình K-means**

# nhập các thư viện cần thiết

from scipy.cluster.vq import kmeans, vq

from sklearn.cluster import KMeans

import pandas as pd

import matplotlib.pyplot as plt

df = pd.read\_csv("dataset.csv") # đọc dữ liệu từ tập dữ liệu dataset.csv

df.head() #hiển thị dữ liệu vừa nhập

**A screenshot of a computer

Description automatically generated with low confidence**

# vẽ biểu đồ phân tán để trực quan hóa dữ liệu

plt.scatter(x=df['x'],y=df['y'])

plt.xlabel('Label X')

plt.ylabel('Label Y')

Chart, scatter chart

Description automatically generated

km = KMeans(n\_clusters=2) # đặt số lượng nhóm

y\_predicted = km.fit\_predict(df[['x','y']]) # tạo dự đoán

y\_predicted



df['cluster']= y\_predicted # thêm 1 cột có tên cluster (cụm) với 0 và 1 là 2 nhóm khác nhau( đối tượng nào có số cluster giống nhau là cùng một nhóm)

df.head()

Graphical user interface, text, application, chat or text message

Description automatically generated

km.cluster\_centers\_ # dùng kmeans xác định tâm của các cụm

****

# vẽ lại biểu đồ và tâm để đễ trực quan

df1 = df[df.cluster==0]

df2 = df[df.cluster==1]

plt.scatter(df1['x'],df1['y'],color='green')

plt.scatter(df2['x'],df2['y'],color='red')

plt.scatter(km.cluster\_centers\_[:,0],km.cluster\_centers\_[:,1],color='purple',marker='\*',label='Tâm cụm')#

plt.xlabel('X') #

plt.ylabel('Y')

plt.legend()

**Chart, scatter chart

Description automatically generated**