TRƯỜNG ĐẠI HỌC
VĂN LANG
Đạo đức - Ý chí - Sáng tạo
KHOA CÔNG NGHỆ THÔNG TIN

THỰC HÀNH NHẬP MÔN HỌC MÁY

THỰC HÀNH – K MEANS

BỘ MÔN KHOA HỌC DỮ LIỆU

THỰC HÀNH – K MEANS

Thuật toán K-means clustering, không biết nhãn -label của từng điểm dữ liệu.

Mục đích là làm thể nào để phân dữ liệu thành các cụm -cluster khác nhau, sao cho *dữ liệu trong cùng một cụm có tính chất giống nhau.*

Ví dụ:

+Một công ty muốn tạo ra những chính sách ưu đãi cho những nhóm khách hàng khác nhau, dựa trên sự tương tác giữa mỗi khách hàng với công ty đó.

+Như số năm khách hàng; số tiền khách hàng đã chi trả cho công ty; độ tuổi; giới tính; thành phố; nghề nghiệp; ...

+Giả sử, công ty đó có rất nhiều dữ liệu của rất nhiều khách hàng nhưng chưa có cách nào chia toàn bộ khách hàng đó thành một số nhóm/cụm khác nhau.

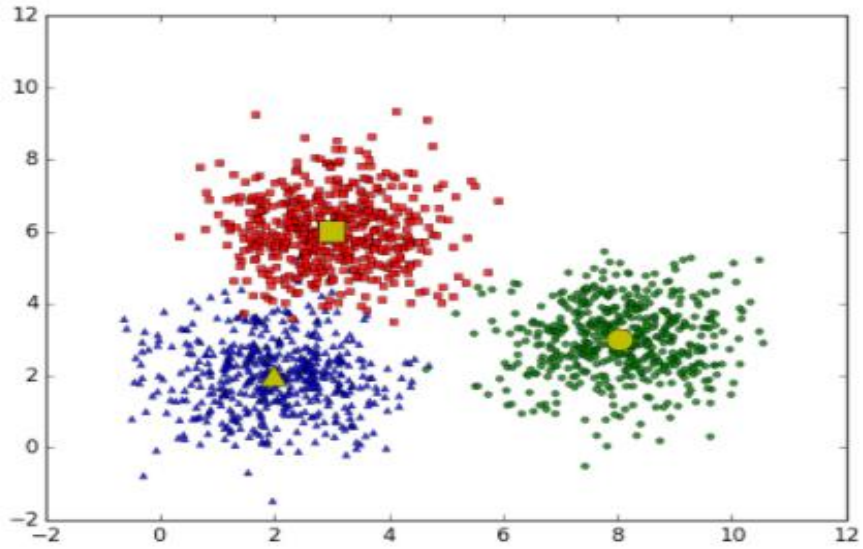
+Phương pháp đầu tiên, ta nghĩ đến dùng K-means Clustering. Vì nó là một trong những thuật toán đầu tiên về Machine Learning.

+Sau khi đã phân ra được từng nhóm, công ty đó có thể lựa chọn ra một vài khách hàng trong mỗi nhóm để quyết định xem mỗi nhóm tương ứng với nhóm khách hàng nào.

THỰC HÀNH – K MEANS

+Ý tưởng đơn giản nhất về cluster là tập hợp các điểm *ở gần nhau trong một không gian nào đó.*

Ví dụ về 3 cụm dữ liệu như hình sau.



+Giả sử mỗi cluster có một điểm đại diện *center* màu vàng.

+Những điểm xung quanh mỗi center thuộc vào cùng nhóm với center đó.

+xét một điểm bất kỳ, xem điểm đó gần với center nào nhất thì nó thuộc về cùng nhóm với center đó.

+Bài toán thú vị như sau - *trên một vùng biển hình vuông lớn có ba đảo hình vuông, tam giác, và tròn màu vàng. Một điểm trên biển được gọi là thuộc lãnh hải của một đảo nếu nó nằm gần đảo này hơn so với hai đảo kia . Hãy xác định ranh giới lãnh hải của các đảo.*

THỰC HÀNH – K MEANS

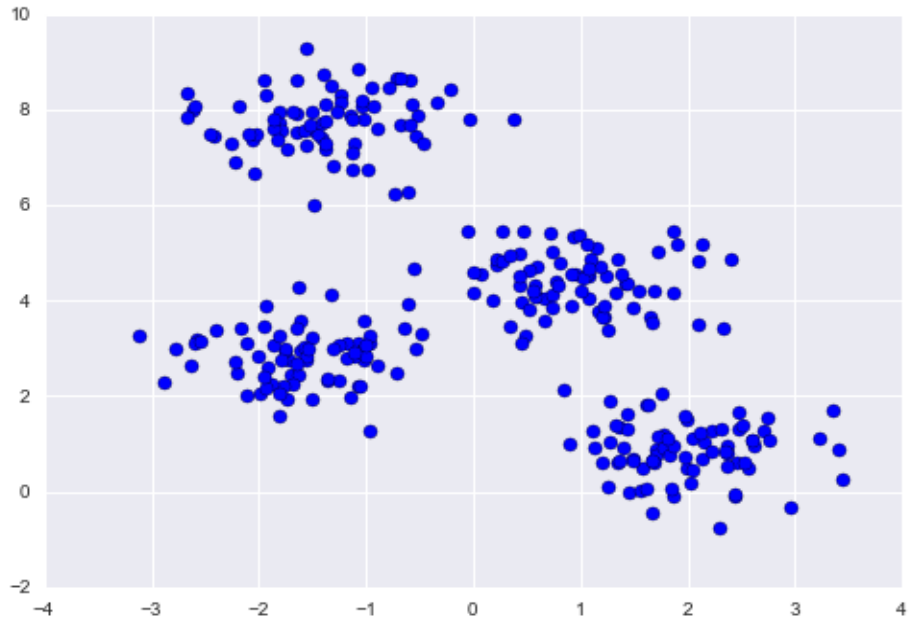
#Thực hành 01

```
>>> import math
>>> from __future__ import print_function
>>> import numpy as np
>>> import matplotlib.pyplot as plt
>>> from scipy.spatial.distance import cdist
>>> np.random.seed(11)
>>> means = [[2, 2], [8, 3], [3, 6]]
>>> cov = [[1, 0], [0, 1]]
>>> N = 500
>>> X0 = np.random.multivariate_normal(means[0], cov, N)
>>> X1 = np.random.multivariate_normal(means[1], cov, N)
>>> X2 = np.random.multivariate_normal(means[2], cov, N)
>>> X = np.concatenate((X0, X1, X2), axis = 0)
>>> K = 3
>>> original_label = np.asarray([0]*N + [1]*N + [2]*N).T
>>> def kmeans_display(X, label):
    K = np.amax(label) + 1
    X0 = X[label == 0, :]
    X1 = X[label == 1, :]
    X2 = X[label == 2, :]
    plt.plot(X0[:, 0], X0[:, 1], 'b^', markersize = 4, alpha = .8)
    plt.plot(X1[:, 0], X1[:, 1], 'go', markersize = 4, alpha = .8)
    plt.plot(X2[:, 0], X2[:, 1], 'rs', markersize = 4, alpha = .8)
    plt.axis('equal')
    plt.plot()
    plt.show()
>>> kmeans_display(X, original_label)
```

THỰC HÀNH – K MEANS

#Thực hành 02

```
import matplotlib.pyplot as plt
import seaborn as sns; sns.set()
import numpy as np
from sklearn.datasets.samples_generator import make_blobs
X, y_true = make_blobs(n_samples=300, centers=4,
                       cluster_std=0.60, random_state=0)
plt.scatter(X[:, 0], X[:, 1], s=50);
```



THỰC HÀNH – K MEANS

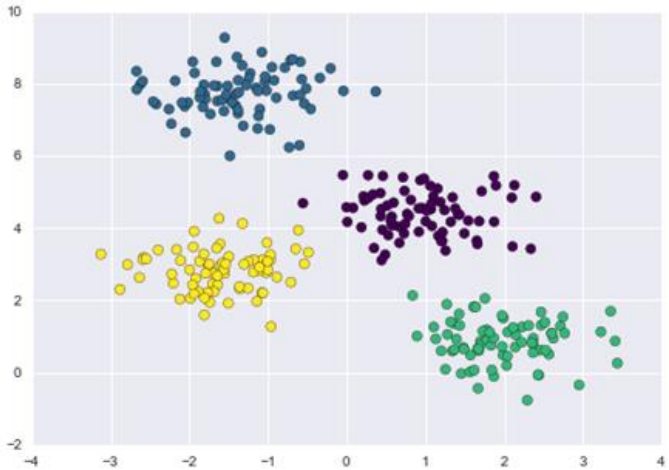
#Thực hành 03

```
from sklearn.cluster import KMeans
kmeans = KMeans(n_clusters=4)
kmeans.fit(X)
y_kmeans = kmeans.predict(X)
plt.scatter(X[:, 0], X[:, 1], c=y_kmeans, s=50, cmap='viridis')

centers = kmeans.cluster_centers_
plt.scatter(centers[:, 0], centers[:, 1], c='black', s=200,alpha=0.5);
```

THỰC HÀNH – K MEANS

#Thực hành 04

```
from sklearn.metrics import pairwise_distances_argmin

def find_clusters(X, n_clusters, rseed=2):
    rng = np.random.RandomState(rseed)
    i = rng.permutation(X.shape[0])[:n_clusters]
    centers = X[i]

    while True:

        labels = pairwise_distances_argmin(X, centers)

        new_centers = np.array([X[labels == i].mean(0)
                                for i in range(n_clusters)])

        if np.all(centers == new_centers):
            break
        centers = new_centers

    return centers, labels

centers, labels = find_clusters(X, 4)
plt.scatter(X[:, 0], X[:, 1], c=labels, s=50, cmap='viridis');
```



THỰC HÀNH – K MEANS

#Thực hành 05

```
centers, labels = find_clusters(X, 4, rseed=0)
plt.scatter(X[:, 0], X[:, 1], c=labels, s=50, cmap='viridis');
```

THỰC HÀNH – K MEANS

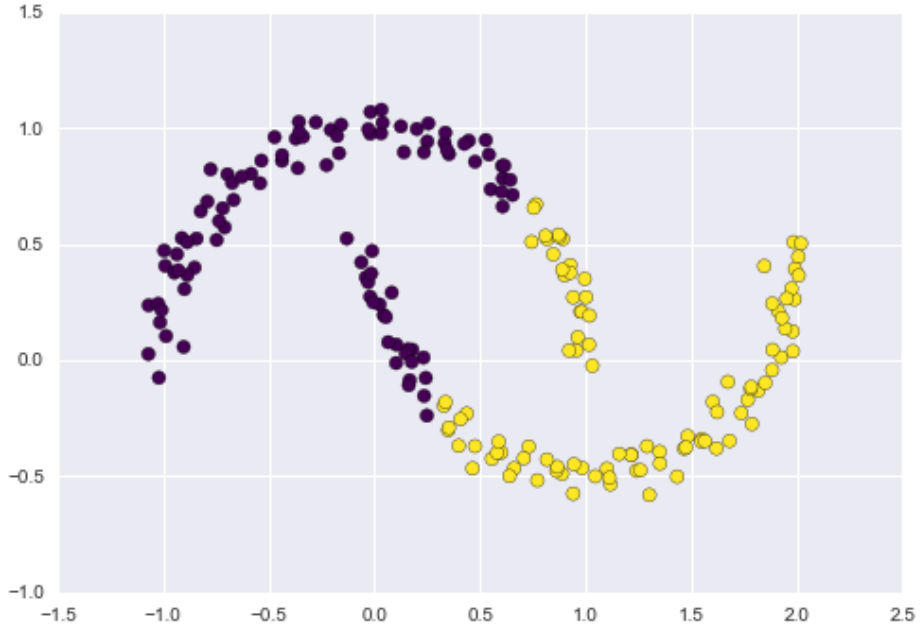
#Thực hành 06

```
labels = KMeans(6, random_state=0).fit_predict(X)
plt.scatter(X[:, 0], X[:, 1], c=labels, s=50, cmap='viridis');
```

THỰC HÀNH – K MEANS

#Thực hành 07

```
from sklearn.datasets import make_moons
X, y = make_moons(200, noise=.05, random_state=0)
labels = KMeans(2, random_state=0).fit_predict(X)
plt.scatter(X[:, 0], X[:, 1], c=labels, s=50, cmap='viridis');
```



TRƯỜNG ĐẠI HỌC
VĂN LANG
Đạo đức - Ý chí - Sáng tạo

KHOA CÔNG NGHỆ THÔNG TIN