KHOA KHOA HỌC MÁY TÍNH

BÁO CÁO BÀI THỰC HÀNH 1

Môn học: Máy học trong Thị giác Máy tính - CS332.I11.KHTN

GVLT: Lê Đình Duy
Mai Tiến Dũng

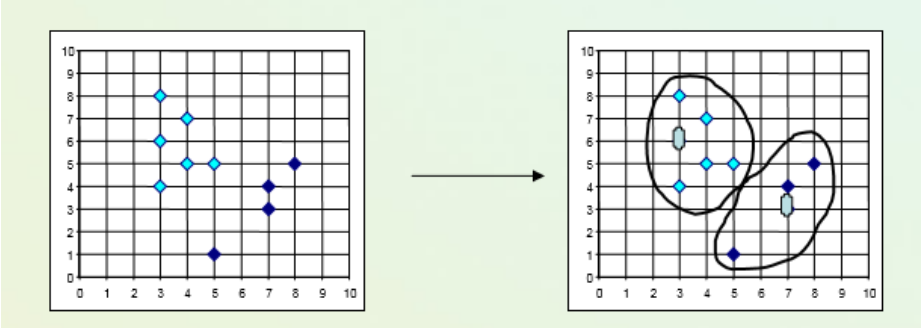
Sinh viên thực hành:
Vũ Thế Dũng 14520205

Contents

I. Bài toán:

- Cluster:
  i. Một cluster là một tập hợp các objects mà các objects này gần giống nhau.
  ii. Tập hợp các đối tượng mà trong đó khoảng cách giữa 2 đối tượng trong 1 cluster nhỏ hơn khoảng cách giữa bất cứ đối tượng nào trong cluster với bất kỳ một đối tượng nào nằm ngoài cluster đó.
  iii.
- Clustering:
  Là quá trình gom nhóm các đối tượng vào các lớp. Mà đối tượng trong các lớp này có thuộc tính giống nhau.

  Giúp người dung hiểu được cấu trúc của bộ dữ liệu
- Good Clustering:
  Tạo ra các cluster mà trong đó:
  Độ giống nhau của các đối tượng trong cluster cao
  Độ giống nhau của các đối tượng thuộc cluster này với cluster khác thấp
  Chất lượng của clustering phụ thuộc vào phương pháp dung để đo độ giống nhau (Khoảng cách) giữa các đối tượng
- Ứng dụng của clustering:
  Economic Science (Phân loại document, web clustering, …)
  Pattern Recognition
  Spatial Data Analysis
  Image Processing



(Nguồn hình: Stefanowski 2008)

Một số các tính khoảng cách:

- Euclidian distance:

$$\sqrt{(\sum_{i=1}^{k}(x_i - y_i)^2)}$$

- Manhattan distance:

$$\sum_{i=1}^{k}|x_i - y_i|$$

- Max of dimensions

$$max_{i=1}^{k}|x_i - y_i|$$

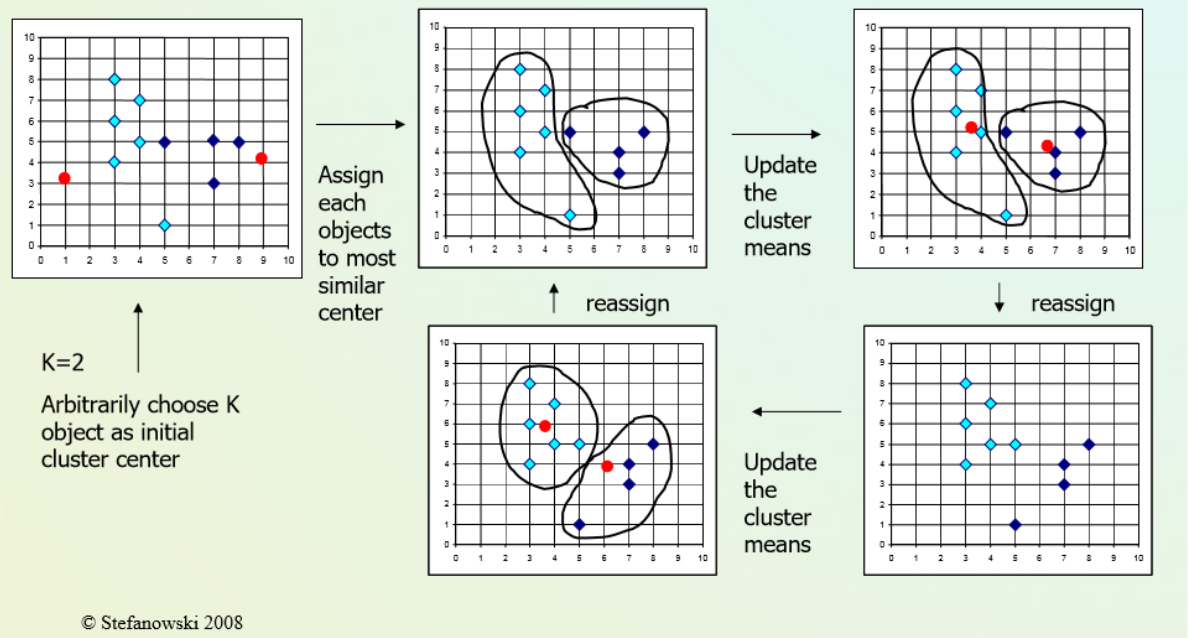
- K means clustering
- Sử dụng K means, Spectral Clustering, DBSCAN, Agglomerative Clustering với tập dữ liệu hand-written digits
- Sử dụng K means, Spectral Clustering, DBSCAN, Agglomerative Clustering với tập dữ liệu face
- Rút trích đặc trưng tiên tiến trên bộ dữ liệu tự chọn

II. Môi trường cài đặt:

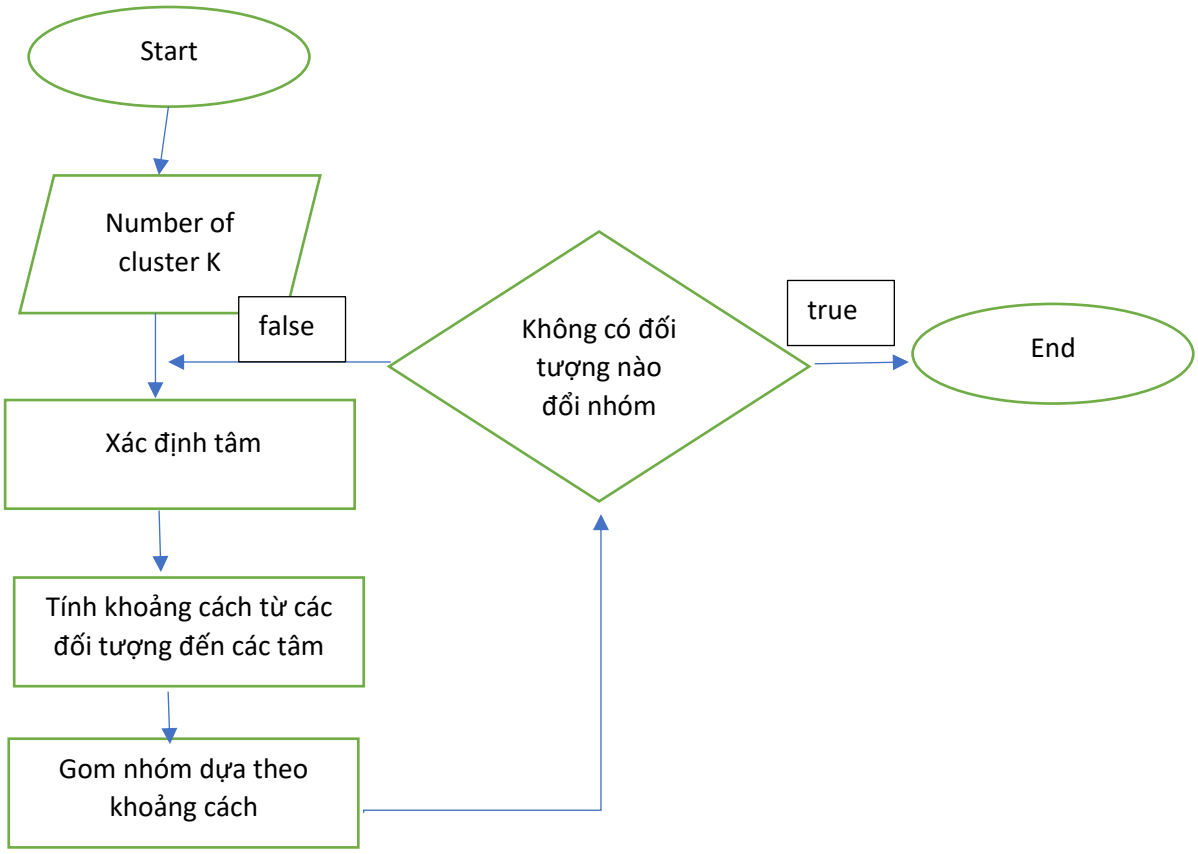
Python 2.7

III. Thực hiện:

Bài tập 1: Thực hiện thuật toán Kmeans với tập dữ liệu được phát sinh ngẫu nhiên gồm 2 Gaussians



- Ý tưởng chính của thuật toán Kmeans là tìm cách phân nhóm các đối tượng đã cho vào K cụm (K là số nguyên, được cho trước) sao cho khoảng các giữa các đối tượng đến tâm nhóm là nhỏ nhất.
- Mô tả thuật toán:



- Khoảng cách giữa objects đến các tâm K thường được tính bằng khoảng cách Euclidean
- Ví dụ:
  i. Tọa độ tâm C(1,1)
  ii. Tọa độ object C1(3, 4)
  iii. Khoảng cách được Euclidian được tính bằng công thức :
  iv. $\sqrt{(3-1)^2+(4-1)^2} = 3.46$

- Thuật toán được cài đặt dựa trên thư viện

```
from sklearn.cluster import KMeans
```

- Sử dụng thư viện numpy và random để generate random dữ liệu thành 2 cụm

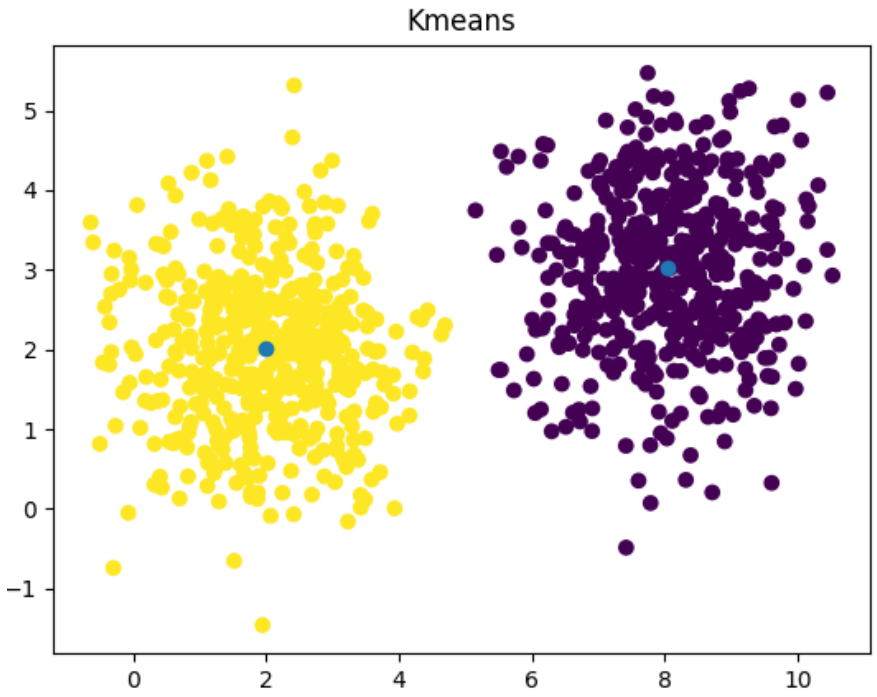
```
X0 = np.random.multivariate_normal(means[0], cov, N)
X1 = np.random.multivariate_normal(means[1], cov, N)
```

- Khởi tại Kmeans và tiến hành cluster

```
cluster = KMeans(n_clusters=2, random_state=0)
result = cluster.fit_predict(X)
```

- Visualize kết quả:

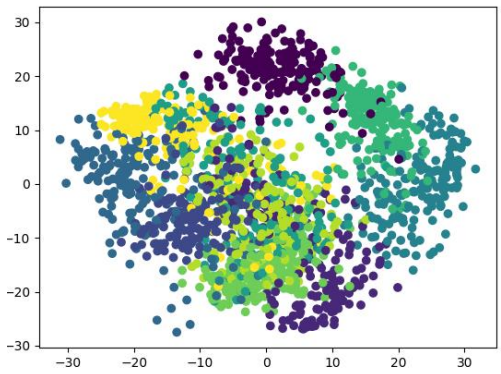
```
centers = cluster.cluster_centers_

plt.scatter(X[:, 0], X[:, 1], c=result)
plt.scatter(centers[:, 0], centers[:, 1])
plt.title("Kmeans")
plt.show()
```



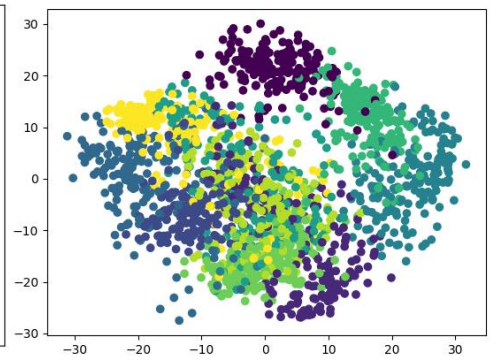
<u>Bài tập 2</u>: Thực kiện thuật toán Kmeans và Spectral clustering trên tập dữ liệu Hand-written digits.

Thực hiện bằng Ngôn ngữ Python dựa trên bộ thư viện Scikit-learn

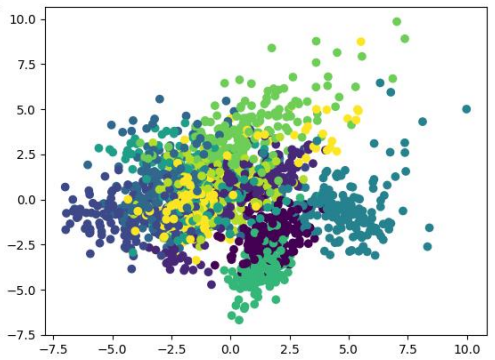
Visualize kết quả:



Kmeans



Spectral



DBSCAN

- So sánh và đánh giá giữa 2 thuật toán Kmeans và Spectral clustering:
  - Kmeans:
    - Ưu điểm: Thuật toán thực hiện nhanh và đơn giản
    - Nhược điểm:
      - Cần phải xác định trước số K
      - Bị ảnh hưởng bởi yếu tố khởi tạo (tâm k)
      - Chỉ có thể cluster những vùng có hình cầu
      - Bị ảnh hưởng bởi các điểm outliers
      - Bị ảnh hưởng bởi công thức tính khoảng cách
  - Spectral clustering: Thực hiện dựa trên ý tưởng của Kmeans

        - Ưu điểm: Có thể thực hiện tìm những cluster khác hình cầu
        - Nhược điểm: Có chung những nhược điểm với Kmeans (trừ yếu tố được khắc phục là cluster tìm được có thể khác hình cầu)
    - DBSCAN: Density Based Spatial Clustering of Applications with Noise
        - Ưu điểm:
            - Có thể cluster được các cụm với hình dạng khác nhau
            - Không bị ảnh hưởng bởi các outliner hay Noise
            - Không cần đặt trước số cluster K
        - Nhược điểm:
            - Phụ thuộc vào cách tính khoảng cách
            - Khó xử lý với data set varying densities
- So sánh giữa các thuật toán:
    - Thuật toán Kmeans thực hiện đơn giản hơn nhưng lại không đem lại hiệu quả bằng thuật các thuật toán còn lại
    - Spectral clustering, DBSCAN có ưu thế hơn trên những ảnh có phân vùng phức tạp.

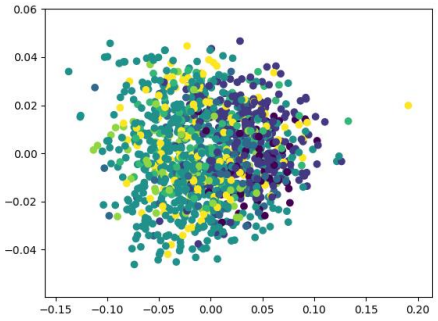
Bài tập 3:

Tương tự bài tập 2 - thay đổi tập dữ liệu là face.
Feature là local binary pattern (LBP).

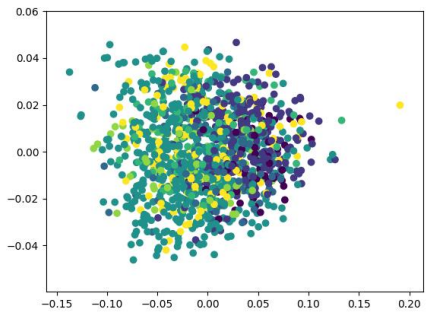
Bộ dữ liệu Face: Sử dụng bộ dữ liệu Face

```
from sklearn.datasets import fetch_lfw_people
lfw_people = fetch_lfw_people(min_faces_per_person=70, resize=0.4)
```

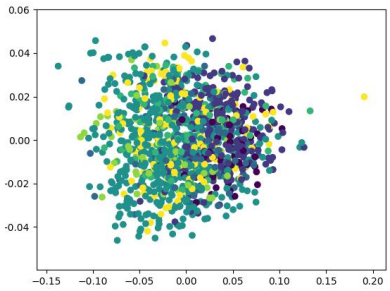
Visualize kết quả:



Kmeans



Spectral



DBSCAN

Bài tập 4:

Chọn một tập dữ liệu cho chính bạn - cùng một phương pháp rút trích đặc trưng tiên tiến.

Bộ dữ liệu 546 file ảnh trong bộ 1360 ảnh:
http://www.robots.ox.ac.uk/~vgg/data/flowers/17/17flowers.tgz

Phương pháp rút trích đặc trưng: HOG (Histogram of oriented gradients):

- Dựa trên sự phân bố về sự phân bố về cường độ ánh sáng của ảnh để trích xuất ra các descriptor
- Hog feature được tính dựa trên histogram của gradient bằng cách sử dụng 1 kernel:

-1
0
1

- Tính độ lớn và hướng của Gradient dựa trên công thức:

$$g = \sqrt{g_x^2 + g_y^2}$$

$$\theta = \arctan \frac{g_y}{g_x}$$

- Chia ảnh thành nhiều của sổ và tính HOG trên những của sổ nhỏ
- Normalization
- Tính HOG feature vector

Phương pháp clustering: Kmeans

Link github: https://github.com/dzungvu/ML

IV. References:

http://scikit-learn.org/stable/modules/generated/sklearn.cluster.KMeans.html
http://scikit-learn.org/stable/modules/generated/sklearn.cluster.SpectralClustering.html
http://scikit-learn.org/stable/modules/generated/sklearn.cluster.DBSCAN.html
http://scikit-learn.org/stable/auto_examples/cluster/plot_kmeans_digits.html#sphx-glr-auto-examples-cluster-plot-kmeans-digits-py

http://opencv-python-tutroals.readthedocs.io/en/latest/py_tutorials/py_feature2d/py_sift_intro/py_sift_intro.html
https://www.learnopencv.com/histogram-of-oriented-gradients/