Лаб. Работа № 4

(продолжительность 2 часа)

Цель работы – на базе данных, полученных в лаб. работе №3, ознакомиться с дополнительными методами кластеризации, произвести сравнительный анализ получаемых с их помощью результатов кластеризации.

Задание для каждого из студентов выдаётся в соответствии с таблицей вариантов (приведена в конце описания лаб. работы).
В каждом из вариантов приведены разновидности кластеризации по паре и тройке полученных в рамках предыдущей лаб. работы признаков, способу преобразования категориального признака в количественный, алгоритму кластеризации.

Требуется:

Произвести кластеризацию пользовательских профилей по парам (p =2) признаков:

- возраст/количество видео (Age/Video)
- возраст/количество фото (Age/Photo)
- возраст/ количество групп (Age/Groups)
- возраст/ количество заметок на странице (Age/Notes)
- возраст/ количество друзей (Age/Friends)
- возраст/место рождения (проживания) (Age/City)*
- место рождения (проживания)/количество видео (City/Video)
- место рождения (проживания)/количество фото (City/Photo)
- место рождения (проживания)/ количество групп (City/Groups)
- место рождения (проживания)/ количество заметок на странице (City/Notes)
- место рождения (проживания)/ количество друзей (City/Friends)

Произвести кластеризацию пользовательских профилей по тройкам (p =3) признаков:

- возраст/место рождения (проживания)/количество видео (Age/City/Video)
- возраст/место рождения (проживания)/количество фото (Age/City/Photo)
- возраст/место рождения (проживания)/ количество групп (Age/City/Groups)
- возраст/место рождения (проживания)/ количество заметок на странице (Age/City/Notes)
- возраст/место рождения (проживания)/ количество друзей (Age/City/Friends)

***Прим:** Так как «место рождения (проживания)» суть категориальный признак, то его **следует перевести** в численную форму способом указанным в таблице, как-то, например:

- взять расстояние от рассматриваемого города до некоторого «центра мира» (Distance). «Центр мира» выбрать по своем усмотрению.
- присвоить индексы городу в зависимости от страны/количества жителей в городе/первой буквы названия на латинице (Index Country + CitizenNumber)
- предложить свой способ (индексы должны быть уникальными, т.е. для разных городов не должны совпадать) (Free Index).

Реализовать кластерный анализ с помощью алгоритмов:

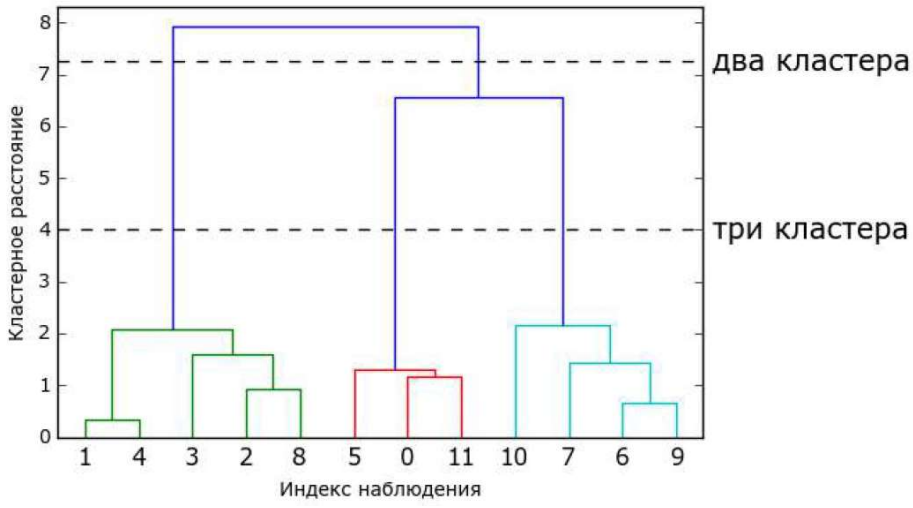
- агломеративная кластеризация (agglomerative clustering), она же - иерархическое дерево или дендограмма, с параметрами: метод Варда (ward), метод средней связи (average), метод полной связи (complete).

**Используем библиотеку from sklearn.cluster import AgglomerativeClustering

*** Для построения результирующего дерева лучше использовать пакет SciPy:

```
# импортируем функцию dendrogram и функцию кластеризации ward из SciPy
from scipy.cluster.hierarchy import dendrogram, ward

X, y = make_blobs(random_state=0, n_samples=12)
# применяем кластеризацию ward к массиву данных X
# функция SciPy ward возвращает массив с расстояниями
# вычисленными в ходе выполнения агломеративной кластеризации
linkage_array = ward(X)

# теперь строим дендрограмму для массива связей, содержащего расстояния
# между кластерами
dendrogram(linkage_array)

# делаем отметки на дереве, соответствующие двум или трем кластерам
ax = plt.gca()
bounds = ax.get_xbound()
ax.plot(bounds, [7.25, 7.25], '--', c='k')
ax.plot(bounds, [4, 4], '--', c='k')

ax.text(bounds[1], 7.25, ' два кластера', va='center', fontdict={'size': 15})
ax.text(bounds[1], 4, ' три кластера', va='center', fontdict={'size': 15})
plt.xlabel("Индекс наблюдения")
plt.ylabel("Кластерное расстояние")
```

Результат:



- DBSCAN (densitybased spatial clustering of applications with noise)

from sklearn.cluster import DBSCAN

Оформить отчёт в электронном виде с примерами результатов кластеризации в виде 2D диаграмм (для p =2) независимо от алгоритма кластеризации. Для экспериментов при p =3 в случае агломеративной кластеризации построить 2D дендограмму, в случае DBSCAN – 3D диаграмму.

Таблица вариантов заданий

# варианта	p=2	p=3	City Index	Clustering Algorithm
1	Age/Photo	Age/City/Video	Distance	agglomerative clustering ward
2	Age/Groups	Age/City/Video	Index Country+CitizenNumber	agglomerative clustering average
3	Age/Notes	Age/City/Video	Free Index	agglomerative clustering complete
4	Age/Friends	Age/City/Video	Distance	DBSCAN
5	City/Photo	Age/City/Video	Index Country+CitizenNumber	agglomerative clustering ward
6	City/Groups	Age/City/Video	Free Index	agglomerative clustering average
7	City/Notes	Age/City/Video	Distance	agglomerative clustering complete
8	City/Friends	Age/City/Video	Index Country+CitizenNumber	DBSCAN
9	Age/Video	Age/City/Photo	Free Index	agglomerative clustering ward
10	Age/Groups	Age/City/Photo	Distance	agglomerative clustering average
11	Age/Notes	Age/City/Photo	Index Country+CitizenNumber	agglomerative clustering complete
12	Age/Friends	Age/City/Photo	Free Index	DBSCAN
13	City/Video	Age/City/Photo	Distance	agglomerative clustering ward
14	City/Groups	Age/City/Photo	Index Country+CitizenNumber	agglomerative clustering average
15	City/Notes	Age/City/Photo	Free Index	agglomerative clustering complete
16	City/Friends	Age/City/Photo	Distance	DBSCAN
17	Age/Photo	Age/City/Groups	Index Country+CitizenNumber	agglomerative clustering ward
18	Age/Video	Age/City/Groups	Free Index	agglomerative clustering average
19	Age/Notes	Age/City/Groups	Distance	agglomerative clustering complete
20	Age/Friends	Age/City/Groups	Index Country+CitizenNumber	DBSCAN
21	City/Photo	Age/City/Groups	Free Index	agglomerative clustering ward
22	City/Video	Age/City/Groups	Distance	agglomerative clustering average
23	City/Notes	Age/City/Groups	Index Country+CitizenNumber	agglomerative clustering complete
24	City/Friends	Age/City/Groups	Free Index	DBSCAN
25	Age/Groups	Age/City/Notes	Distance	agglomerative clustering ward
26	Age/Photo	Age/City/Notes	Index Country+CitizenNumber	agglomerative clustering average
27	Age/Video	Age/City/Notes	Free Index	agglomerative clustering complete
28	Age/Friends	Age/City/Notes	Distance	DBSCAN
29	City/Photo	Age/City/Notes	Index Country+CitizenNumber	agglomerative clustering ward
30	City/Video	Age/City/Notes	Free Index	agglomerative clustering average
31	City/Friends	Age/City/Notes	Distance	agglomerative clustering complete
32	City/Groups	Age/City/Notes	Index Country+CitizenNumber	DBSCAN
33	Age/Groups	Age/City/Friends	Free Index	agglomerative clustering ward
34	Age/Photo	Age/City/Friends	Distance	agglomerative clustering average

35	Age/Video	Age/City/Friends	Index Country+CitizenNumber	agglomerative clustering complete
36	Age/Notes	Age/City/Friends	Free Index	DBSCAN
37	City/Photo	Age/City/Friends	Distance	agglomerative clustering ward
38	City/Video	Age/City/Friends	Index Country+CitizenNumber	agglomerative clustering average
39	City/Notes	Age/City/Friends	Free Index	agglomerative clustering complete
40	City/Groups	Age/City/Friends	Distance	DBSCAN
41	Age/City	City/Friends/Notes	Index Country+CitizenNumber	agglomerative clustering ward
42	Age/City	City/Friends/Groups	Free Index	agglomerative clustering average
43	Age/City	City/Friends/Photos	Distance	agglomerative clustering complete
44	Age/City	City/Friends/Video	Index Country+CitizenNumber	DBSCAN
45	Age/Groups	Age/City/Notes	Distance	agglomerative clustering ward
46	Age/Photo	Age/City/Notes	Index Country+CitizenNumber	DBSCAN
47	Age/Video	Age/City/Notes	Free Index	agglomerative clustering ward
48	Age/Friends	Age/City/Notes	Distance	agglomerative clustering average
49	City/Photo	Age/City/Notes	Index Country+CitizenNumber	agglomerative clustering complete
50	City/Video	Age/City/Notes	Free Index	DBSCAN

Отследить корреляцию между возрастом, местом проживания и активностью профиля/ или между активностью профиля и группами, в которых пользователь состоит.

Face Labels,
OpenCV – face detection and comparison

To be continued …