Информатика. Задания на кластеризацию. Повторение материала.

1 Задача

Учёный решил провести кластеризацию некоторого множества звёзд по их расположению на карте звёздного неба. Кластер звёзд – это набор звёзд (точек) на графике, каждая из которых находится от хотя бы одной другой звезды на расстоянии не более R условных единиц. Каждая звезда обязательно принадлежит только одному из кластеров. Истинный центр кластера, или центроид, – это одна из звёзд на графике, сумма расстояний от которой до всех остальных звёзд кластера минимальна. Под расстоянием понимается расстояние Евклида между двумя точками $A(x1, y1)$ и $B(x2, y2)$ на плоскости, которое вычисляется по формуле:

$$d(A, B) = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2}$$

Аномалиями назовём точки, находящиеся на расстоянии более одной условной единицы от точек кластеров. При расчётах аномалии учитывать не нужно.

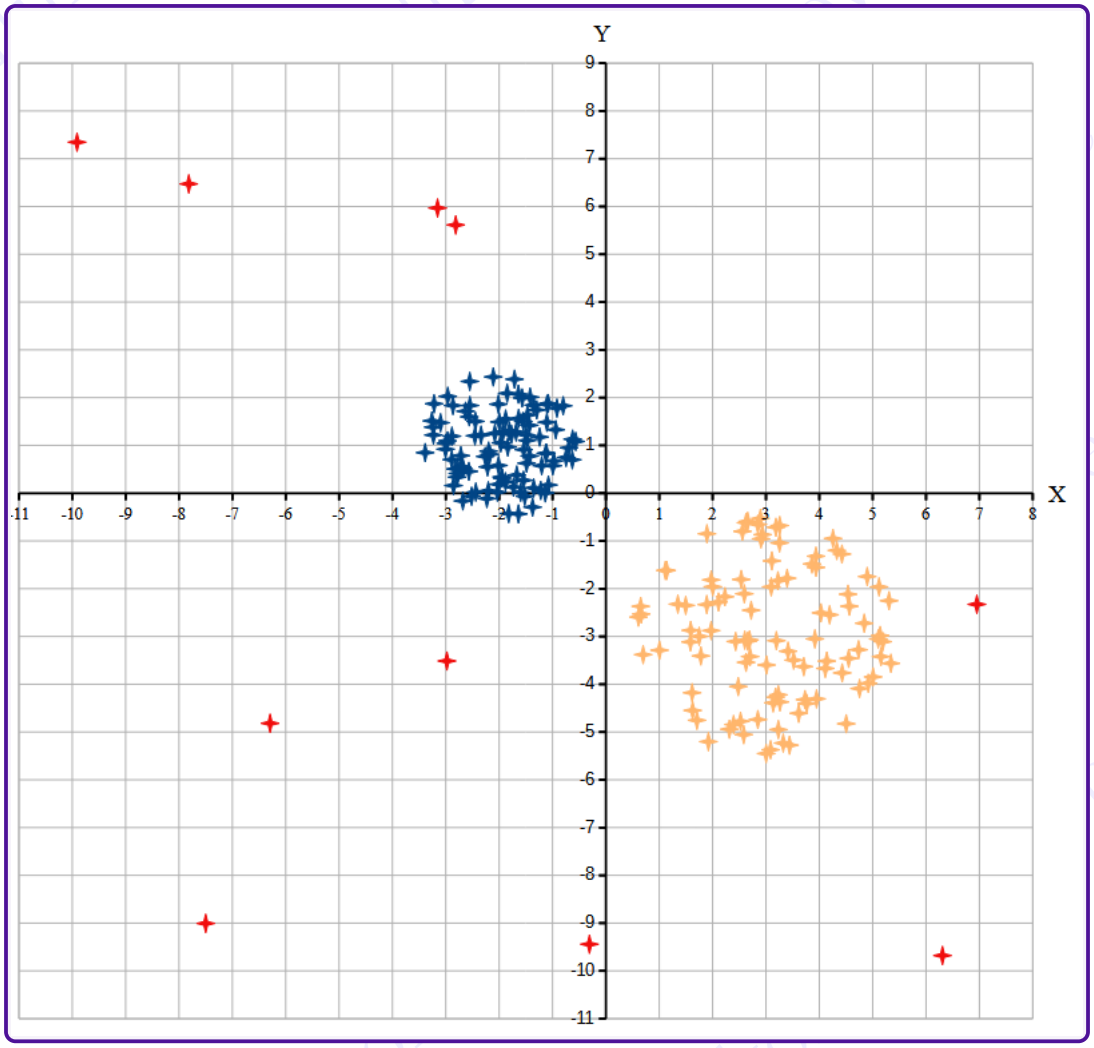
В файле А хранятся данные о звёздах двух кластеров, где $R = 0,5$ для каждого кластера. В каждой строке записана информация о расположении на карте одной звезды: сначала координата x, затем координата y. Значения даны в условных единицах, которые представлены вещественными числами. Известно, что количество звёзд не превышает 2500.

В файле Б хранятся данные о звёздах четырех кластеров, где $R = 0,2$ для каждого кластера. Известно, что количество звёзд не превышает 10 000. Структура хранения информации о звездах в файле Б аналогична файлу А.

Для каждого файла определите координаты центра каждого кластера, затем вычислите два числа: P_x — среднее арифметическое абсцисс центров кластеров, и P_y – среднее арифметическое ординат центров кластеров.

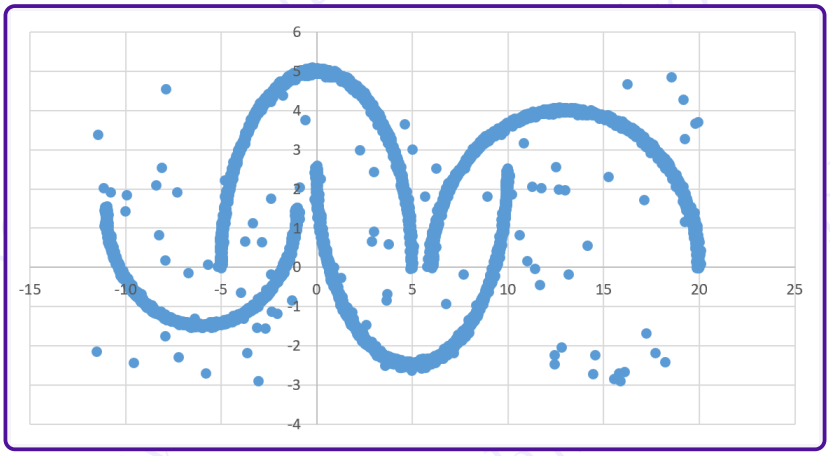
В ответе запишите четыре числа через пробел: сначала целую часть произведения $P_x \cdot 100$ для файла А, затем $P_y \cdot 100$ для файла А, далее целую часть $P_x \cdot 100$ для файла Б и $P_y \cdot 100$ для файла Б. Возможные данные одного из файлов иллюстрированы графиком.

Внимание! График приведён в иллюстративных целях для произвольных значений, не имеющих отношения к заданию. Для выполнения задания используйте данные из прилагаемого файла.



Решение:

Открываем файл Excel, выделяем полностью два столбца X и Y, переходим в раздел «Вставка», выбираем точечную диаграмму. Получили кластеры:



Как будет работать программа?

Выбираем случайным образом звезду и начинаем искать близлежащие к ней, далее ищем все ближайшие к этим звездам, и так далее. Каждая звезда будет принадлежать какому-либо кластеру (если ближайших звезд нет, кластер состоит из одной звезды). Если кластер состоит меньше, чем из двух звезд, он рассматриваться не будет.

```python
from math import dist

f = open('27-b.txt') # Открываем файл
s = f.readline() # Считываем первую строку с названием столбцов
# Записываем координаты звезд
a = [list(map(float, i.replace(',','.').split())) for i in f]
cl = [] # Создаем массив для кластеров
R = 0.2
while a: # Пока в списке 'a' есть звезды
    cl.append([a.pop(0)]) # Создаем новый кластер
    for i in cl[-1]:
        for j in a:
            if dist(i, j) <= R: # Если рядом есть звезда
                cl[-1].append(j) # Добавляем звезду в кластер
                a.remove(j) # Удаляем ее из исходного списка
star_x = star_y = 0 # Искомые средн. арифметические
for i in cl: # Проход по кластерам
    if len(i) > 2: # Если элементов больше, чем 2 (не выброс)
        mn = 10**100 # Минимальное расстояние
        for k in i: # Для каждой звезды кластера
            star = k # Координаты звезды, от которой ищем расстояние
            sm = 0 # Сумма расстояний
            for j in i:
                sm += dist(star, j) # Увеличиваем сумму
            if sm < mn: # Если сумма расстояний меньше минимума
                mn = sm # Обновляем mn
                center = star # Обновляем центр
        star_x += center[0] # Добавляем координату 'x' найденного центроида
        star_y += center[1] # Добавляем координату 'y' найденного центроида
print(int((star_x / 4) * 100), int((star_y / 4) * 100))
```

Ответ для файла *B*: 315 127