27 (высокий уровень, время – 35 мин)

Тема: Анализ данных. Кластеризация.

Что проверяется:

Умение выполнять последовательность решения задач анализа данных: сбор первичных данных, очистка и оценка качества данных, выбор и построение модели, преобразование данных, визуализация данных, интерпретация результатов.

4.1. Анализ данных. Основные задачи анализа данных: прогнозирование, классификация, кластеризация, анализ отклонений. Последовательность решения задач анализа данных: сбор первичных данных, очистка и оценка качества данных, выбор и/или построение модели, преобразование данных, визуализация данных, интерпретация результатов. Программные средства и интернет-сервисы для обработки и представления данных. Большие данные. Машинное обучение.

2.2. Умение классифицировать основные задачи анализа данных (прогнозирование, классификация, кластеризация, анализ отклонений); понимать последовательность решения задач анализа данных: сбор первичных данных, очистка и оценка качества данных, выбор и/или построение модели, преобразование данных, визуализация данных, интерпретация результатов.

Что нужно знать:

- как прочитать данные из файла
- принципы кластеризации

Пример задания (И. Воропаев):

P-01. Условие полностью совпадает с условием задачи P-00 (из демо-варианта 2025 года), но «расстояние» между точками вычисляется как **квадрат обычного евклидового расстояния**

(здесь нет квадратного корня!!!).

В этом случае можно организовать перебор так, чтобы он имел **линейную сложность**, а не квадратичную. Это может быть важно, если точек очень много (скажем, 1 000 000) и алгоритм с квадратичной сложностью не закончится за разумное время.

Решение (И. Воропаев, К. Поляков):

1) Предположим, что мы уже разделили все точки на независимые кластеры. Рассмотрим поиск центроида для одного кластера.
2) Пусть точки кластера имеют координаты (x_i, y_i) для $i = 1, \ldots, N$. Вычислим сумму «расстояний» от каждой из этих точек до первой точки с координатами (x_i, y_i).

$$D_1 = \sum_{i=1}^{N} (x_1 - x_i)^2 + (y_1 - y_i)^2 .$$

3) Раскроем скобки

$$D_1 = \sum_{i=1}^{N} (x_1^2 - 2x_1 x_i + x_i^2 + y_1^2 - 2y_1 y_i + y_i^2)$$

4) Разделим сумму на части и вынесем за знаки суммы сомножители, которые не зависят от индекса i:

$$D_1 = \sum_{i=1}^{N} x_1^2 - 2x_1 \sum_{i=1}^{N} x_i + \sum_{i=1}^{N} x_i^2 + \sum_{i=1}^{N} y_1^2 - 2y_1 \sum_{i=1}^{N} y_i + \sum_{i=1}^{N} y_i^2$$

5) Перегруппируем слагаемые:

$$D_1 = \sum_{i=1}^{N} (x_1^2 + y_1^2) - 2\left(x_1 \sum_{i=1}^{N} x_i + y_1 \sum_{i=1}^{N} y_i \right) + \sum_{i=1}^{N} (x_i^2 + y_i^2)$$

6) Чтобы упростить запись, введём обозначения

$$A_1 = \sum_{i=1}^{N}(x_1^2 + y_1^2) = N(x_1^2 + y_1^2)$$

$$S_x = \sum_{i=1}^{N} x_i, \quad S_y = \sum_{i=1}^{N} y_i$$

$$Q = \sum_{i=1}^{N}(x_i^2 + y_i^2)$$

7) при этом получается

$$D_1 = A_1 - 2\left(x_1 S_x + y_1 S_y\right) + Q$$

8) Заметим следующее:
   а) для вычисления A_1 не требуется цикл;
   б) значение S_x, S_y и Q не зависят от номера точки, до которой вычисляется суммарное расстояние; т. е. они могут быть вычислены заранее за линейное время;
   в) при поиске центроида значение Q можно не учитывать, оно добавляется как слагаемое ко всем суммарным расстояниям

9) Функция на Python, которая находит центроид для кластера (К. Поляков):

```
def find_centre(cluster):
  Sx = sum( x for x, y in cluster )
  Sy = sum( y for x, y in cluster )
  Q =  sum( x*x + y*y for x, y in clusters )
  N = len( cluster )
  minSumDist = float('inf')
  for pCenter in cluster:
    xc, yc = pCenter
    Ac = N*(xc*xc + yc*yc)
    sumDist = Ac - 2*(xc*Sx + yc*Sy)
    if sumDist < minSumDist:
      minSumDist = sumDist
      center = pCenter
  return center
```

10) Функция на Python, которая находит центроид для кластера (И. Воропаев):

```
def find_centre(cluster):
    min_sum = 10 ** 10
    centre = [-1, -1]
    n = len(cluster)
    sum_x = sum([x for x, y in cluster])
    sum_y = sum([y for x, y in cluster])
    sum_kv_x_y = sum([x**2 + y**2 for x, y in cluster])
    for x, y in cluster:
        cur_sum = ((x**2+y**2) * n) + sum_kv_x_y - 2*x*sum_x - 2*y*sum_y
        if cur_sum < min_sum:
            min_sum = cur_sum
            centre = [x, y]
    return centre
```

Ещё пример задания:

P-00. (**демо-2025**) *Учёный решил провести кластеризацию некоторого множества звёзд по их*

расположению на карте звёздного неба. Кластер звёзд – это набор звёзд (точек) на графике, лежащий внутри прямоугольника высотой H и шириной W. 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(x_1, y_1) и B(x_2, y_2) на плоскости, которое вычисляется по формуле:* $d(A,B) = \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2}$.

Входные данные

В файле А хранятся данные о звёздах двух кластеров, где H=3, W=3 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. Известно, что количество звёзд не превышает 1000.

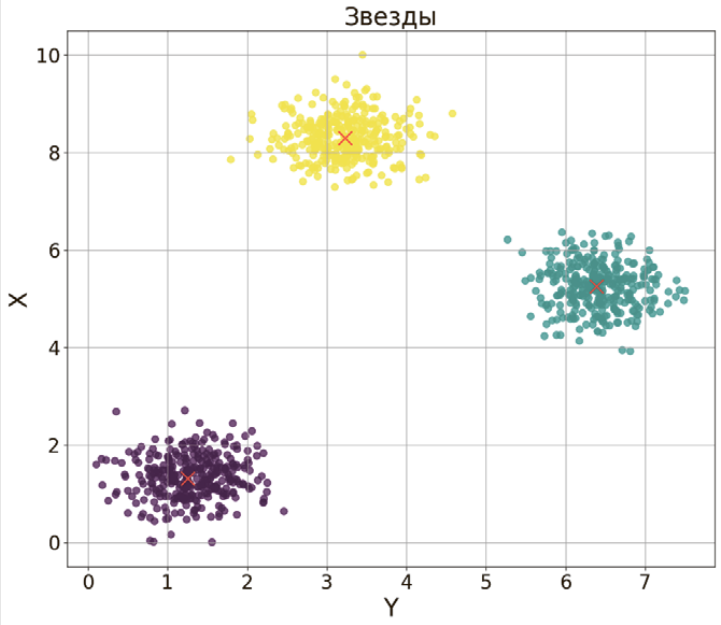
В файле Б хранятся данные о звёздах трёх кластеров, где H=3, W=3 для каждого кластера. Известно, что количество звёзд не превышает 10 000. Структура хранения информации о звездах в файле Б аналогична файлу А.

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

Выходные данные

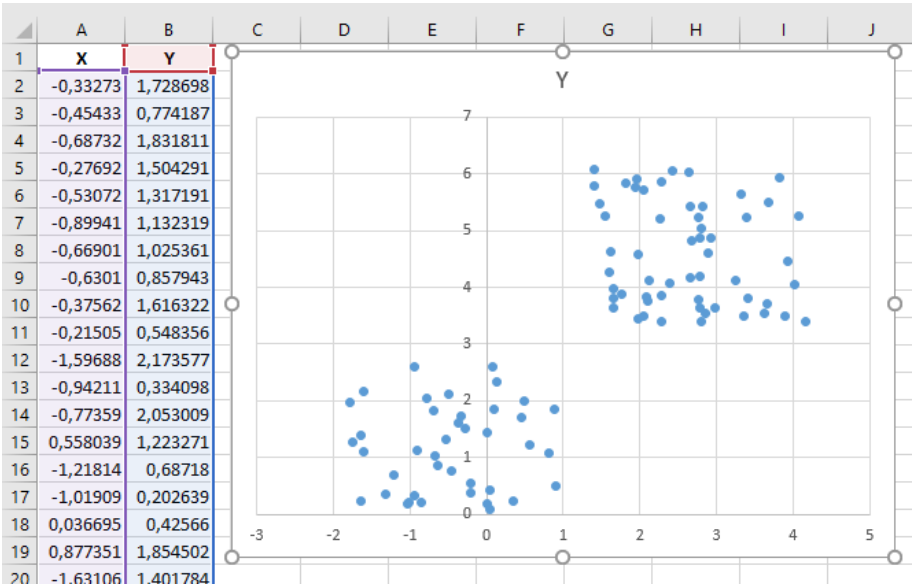
В ответе запишите четыре числа: в первой строке сначала целую часть произведения Px × 10 000, затем целую часть произведения Py × 10 000 для файла А, во второй строке – аналогичные данные для файла Б.

Возможные данные одного из файлов иллюстрированы графиком.



Решение:

11) Для каждой задачи (А и Б) даны два файла: один текстовый, второй – с теми же данными в виде электронной таблицы. Используя электронную таблицу, мы можем сразу определить границы кластеров (визуально).
12) Откроем электронную таблицу для задачи А и построим точечную диаграмму.



	X	Y
2	-0,33273	1,728698
3	-0,45433	0,774187
4	-0,68732	1,831811
5	-0,27692	1,504291
6	-0,53072	1,317191
7	-0,89941	1,132319
8	-0,66901	1,025361
9	-0,6301	0,857943
10	-0,37562	1,616322
11	-0,21505	0,548356
12	-1,59688	2,173577
13	-0,94211	0,334098
14	-0,77359	2,053009
15	0,558039	1,223271
16	-1,21814	0,68718
17	-1,01909	0,202639
18	0,036695	0,42566
19	0,877351	1,854502
20	-1,63106	1,401784

Видно, что здесь можно выделить два кластера, которые разделяются прямой $x = 1$. Все точки слева от этой прямой нужно отнести к кластеру 0 (в программе нам удобнее сделать нумерацию кластеров с нуля), а все точки с $x > 1$ – к кластеру 1. Напишем функцию, которая будет определяет номер кластера по координатам:

```
def findClusterNo( x, y ):
  return 0 if x < 1 else 1
```

13) Обозначим через K количество кластеров (K = 2 для задачи А и K = 3 для задачи Б).

```
K = 2
```

14) Данные о кластерах будет хранить в списке:

```
clusters = [ [], [] ]
# или clusters = [ [] for i in range(K) ]
```

15) Прочитаем данные из файла, сразу распределяя их по кластерам. При этом нужно учесть, что первая строка файла содержит заголовки (буквы X и Y), а в качестве разделителя целой и дробной части чисел использована запятая, которую нужно заменить на точку для нормальной работы функции `float`:

```
with open('27-A.txt') as F:
  F.readline()  # читаем и пропускаем заголовки
  for s in F:
    x, y = s.replace(',','.').split()
    x, y = float(x), float(y)
    clusterNo = findClusterNo( x, y )
    clusters[clusterNo].append( (x, y) )
```

Теперь `clusters[0]` – это список пар `(x, y)` (кортежей), в которых записаны координаты всех точек кластера 0, а `clusters[1]` – такие же пары для точек кластера 1.

16) Следующий этап – в каждом кластере найти *центроид* – точку, для которой сумма расстояний до других точек кластера минимальна. Напишем функцию, которая вычисляет расстояние между точками, данные о которых передаются в виде двух кортежей:

```
def dist( p1, p2 ):
  return ((p1[0] - p2[0])**2 + (p1[1] - p2[1])**2) ** 0.5
```

или так

```
import math
def dist( p1, p2 ):
  return math.hypot( p1[0] - p2[0], p1[1] - p2[1] )
```

17) Создаём пустой список для хранения координат центров кластеров:

```
centers = []
```

18) Перебираем в цикле все кластеры, для каждого определяем центроид и добавляем его координаты в список `centers`:

```
for k in range(K):
  ...
  centers.append( center )
```

19) В теле цикла перебираем все точки кластера; для каждой из них находим сумму расстояний до остальных точек `sumDist` и выбираем такую точку `center`, для которой это расстояние наименьшее

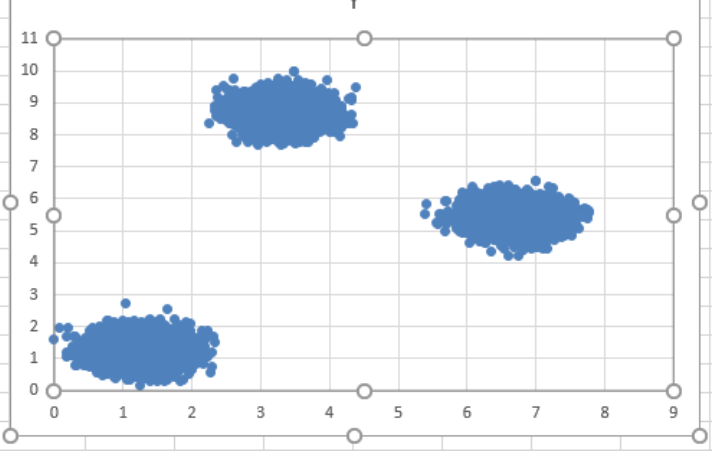
```
  minSumDist = float('inf')
  for pCenter in clusters[k]:
    sumDist = sum( dist(pCenter,p)
                   for p in clusters[k] )
    if sumDist < minSumDist:
      minSumDist = sumDist
      center = pCenter
```

20) Остаётся вычислить среднее значение координат центроидов (по каждой оси), умножить их на 10000 и найти целые части этих чисел:

```
sumX, sumY = 0, 0
for k in range(K):
  sumX += centers[k][0]
  sumY += centers[k][1]
print( int(sumX/K*10000), int(sumY/K*10000) )
```

21) Для решения задачи Б построим точечную диаграмму по данным второго файла:

	A	B
1	X	Y
2	3,161941	8,358027
3	6,177517	4,681664
4	3,51402	8,835359
5	6,3882	4,800801
6	6,892524	5,650152
7	6,623268	5,909666
8	3,334963	8,991917
9	3,594275	8,635256
10	1,41738	1,339414
11	6,784158	5,087835
12	3,965927	8,439298
13	6,521812	5,729428
14	6,522029	5,409555
15	6,629941	5,323633
16	1,464996	1,376862
17	6,763223	5,396183

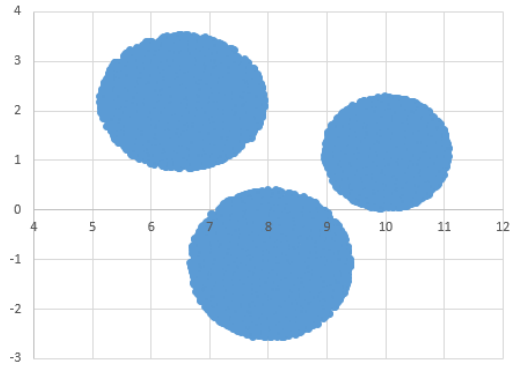


22) Единственная сложность – написать условия, разделяющие эти кластеры. Заметим, что самый левый кластер отделяется от двух других прямой $y = 3$, а два правых разделяются прямой $x = 5$. Таким образом, в программе нужно только изменить значение K и функцию `findClusterNo`:

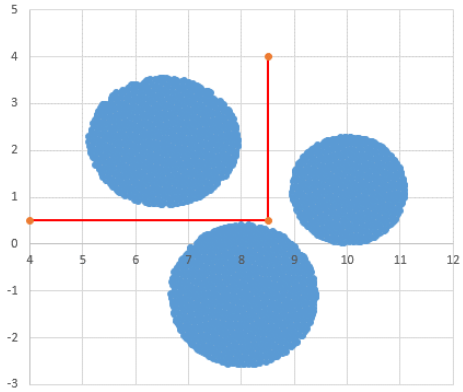
```
K = 3  # количество кластеров
def findClusterNo( x, y ):
  return 0 if y < 3 else \
         1 if x < 5 else \
         2
```

Остальные части программы не изменяются.

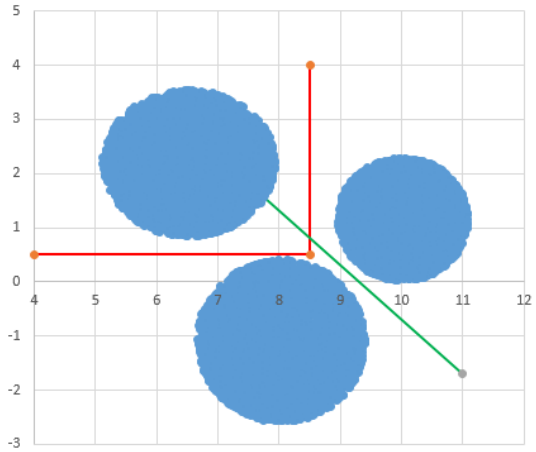
23) К сожалению, не всегда удаётся разделить кластеры, используя только одну координату (т. е. только горизонтальные и вертикальные прямые). Предположим, что распределение точек на плоскости выглядит примерно так:



Здесь легче всего сначала отделить кластер, который находится на диаграмме слева вверху – это можно сделать прямыми $x = 8{,}5$ и $y = 0{,}5$.



24) Затем разделяем оставшиеся два кластера прямой $y = -x + 9{,}3$ (отрезок этой прямой показан зелёным цветом):



Коэффициенты уравнения прямой подбираются методом проб и ошибок в табличном процессоре.

25) В результате функция, определяющая номер кластера для точки с координатами (x, y) выглядит так:

```
def findClusterNo( x, y ):
  return 0 if y > 0.5 and x < 8.5 else \
         1 if y > -x+9.3 else \
         2
```

Решение с помощью PascalABC.NET (М. Крючков):

Для решения этой задачи требуется установить последнюю версию PascalABC.Net.

1) Открыть файл 27-A.txt в текстовом редакторе (типа Блокнота). **Удалить верхнюю строку** заголовков, т.к. она нарушает структуру хранения данных. Сохранить файл с именем **27A1.txt**, используя английскую букву "A".
2) Подключить модуль turtle. Он нужен для визуального анализа данных.

```
## uses Turtle;
```

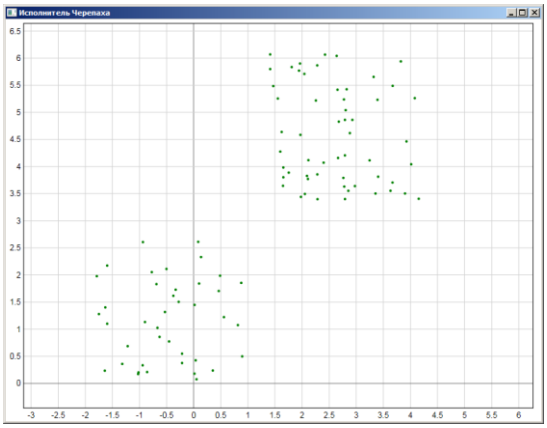
3) Загрузить все координаты в матрицу.

   Разделитель в файлах - десятичная запятая. Будем указывать ее для преобразования текста в вещественные числа.

```
var t := Matrbyrow(ReadAllText('27a1.txt').ToReals(',').Batch(2));
```

4) Вывести на экран все точки.

```
DrawPoints(t.col(0), t.col(1));
```



5) Замечаем, что на графике два кластера, которые разделены по абсциссе координатой 3. Можно использовать колесико мыши для изменения масштаба. Зажатая левая кнопка мыши позволяет двигать сетку. Это гораздо удобнее, чем в табличном редакторе Excel.

6) Сохранить два кластера в отдельные запросы.

```
var KL1 := t.Rows.Where(\(x,y) -> y>3);
var KL2 := t.Rows.Where(\(x,y) -> y<3);
```

7) Для первого кластера находим центроид с помощью обычного цикла.

```
var (k1x, k1y):=(0.0, 0.0);
var sumD := real.MaxValue;
foreach var (cx,cy) in KL1 do begin
    var sum1 := KL1.Sum(\(x,y) -> (cx-x)**2 + (cy-y)**2);
    if sum1 < sumD then
       (sumD, k1x, k1y):=(sum1, cx, cy);
end;
```

8) Для второго кластера другой способ – с помощью LINQ.

```
var (k2x,k2y) := KL2.OrderBy(\(cx,cy) ->
                      KL2.Sum(\(x,y) -> (cx-x)**2+(cy-y)**2)).First;
```

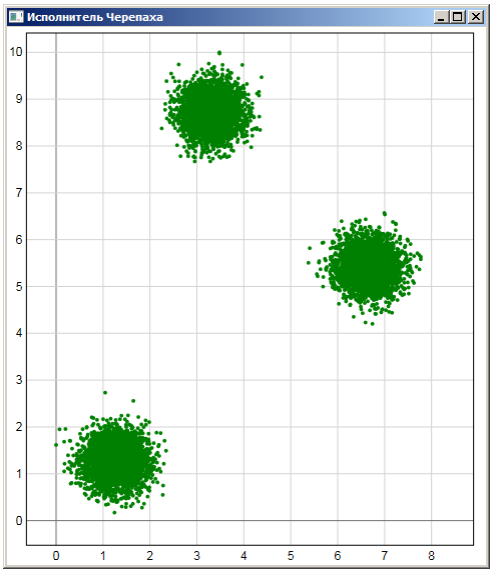
   Всего лишь одна строка вместо шести.

9) Вывести ответ к задаче.

```
Println( floor((k1x+k2x)/2*10000), floor((k1y+k2y)/2*10000) );
```

10) Для файла Б поступаем аналогично: изменить и сохранить в файл 27B1.txt.

```
## uses Turtle;
var t := Matrbyrow(ReadAllText('27b1.txt').ToReals(',').Batch(2));
DrawPoints(t.col(0), t.col(1));
```



11) Три кластера определяем визуально по значениям абсцисс. Первый – от 0 до 3, второй – от 4 до 7, третий – от 7 до 10. В случае более сложной пересекающейся структуры запросы корректируются несложно с указанием обеих координат. Ни одна точка не лежит на прямой x = 7. Поэтому указываем 7 в двух местах без проблем. Сохраняем три кластера. Будем нумеровать их от ноля, чтобы дальше не напутать с индексацией массивов.

```
var KL0 := t.Rows.Where(\(x,y) -> y in 0..3);
var KL1 := t.Rows.Where(\(x,y) -> y in 4..7);
var KL2 := t.Rows.Where(\(x,y) -> y in 7..10);
```

12) Находим все центроиды. Заводить массивы ради трёх координат или сохранить в отдельные переменные – дело вкуса программиста. В этом варианте сделаем массивы вещественных чисел.

```
var (kx, ky) := (|0.0|*3, |0.0|*3);
(kx[0],ky[0]) := KL0.OrderBy(\(cx,cy) ->
                 KL0.Sum(\(x,y) -> sqrt((cx-x)**2+(cy-y)**2))).First;
(kx[1],ky[1]) := KL1.OrderBy(\(cx,cy) ->
                 KL1.Sum(\(x,y) -> sqrt((cx-x)**2+(cy-y)**2))).First;
(kx[2],ky[2]) := KL2.OrderBy(\(cx,cy) ->
                 KL2.Sum(\(x,y) -> sqrt((cx-x)**2+(cy-y)**2))).First;
```

13) Вывести ответ для массивов проще.

```
Println(floor(kx.Average*10000), floor(ky.Average*10000));
```

Это **медленный алгоритм** ($O(n^2)$), т.к. для каждой точки в кластере нужно найти сумму расстояний до остальных в ее кластере. Для файла Б на 10000 точек с тремя кластерами на старом ноутбуке получилось следующее время выполнения программ: Python – 53 секунды, PascalABC.Net – 24 секунды. Алгоритм полного перебора для 100000 «звезд» будет работать больше 48 минут.

Задачи для тренировки:

1) Учёный решил провести кластеризацию некоторого множества звёзд по их расположению на карте звёздного неба… (см. условие задачи из демо-варианта 2025 года).
   Исходные данные находятся в файлах `27-1a.txt` и `27-1b.txt`.
2) Учёный решил провести кластеризацию некоторого множества звёзд по их расположению на карте звёздного неба… (см. условие задачи из демо-варианта 2025 года).
   Исходные данные находятся в файлах `27-2a.txt` и `27-2b.txt`.
3) Учёный решил провести кластеризацию некоторого множества звёзд по их расположению на карте звёздного неба… (см. условие задачи из демо-варианта 2025 года).
   Исходные данные находятся в файлах `27-3a.txt` и `27-3b.txt`.
4) Учёный решил провести кластеризацию некоторого множества звёзд по их расположению на карте звёздного неба… (см. условие задачи из демо-варианта 2025 года).
   Исходные данные находятся в файлах `27-4a.txt` и `27-4b.txt`.
5) Учёный решил провести кластеризацию некоторого множества звёзд по их расположению на карте звёздного неба… (см. условие задачи из демо-варианта 2025 года).
   Исходные данные находятся в файлах `27-5a.txt` и `27-5b.txt`.
6) Учёный решил провести кластеризацию некоторого множества звёзд по их расположению на карте звёздного неба… (см. условие задачи из демо-варианта 2025 года).
   Исходные данные находятся в файлах `27-6a.txt` и `27-6b.txt`.
7) Учёный решил провести кластеризацию некоторого множества звёзд по их расположению на карте звёздного неба… (см. условие задачи из демо-варианта 2025 года).
   Исходные данные находятся в файлах `27-7a.txt` и `27-7b.txt`.
8) Учёный решил провести кластеризацию некоторого множества звёзд по их расположению на карте звёздного неба… (см. условие задачи из демо-варианта 2025 года).
   Исходные данные находятся в файлах `27-8a.txt` и `27-8b.txt`.
9) Учёный решил провести кластеризацию некоторого множества звёзд по их расположению на карте звёздного неба… (см. условие задачи из демо-варианта 2025 года).
   Исходные данные находятся в файлах `27-9a.txt` и `27-9b.txt`.
10) Учёный решил провести кластеризацию некоторого множества звёзд по их расположению на карте звёздного неба… (см. условие задачи из демо-варианта 2025 года).
    Исходные данные находятся в файлах `27-10a.txt` и `27-10b.txt`.
11) Учёный решил провести кластеризацию некоторого множества звёзд по их расположению на карте звёздного неба… (см. условие задачи из демо-варианта 2025 года).
    Исходные данные находятся в файлах `27-11a.txt` и `27-11b.txt`.
12) Учёный решил провести кластеризацию некоторого множества звёзд по их расположению на карте звёздного неба… (см. условие задачи из демо-варианта 2025 года).
    Исходные данные находятся в файлах `27-12a.txt` и `27-12b.txt`.
13) Учёный решил провести кластеризацию некоторого множества звёзд по их расположению на карте звёздного неба… (см. условие задачи из демо-варианта 2025 года).
    Исходные данные находятся в файлах `27-13a.txt` и `27-13b.txt`.
14) Учёный решил провести кластеризацию некоторого множества звёзд по их расположению на карте звёздного неба… (см. условие задачи из демо-варианта 2025 года).
    Исходные данные находятся в файлах `27-14a.txt` и `27-14b.txt`.
15) Учёный решил провести кластеризацию некоторого множества звёзд по их расположению на карте звёздного неба… (см. условие задачи из демо-варианта 2025 года).
    Исходные данные находятся в файлах `27-15a.txt` и `27-15b.txt`.

16) Учёный решил провести кластеризацию некоторого множества звёзд по их расположению на карте звёздного неба… (см. условие задачи из демо-варианта 2025 года).
Исходные данные находятся в файлах `27-16a.txt` и `27-16b.txt`.

17) Учёный решил провести кластеризацию некоторого множества звёзд по их расположению на карте звёздного неба… (см. условие задачи из демо-варианта 2025 года).
Исходные данные находятся в файлах `27-17a.txt` и `27-17b.txt`.

18) Учёный решил провести кластеризацию некоторого множества звёзд по их расположению на карте звёздного неба… (см. условие задачи из демо-варианта 2025 года).
Исходные данные находятся в файлах `27-18a.txt` и `27-18b.txt`.

19) Учёный решил провести кластеризацию некоторого множества звёзд по их расположению на карте звёздного неба… (см. условие задачи из демо-варианта 2025 года).
Исходные данные находятся в файлах `27-19a.txt` и `27-19b.txt`.

20) Учёный решил провести кластеризацию некоторого множества звёзд по их расположению на карте звёздного неба… (см. условие задачи из демо-варианта 2025 года).
Исходные данные находятся в файлах `27-20a.txt` и `27-20b.txt`.

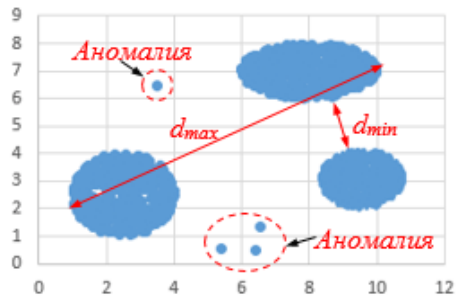
21) (**В. Шубинкин**) При проведении эксперимента заряженные частицы попадают на чувствительный экран размером 12 на 9 условных единиц. При попадании каждой частицы на экран в протоколе фиксируются координаты попадания в условных единицах. При анализе результатов выделяют кластеры – группы точек на экране, в которые попали частицы. Размер каждого кластера – не более W условных единиц в ширину и не более H условных единиц в высоту. Каждая точка принадлежит только одному кластеру. Минимальное (максимальное) расстояние между кластерами – это минимальное (максимальное) расстояние между двумя точками, одна из которых принадлежит одному кластеру, а вторая – другому. Расстояние между двумя точками A(x_1,y_1) и B(x_2,y_2) вычисляется по формуле $d(A,B)=\sqrt{(x_1-x_2)^2+(y_1-y_2)^2}$.

Аномалиями назовём точки, находящиеся на расстоянии более одной условной единицы от точек кластеров. Аномалии следует исключить при проведении расчётов.

В файле А хранятся данные о точках **двух** кластеров, где W=4, H=4 для каждого кластера. В каждой строке записана информация о расположении одной точки: сначала координата x, затем координата y. Значения даны в условных единицах. Известно, что общее количество точек не превышает 1000.

В файле Б, который имеет ту же структуру, что и файл А, хранятся данные о точках **трёх** кластеров, где W=3, H=3 для каждого кластера. Известно, что общее количество точек не превышает 10 000.

Для каждого файла определите минимальное d_{min} и максимальное d_{max} расстояния между двумя кластерами. В ответ запишите 4 числа: в первой строке целую часть произведения d_{min} × 10 000, затем целую часть произведения d_{max} × 10 000 для файла А, во второй строке – аналогичные данные для файла Б.



Исходные данные находятся в файлах `27-21a.txt` и `27-21b.txt`.

22) (**В. Шубинкин**) В ходе эксперимента были зафиксированы очаги радиации. Чтобы изучить данное явление, решили провести кластеризацию источников излучения. Кластер – это набор источников (точек) на графике, лежащий внутри прямоугольника высотой H и шириной W. Каждая точка

обязательно принадлежит только одному из кластеров. Истинный центр кластера, или **центроид**, – это одна из точек на графике, сумма расстояний от которой до всех остальных звёзд кластера минимальна. Под расстоянием понимается расстояние Евклида между двумя точками $A(x_1, y_1)$ и $B(x_2, y_2)$ на плоскости, которое вычисляется по формуле: $d(A,B) = \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2}$.

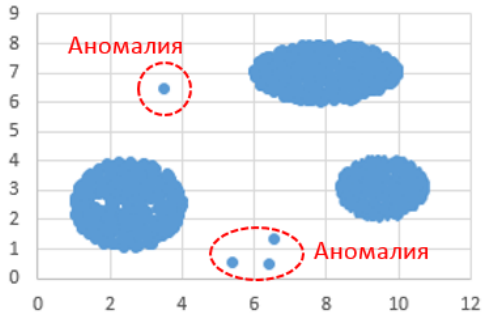
Аномалиями назовём точки, находящиеся на расстоянии более одной условной единицы от точек кластеров. Аномалии следует исключить при проведении расчётов.

В файле А хранятся данные о точках **двух** кластеров, где H=3, W=3 для каждого кластера. В каждой строке записаны координаты одной точки в условных единицах: сначала x, затем y. Известно, что количество точек не превышает 1000.

В файле Б той же структуры хранятся данные о точках **трёх** кластеров, где H=3, W=3 для каждого кластера; количество точек не превышает 10 000.

Для каждого файла определите координаты центра каждого кластера, затем вычислите два числа: Px – среднее арифметическое абсцисс центров кластеров, и Py – среднее арифметическое ординат центров кластеров. В ответе запишите четыре числа: в первой строке сначала целую часть произведения Px × 100 000, затем целую часть произведения Py × 100 000 для файла А, во второй строке – аналогичные данные для файла Б.

Возможные данные одного из файлов иллюстрированы графиком.



23) (**В. Ланская, Р. Ягафаров**) Учёный решил провести кластеризацию некоторого множества звёзд по их расположению на карте звёздного неба… (см. условие задачи из демо-варианта 2025 года). В файле Б хранятся данные о звёздах **четырех** кластеров, …

Исходные данные находятся в файлах `27-23a.txt` и `27-23b.txt`.

24) (**В. Ланская, Р. Ягафаров**) Шёл 2077 год. Ученому необходимо провести кластеризацию населенных пунктов двух больших районов на картах планет Информатикус и Алгоритмикус. Район (кластер) – это группа населенных пунктов, которые находятся внутри прямоугольника высотой H и шириной W. Каждый населенный пункт обязательно принадлежит только одному району. Столица района (или центроид) – это такой населенный пункт, сумма манхэттенских расстояний от которого до всех других населённых пунктов в кластере минимальна. Манхэттенское расстояние между двумя точками $A(x_1, y_1)$ и $B(x_2, y_2)$ вычисляется как сумма модулей разностей их координат: $d(A, B) = |x_2 - x_1| + |y_2 - y_1|$.

В файле А хранятся данные о населенных пунктах двух районов (кластеров) планеты Информатикус, где H = 3, W = 3 для каждого кластера. В каждой строке записаны координаты одного населенного пункта в условных единицах: сначала x, затем y. Известно, что количество звёзд не превышает 1000. В файле Б той же структуры хранятся данные о населенных пунктах трёх кластеров планеты Алгоритмикус, где H = 3, W = 3 для каждого кластера. Известно, что количество населенных пунктов не превышает 10 000.

Для каждого файла определите координаты центра каждого кластера, затем вычислите два числа: Px – среднее арифметическое абсцисс центров кластеров, и Py – среднее арифметическое ординат центров кластеров. В ответе запишите четыре числа: в первой строке сначала целую часть

произведения Px × 10 000, затем целую часть произведения Py × 10 000 для файла A, во второй строке – аналогичные данные для файла Б.

Исходные данные находятся в файлах 27-24a.txt и 27-24b.txt.

25) (**В. Ланская, Р. Ягафаров**) В городе X тестируется проект по оптимизации размещения кранов на складах. Оптимальное местоположение для крана (или центроид) будет таким, при котором сумма расстояний Чебышева от этого места до всех других точек на складе была минимальной. Расстояние Чебышева между двумя точками $A(x_1, y_1)$ и $B(x_2, y_2)$ вычисляется как максимум модулей разностей их координат: $d(A, B) = \max(|x_2 - x_1|, |y_2 - y_1|)$.

В файле A хранятся данные о двух складских комплексах (кластерах). Каждый комплекс имеет форму прямоугольника. Каждая строка файла содержит координаты одной точки на складе: сначала x, затем y. Количество точек в каждом комплексе не превышает 1000. В файле Б той же структуры хранятся данные о трёх кластерах, каждый из которых имеет вид прямоугольника размером H = 6 и W = 8. Количество точек в каждом комплексе не превышает 10 000.

Для каждого файла определите координаты центра каждого кластера, затем вычислите два числа: Px – среднее арифметическое абсцисс центров кластеров, и Py – среднее арифметическое ординат центров кластеров. В ответе запишите четыре числа: в первой строке сначала целую часть произведения Px × 10 000, затем целую часть произведения Py × 10 000 для файла A, во второй строке – аналогичные данные для файла Б.

Исходные данные находятся в файлах 27-25a.txt и 27-25b.txt.

26) (**М. Крючков**) В лесу выделено несколько мест (кластеров), где растёт много деревьев, предназначенных для вырубки После спиливания дерева его нужно доставить в точку сбора, которая совпадает с одним из деревьев кластера. Стоимость доставки определяется как расстояние от дерева до точки сбора, умноженное на высоту дерева. Под расстоянием понимается расстояние Евклида между двумя точками $A(x_1, y_1)$ и $B(x_2, y_2)$ на плоскости, которое вычисляется по формуле: $d(A,B) = \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2}$. В каждом кластере нужно найти оптимальную точку сбора (центроид), такую что суммарная стоимость доставки в это место всех спиленных деревьев данного кластера минимальна. **Аномалиями** назовём точки, находящиеся на расстоянии более 30 м от точек кластеров. При расчётах аномалии учитывать не нужно.

В файле A хранятся данные о двух кластерах. Каждый кластер имеет форму прямоугольника размером 100 × 100 м. Каждая строка файла содержит три характеристики одного дерева: координату x, затем координату y и затем высоту дерева. Количество деревьев в каждом кластере не превышает 1000. В файле Б той же структуры хранятся данные о трёх кластерах, каждый из которых имеет вид прямоугольника размером не более 100 × 200 м. Количество точек в каждом кластере не превышает 10 000.

Для каждого файла определите координаты центра каждого кластера, затем вычислите два числа: Px – среднее арифметическое абсцисс центров кластеров, и Py – среднее арифметическое ординат центров кластеров. В ответе запишите четыре числа: в первой строке сначала целую часть произведения Px × 100 000, затем целую часть произведения Py × 100 000 для файла A, во второй строке – аналогичные данные для файла Б.

Исходные данные находятся в файлах 27-26a.txt и 27-26b.txt.

27) Учёный решил провести кластеризацию некоторого множества звёзд по их расположению на карте звёздного неба… (см. условие задачи из демо-варианта 2025 года). **Аномалиями** назовём точки, находящиеся на расстоянии более одной условной единицы от точек кластеров. При расчётах аномалии учитывать не нужно.

Для каждого файла определите координаты центра каждого кластера, затем вычислите два числа: Px – среднее арифметическое абсцисс центров кластеров, и Py – среднее арифметическое ординат центров кластеров. В ответе запишите четыре числа: в первой строке сначала целую часть

произведения Px × 100 000, затем целую часть произведения Py × 100 000 для файла А, во второй строке – аналогичные данные для файла Б.

Исходные данные находятся в файлах 27-27a.txt и 27-27b.txt.

28) Учёный решил провести кластеризацию некоторого множества звёзд по их расположению на карте звёздного неба… (см. условие задачи из демо-варианта 2025 года). **Аномалиями** назовём точки, находящиеся на расстоянии более одной условной единицы от точек кластеров. При расчётах аномалии учитывать не нужно.

Для каждого файла определите координаты центра каждого кластера, затем вычислите два числа: Px – среднее арифметическое абсцисс центров кластеров, и Py – среднее арифметическое ординат центров кластеров. В ответе запишите четыре числа: в первой строке сначала целую часть произведения Px × 100 000, затем целую часть произведения Py × 100 000 для файла А, во второй строке – аналогичные данные для файла Б.

Исходные данные находятся в файлах 27-28a.txt и 27-28b.txt.

29) Учёный решил провести кластеризацию некоторого множества звёзд по их расположению на карте звёздного неба… (см. условие задачи из демо-варианта 2025 года). **Аномалиями** назовём точки, находящиеся на расстоянии более одной условной единицы от точек кластеров. При расчётах аномалии учитывать не нужно.

Для каждого файла определите координаты центра каждого кластера, затем вычислите два числа: Px – среднее арифметическое абсцисс центров кластеров, и Py – среднее арифметическое ординат центров кластеров. В ответе запишите четыре числа: в первой строке сначала целую часть произведения Px × 100 000, затем целую часть произведения Py × 100 000 для файла А, во второй строке – аналогичные данные для файла Б.

Исходные данные находятся в файлах 27-29a.txt и 27-29b.txt.

30) Учёный решил провести кластеризацию некоторого множества звёзд по их расположению на карте звёздного неба… (см. условие задачи из демо-варианта 2025 года). **Аномалиями** назовём точки, находящиеся на расстоянии более одной условной единицы от точек кластеров. При расчётах аномалии учитывать не нужно.

Для каждого файла определите координаты центра каждого кластера, затем вычислите два числа: Px – среднее арифметическое абсцисс центров кластеров, и Py – среднее арифметическое ординат центров кластеров. В ответе запишите четыре числа: в первой строке сначала целую часть произведения Px × 100 000, затем целую часть произведения Py × 100 000 для файла А, во второй строке – аналогичные данные для файла Б.

Исходные данные находятся в файлах 27-30a.txt и 27-30b.txt.

31) Учёный решил провести кластеризацию некоторого множества звёзд по их расположению на карте звёздного неба… (см. условие задачи из демо-варианта 2025 года). **Аномалиями** назовём точки, находящиеся на расстоянии более одной условной единицы от точек кластеров. При расчётах аномалии учитывать не нужно.

Для каждого файла определите координаты центра каждого кластера, затем вычислите два числа: Px – среднее арифметическое абсцисс центров кластеров, и Py – среднее арифметическое ординат центров кластеров. В ответе запишите четыре числа: в первой строке сначала целую часть произведения Px × 100 000, затем целую часть произведения Py × 100 000 для файла А, во второй строке – аналогичные данные для файла Б.

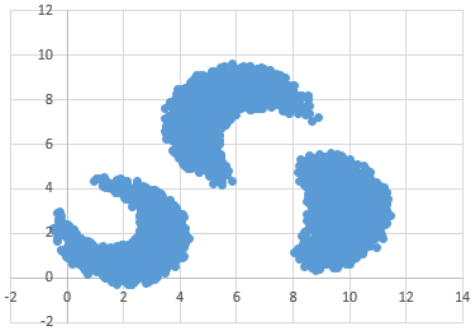
Исходные данные находятся в файлах 27-31a.txt и 27-31b.txt.

32) Учёный решил провести кластеризацию некоторого множества звёзд по их расположению на карте звёздного неба… (см. условие задачи из демо-варианта 2025 года). **Аномалиями** назовём точки, находящиеся на расстоянии более одной условной единицы от точек кластеров. При расчётах аномалии учитывать не нужно.

Для каждого файла определите координаты центра каждого кластера, затем вычислите два числа: Px – среднее арифметическое абсцисс центров кластеров, и Py – среднее арифметическое ординат центров кластеров. В ответе запишите четыре числа: в первой строке сначала целую часть произведения Px × 100 000, затем целую часть произведения Py × 100 000 для файла A, во второй строке – аналогичные данные для файла Б.

Исходные данные находятся в файлах `27-32a.txt` и `27-32b.txt`.

33) Учёный решил провести кластеризацию некоторого множества звёзд по их расположению на карте звёздного неба… (см. условие задачи из демо-варианта 2025 года). Кластеры имеют форму «рогалика». Для каждого файла определите координаты центра каждого кластера, затем вычислите два числа: Px – среднее арифметическое абсцисс центров кластеров, и Py – среднее арифметическое ординат центров кластеров. В ответе запишите четыре числа: в первой строке сначала целую часть произведения Px × 100 000, затем целую часть произведения Py × 100 000 для файла A, во второй строке – аналогичные данные для файла Б.



Исходные данные находятся в файлах `27-33a.txt` и `27-33b.txt`.

34) Учёный решил провести кластеризацию некоторого множества звёзд по их расположению на карте звёздного неба… (см. условие задачи из демо-варианта 2025 года). Кластеры имеют форму «рогалика». Для каждого файла определите координаты центра каждого кластера, затем вычислите два числа: Px – среднее арифметическое абсцисс центров кластеров, и Py – среднее арифметическое ординат центров кластеров. В ответе запишите четыре числа: в первой строке сначала целую часть произведения Px × 100 000, затем целую часть произведения Py × 100 000 для файла A, во второй строке – аналогичные данные для файла Б.

Исходные данные находятся в файлах `27-34a.txt` и `27-34b.txt`.

35) Учёный решил провести кластеризацию некоторого множества звёзд по их расположению на карте звёздного неба… (см. условие задачи из демо-варианта 2025 года). Кластеры имеют форму «рогалика». Для каждого файла определите координаты центра каждого кластера, затем вычислите два числа: Px – среднее арифметическое абсцисс центров кластеров, и Py – среднее арифметическое ординат центров кластеров. В ответе запишите четыре числа: в первой строке сначала целую часть произведения Px × 100 000, затем целую часть произведения Py × 100 000 для файла A, во второй строке – аналогичные данные для файла Б.

Исходные данные находятся в файлах `27-35a.txt` и `27-35b.txt`.

36) Учёный решил провести кластеризацию некоторого множества звёзд по их расположению на карте звёздного неба… (см. условие задачи из демо-варианта 2025 года). Кластеры имеют форму «рогалика». Для каждого файла определите координаты центра каждого кластера, затем вычислите два числа: Px – среднее арифметическое абсцисс центров кластеров, и Py – среднее арифметическое ординат центров кластеров. В ответе запишите четыре числа: в первой строке сначала целую часть произведения Px × 100 000, затем целую часть произведения Py × 100 000 для файла A, во второй строке – аналогичные данные для файла Б.

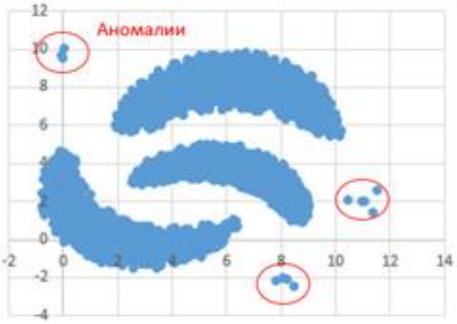
Исходные данные находятся в файлах `27-36a.txt` и `27-36b.txt`.

37) Учёный решил провести кластеризацию некоторого множества звёзд по их расположению на карте звёздного неба… (см. условие задачи из демо-варианта 2025 года). Кластеры имеют форму «рогалика». Для каждого файла определите координаты центра каждого кластера, затем вычислите два числа: Px – среднее арифметическое абсцисс центров кластеров, и Py – среднее арифметическое ординат центров кластеров. В ответе запишите четыре числа: в первой строке

сначала целую часть произведения Px × 100 000, затем целую часть произведения Py × 100 000 для файла A, во второй строке – аналогичные данные для файла Б.

Исходные данные находятся в файлах `27-37a.txt` и `27-37b.txt`.

38) Учёный решил провести кластеризацию некоторого множества звёзд по их расположению на карте звёздного неба… (см. условие задачи из демо-варианта 2025 года). Кластеры имеют форму «рогалика». **Аномалиями** назовём точки, находящиеся на расстоянии более одной условной единицы от точек кластеров. При расчётах аномалии учитывать не нужно. Для каждого файла определите координаты центра каждого кластера, затем вычислите два числа: Px – среднее арифметическое абсцисс центров кластеров, и Py – среднее арифметическое ординат центров кластеров. В ответе запишите четыре числа: в первой строке сначала целую часть произведения Px × 100 000, затем целую часть произведения Py × 100 000 для файла A, во второй строке – аналогичные данные для файла Б.



Исходные данные находятся в файлах `27-38a.txt` и `27-38b.txt`.

39) Учёный решил провести кластеризацию некоторого множества звёзд по их расположению на карте звёздного неба… (см. условие задачи из демо-варианта 2025 года). Кластеры имеют форму «рогалика». **Аномалиями** назовём точки, находящиеся на расстоянии более одной условной единицы от точек кластеров. При расчётах аномалии учитывать не нужно. Для каждого файла определите координаты центра каждого кластера, затем вычислите два числа: Px – среднее арифметическое абсцисс центров кластеров, и Py – среднее арифметическое ординат центров кластеров. В ответе запишите четыре числа: в первой строке сначала целую часть произведения Px × 100 000, затем целую часть произведения Py × 100 000 для файла A, во второй строке – аналогичные данные для файла Б.

Исходные данные находятся в файлах `27-39a.txt` и `27-39b.txt`.

40) Учёный решил провести кластеризацию некоторого множества звёзд по их расположению на карте звёздного неба… (см. условие задачи из демо-варианта 2025 года). Кластеры имеют форму «рогалика». **Аномалиями** назовём точки, находящиеся на расстоянии более одной условной единицы от точек кластеров. При расчётах аномалии учитывать не нужно. Для каждого файла определите координаты центра каждого кластера, затем вычислите два числа: Px – среднее арифметическое абсцисс центров кластеров, и Py – среднее арифметическое ординат центров кластеров. В ответе запишите четыре числа: в первой строке сначала целую часть произведения Px × 100 000, затем целую часть произведения Py × 100 000 для файла A, во второй строке – аналогичные данные для файла Б.

Исходные данные находятся в файлах `27-40a.txt` и `27-40b.txt`.

41) Учёный решил провести кластеризацию некоторого множества звёзд по их расположению на карте звёздного неба… (см. условие задачи из демо-варианта 2025 года). Кластеры имеют форму «рогалика». **Аномалиями** назовём точки, находящиеся на расстоянии более одной условной единицы от точек кластеров. При расчётах аномалии учитывать не нужно. Для каждого файла определите координаты центра каждого кластера, затем вычислите два числа: Px – среднее арифметическое абсцисс центров кластеров, и Py – среднее арифметическое ординат центров кластеров. В ответе запишите четыре числа: в первой строке сначала целую часть произведения Px × 100 000, затем целую часть произведения Py × 100 000 для файла A, во второй строке – аналогичные данные для файла Б.

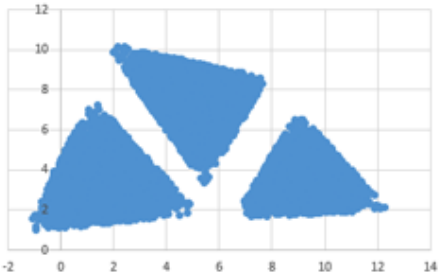
Исходные данные находятся в файлах `27-41a.txt` и `27-41b.txt`.

42) Учёный решил провести кластеризацию некоторого множества звёзд по их расположению на карте звёздного неба… (см. условие задачи из демо-варианта 2025 года). Кластеры имеют форму

«рогалика». **Аномалиями** назовём точки, находящиеся на расстоянии более одной условной единицы от точек кластеров. При расчётах аномалии учитывать не нужно. Для каждого файла определите координаты центра каждого кластера, затем вычислите два числа: Px – среднее арифметическое абсцисс центров кластеров, и Py – среднее арифметическое ординат центров кластеров. В ответе запишите четыре числа: в первой строке сначала целую часть произведения Px × 100 000, затем целую часть произведения Py × 100 000 для файла А, во второй строке – аналогичные данные для файла Б.

Исходные данные находятся в файлах `27-42a.txt` и `27-42b.txt`.

43) Учёный решил провести кластеризацию некоторого множества звёзд по их расположению на карте звёздного неба… (см. условие задачи из демо-варианта 2025 года). Кластеры имеют треугольную форму. Для каждого файла определите координаты центра каждого кластера, затем вычислите два числа: Px – среднее арифметическое абсцисс центров кластеров, и Py – среднее арифметическое ординат центров кластеров. В ответе запишите четыре числа: в первой строке сначала целую часть произведения Px × 100 000, затем целую часть произведения Py × 100 000 для файла А, во второй строке – аналогичные данные для файла Б.



Исходные данные находятся в файлах `27-43a.txt` и `27-43b.txt`.

44) Учёный решил провести кластеризацию некоторого множества звёзд по их расположению на карте звёздного неба… (см. условие задачи из демо-варианта 2025 года). Кластеры имеют треугольную форму. Для каждого файла определите координаты центра каждого кластера, затем вычислите два числа: Px – среднее арифметическое абсцисс центров кластеров, и Py – среднее арифметическое ординат центров кластеров. В ответе запишите четыре числа: в первой строке сначала целую часть произведения Px × 100 000, затем целую часть произведения Py × 100 000 для файла А, во второй строке – аналогичные данные для файла Б.

Исходные данные находятся в файлах `27-44a.txt` и `27-44b.txt`.

45) Учёный решил провести кластеризацию некоторого множества звёзд по их расположению на карте звёздного неба… (см. условие задачи из демо-варианта 2025 года). Кластеры имеют треугольную форму. Для каждого файла определите координаты центра каждого кластера, затем вычислите два числа: Px – среднее арифметическое абсцисс центров кластеров, и Py – среднее арифметическое ординат центров кластеров. В ответе запишите четыре числа: в первой строке сначала целую часть произведения Px × 100 000, затем целую часть произведения Py × 100 000 для файла А, во второй строке – аналогичные данные для файла Б.

Исходные данные находятся в файлах `27-45a.txt` и `27-45b.txt`.

46) Учёный решил провести кластеризацию некоторого множества звёзд по их расположению на карте звёздного неба… (см. условие задачи из демо-варианта 2025 года). Кластеры имеют треугольную форму. Для каждого файла определите координаты центра каждого кластера, затем вычислите два числа: Px – среднее арифметическое абсцисс центров кластеров, и Py – среднее арифметическое ординат центров кластеров. В ответе запишите четыре числа: в первой строке сначала целую часть произведения Px × 100 000, затем целую часть произведения Py × 100 000 для файла А, во второй строке – аналогичные данные для файла Б.

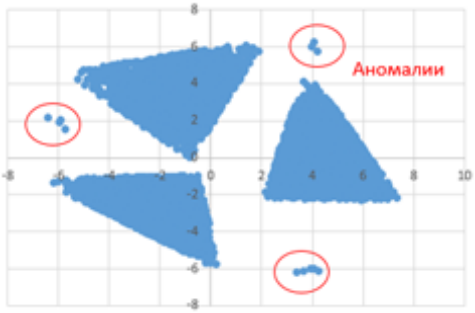
Исходные данные находятся в файлах `27-46a.txt` и `27-46b.txt`.

47) Учёный решил провести кластеризацию некоторого множества звёзд по их расположению на карте звёздного неба… (см. условие задачи из демо-варианта 2025 года). Кластеры имеют треугольную форму. Для каждого файла определите координаты центра каждого кластера, затем вычислите два числа: Px – среднее арифметическое абсцисс центров кластеров, и Py – среднее арифметическое ординат центров кластеров. В ответе запишите четыре числа: в первой строке

сначала целую часть произведения Px × 100 000, затем целую часть произведения Py × 100 000 для файла А, во второй строке – аналогичные данные для файла Б.

Исходные данные находятся в файлах 27-47a.txt и 27-47b.txt.

48) Учёный решил провести кластеризацию некоторого множества звёзд по их расположению на карте звёздного неба… (см. условие задачи из демо-варианта 2025 года). Кластеры имеют треугольную форму. **Аномалиями** назовём точки, находящиеся на расстоянии более одной условной единицы от точек кластеров. При расчётах аномалии учитывать не нужно. Для каждого файла определите координаты центра каждого кластера, затем вычислите два числа: Px – среднее арифметическое абсцисс центров кластеров, и Py – среднее арифметическое ординат центров кластеров. В ответе запишите четыре числа: в первой строке сначала целую часть произведения Px × 100 000, затем целую часть произведения Py × 100 000 для файла А, во второй строке – аналогичные данные для файла Б.



Исходные данные находятся в файлах 27-48a.txt и 27-48b.txt.

49) Учёный решил провести кластеризацию некоторого множества звёзд по их расположению на карте звёздного неба… (см. условие задачи из демо-варианта 2025 года). Кластеры имеют треугольную форму. **Аномалиями** назовём точки, находящиеся на расстоянии более одной условной единицы от точек кластеров. При расчётах аномалии учитывать не нужно. Для каждого файла определите координаты центра каждого кластера, затем вычислите два числа: Px – среднее арифметическое абсцисс центров кластеров, и Py – среднее арифметическое ординат центров кластеров. В ответе запишите четыре числа: в первой строке сначала целую часть произведения Px × 100 000, затем целую часть произведения Py × 100 000 для файла А, во второй строке – аналогичные данные для файла Б.

Исходные данные находятся в файлах 27-49a.txt и 27-49b.txt.

50) Учёный решил провести кластеризацию некоторого множества звёзд по их расположению на карте звёздного неба… (см. условие задачи из демо-варианта 2025 года). Кластеры имеют треугольную форму. **Аномалиями** назовём точки, находящиеся на расстоянии более одной условной единицы от точек кластеров. При расчётах аномалии учитывать не нужно. Для каждого файла определите координаты центра каждого кластера, затем вычислите два числа: Px – среднее арифметическое абсцисс центров кластеров, и Py – среднее арифметическое ординат центров кластеров. В ответе запишите четыре числа: в первой строке сначала целую часть произведения Px × 100 000, затем целую часть произведения Py × 100 000 для файла А, во второй строке – аналогичные данные для файла Б.

Исходные данные находятся в файлах 27-50a.txt и 27-50b.txt.

51) Учёный решил провести кластеризацию некоторого множества звёзд по их расположению на карте звёздного неба… (см. условие задачи из демо-варианта 2025 года). Кластеры имеют треугольную форму. **Аномалиями** назовём точки, находящиеся на расстоянии более одной условной единицы от точек кластеров. При расчётах аномалии учитывать не нужно. Для каждого файла определите координаты центра каждого кластера, затем вычислите два числа: Px – среднее арифметическое абсцисс центров кластеров, и Py – среднее арифметическое ординат центров кластеров. В ответе запишите четыре числа: в первой строке сначала целую часть произведения Px × 100 000, затем целую часть произведения Py × 100 000 для файла А, во второй строке – аналогичные данные для файла Б.

Исходные данные находятся в файлах 27-51a.txt и 27-51b.txt.

52) Учёный решил провести кластеризацию некоторого множества звёзд по их расположению на карте звёздного неба… (см. условие задачи из демо-варианта 2025 года). Кластеры имеют треугольную форму. **Аномалиями** назовём точки, находящиеся на расстоянии более одной условной единицы от точек кластеров. При расчётах аномалии учитывать не нужно. Для каждого файла определите координаты центра каждого кластера, затем вычислите два числа: Px – среднее арифметическое абсцисс центров кластеров, и Py – среднее арифметическое ординат центров кластеров. В ответе запишите четыре числа: в первой строке сначала целую часть произведения Px × 100 000, затем целую часть произведения Py × 100 000 для файла А, во второй строке – аналогичные данные для файла Б.
Исходные данные находятся в файлах `27-52a.txt` и `27-52b.txt`.