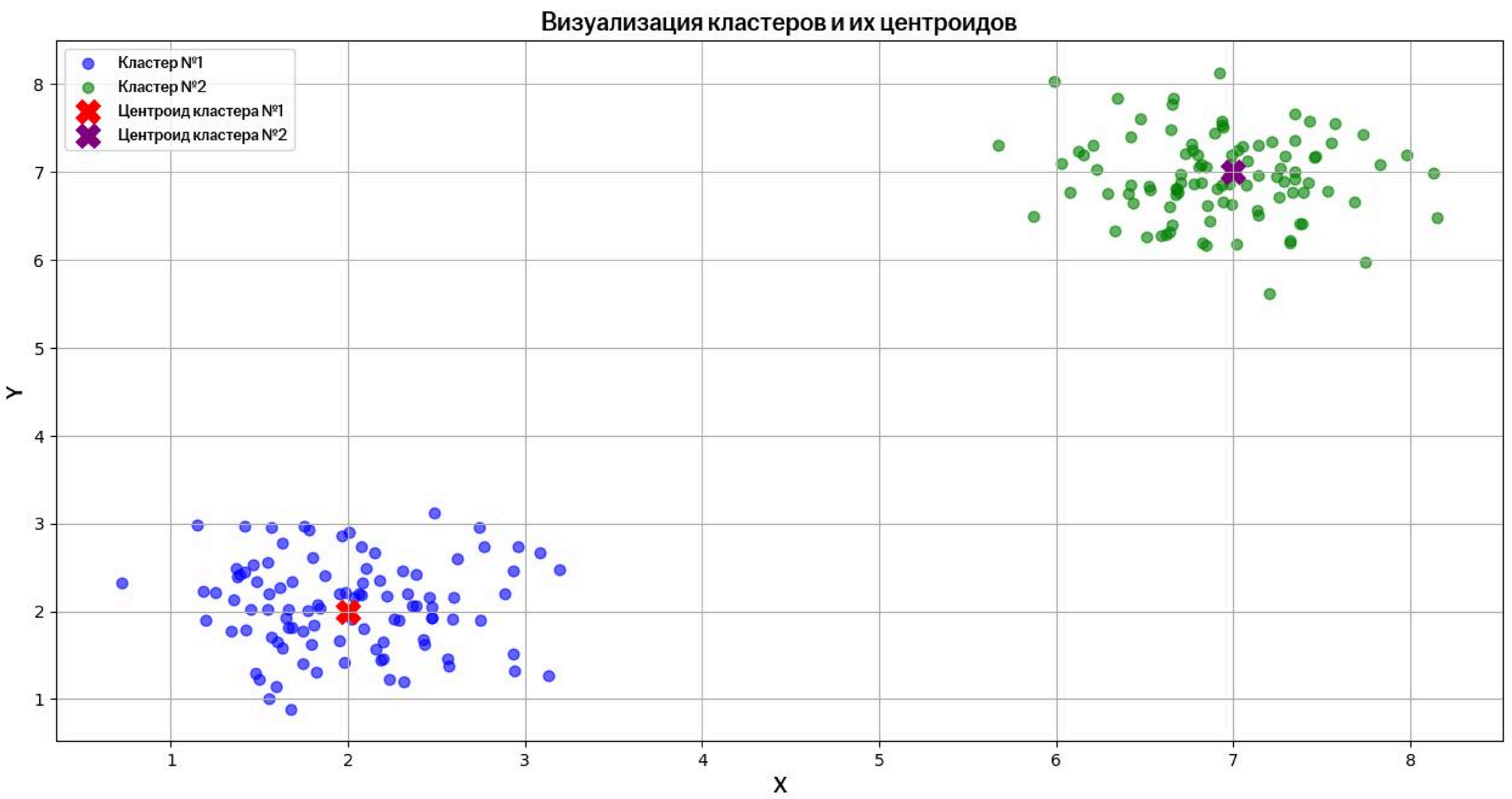
Атрибуты. Кластеризация

Кластер – это группа объектов (например, звёзд, документов, клиентов и т.д.), которые находятся в пределах определённой области и схожи по каким-то признакам. Кластер может быть представлен в виде набора точек на плоскости или в многомерном пространстве, которые сгруппированы на основе расстояний или других мер схожести. В контексте задачи о звёздах кластером считается область, содержащая звёзды, расположенные близко друг к другу в рамках заданного прямоугольника.

Центроид кластера – это центральная точка, которая характеризует данный кластер. В случае задачи о звёздах центроид выбирается среди самих звёзд, и для его нахождения вычисляется точка, от которой сумма расстояний до всех остальных звёзд кластера минимальна.



Кластеризация – это метод машинного обучения, используемый для группировки данных в подмножества (кластеры), основываясь на схожести объектов внутри каждого кластера.

Основная цель кластеризации – разделить набор данных так, чтобы объекты внутри одного кластера были схожи, а между кластерами – максимально различались. В отличие от классификации, кластеризация относится к методам обучения без учителя, так как заранее не предполагается, что данные уже размечены (разделены на классы).

Атрибуты. Кластеризация

После выполнения кластеризации можно анализировать различные атрибуты кластеров, которые помогают извлечь дополнительную информацию и использовать её для принятия решений в дальнейших действиях.

Как считывать точки с атрибутами в процессе кластеризации?

```
f = open('task27_7/27B.txt')
f.readline()
roots = [list(map(float, s.replace(",", ".").split())) for s in f]
clusters = [[], [], []]
for x, y, value in roots:
    if x < 4:
        clusters[0].append((x, y, value))
    elif x > 4 and y > 2:
        clusters[1].append((x, y, value))
    else:
        clusters[2].append((x, y, value))
```

Как считывать точки с атрибутами в процессе кластеризации?

После кластеризации можно вычислять различные математические характеристики, которые помогают:

- Анализировать структуру кластеров
- Определять их плотность, разброс и аномалии
- Использовать эти данные для предсказаний

К примеру для каждого файла нужно найти сумму особых точек.

Особой считается звезда, если её яркость отличается от среднего значения яркости в кластере более чем на 1.5 стандартных отклонений (это корень квадратный из дисперсии)

Дисперсия: $\sigma = \frac{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \cdots + (x_n - \bar{x})^2}{n}$ Стандартное отклонение: $\alpha = \sqrt{\sigma}$

```
deviations = []

for i in range (len(clusters)):
    min_dist = 10**10
    disp = []
    for x1, y1, value1 in clusters[i]:
        dist = 0
        for x2, y2, value2 in clusters[i]:
            euclidian_dist = ((x2 - x1)**2 + (y2 - y1)**2)**0.5
            dist += euclidian_dist
        if dist < min_dist:
            min_dist = dist
            best_centroids[i] = [x1, y1]
        disp.append((value1 - mean)**2)
    disp = sum(disp) / (len(disp) - 1)
    deviations.append(disp ** 0.5)

spec_points = 0
```

Атрибуты. Кластеризация

Далее уже рассчитаем количество особых точек и получим ответ к нашей задаче!

```
for i in range(len(clusters)):
    mean = sum(value for x, y, value in clusters[i]) / len(clusters[i])
    for x, y, value in clusters[i]:
        if abs(value - mean) > 1.5 * deviations[i]:
            spec_points += 1

print(anomalies)
```

Заметки