

K-means. Метод кластеризации

Дисциплина: Основы машинного обучения.

гр. 5030102/20201

Смирнова А. П.

Грушин А. Д.

Введение в кластеризацию

Кластеризация

— это задача разделения набора объектов на группы (кластеры) таким образом, чтобы объекты внутри одной группы были похожи между собой, а объекты из разных групп — как можно более различны.

Кластеризацию полезно использовать для выявления скрытых закономерностей в данных, сегментации рынка, обработки изображений и в других областях.

Введение в k-means

Метод k-means

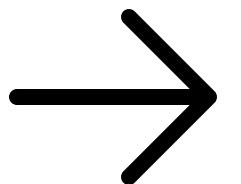
— это алгоритм кластеризации, направленный на разбиение набора данных на k кластеров, таких, что каждый объект принадлежит кластеру с ближайшим к нему центром.

Цель алгоритма — минимизировать сумму квадратов расстояний точек кластеров от их центров.

Алгоритм k-means

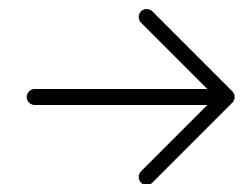
Шаг 1

Выбираются k случайных центров кластеров для набора данных. Эти центры могут быть выбраны случайным образом или используя специальные методы инициализации.



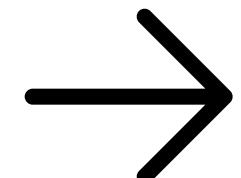
Шаг 2

Каждая точка данных присваивается к ближайшему центру кластера. Это делается путем расчета расстояния (обычно евклидово) между каждой точкой данных и каждым центром.

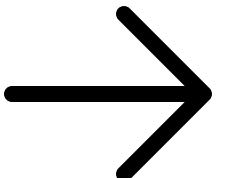


Алгоритм k-means

Шаг 3



Центры кластеров обновляются как среднее значение всех точек, принадлежащих каждому кластеру.



Шаг 4

Шаги 2 и 3 повторяются до тех пор, пока центры не перестанут меняться или до достижения заданного критерия сходимости.

Алгоритм k-means на примере

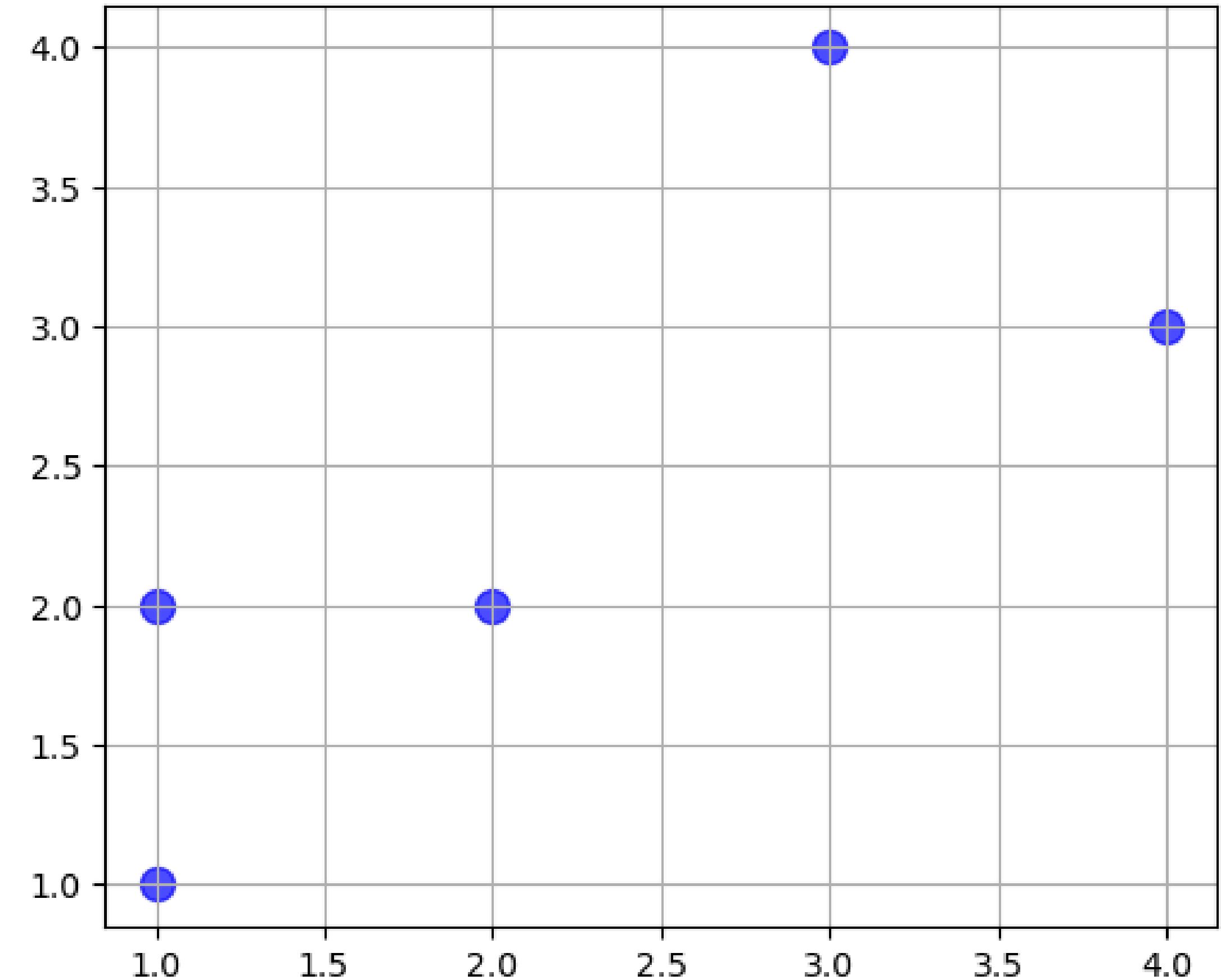
Пример

Рассмотрим набор из 5 точек:

$(1, 1), (1, 2), (2, 2), (3, 4), (4, 3)$

По графику можно предположить, что
число 2 — это оптимальное
количество кластеров. Так и сделаем.
Разобьем наши данные на 2 кластера.

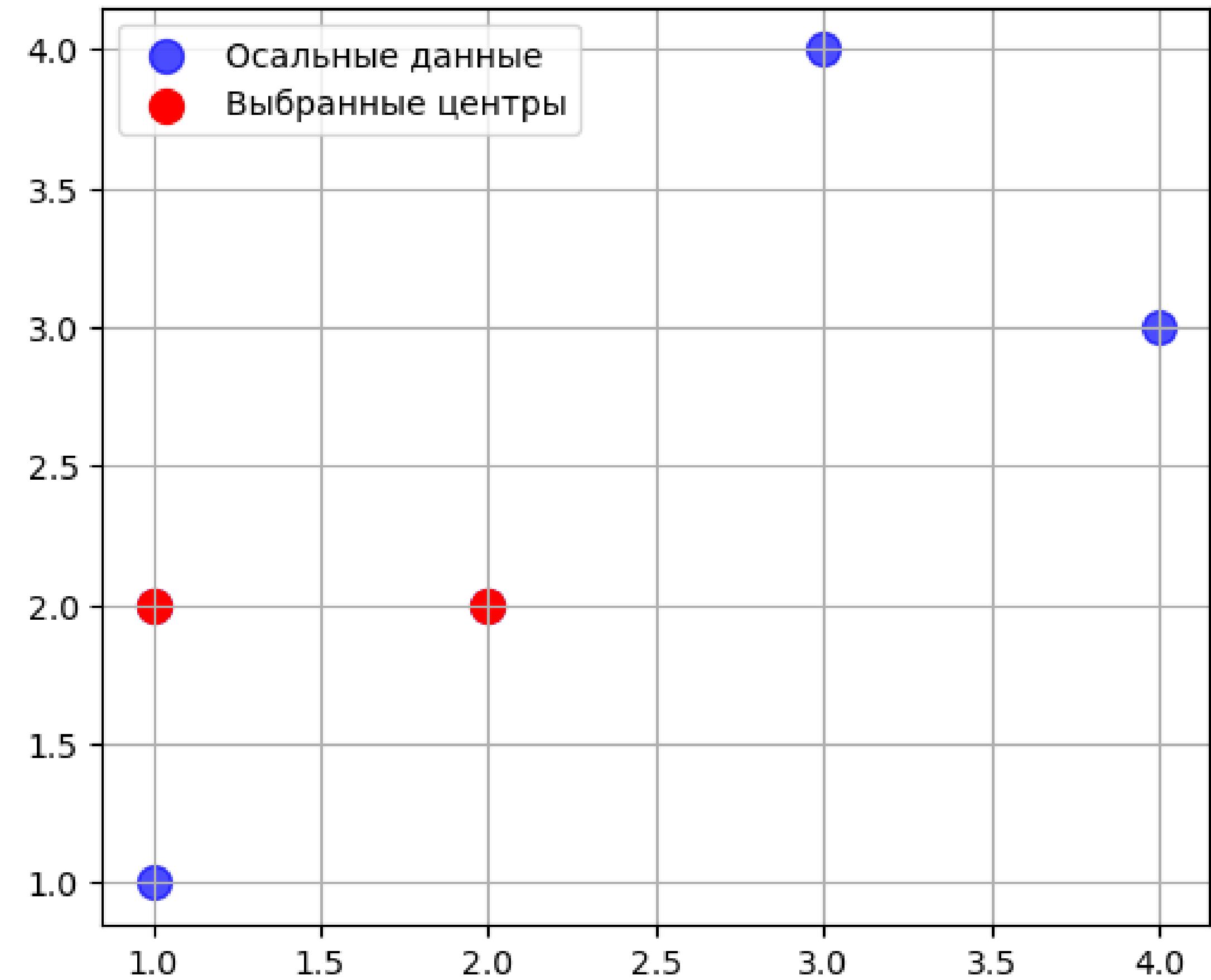
Далее в презентации проверим, были
мы правы или нет.



Алгоритм k-means на примере

Шаг 1

Случайным образом выберем 2 центра для кластеров: (1, 2) и (2, 2)



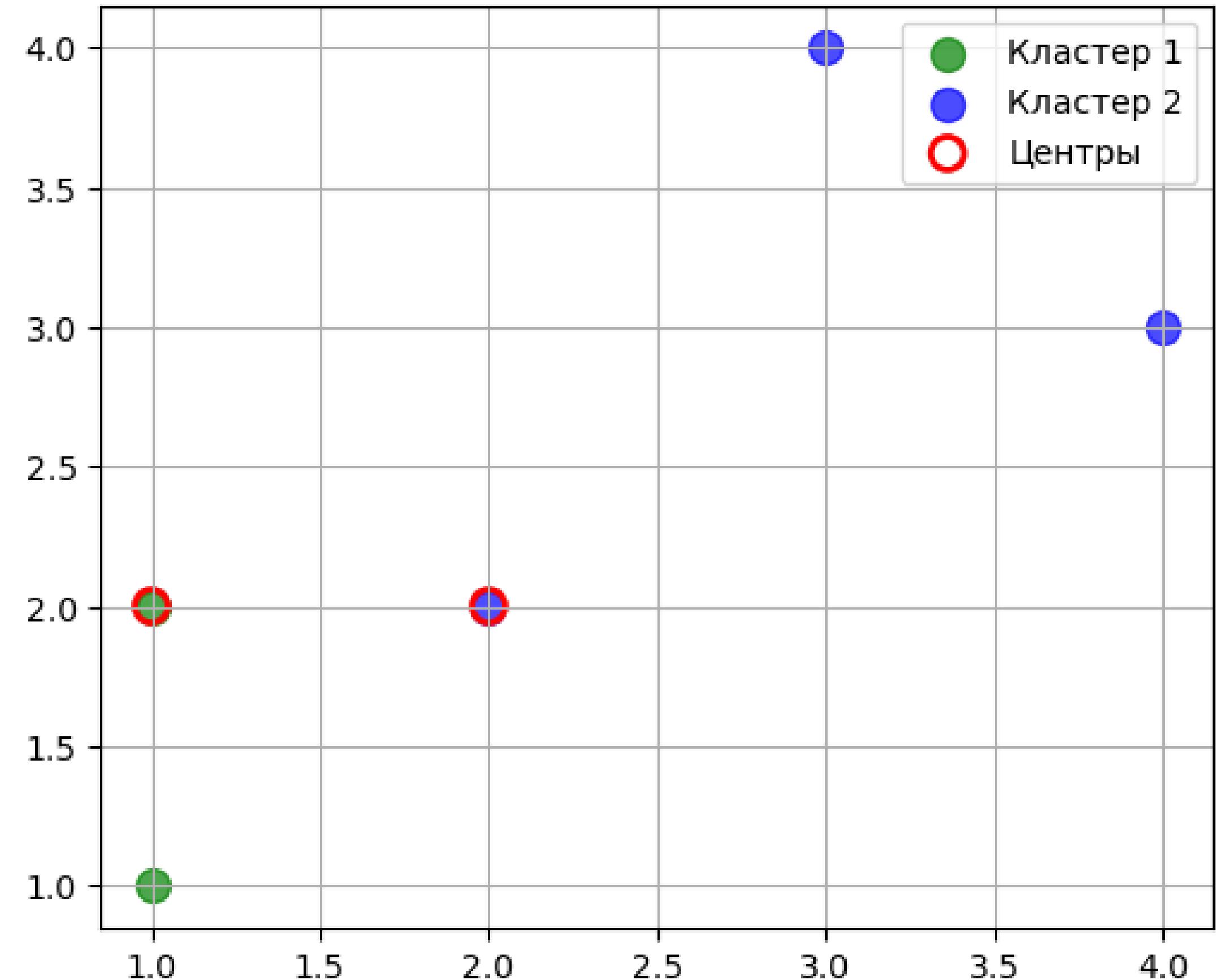
Алгоритм k-means на примере

Шаг 2

Для каждой точки вычислим евклидово расстояние до каждого центра по следующей формуле:

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

| | (1, 1) | (1, 2) | (2, 2) | (3, 4) | (4, 3) |
|--------|--------|--------|--------|--------|--------|
| (1, 2) | 1 | 0 | 1 | ~2,8 | ~3,2 |
| (2, 2) | ~1,4 | 1 | 0 | ~2,2 | ~2,2 |



Алгоритм k-means на примере

Шаг 3

Пересчитаем центры кластеров как среднее значение всех точек, принадлежащих кластеру.

По формуле:

$$c = \left(\frac{1}{n} \sum_{i=1}^n x_i, \frac{1}{n} \sum_{i=1}^n y_i \right)$$

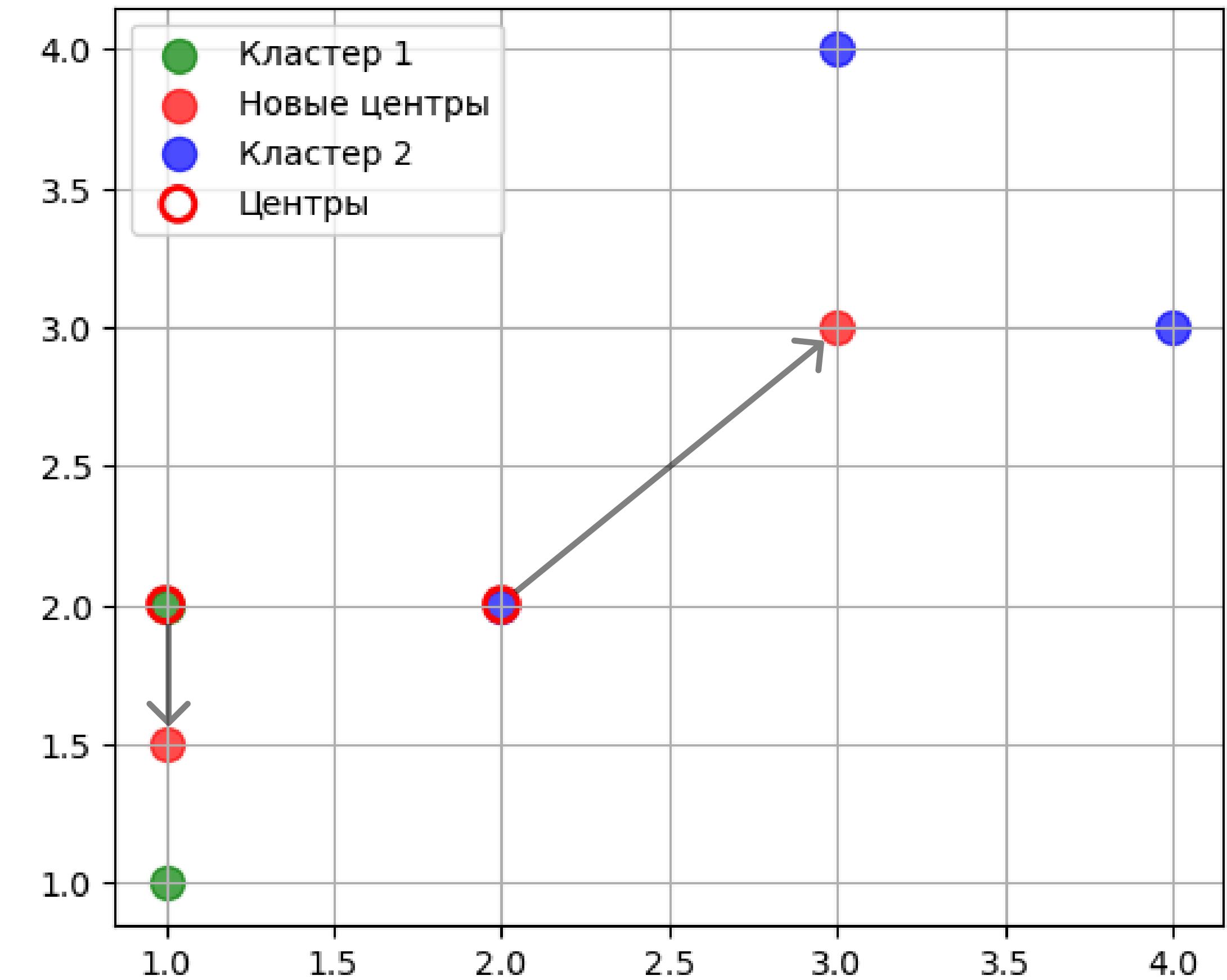
где с - центр кластера, в котором n элементов, x_i и y_i это координаты элементов этого кластера.

1 кластер: (1, 1), (1, 2)

центр: (1, 2) → (1, 1,5)

2 кластер: (2, 2), (3, 4), (4, 3)

центр: (2, 2) → (3, 3)

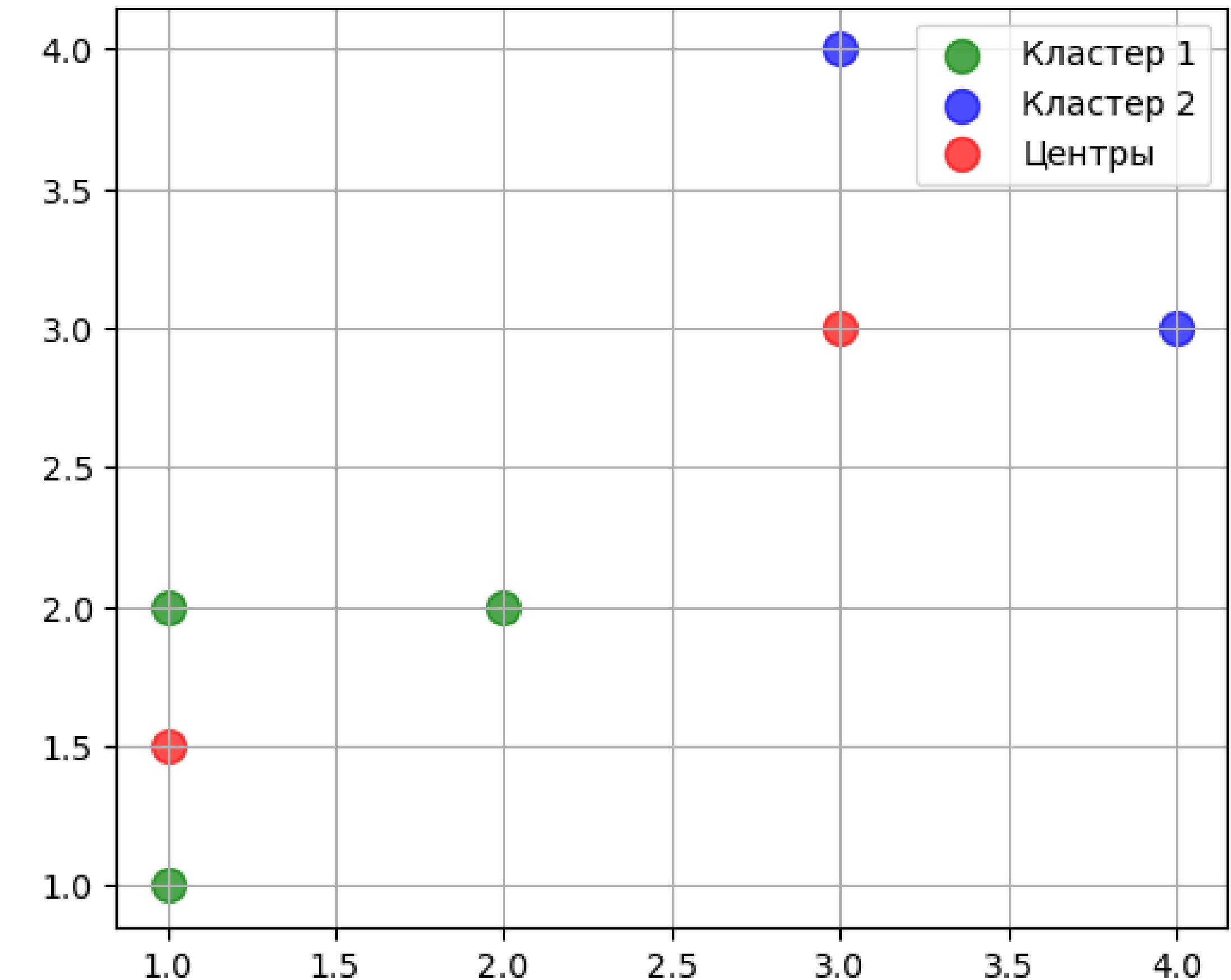


Алгоритм k-means на примере

Шаг 2 (2-е повторение)

Для каждой точки вычислим евклидово расстояние до каждого нового центра.

| | (1, 1) | (1, 2) | (2, 2) | (3, 4) | (4, 3) |
|-----------|--------|--------|--------|--------|--------|
| (1, 1, 5) | 0,5 | 0,5 | ~1,1 | ~3,2 | ~3,4 |
| (3, 3) | ~2,8 | ~2,2 | ~1,4 | 1 | 1 |



Алгоритм k-means на примере

Шаг 3 (2-е повторение)

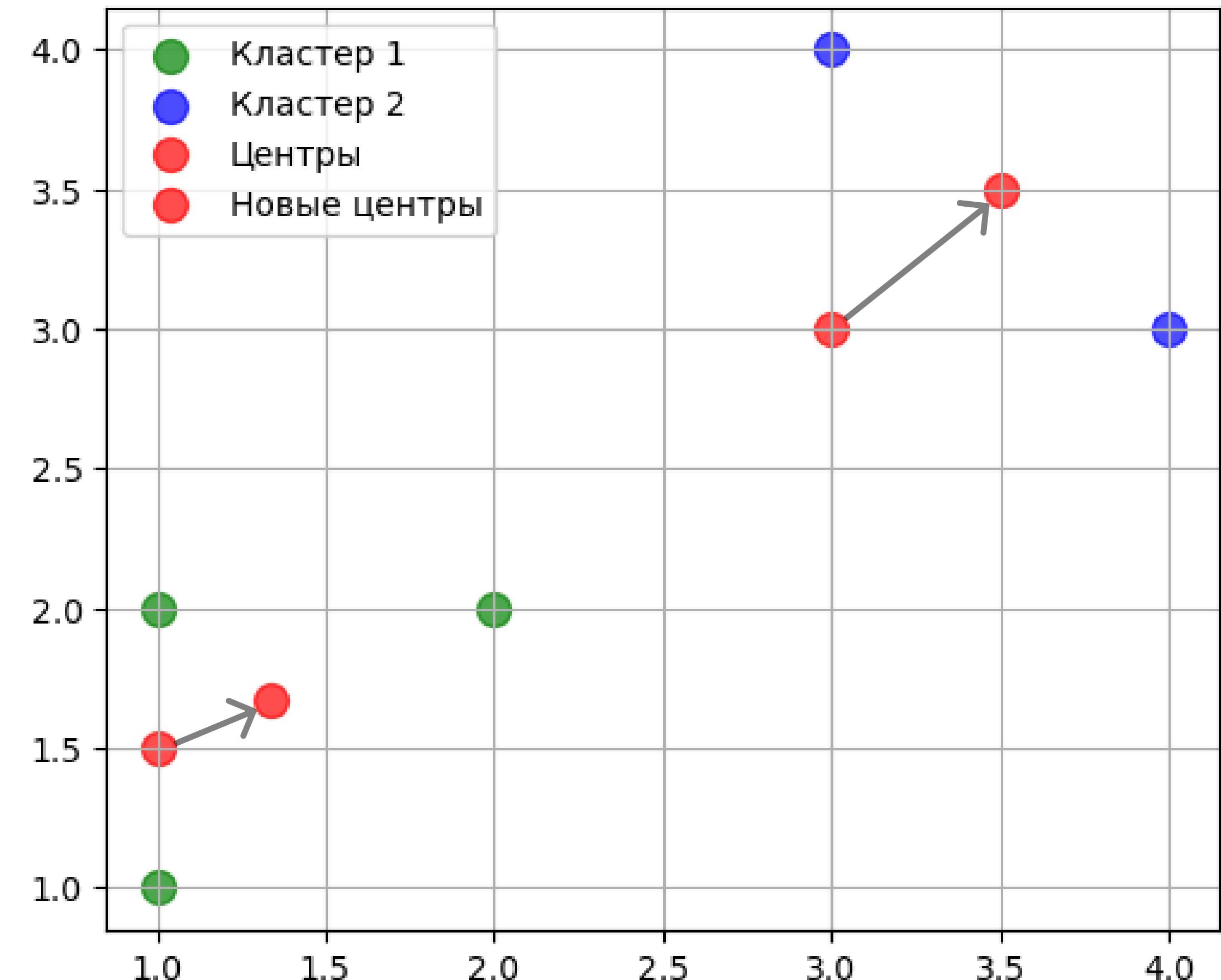
Пересчитаем центры кластеров как среднее значение всех точек, принадлежащих кластеру.

1 кластер: $(1, 1), (1, 2), (2, 2)$

центр: $(1, 1, 1) \rightarrow (1, 1, 1)$

2 кластер: $(3, 4), (4, 3)$

центр: $(3, 3) \rightarrow (3, 3, 3)$

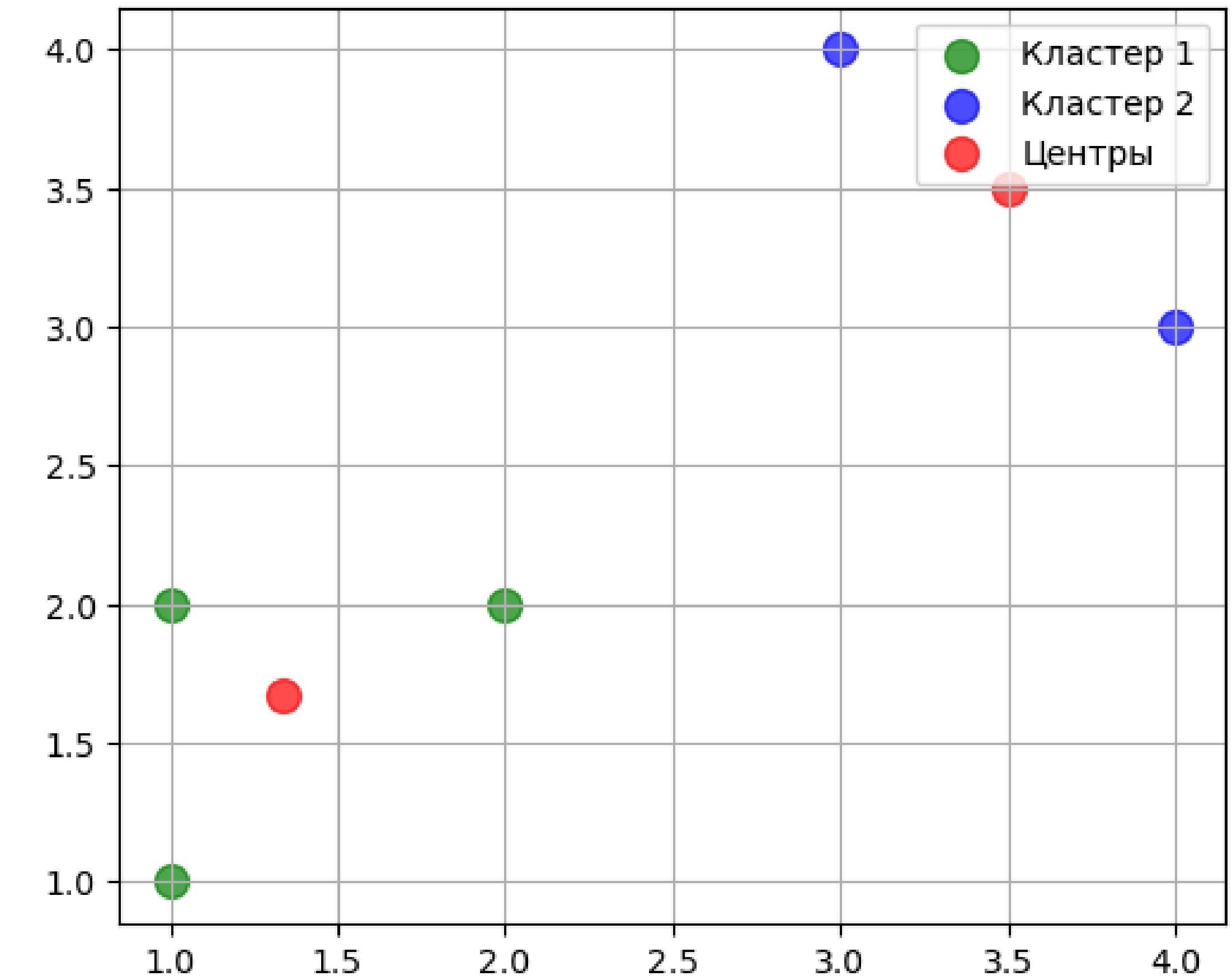


Алгоритм k-means на примере

Шаг 2 (3-е повторение)

Для каждой точки вычислим евклидово расстояние до каждого нового центра.

| | (1, 1) | (1, 2) | (2, 2) | (3, 4) | (4, 3) |
|--------------|--------|--------|--------|--------|--------|
| (1, 3, 1, 6) | ~0,7 | ~0,5 | ~0,7 | ~2,9 | ~3 |
| (3, 5, 3, 5) | ~3,5 | ~2,9 | ~2,1 | ~0,7 | ~0,7 |



Алгоритм k-means на примере

Шаг 3 (3-е повторение)

Пересчитаем центры кластеров как среднее значение всех точек, принадлежащих кластеру.

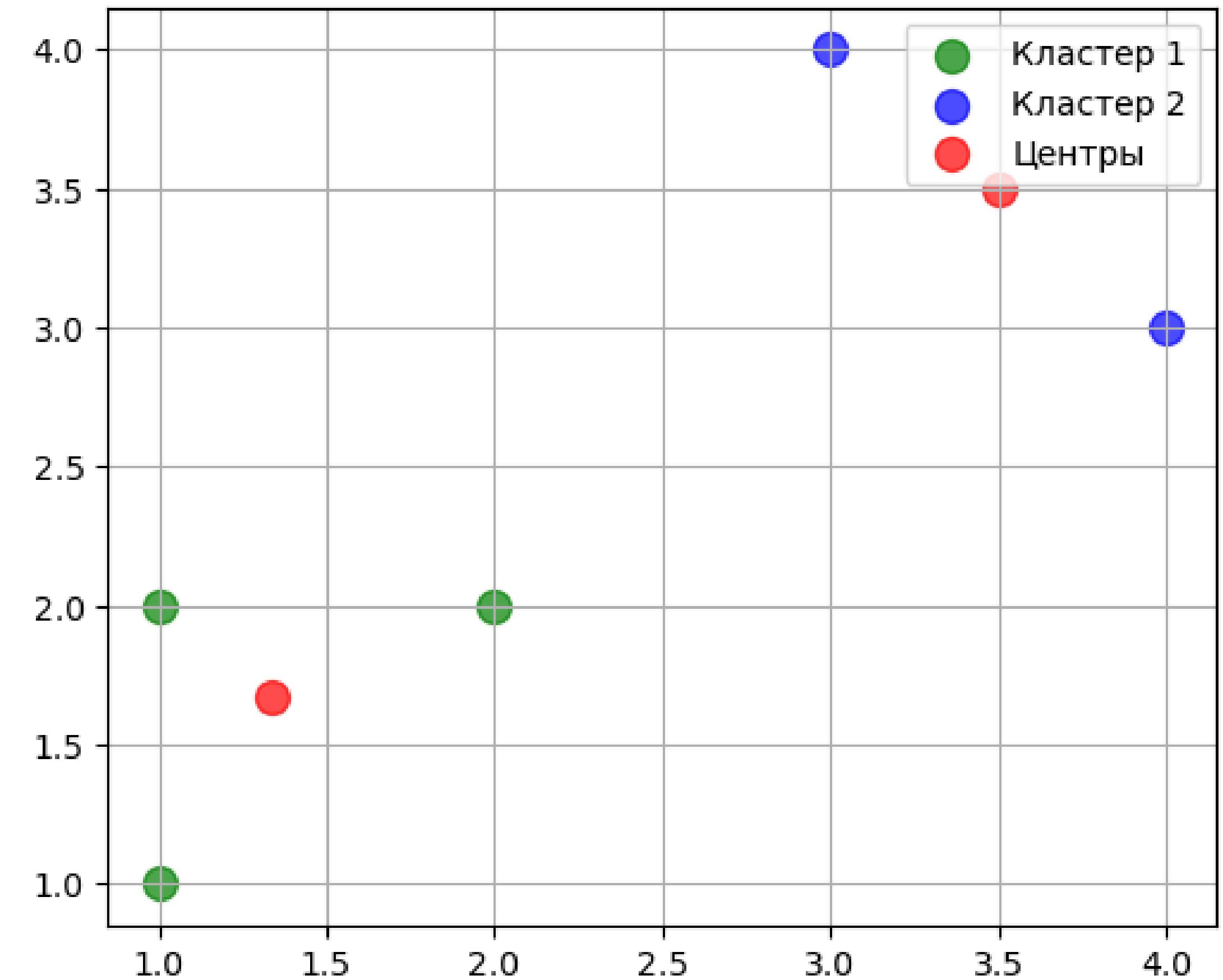
1 кластер: $(1, 1), (1, 2), (2, 2)$

центр: $(1, 3, 1, 6) \rightarrow (1, 3, 1, 6)$

2 кластер: $(3, 4), (4, 3)$

центр: $(3, 5, 3, 5) \rightarrow (3, 5, 3, 5)$

Центры не изменились \Rightarrow алгоритм завершается



Количество кластеров

Алгоритм k-means требует заранее заданное значение k — количество кластеров. Важно выбрать правильное значение, ведь это влияет на качество кластеризации:

1. Если k слишком маленькое, то данные будут сильно обобщены, и точная структура данных будет утрачена.
2. Если k слишком большое, кластеры будут слишком раздроблены, и значимые закономерности могут быть скрыты.

Методы выбора числа кластеров

Метод локтя (Elbow method)

Строится график зависимости суммы квадратов расстояний от числа кластеров k .

По мере увеличения k , ошибка уменьшается, но после некоторого момента улучшения становятся незначительными.

Точку «излома» (локоть на графике) и считают оптимальным числом кластеров.

Метод силуэта (Silhouette method)

Оценивается качество кластеризации для каждой точки:

$$s = \frac{b - a}{\max(a, b)}$$

где a — среднее внутрикластерное расстояние, b — среднее расстояние до ближайшего кластера

Вычисляется значение силуэта как среднее значение коэффициента для всех объектов.

Выбирается k , при котором значение силуэта максимально.

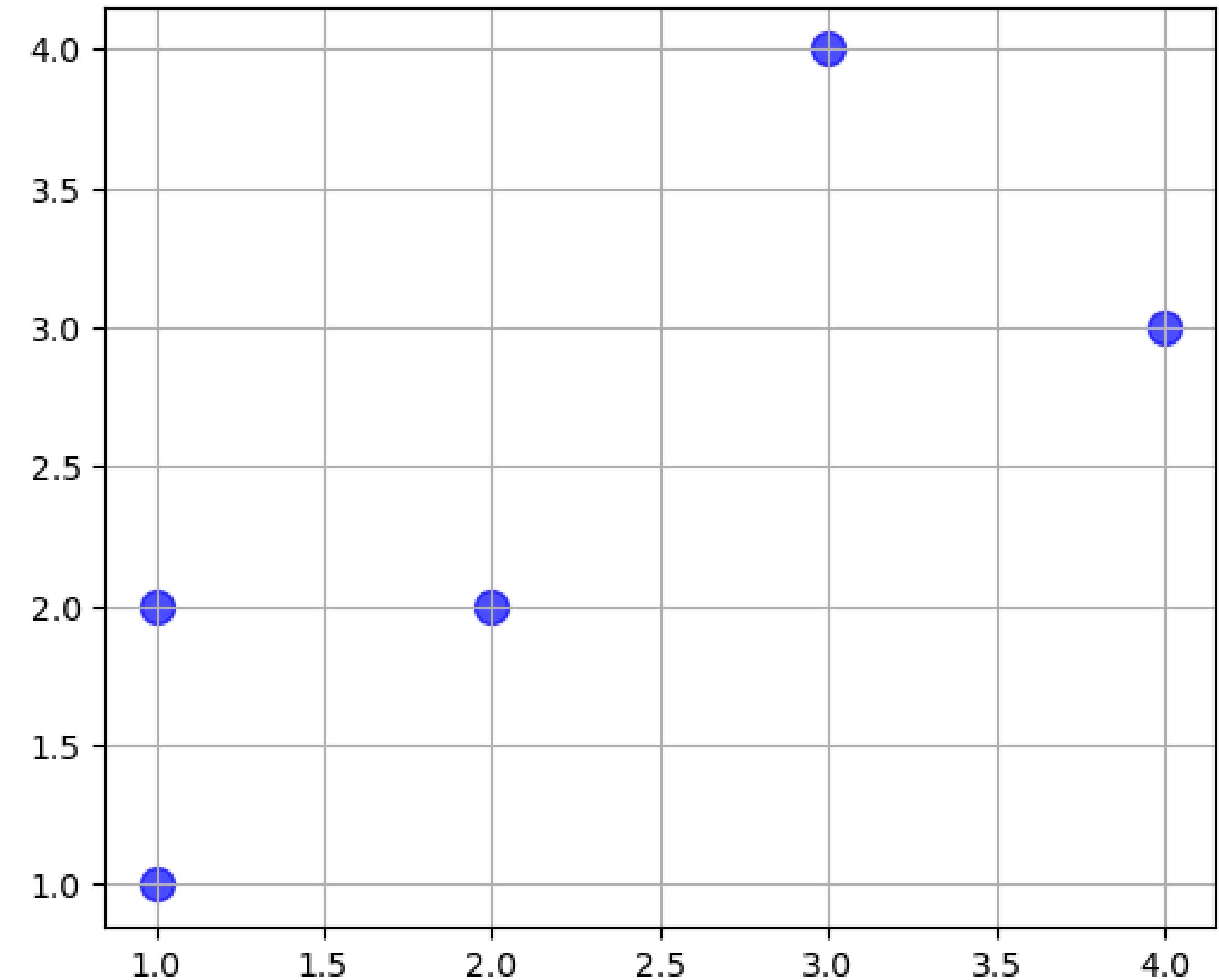
Метод локтя на примере

Пример

Рассмотрим набор из 5 точек:
 $(1, 1), (1, 2), (2, 2), (3, 4), (4, 3)$

Ранее мы предположили, что число 2 — это оптимальное количество кластеров.
Проверим это методом локтя.

Построим график зависимости суммы квадратов расстояний от точек до центров кластеров от числа кластеров.

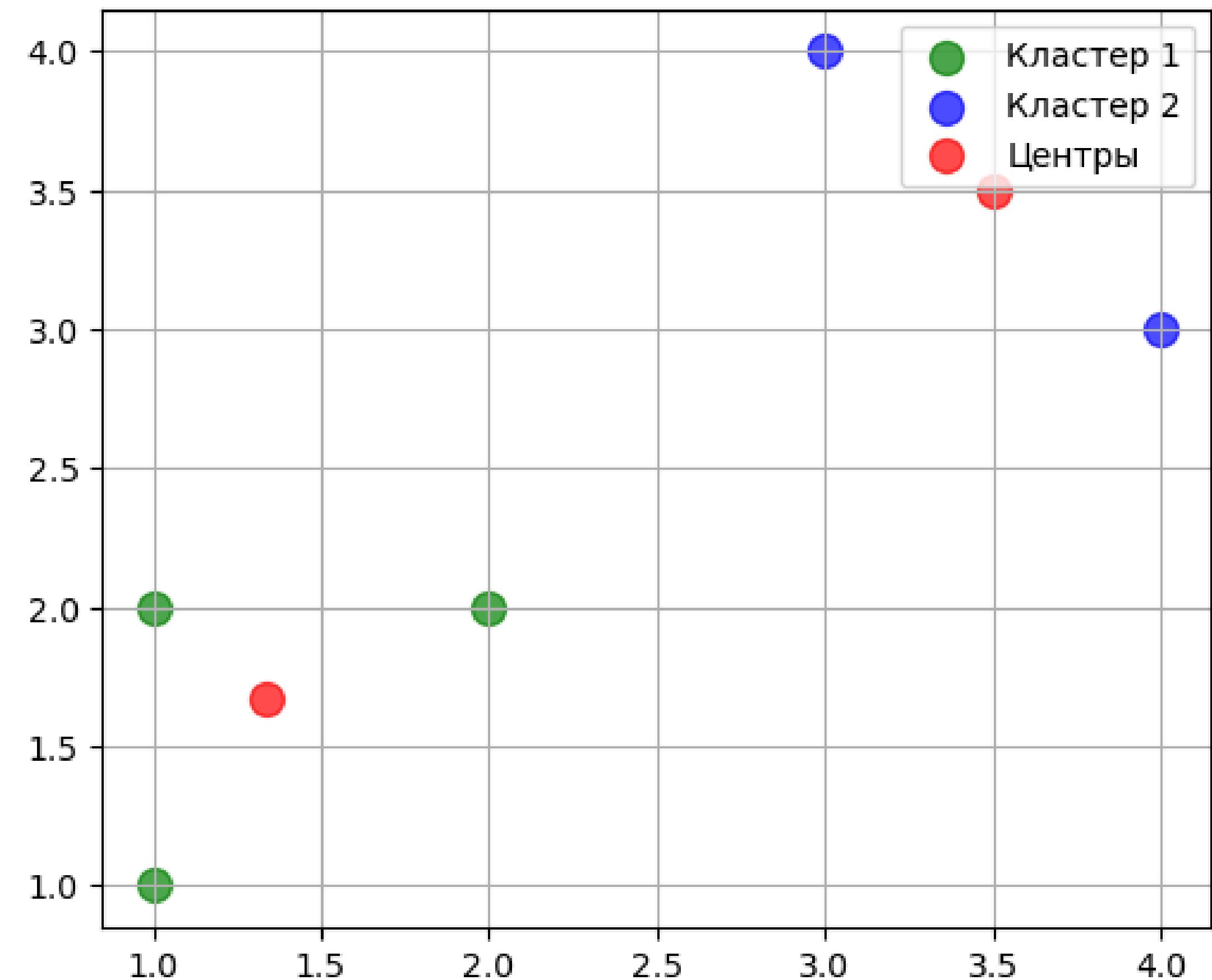


Метод локтя на примере

Сумму квадратов расстояний от точек до центров считаем по формуле:

$$S = \sum_{i=1}^k \sum_{j=1}^{n_k} ((x_i - x_j)^2 + (y_i - y_j)^2)$$

где k - это количество кластеров, в каждом из них количество элементов - n_k
 x_i и y_i - центры кластеров
 x_j и y_j - координаты элементов соответственных кластеров



Метод локтя на примере

Для 2-х кластеров, которые мы получили ранее:

1 кластер с центром $(1,3, 1,6)$:

$(1, 1), (1, 2), (2, 2)$

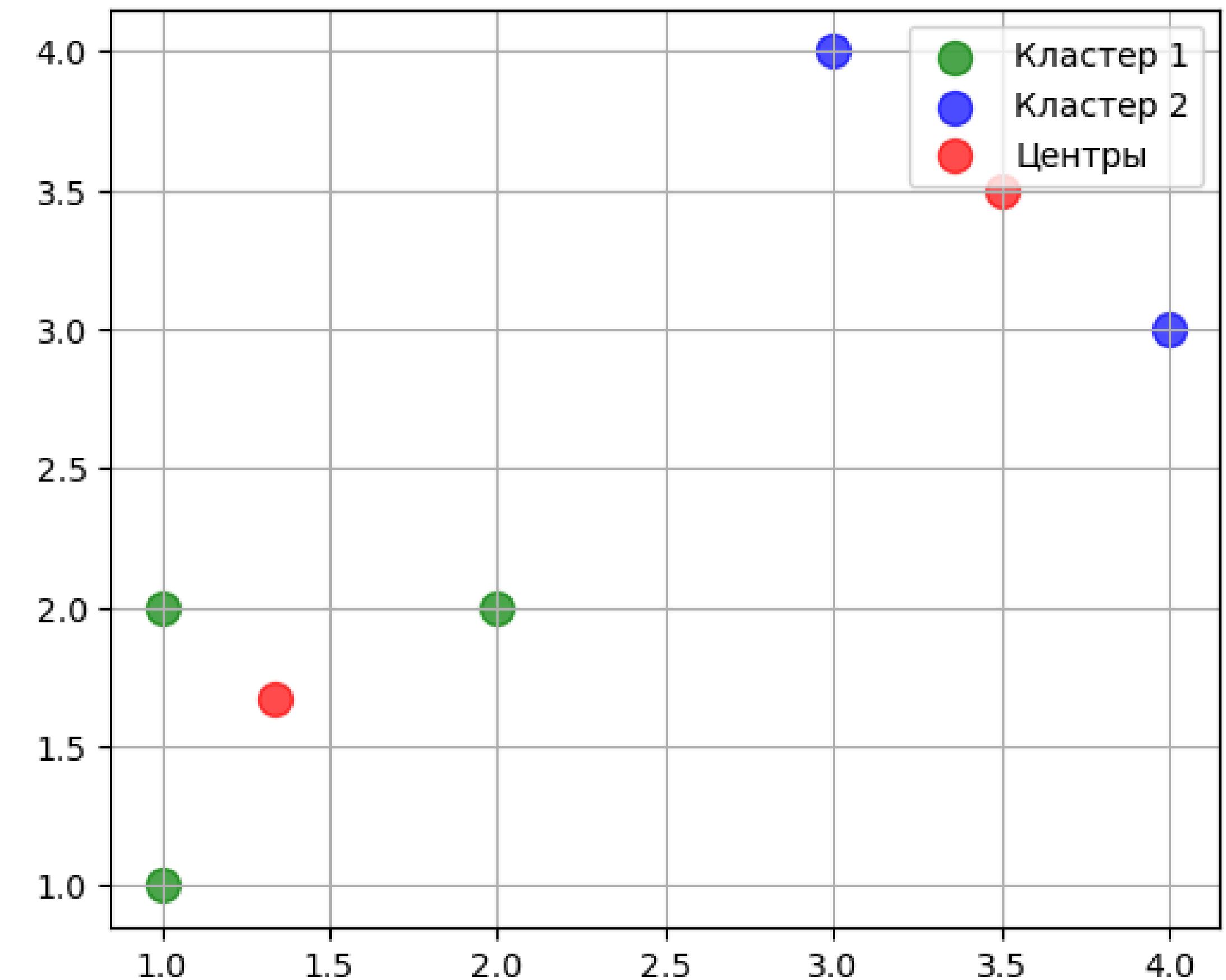
2 кластер с центром $(3,5, 3,5)$:

$(3, 4), (4, 3)$

Сумма квадратов расстояний для первого кластера $\sim 1,35$

Сумма квадратов расстояний для второго кластера = 1

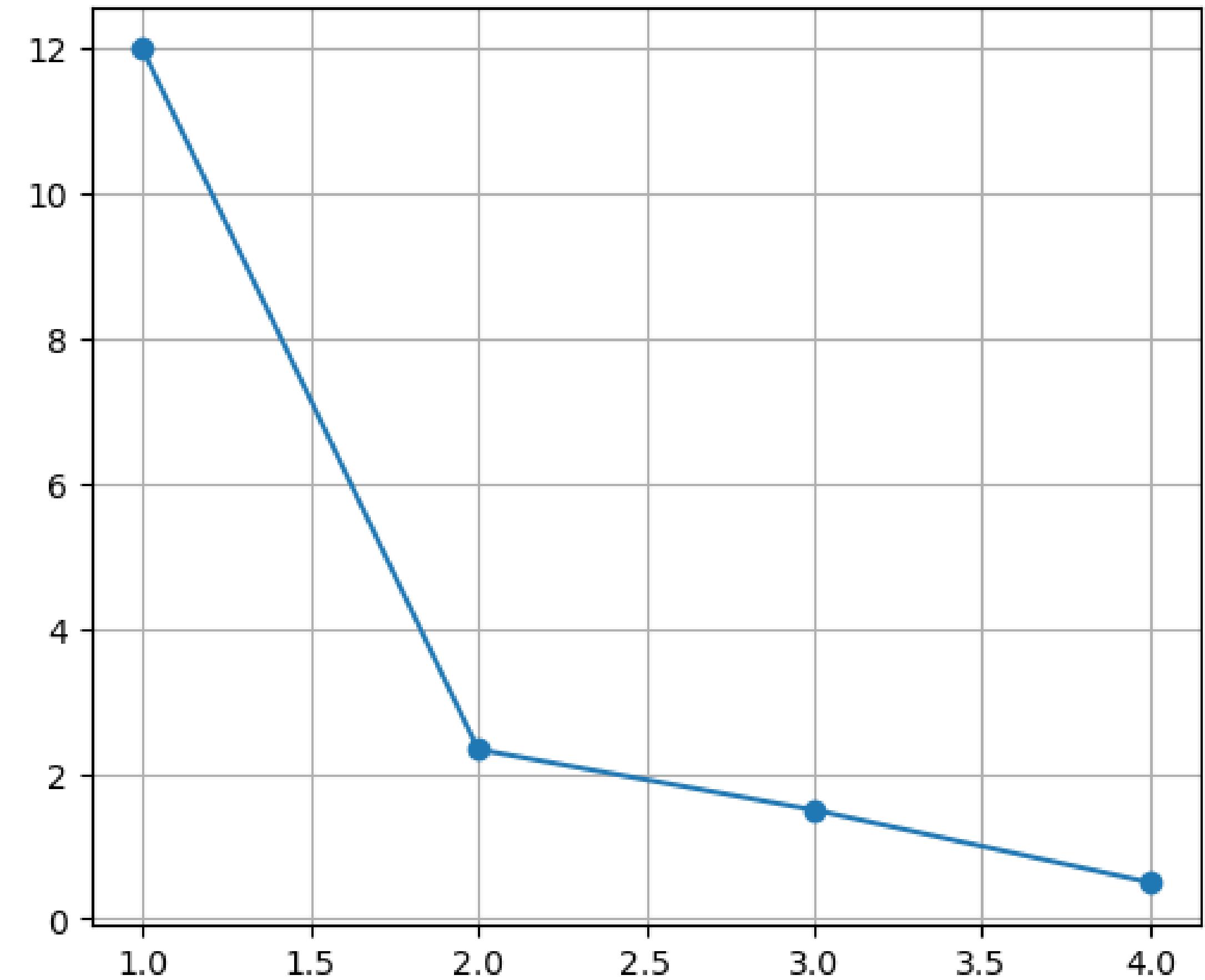
Общая сумма квадратов расстояний для примера $\sim 2,35$



Метод локтя на примере

Теперь аналогично высчитаем расстояние для другого количества кластеров и построим график зависимости квадрата расстояний от числа кластеров.

По графику можно заметить, что после 2-х изменения незначительны. Поэтому можно сделать вывод, что наше предположение ранее оказалось верным и 2 действительно является оптимальным числом кластеров.



Метод силуэта на примере

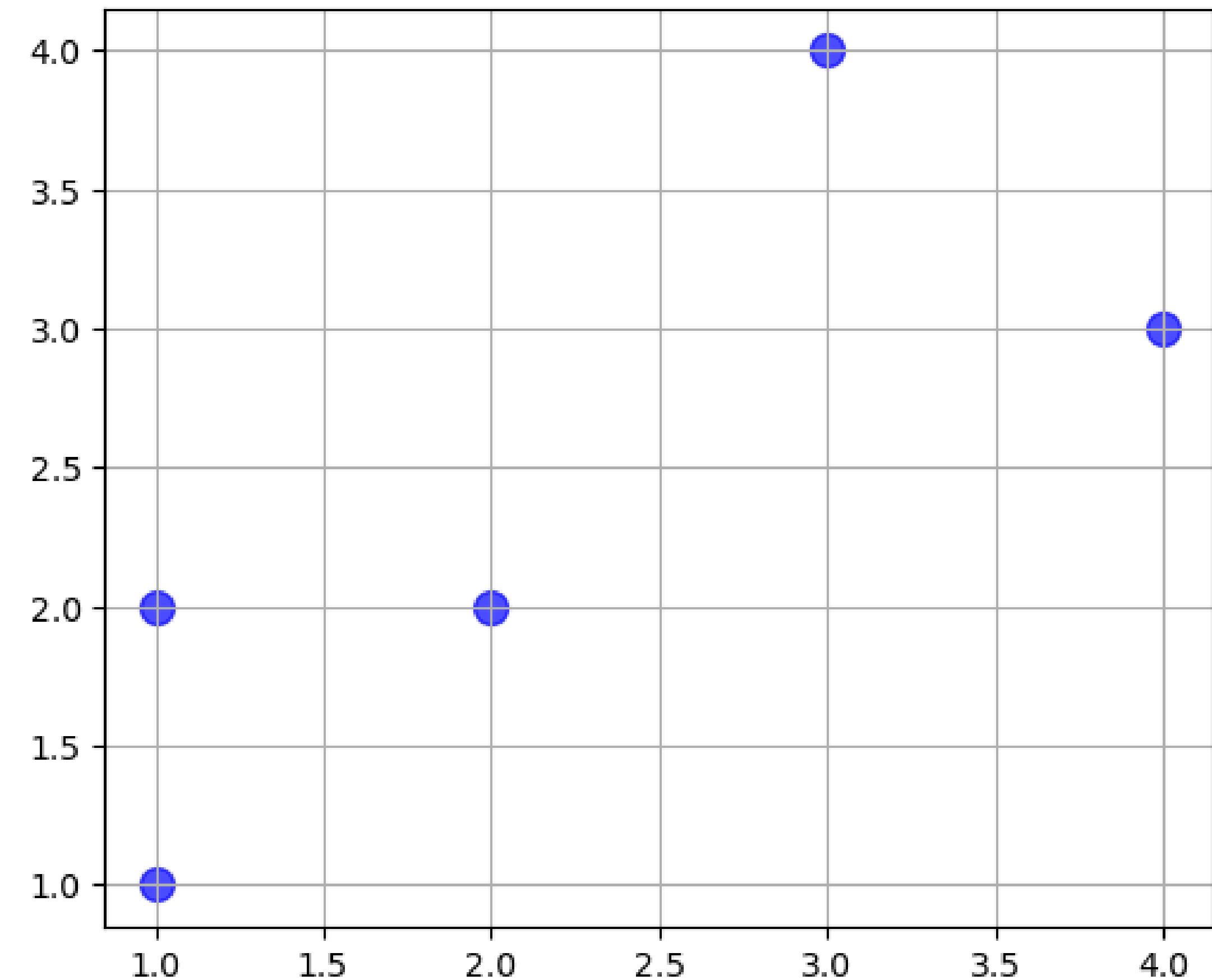
Пример

Рассмотрим тот же набор из 5 точек:

$(1, 1), (1, 2), (2, 2), (3, 4), (4, 3)$

Проверим методом силуэта какое число кластеров оптимально для наших данных.

Построим график зависимости значения силуэта от числа кластеров.



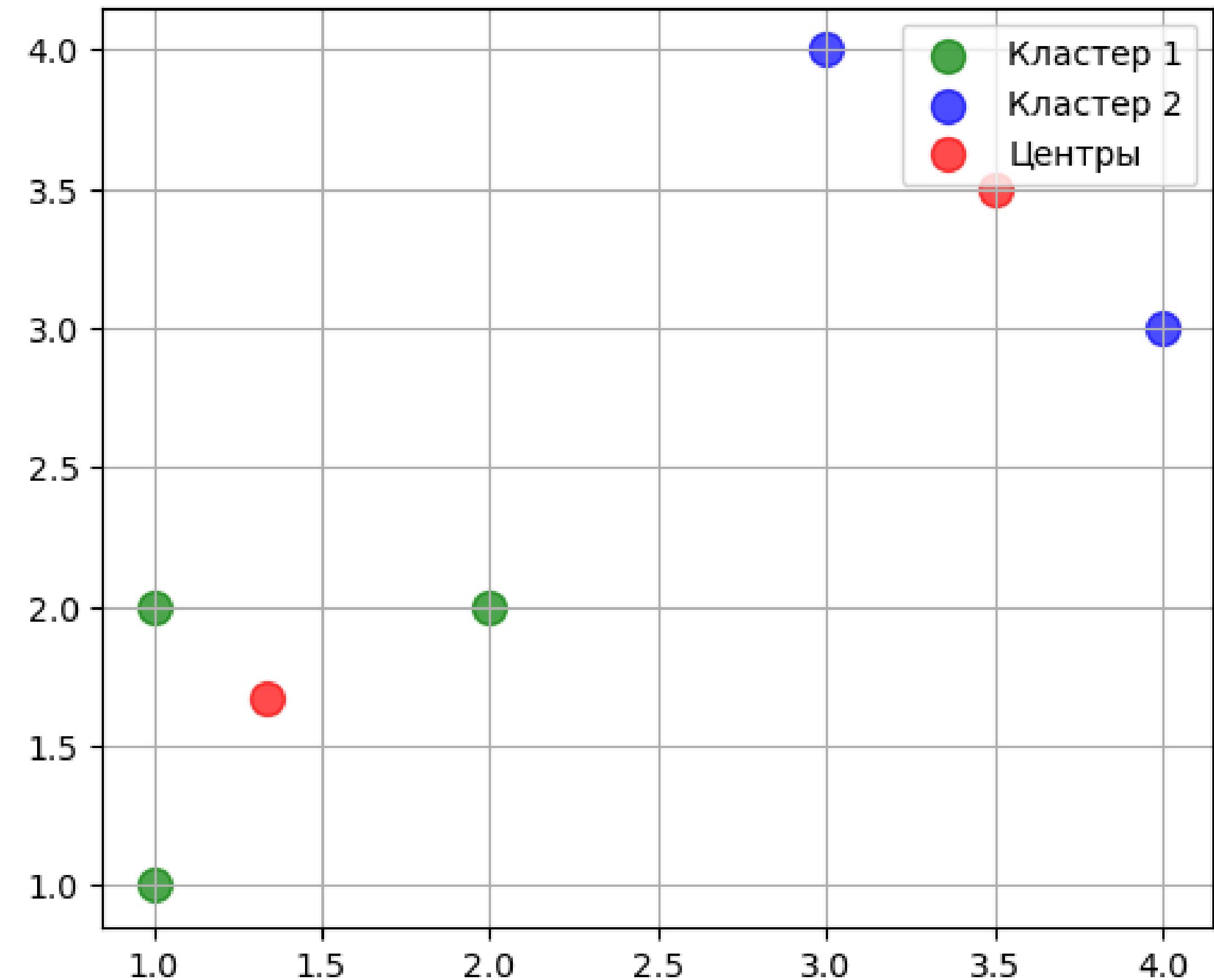
Метод силуэта на примере

Вычислим значение силуэта для 2-х кластеров, которые мы получили ранее.

Для начала оценим качество кластеризации для каждой точки по формуле:

$$s = \frac{b - a}{\max(a, b)}$$

где a — среднее внутрикластерное расстояние,
 b — среднее расстояние до ближайшего кластера



Метод силуэта на примере

1 кластер с центром (1,3, 1,6):

(1, 1), (1, 2), (2, 2)

2 кластер с центром (3,5, 3,5):

(3, 4), (4, 3)

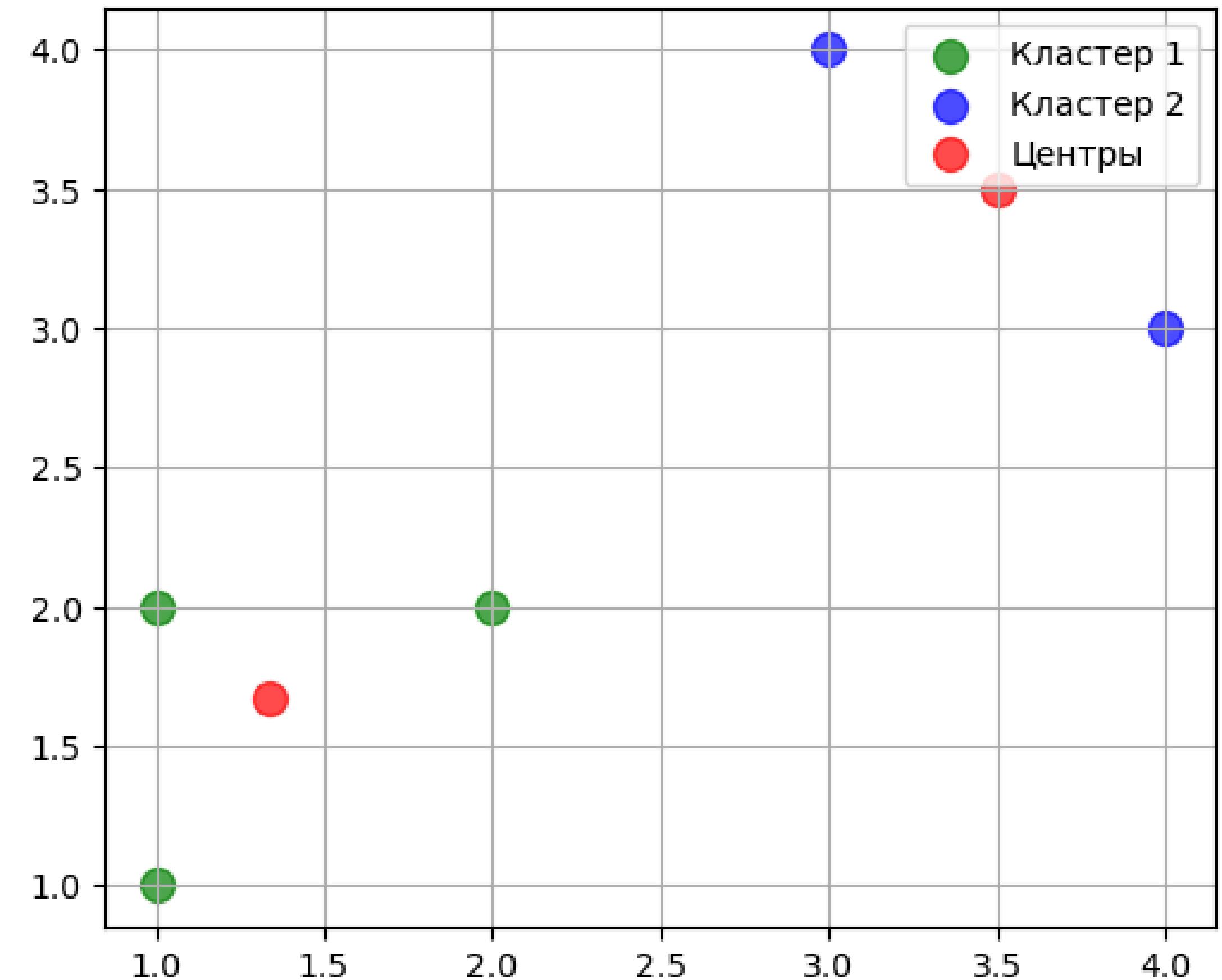
Оценим качество кластеризации для точки (1, 1).

Точка находится в кластере 1, рассчитываем среднее расстояние от этой точки до всех остальных точек в кластере:

$$d((1, 1), (1, 2)) = 1$$

$$d((1, 1), (2, 2)) \sim 1,4$$

Среднее внутрикластерное расстояние ~1,2



Метод силуэта на примере

Теперь рассчитываем среднее расстояние от точки до ближайшего соседнего кластера:

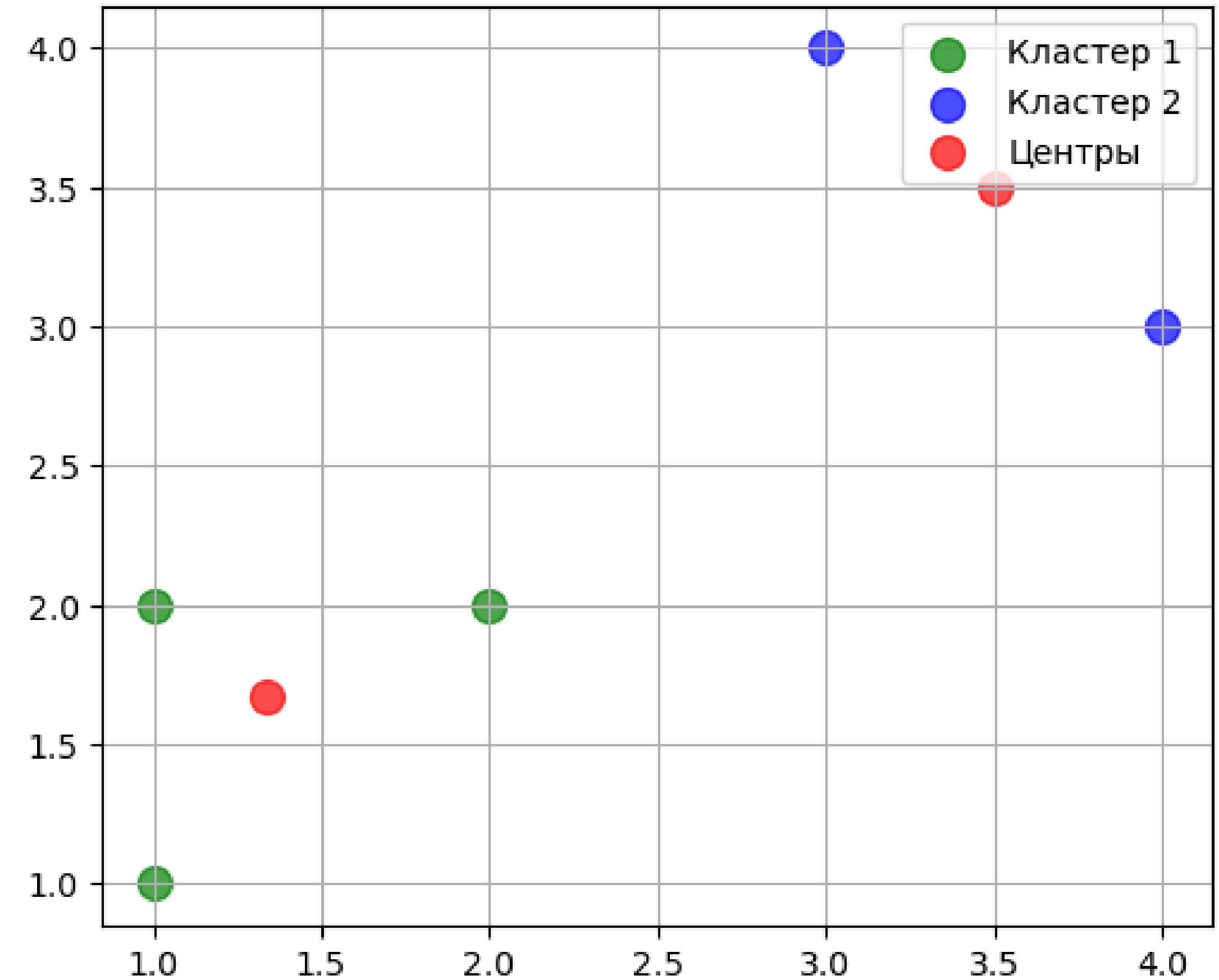
$$d((1, 1), (3, 4)) \sim 3,6$$

$$d((1, 1), (4, 3)) \sim 3,6$$

Среднее межкластерное расстояние $\sim 3,6$

Вычислим силуэт для точки (1, 1):

$$s = \frac{3,6 - 1,2}{\max(3,6, 1,2)} = \frac{2,4}{3,6} \simeq 0,67$$



Метод силуэта на примере

Рассчитаем аналогично значения силуэта для каждой точки:

(1, 1) ~ 0,67

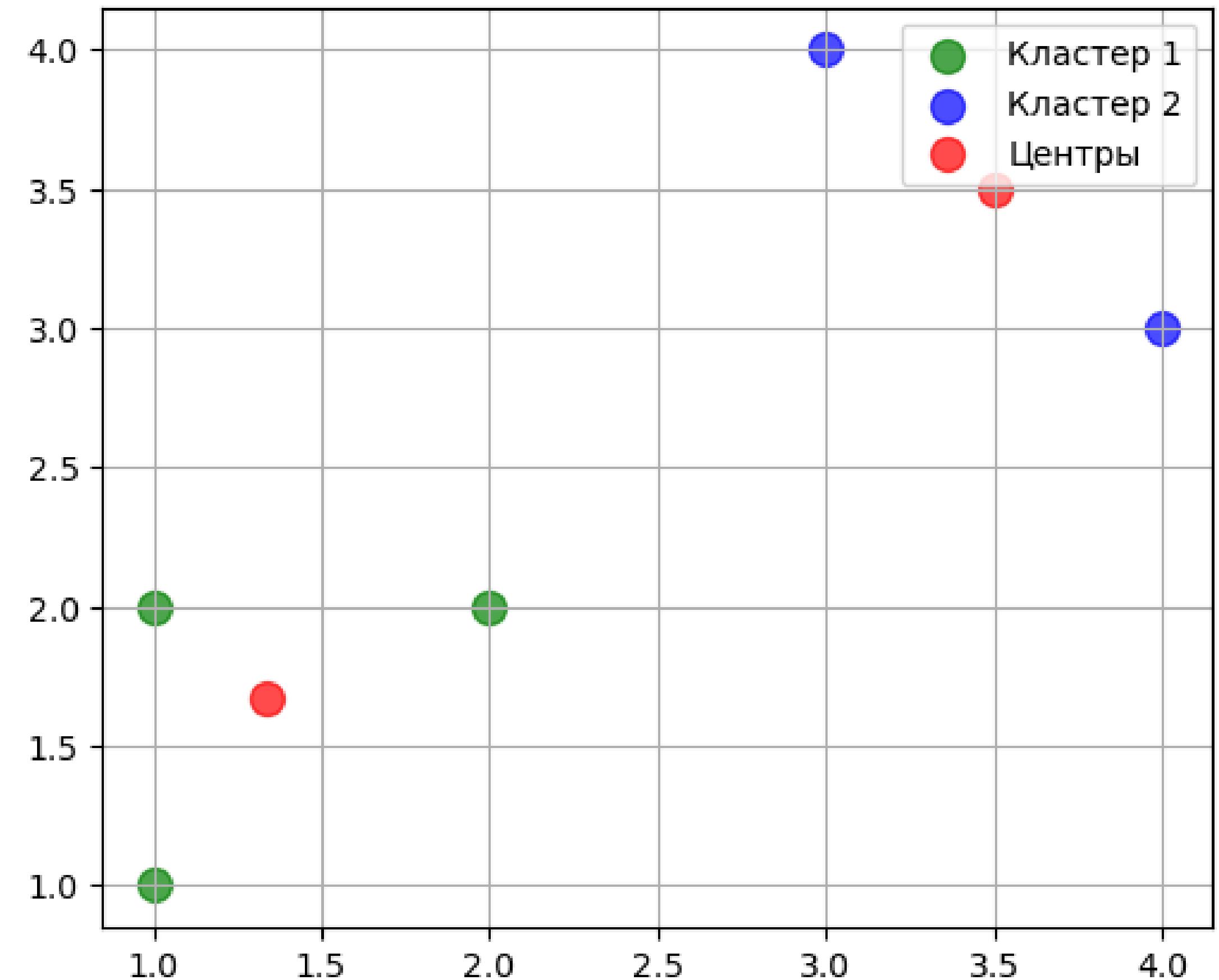
(1, 2) ~ 0,67

(2, 2) ~ 0,45

(3, 4) ~ 0,51

(4, 3) ~ 0,47

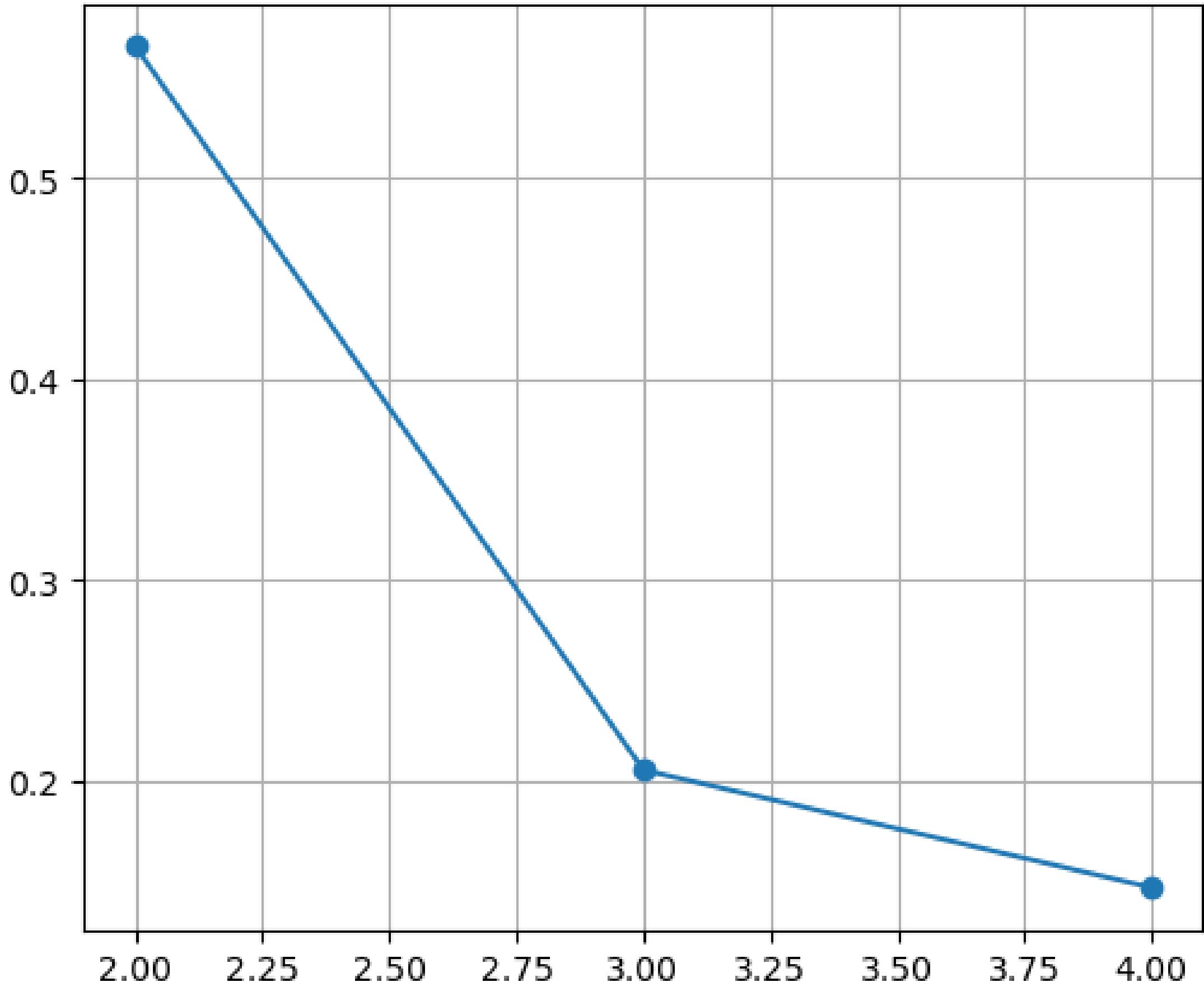
Общее значение силуэта вычисляем как среднее для всех точек ~0,554



Метод силуэта на примере

Теперь аналогично высчитаем значение силуэта для другого количества кластеров и строим график зависимости силуэта от числа кластеров.

По графику видно, что наибольшее значение силуэта при числе кластеров равном двум. Что также подтверждает наше предположение, что 2 - это оптимальное количество кластеров для нашего примера.



Пример 2. Программа на Python

Рассмотрим набор из 20 зафиксированных точек. С помощью метода локтя и метода силуэта выберем оптимальное число кластеров и разделим точки на полученное количество кластеров с помощью метода k-means.



Ограничения метода

Локальные минимумы

Алгоритм K-means может застрять в локальном минимуме, зависящем от начальной инициализации центров.

Количество кластеров

Необходимо заранее определить количество кластеров k , которое нужно найти в данных.

Начальная инициализация

Разные начальные центры могут привести к разным кластерам.

Преимущества метода

Простота

Алгоритм K-means относительно прост в реализации и понимании.

Скорость

K-means может работать с огромными наборами данных, он может быть использован для решения многих сложных задач машинного обучения.

Широкая поддержка

Алгоритм K-means реализован в различных фреймворках машинного обучения, что облегчает его использование.

Заключение

Метод k-means

— это простой и эффективный способ кластеризации данных. Он активно используется в различных областях.

Основная его проблема — правильный выбор числа кластеров k и чувствительность к начальным условиям, но его простота и скорость делают его очень популярным в практике машинного обучения.

Литература

При подготовке презентации использовались:

1. Статья на Habr “Алгоритм k-means и метод локтя: кластеризация данных с примерами на Python” (<https://habr.com/ru/companies/skillfactory/articles/877684/>)
2. Статья на SkyPro “Что такое k-means: принцип работы и применение алгоритма кластеризации” (<https://sky.pro/wiki/analytics/chto-takoe-k-means-princip-raboty-i-primenenie-algoritma-klasterizatsii/>)