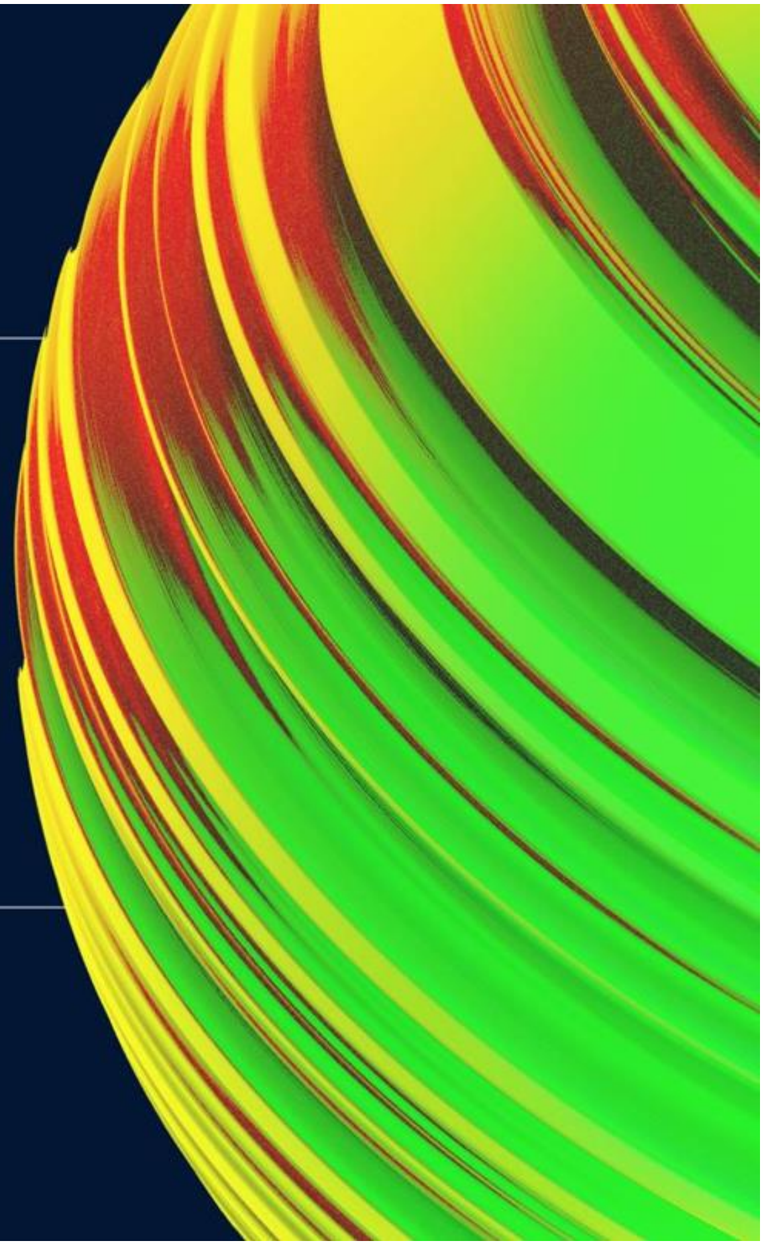
УНИВЕРСИТЕТ ИННОПОЛИС

k-means кластеризация



Корнеева Елена
e.korneeva@innopolis.ru
https://t.me/Allyonzy

Елена Корнеева

Преподаватель на курсе Аналитика данных и машинное обучение

Интересы:
- программирование;
- data mining,
- text mining (NLP, компьютерная лингвистика),
- machine learning,
- компьютерное зрение (распознавание документов, сегментация фотографий и распознавание объектов)

Общая информация:
- Преподаватель и наставник с 2019 года
- Занимаюсь машинным обучением с 2015 года
- Разработчик ПО с 2014 года

План занятия (лекция + семинар)

★ Обучение без учителя. Что такое кластеризация?

★ Алгоритм кластеризации Ллойда или k-means

★ Как выбрать оптимальное число кластеров? Методы локтя и силуэта

Классическое Обучение

Данные заранее категоризированы или численные

Данные никак не размечены

С учителем

Предсказать категорию

Предсказать значение

Классификация

«Разложи носки по цвету»

Регрессия

«Разложи галстуки по длине»

Без учителя

Разделить по схожести

Найти зависимости

Выявить последовательности

Кластеризация

«Разложи похожие вещи по кучкам»

Ассоциация

«Найди какие вещи я часто ношу вместе»

Уменьшение Размерности (обобщение)

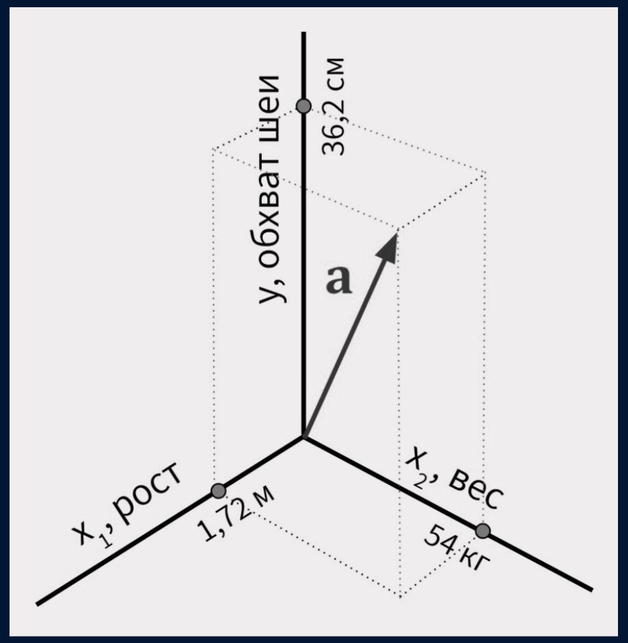
«Собери из вещей лучшие наряды»

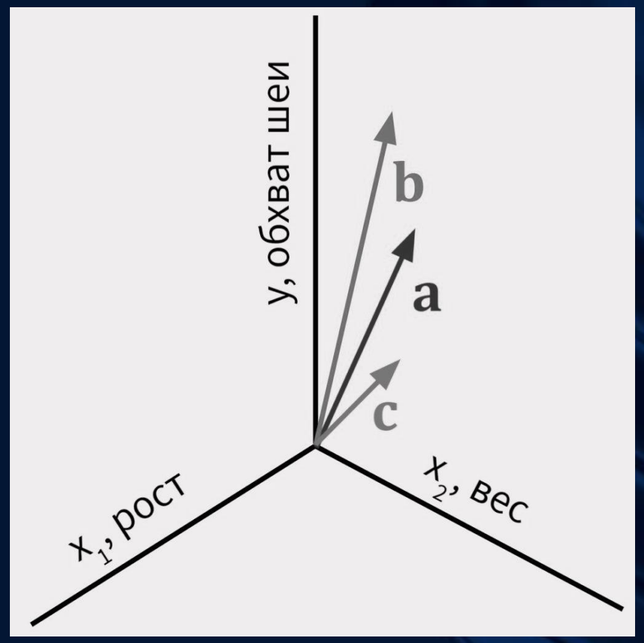
Данные как векторы и матрицы

- ✓ Данные многомерны (n-мерны)
- ✓ Можем представить данные на графике (до n=3), в форме вектора.

Вектор — набор чисел, описывающих координаты данных. Если взять несколько точек и, соответственно, несколько векторов, то получится набор чисел, называемых матрицей.
Вектор — матрица, в которой один столбец или одна строка.
Кластерный анализ использует подход по сравнению векторов данных (оценку схожести). Измеряем расстояние между точками (функция расстояния). Принимаем решение к какому кластеру отнести тот или иной объект по расстоянию.

$$\mathbf{a} = \begin{vmatrix} 1{,}72 & 54 & 36{,}2 \end{vmatrix}$$



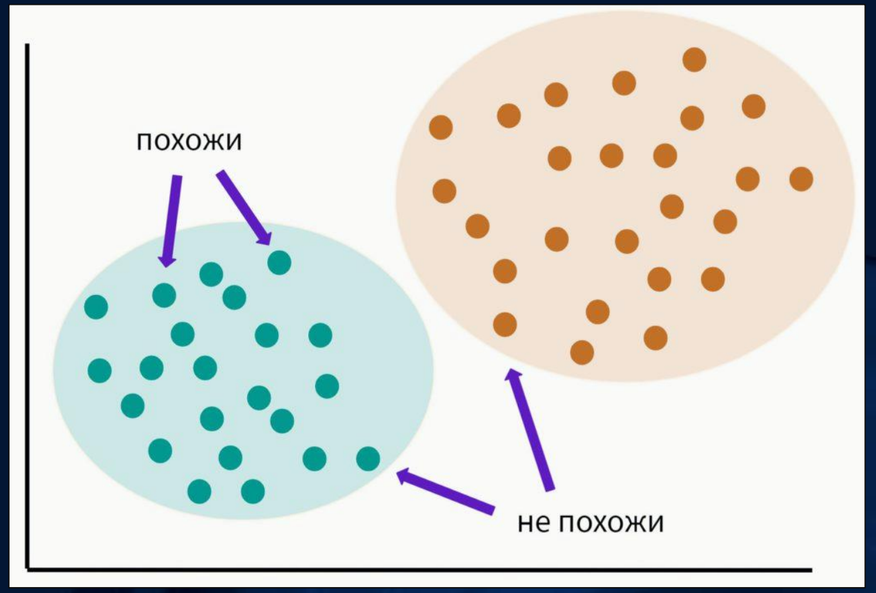


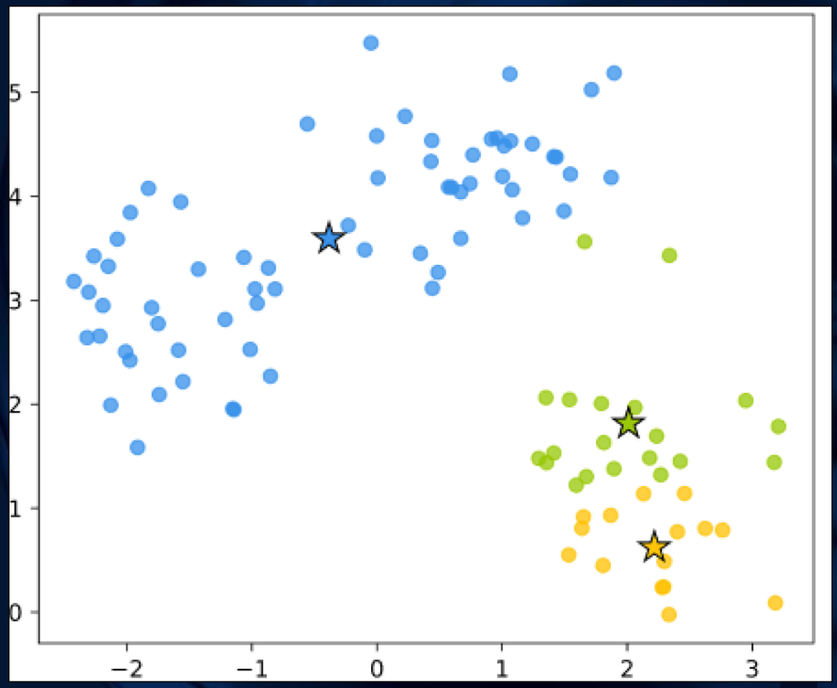
$$\begin{vmatrix} 1{,}72 & 54 & 36{,}2 \\ 1{,}74 & 58 & 36{,}3 \\ 1{,}68 & 52 & 32{,}9 \end{vmatrix}$$

Идея кластерного анализа

Основная идея **кластерного анализа** (clustering, cluster analysis) заключается в том, чтобы разбить объекты на группы или кластеры таким образом, чтобы внутри группы эти наблюдения были более похожи друг на друга, чем на объекты другого кластера.

Алгоритм без учителя (Unsupervised Learning). Обучаем модель на неразмеченных данных (unlabeled data), то есть без целевой переменной, компонента y. Заранее не известно на какие кластеры разбить данные





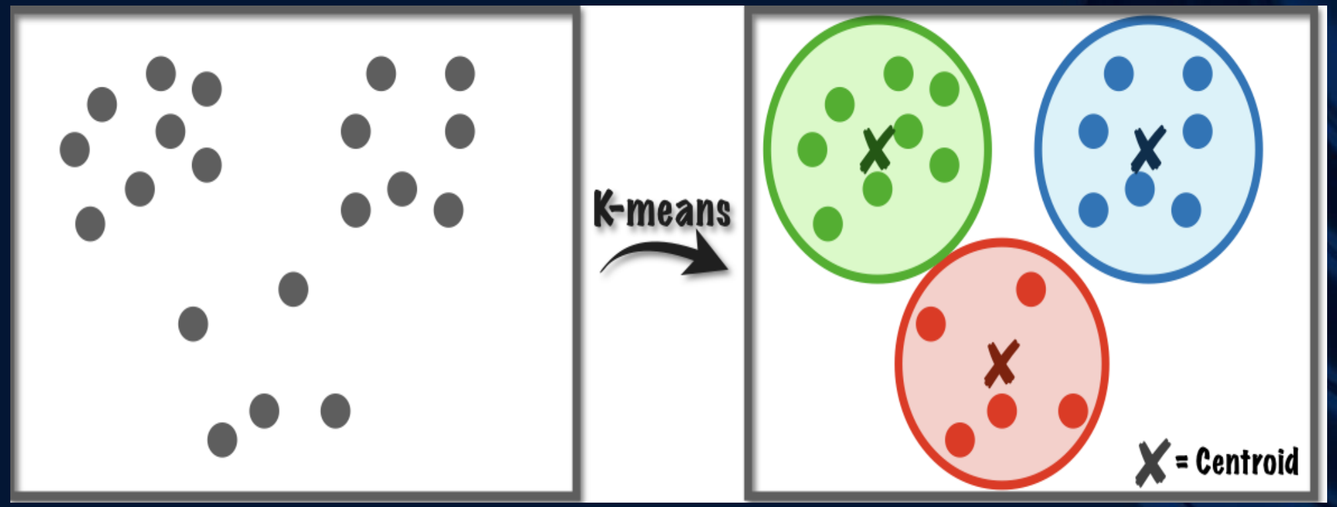
Постановка задачи

Задача кластеризации (обучения без учителя) заключается в следующем:

- ✓ Имеется обучающая выборка $X\ell = \{x1, \ldots, x\ell\} \subset X$ и функция расстояния между объектами $\rho(x, x')$.
- ✓ Требуется разбить выборку на непересекающиеся подмножества, называемые кластерами, так, чтобы каждый кластер состоял из объектов, близких по метрике ρ, а объекты разных кластеров существенно отличались.
- ✓ При этом каждому объекту $xi \in X\ell$ приписывается метка (номер) кластера yi.

Алгоритм кластеризации — это функция $a: X \to Y$, которая любому объекту $x \in X$ ставит в соответствие метку кластера $y \in Y$. Множество меток Y в некоторых случаях не известно заранее.

Неразмеченные данные (unlabelled data)



Размеченные кластеры (labelled clusters)

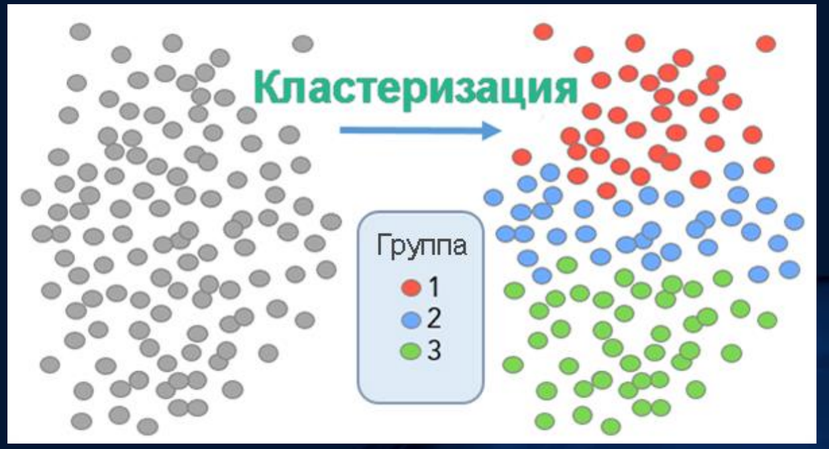
Цели кластеризации

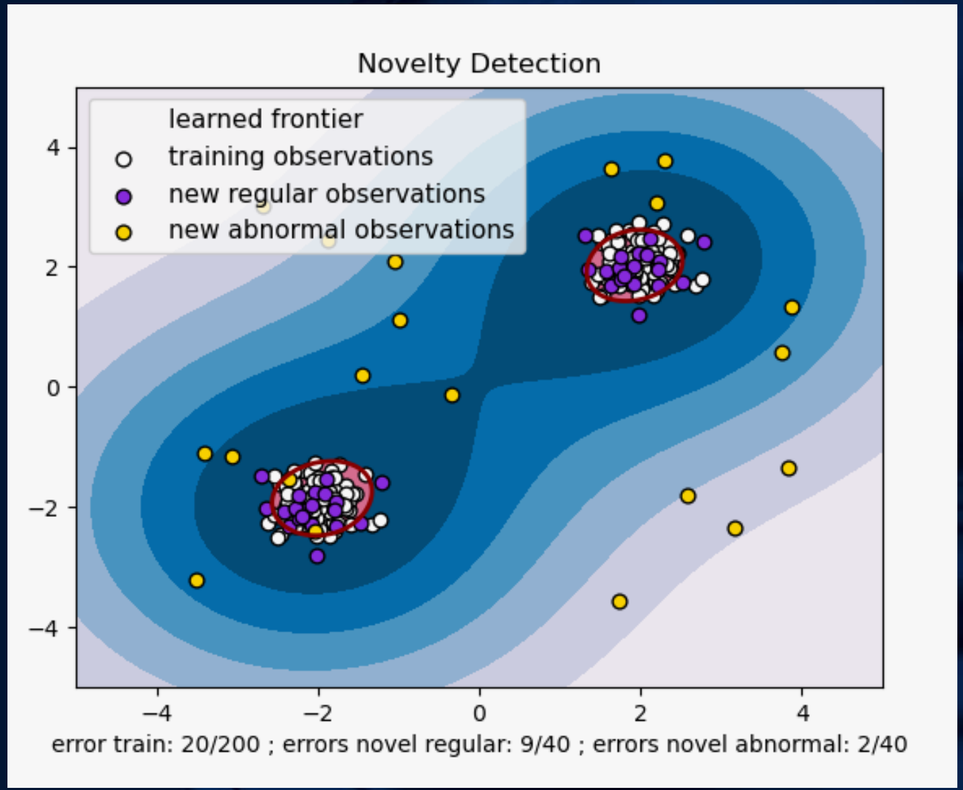
1. Разбить множество объектов Xℓ на группы и понять структуру множества.

2. Упростить дальнейшую обработку данных и принятия решений, работая с каждым кластером по отдельности (стратегия «разделяй и властвуй»).

3. Сократить объём хранимых данных (снижение размерности) в случае сверхбольшой выборки Xℓ, оставив по одному наиболее типичному представителю от каждого кластера.

4. Выделить нетипичные объекты, которые не подходят ни к одному из кластеров. Эту задачу называют одноклассовой классификацией, обнаружением нетипичности или новизны (novelty detection).





Области применения кластерного анализа данных

Маркетинг

Для сегментирования рынка:
- выявление закономерностей в покупках, совершаемых клиентами;
- выделение групп потребителей со схожими стереотипами поведения и т.п.).

Банковское дело

Для определения типичных групп (профилей) добросовестных и неблагонадёжных заёмщиков.

Страховой бизнес

Для получения профилей клиентов с целью определения услуг страхования, обеспечивающих наименьшие для компании риски.

Медицина

Для выявления типичных клинических случаев и классификации медико-биологических объектов.

Телекоммуникации

Для поиска родственных групп клиентов с похожими типами пользования услугами (с целью разработки привлекательных наборов цен и услуг).

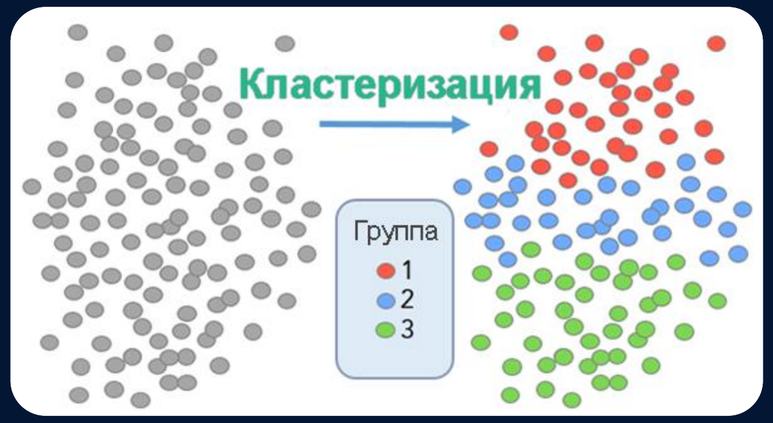
Социология

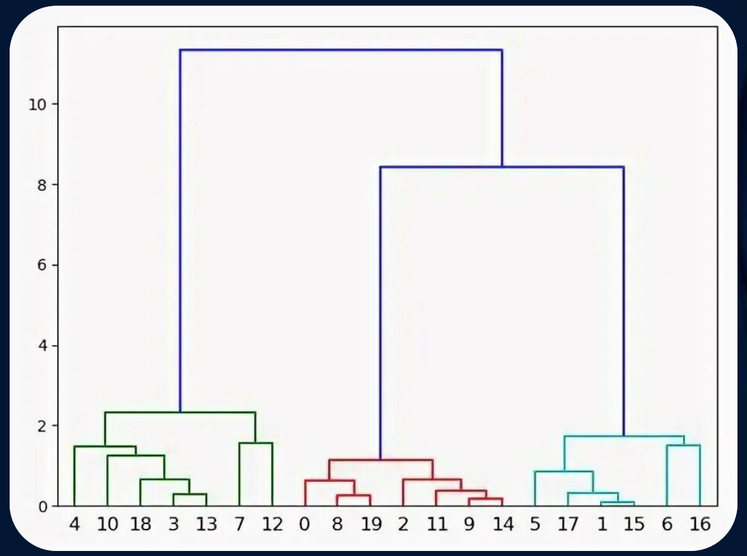
Для обработки результатов опросов общественного мнения.

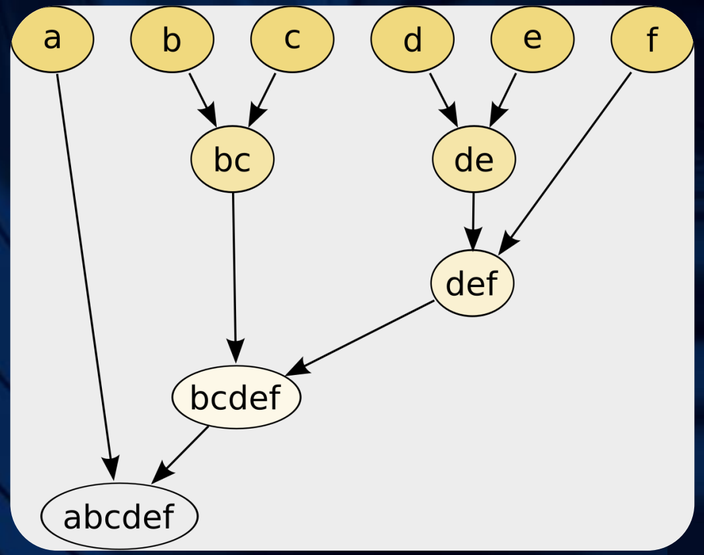
Типы кластеризации

Типы задач кластерного анализа данных:

- число кластеров априори задано (k-means, k-medoids, gaussian mixture models (GMM), spectral clustering);
- число кластеров неизвестно и подлежит определению (DBSCAN, OPTICS, mean shift, агломеративная иерархическая кластеризация (таксономия));
- число кластеров неизвестно, но его определение не является условием решения задачи, а необходимо построить иерархическое дерево (дендрограмму) разбиения анализируемой совокупности объектов на кластеры (divisive anomaly detection, hierarchical density-based clustering, top-down clustering, g-means).







Функционалы качества кластеризации

Задачу кластеризации можно ставить как задачу дискретной оптимизации: необходимо так приписать номера кластеров y_i объектам x_i, чтобы значение выбранного функционала качества приняло наилучшее значение.

Среднее внутрикластерное расстояние должно быть как можно меньше:

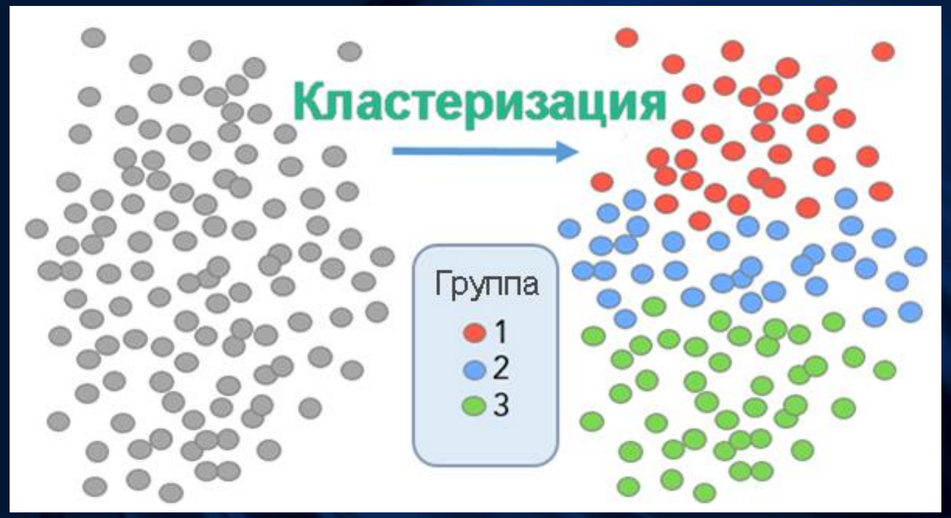
$$F_0 = \frac{\sum_{i<j}[y_i = y_j]\rho(x_i, x_j)}{\sum_{i<j}[y_i = y_j]} \to \min$$

Среднее межкластерное расстояние должно быть как можно больше:

$$F_1 = \frac{\sum_{i<j}[y_i \neq y_j]\rho(x_i, x_j)}{\sum_{i<j}[y_i \neq y_j]} \to \max.$$

Этапы кластерного анализа

- Определение множества переменных, по которым будут оцениваться объекты.
- Нормализация значений переменных.
- Выбор метода кластеризации и вида метрики для вычисления значений меры сходства между объектами.
- Создания групп сходных объектов (кластеров).
- Анализ результатов, корректировка выбранной метрики и метода кластеризации до получения желаемого результата.

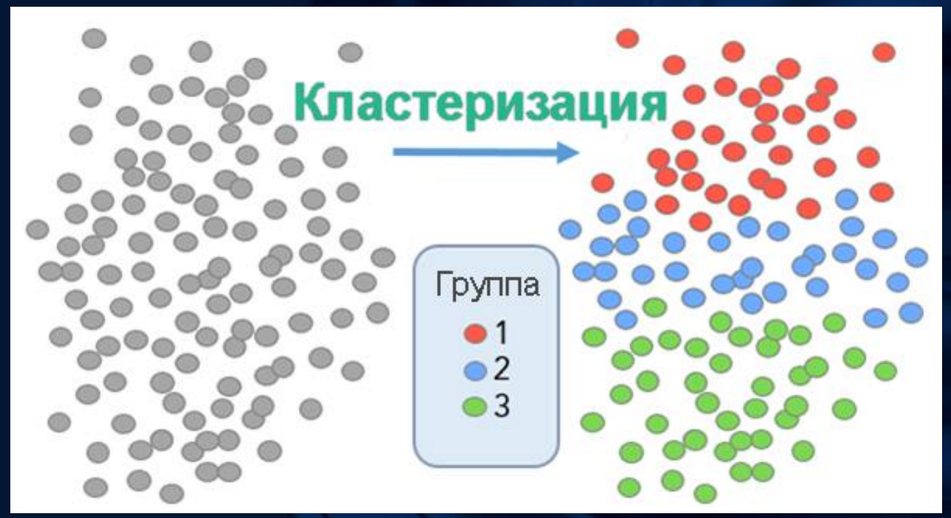


Как происходит разбиение

Чтобы сравнить два объекта, необходимо иметь критерий, на основании которого будет происходить сравнение.

Критерий - расстояние между объектами, мера близости (метрика)

- $d(\vec{X}_i, \vec{X}_j) \geq 0$
- $d(\vec{X}_i, \vec{X}_j) = 0, \rightarrow \vec{X}_i = \vec{X}_j$
- $d(\vec{X}_i, \vec{X}_j) = d(\vec{X}_j, \vec{X}_i)$
- $d(\vec{X}_i, \vec{X}_k) \leq d(\vec{X}_i, \vec{X}_j) + d(\vec{X}_j, \vec{X}_k)$



Мера близости

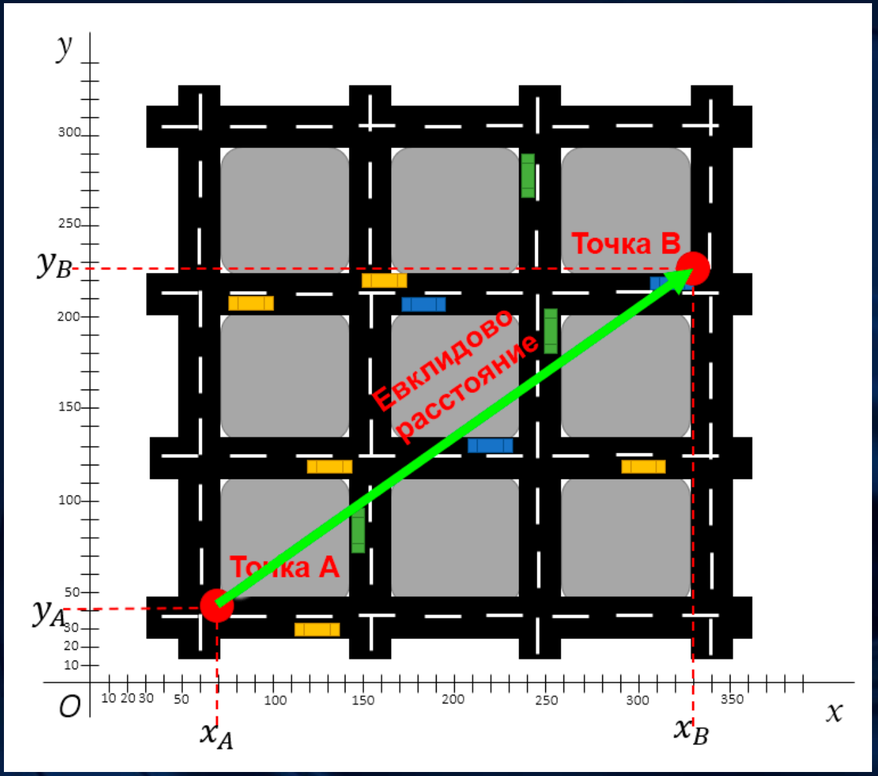
Евклидово расстояние - наиболее общий тип расстояния. Является геометрическим расстоянием между точками в многомерном пространстве. Евклидово расстояние характеризуется прямой линией.

$$d(A,B) = \sqrt{\sum_{k=1}^{n} (A_k - B_k)^2}$$

$$\rho_{ij} = \left[\sum_{k} (x_{ik} - x_{jk})^2 \right]^{1/2}$$

Квадрат евклидова расстояния – используется, чтобы придать большие веса более отдаленным друг от друга объектам

$$\rho_{ij} = \left[\sum_{k} (x_{ik} - x_{jk})^2 \right]$$



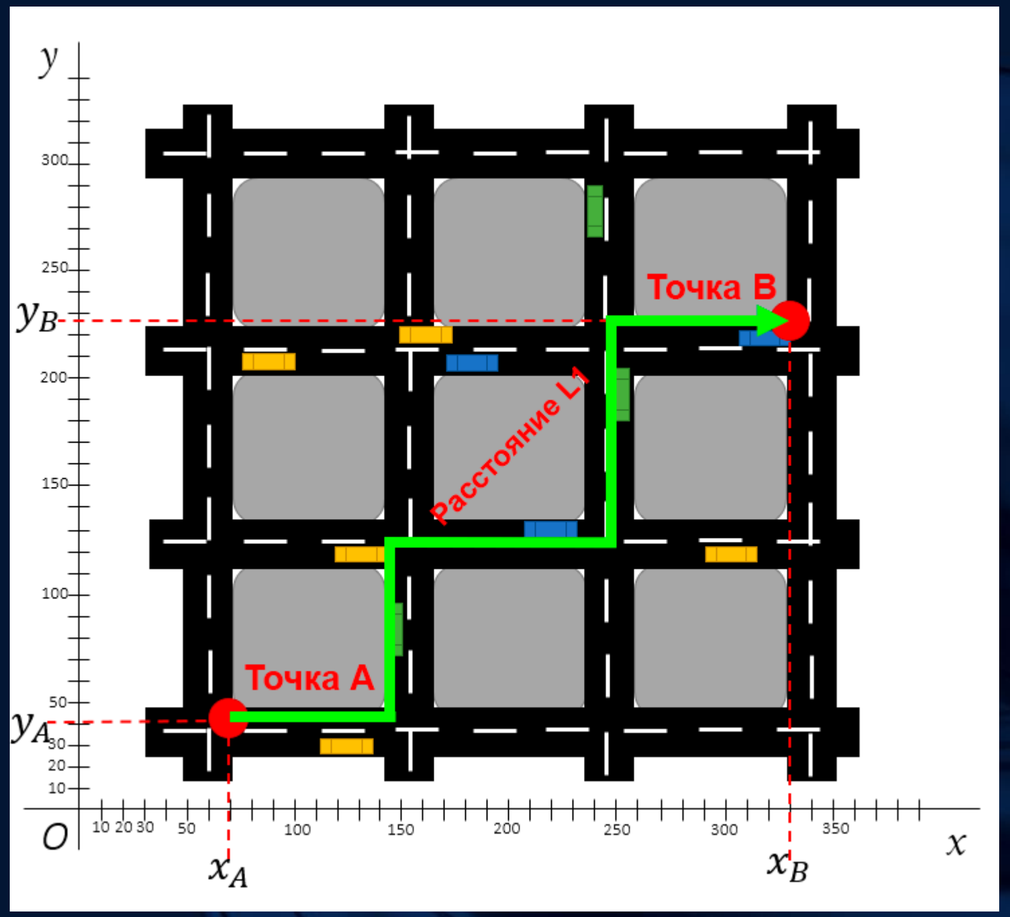
https://tproger.ru/translations/3-basic-distances-in-data-science

Мера близости

Расстояние city-block (городских кварталов, L1) или манхэттенское расстояние - по сравнением с евклидовым расстоянием влияние отдельных больших разностей (выбросов) уменьшается, так как они не возводятся в квадрат:

$$d(A,B) = \sum_{k=1}^{n} |A_k - B_k|$$

$$\rho_{ij} = \sum_{k} | x_{ik} - x_{jk} |$$



Кроме показанного пути существует несколько альтернативных способов. Например, от точки А можно подняться на два блока вверх, а потом на три блока вправо, либо же на три блока вправо и два блока вверх.

Расстояние Минковского: Обобщение Евклидова

$$\rho_{ij} = \left[\sum_{k} \left| x_{ik} - x_{jk} \right|^{p} \right]$$

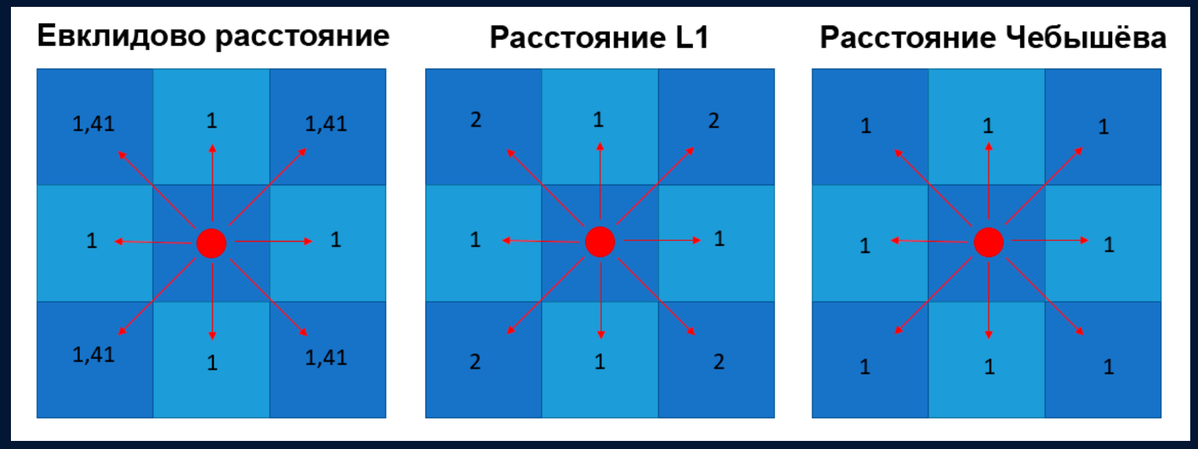
https://tproger.ru/translations/3-basic-distances-in-data-science

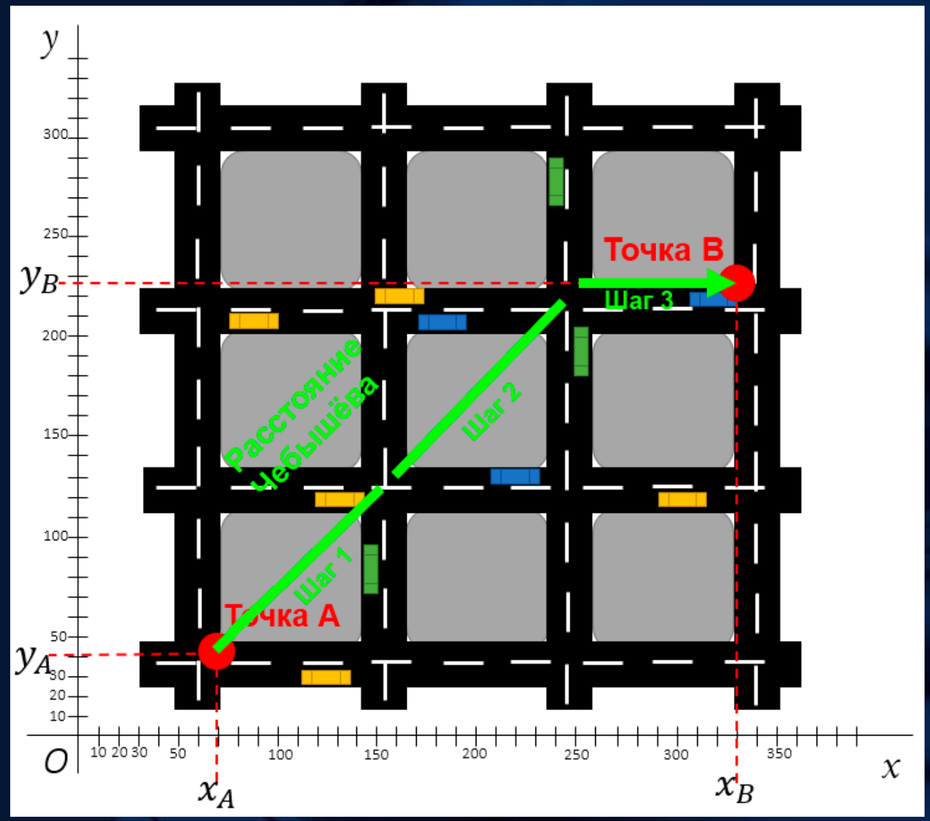
Мера близости

Расстояние Чебышёва (метрика шахматной доски) — представить короля на шахматной доске — он может ходить во всех направлениях: вперёд, назад, влево, вправо и по диагонали.

Разница расстояния L1 и расстояния Чебышёва в том, что при переходе на одну клетку по диагонали в первом случае засчитывается два хода (например вверх и влево), а во втором случае засчитывается всего один ход.
Отличаются от Евклидового расстояния тем,
что у Евклидового используется теорема Пифагора.

$$d(A,B) = \max_k |A_k - B_k|$$





https://tproger.ru/translations/3-basic-distances-in-data-science

Алгоритм K-means

Алгоритм стремится минимизировать среднеквадратичное отклонение от центра для элементов каждого кластера.

- Произвольно выбираются центры кластеров k точек
- Для каждого объекта определяется к какому кластеру он принадлежит (по умолчанию Евклидово расстояние между каждой точкой и центром кластера)
- Пересчитывается центр каждого кластера (как среднее для всех элементов кластера)

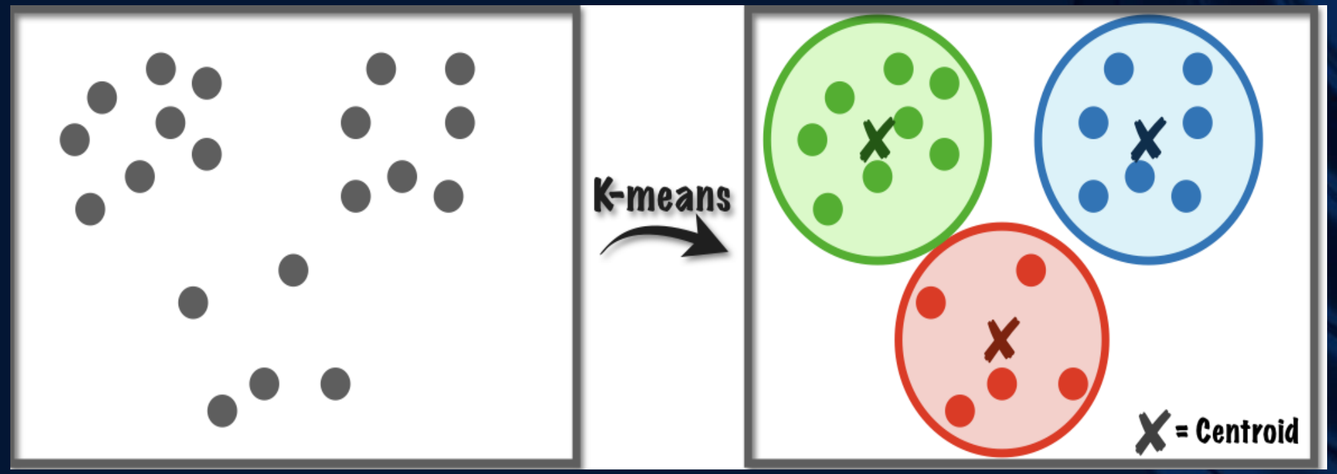
Кол-во кластеров
Кол-во наблюдений
i-ое наблюдение
центроид j-ого кластера

$$J = \sum_{j=1}^{k} \sum_{i=1}^{n} \min(||x_i^{(j)} - c_j||)^2$$

Функция расстояния

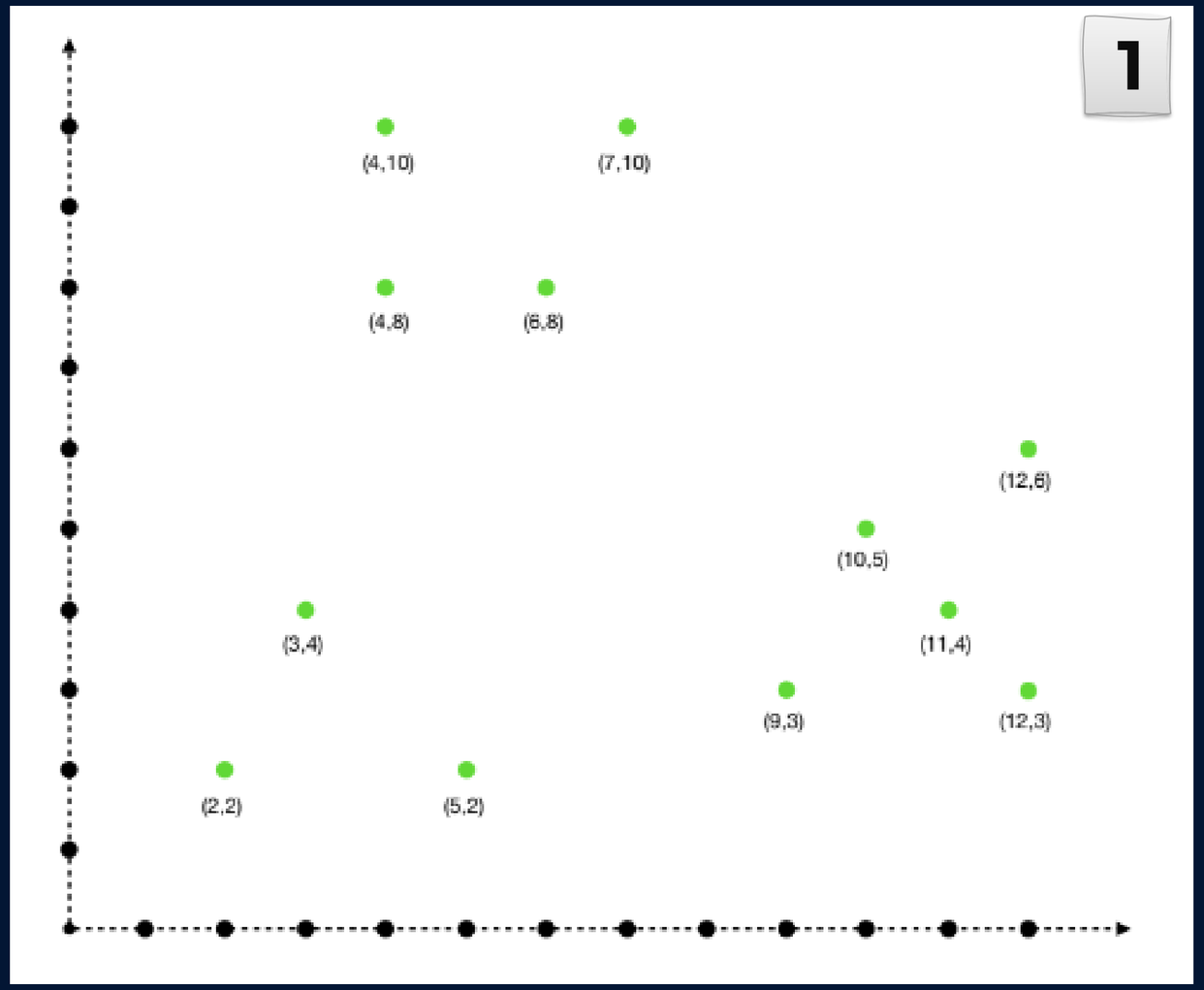
Функция потерь
(еще говорят целевая функция, objective function)

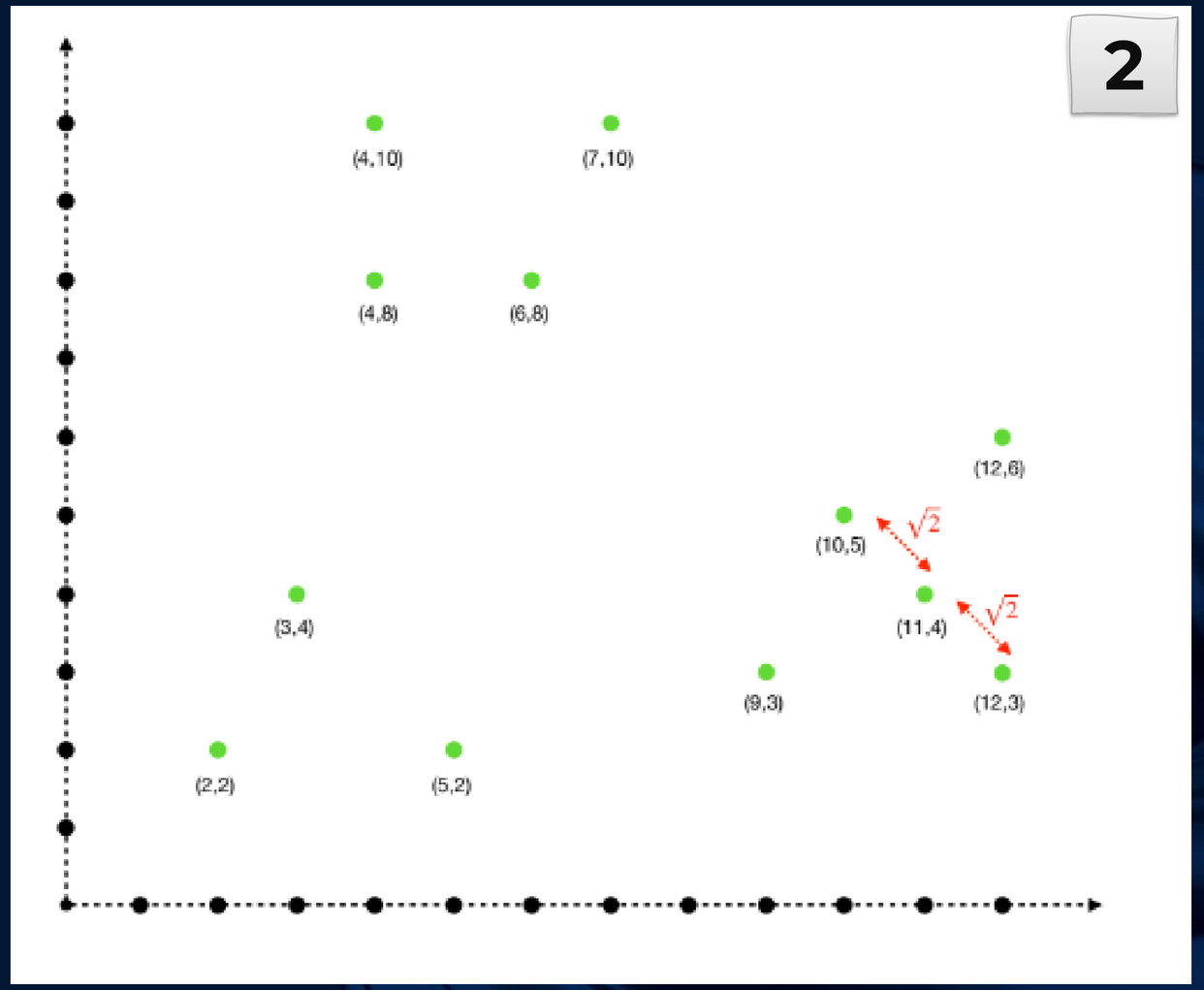
Неразмеченные данные (unlabelled data)



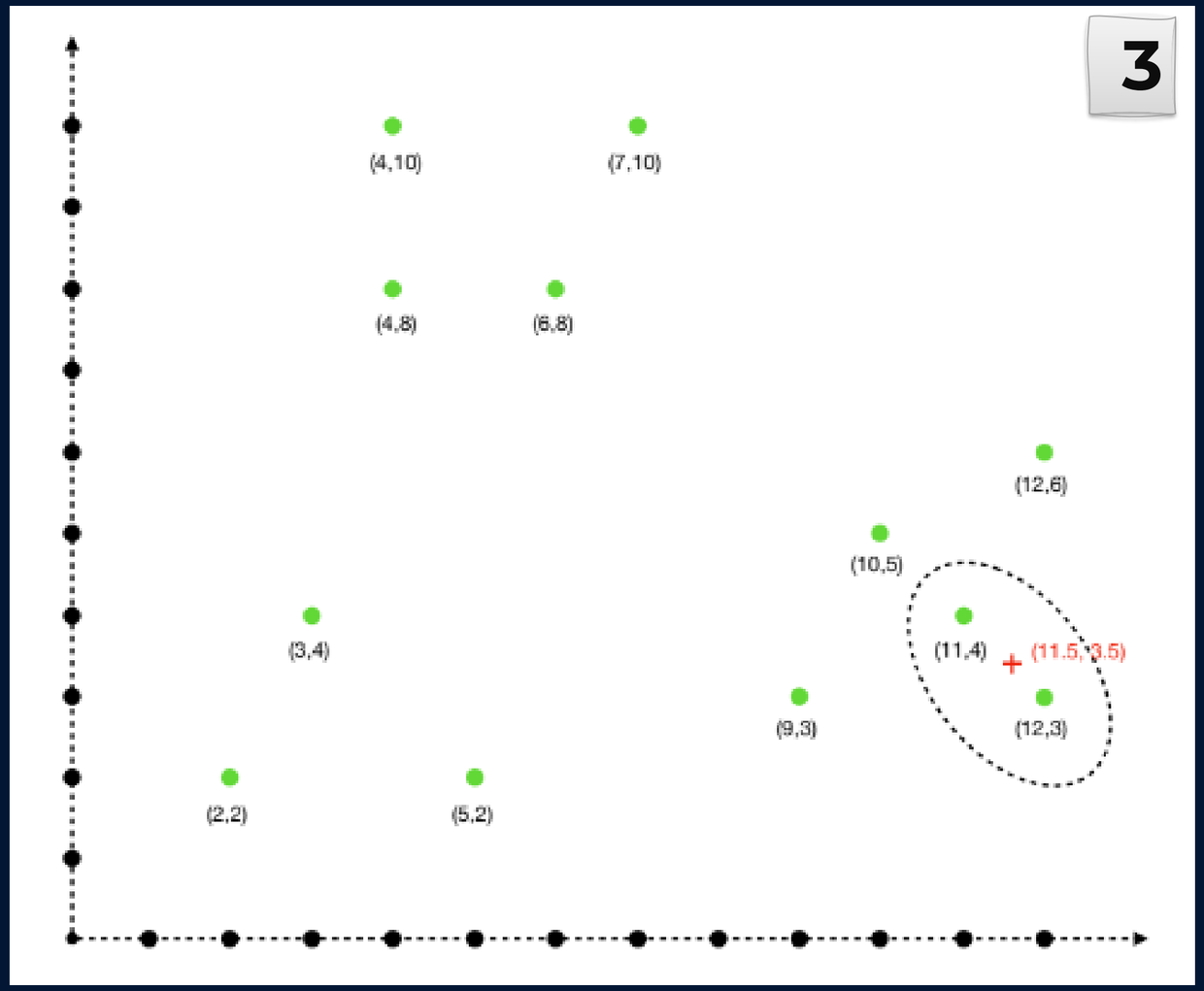
Размеченные кластеры (labelled clusters)

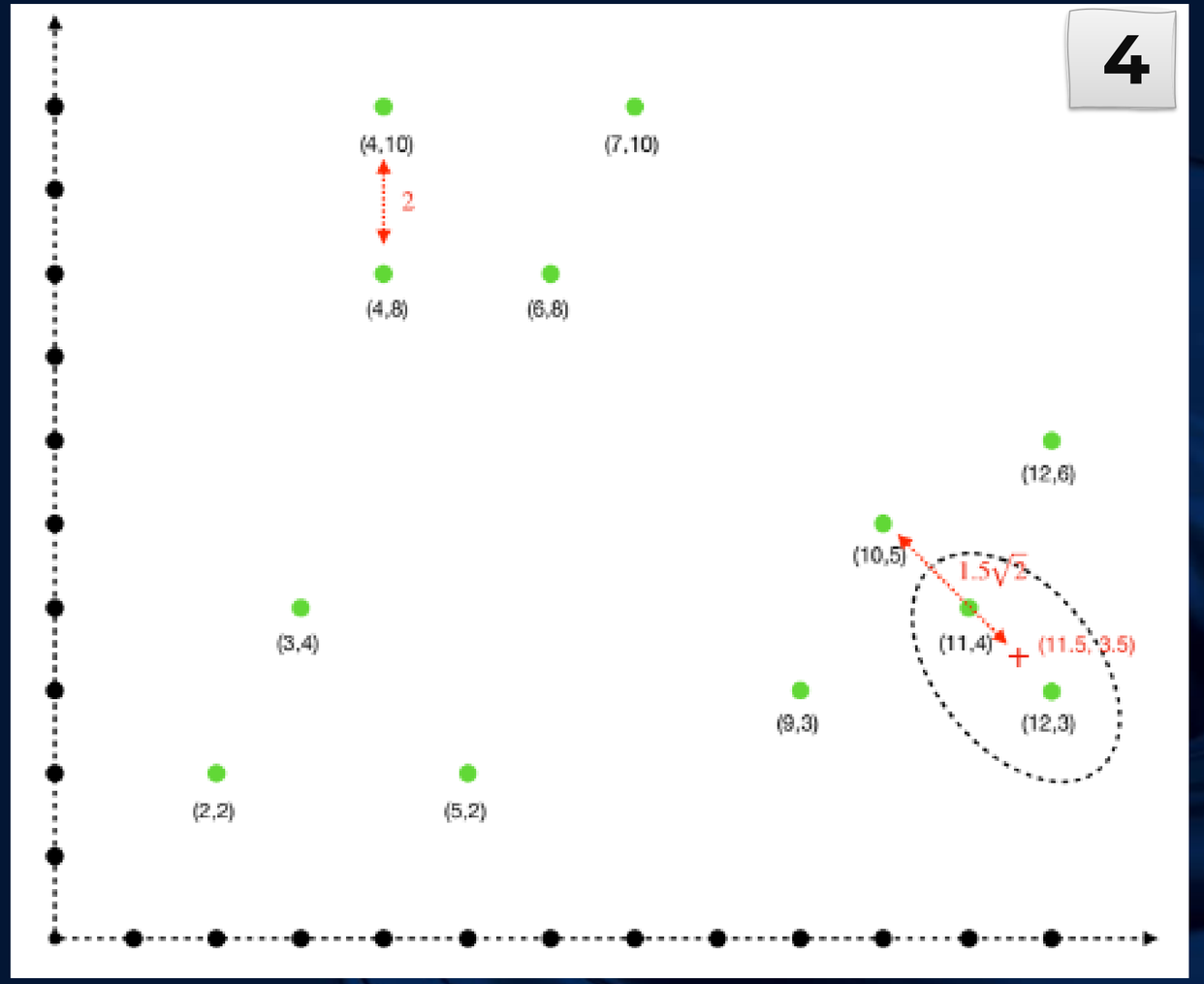
Алгоритм K-means



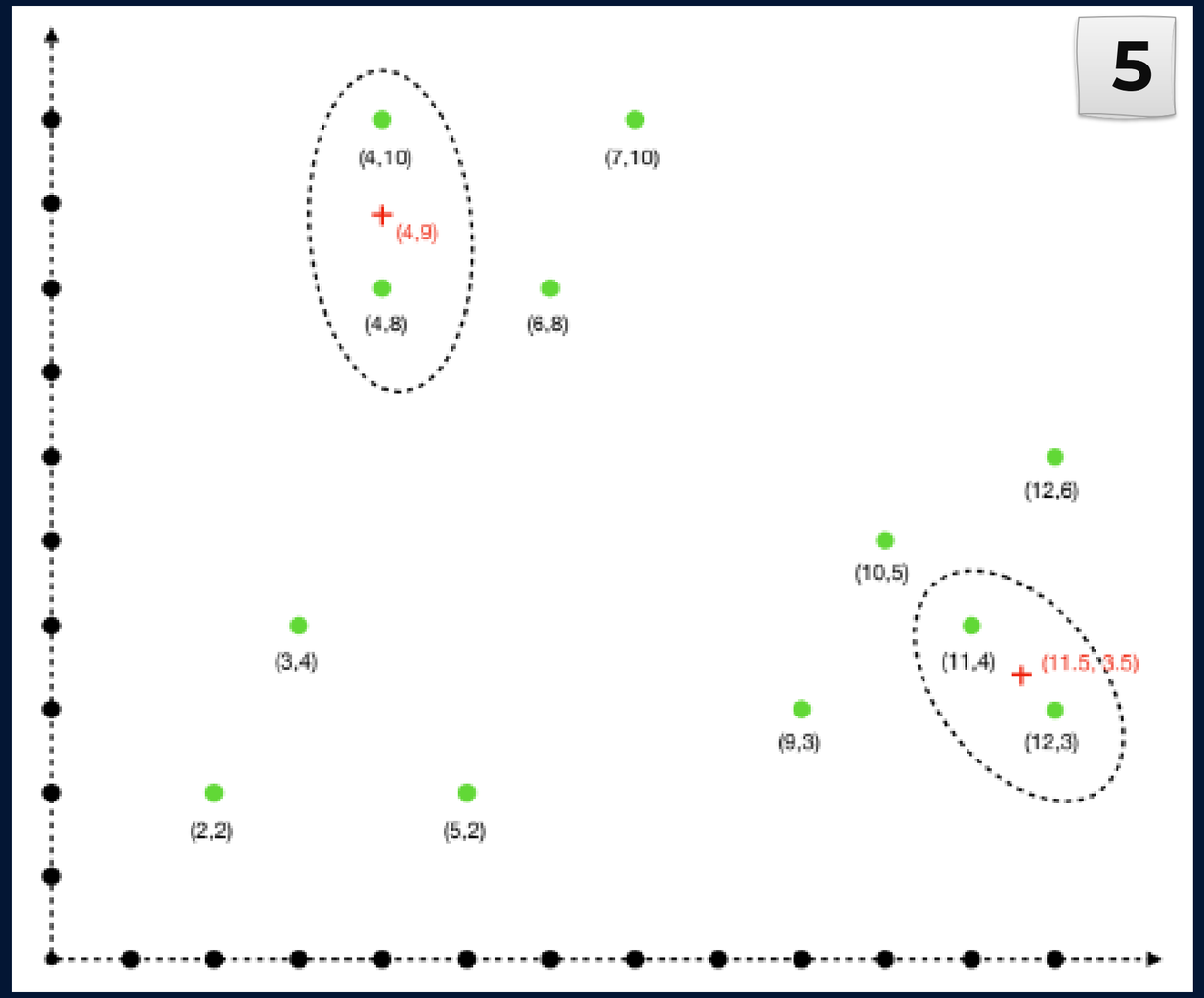


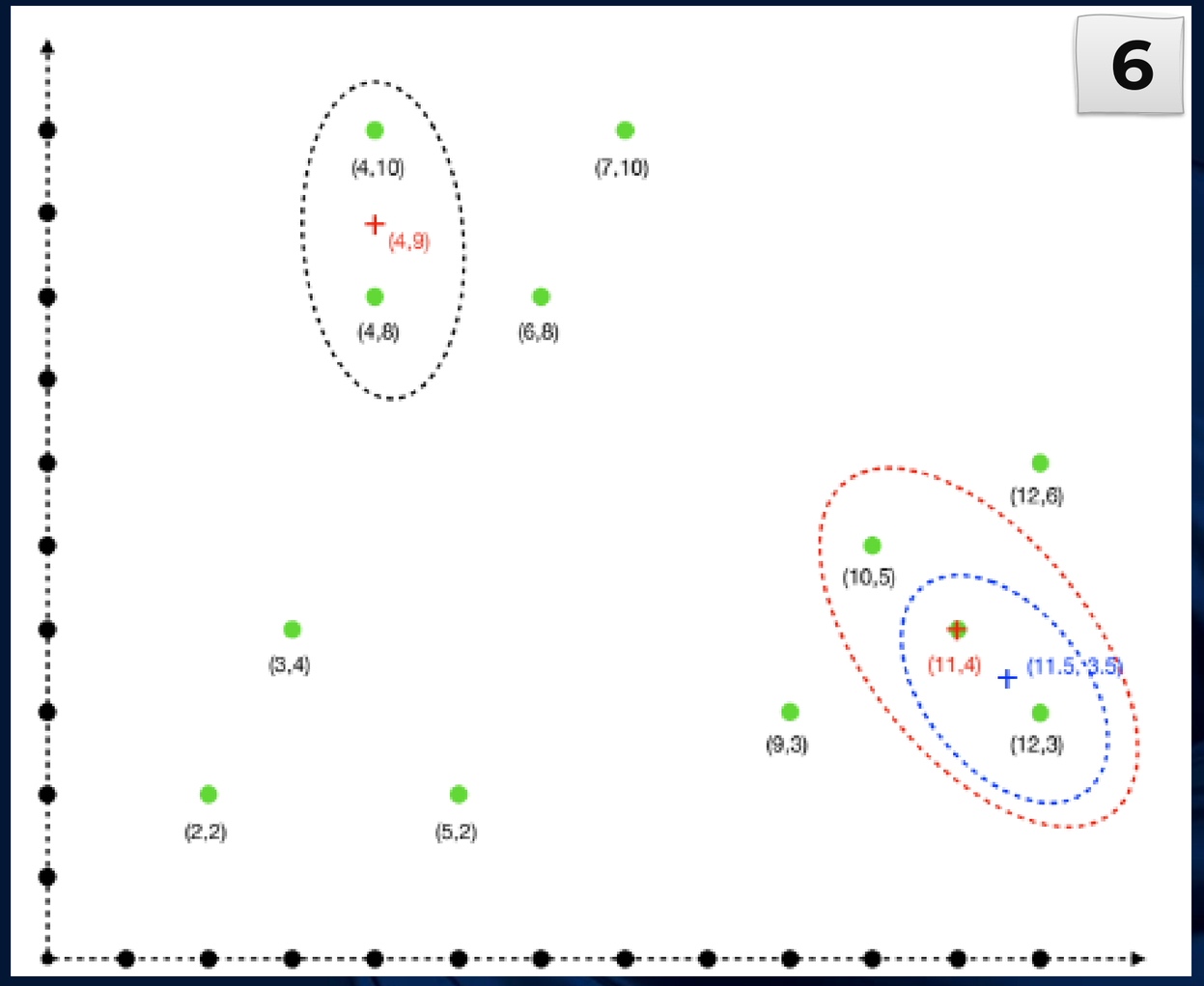
Алгоритм K-means



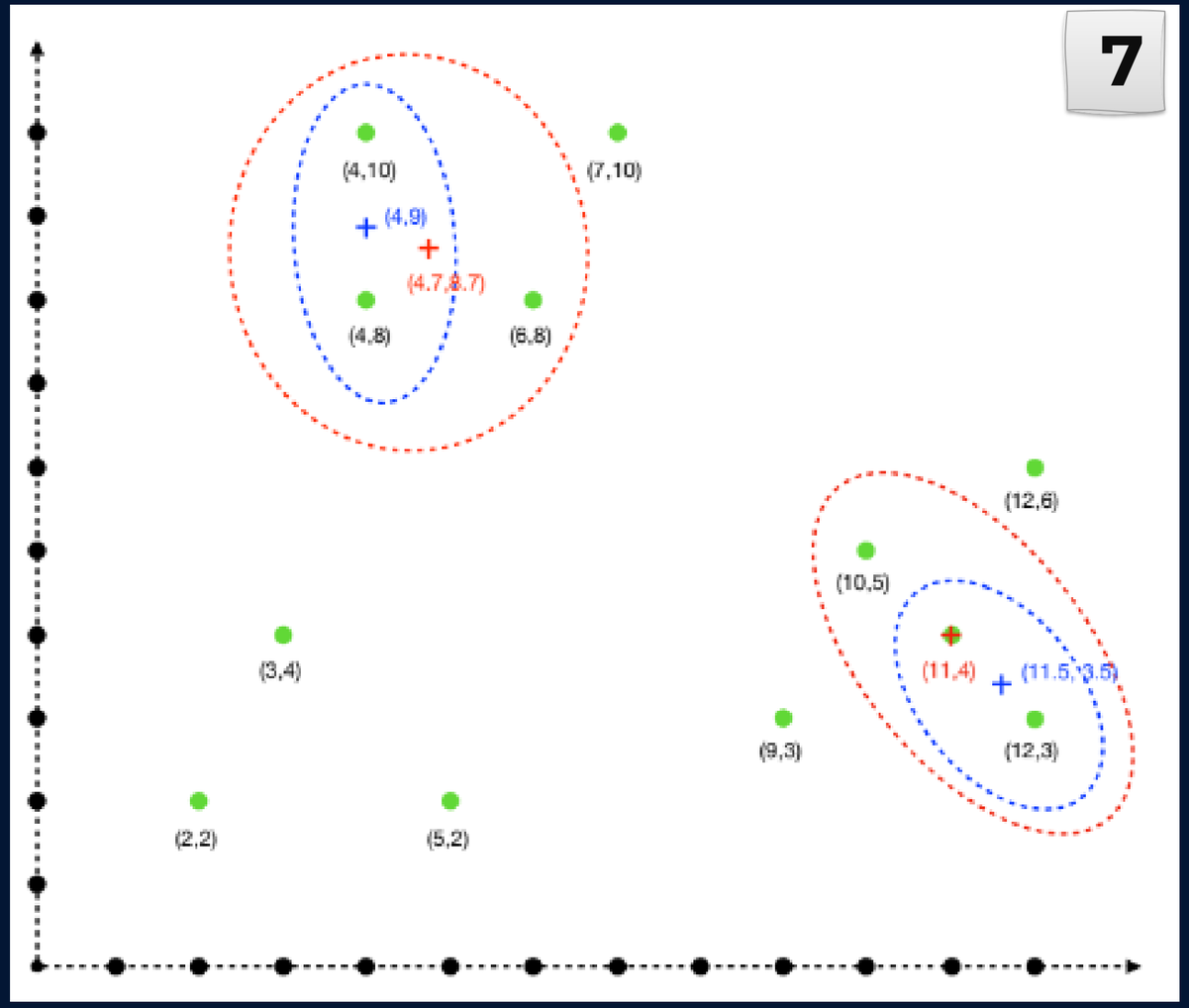


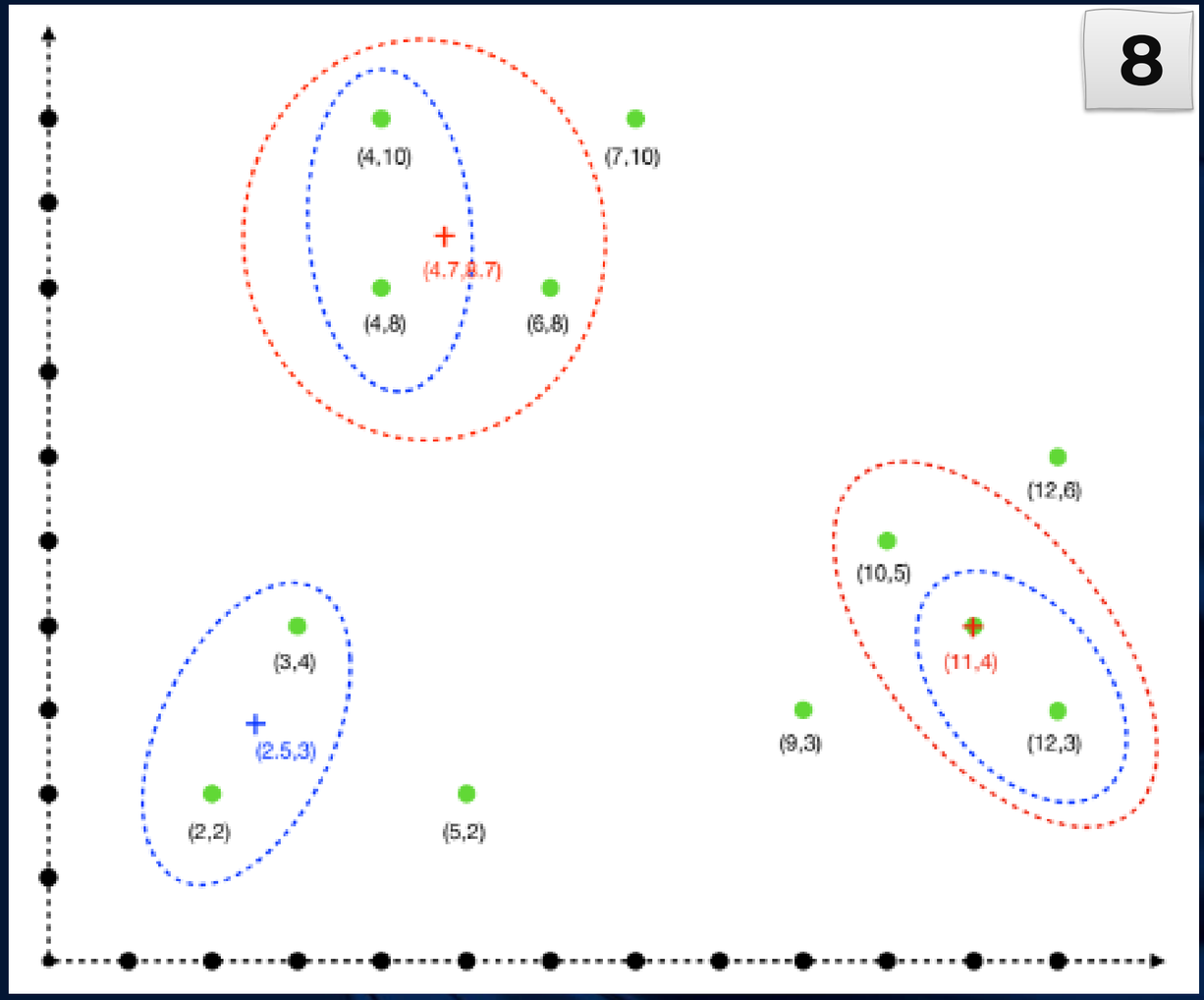
Алгоритм K-means



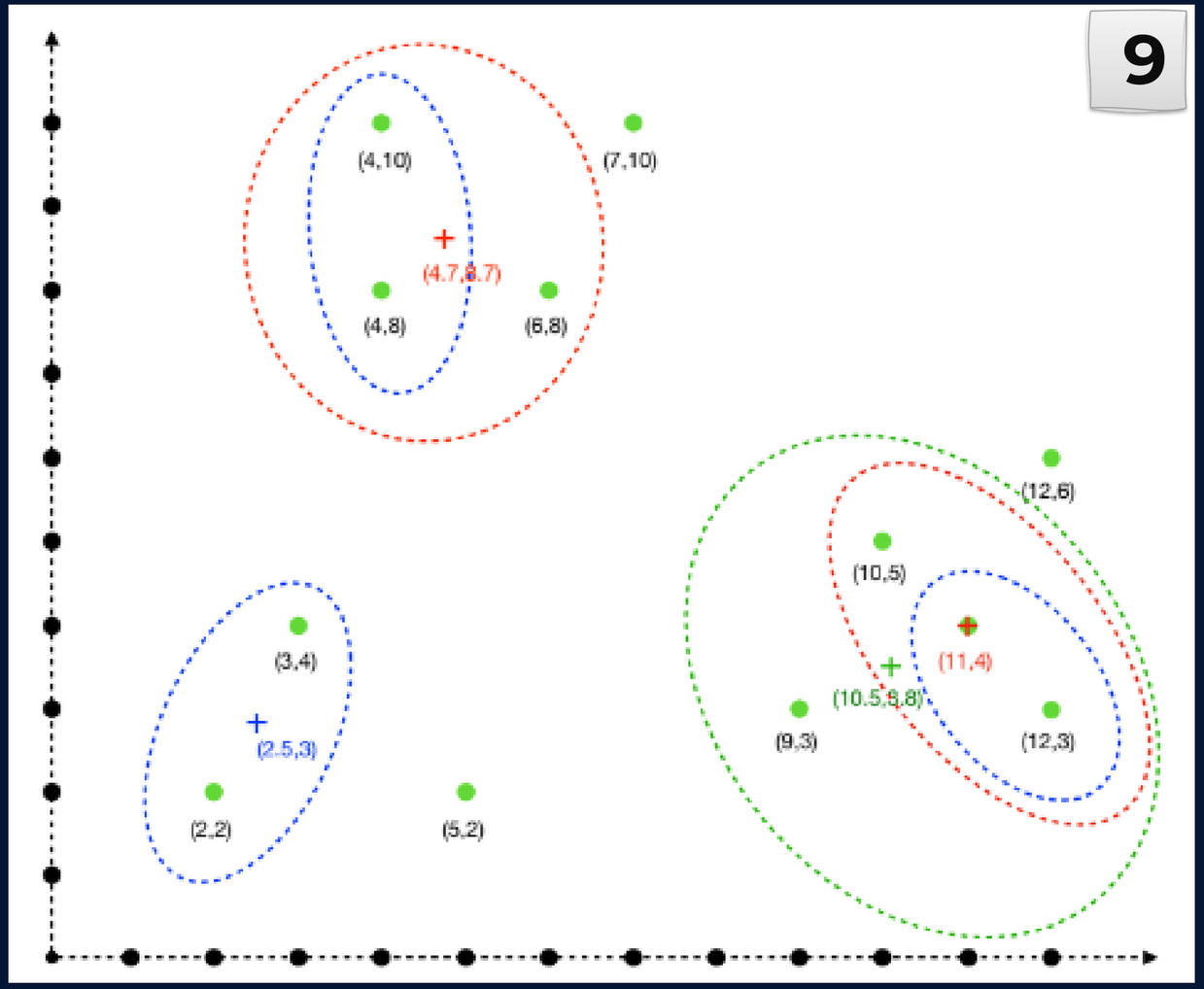


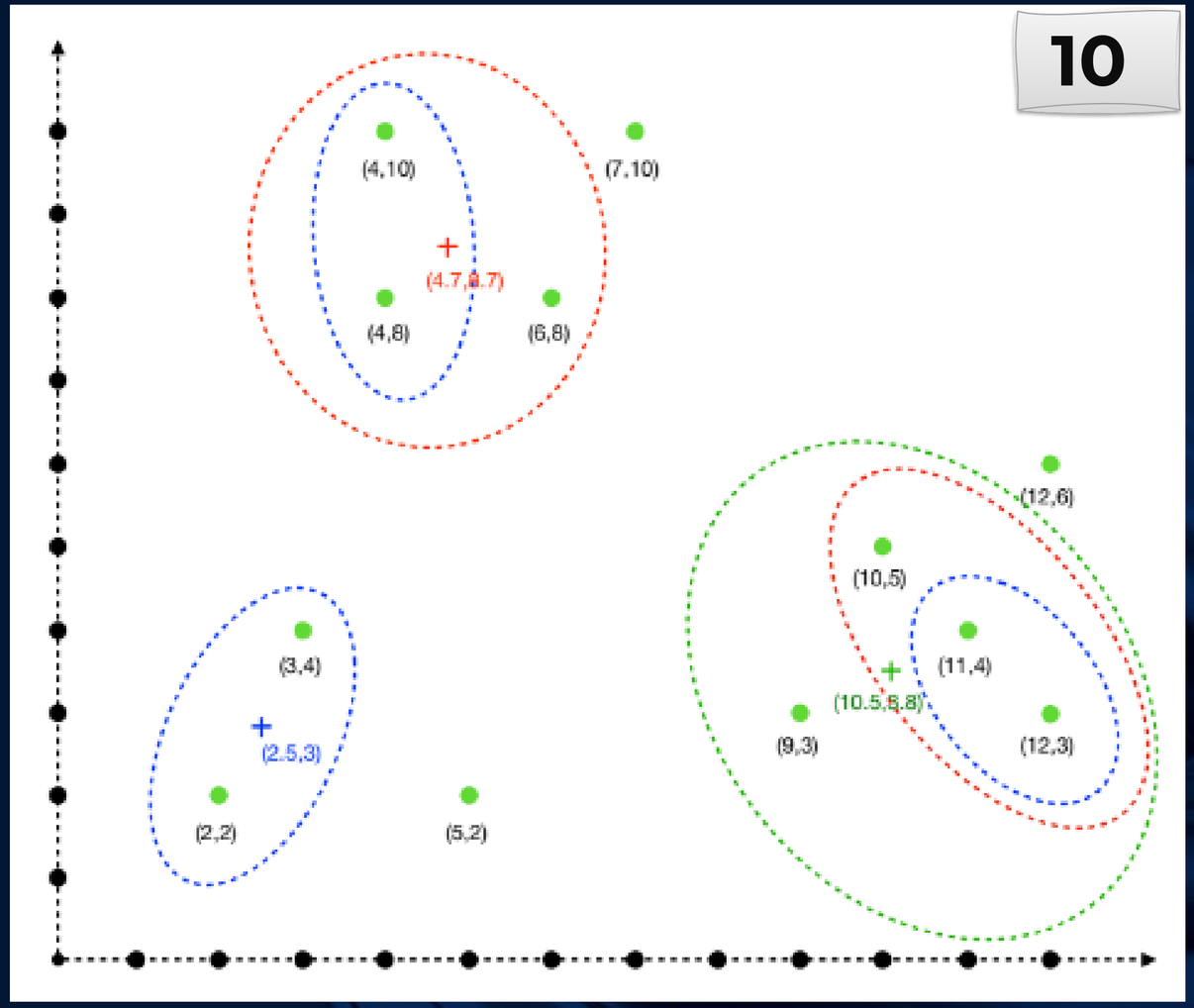
Алгоритм K-means





Алгоритм K-means



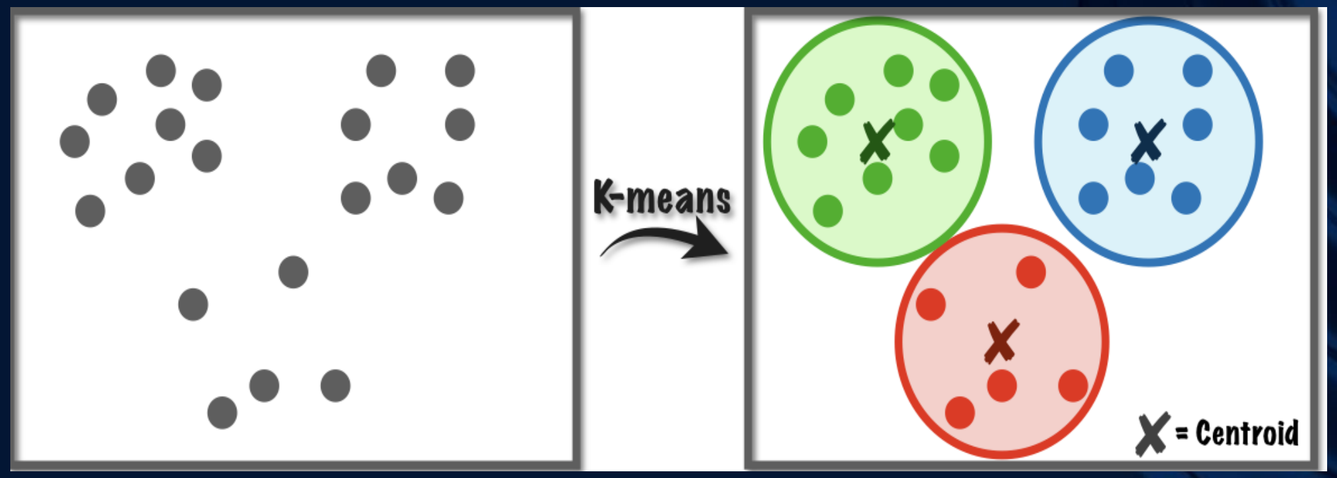


Недостатки k-means

- Алгоритм не всегда находит глобальный минимум, соответствующий целевой функции
- Алгоритм очень чувствителен к начальному определению центроидов (поэтому в сомнительных ситуациях рекомендуется задать начальные центроиды несколько раз).
- Алгоритм очень чувствителен к количеству определяемых кластеров.

Тем не менее, данный алгоритм хорошо адаптирован для многих предметных областей и дает хороший результат при правильном использовании

Неразмеченные данные (unlabelled data)



Размеченные кластеры (labelled clusters)

k-means улучшения

Одним из ключевых вопросов при использовании K-Means является выбор начальных центроидов, поскольку от них зависит качество и скорость сходимости алгоритма. Способы:

1. **Случайный выбор:** выбирается k случайных точек из данных в качестве начальных центроидов. Такой метод прост, но может привести к плохим результатам, если начальные центроиды слишком близки друг к другу или к краям распределения.

2. **K-Means++:** первый центроид выбирается случайно, а затем выбираются остальные центроиды с вероятностью, пропорциональной квадрату расстояния до ближайшего уже выбранного центроида. Данный метод улучшает качество кластеризации, уменьшая вероятность попадания в локальный минимум, но требует дополнительных вычислений.

3. **Greedy K-Means++** — модификация K-Means++, которая ускоряет сходимость и улучшает качество кластеризации за счёт того, что на каждом шаге при выборе центра кластера производится несколько попыток и выбирается лучший (тот, который минимизирует суммарное квадратичное отклонение точек от центров кластеров).

Алгоритм K-means++ как улучшение

У алгоритмов отличается только этап number 1. Доработка классического алгоритма производилась из-за нестабильности выборки начальных отправных точек.

Этапы:

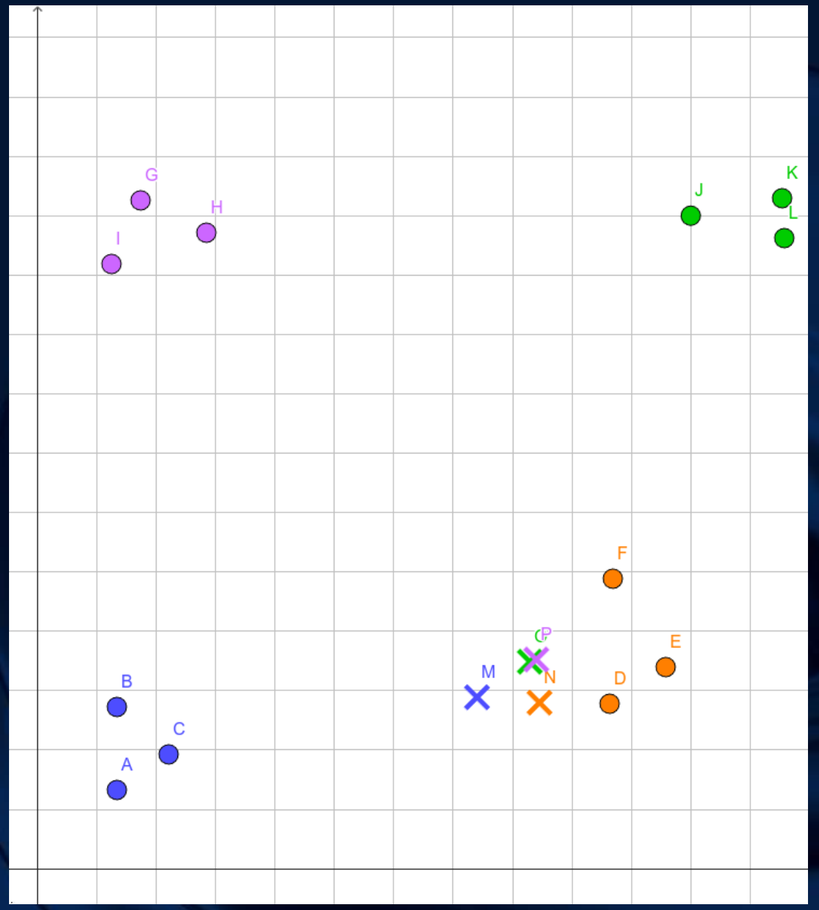
1.1 Выбираем из всех точек на плоскости, случайный первый центроид;

1.2. Находим квадрат расстояния от каждой точки на плоскости до выбранного центроида. Параллельно считаем их общую сумму;

1.3. Далее случайно указываем на число из интервала [0; random(0.0, 1.0) * sum];

1.4. Начинаем снова подсчитывать сумму квадратов расстояний (этапа 1.2) точек, пока сумма не превысит границу выбранного интервала. Берем точку, на которой подсчет был приостановлен;

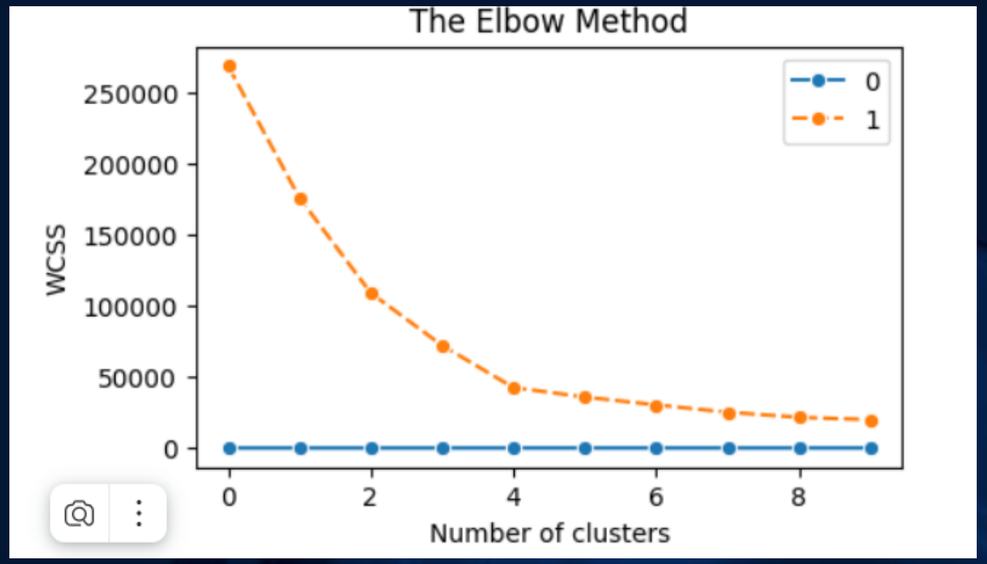
1.5. Повтор шагов 2-4, пока не будут найдены все центроиды. Далее k-means

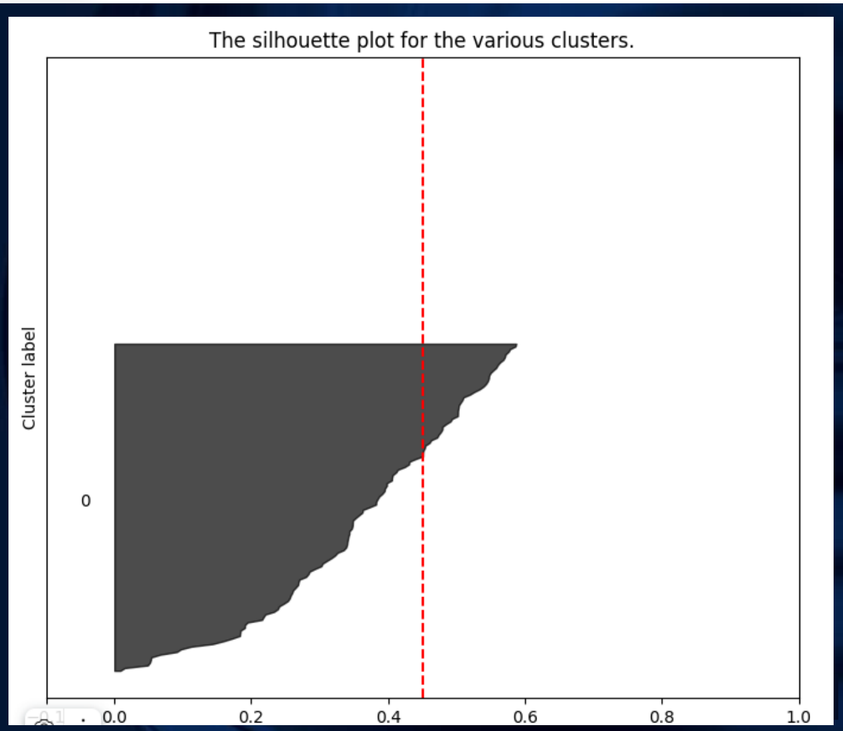


Выбор (оптимального) числа кластеров

Если алгоритм не поддерживает автоматическое определение оптимального количества кластеров, то здесь есть несколько эмпирических правил при условии, что каждый кластер будет в дальнейшем подвергаться содержательной интерпретации аналитиком:

- ✓ Два или три кластера, как правило, не достаточно – кластеризация будет слишком грубой, приводящей к потере информации об индивидуальных свойствах объектов.
- ✓ Больше десяти кластеров не укладывается в известное «числа Миллера 7÷2»: аналитику трудно держать в кратковременной памяти столько кластеров.
- ✓ Поэтому в подавляющем большинстве случаев число кластеров варьируется от 4 до 9.





Выбор (оптимального) числа кластеров: метод локтя

Способы выбора количества кластеров:

Экспертный метод. Выбор количества кластеров будет зависеть от знания о предметной области (domain knowledge)

Метод локтя (elbow method). (1) обучить модель используя несколько вариантов количества кластеров, (2) измерить сумму квадратов внутрикластерных расстояний и (3) выбрать тот вариант, при котором данное расстояние перестанет существенно уменьшаться.

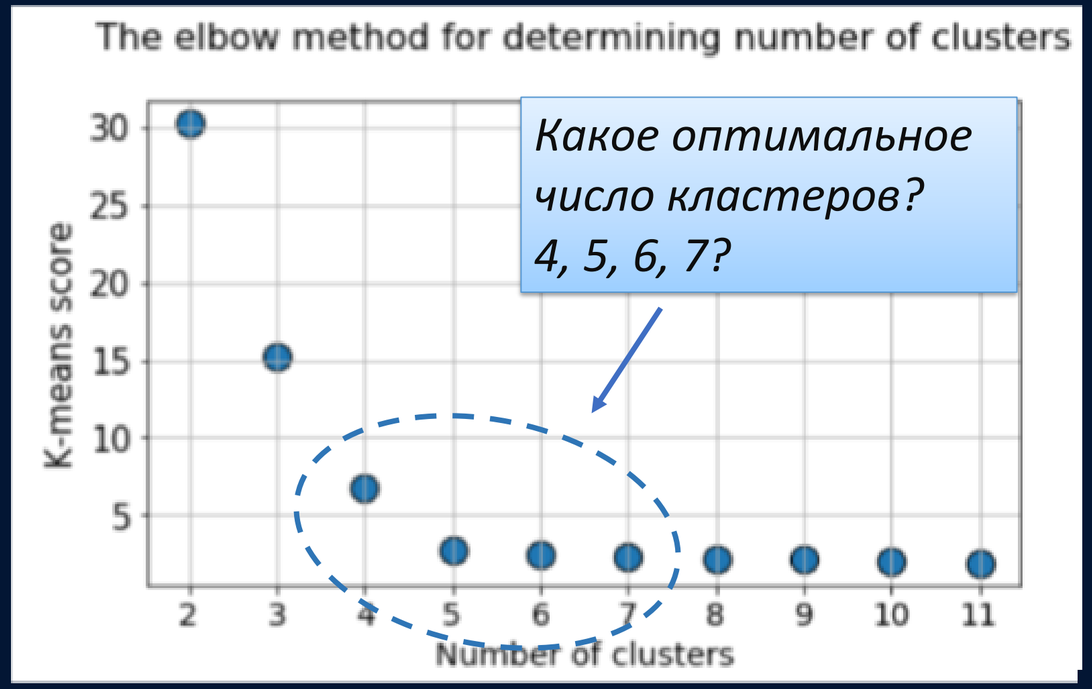
После того как количество кластеров достигает трех, сумма квадратов внутрикластерных расстояний перестает существенно уменьшаться. Три кластера и будет оптимальным значением.

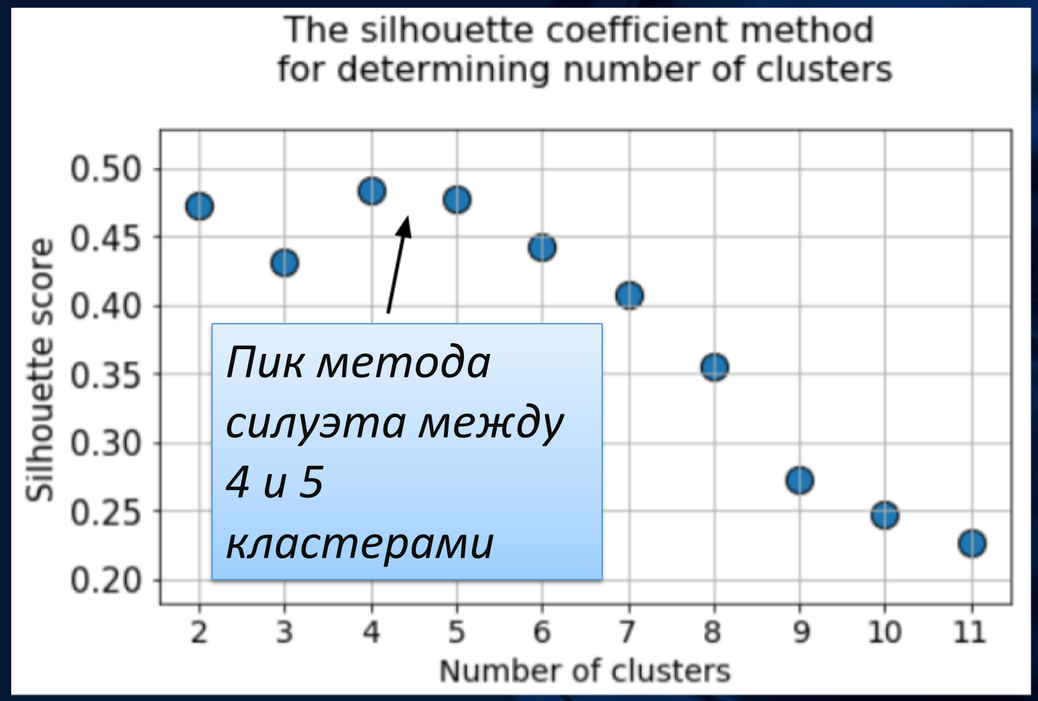
* Сумма квадратов внутрикластерных расстояний (within-cluster sum of squares, WCSS, наша функция потерь)

Выбор (оптимального) числа кластеров: оценка силуэта

Оценка силуэта - это еще один показатель, позволяющий проверить компактность кластера, чтобы определить, является ли кластер хорошим.
Оценка силуэта вычисляет среднее расстояние внутри кластера, как метод локтя, а также среднее значение ближайшего кластера.

Показатель силуэта измеряет как компактность одного и того же кластера, так и разделение с другими кластерами. График силуэта имеет пиковый характер, в отличие от мягко изогнутого графика при использовании метода локтя. Его проще визуализировать и обосновать.

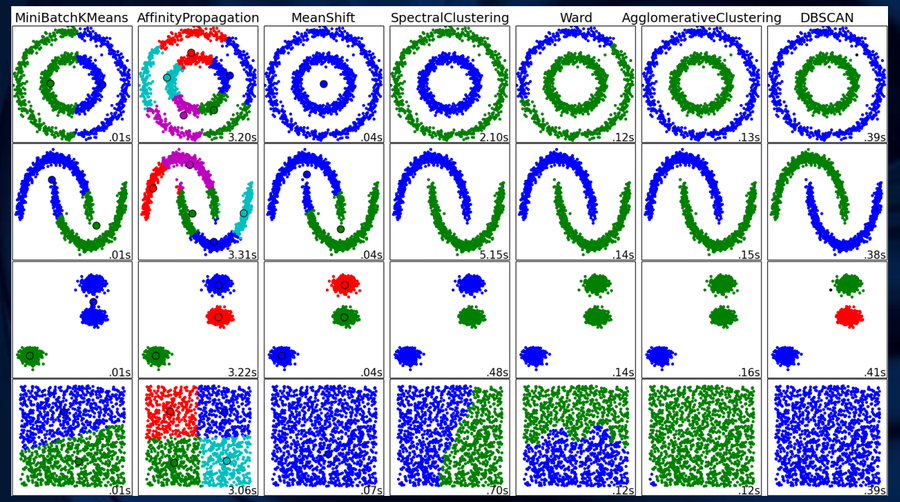




ДЕМОНСТРАЦИЯ

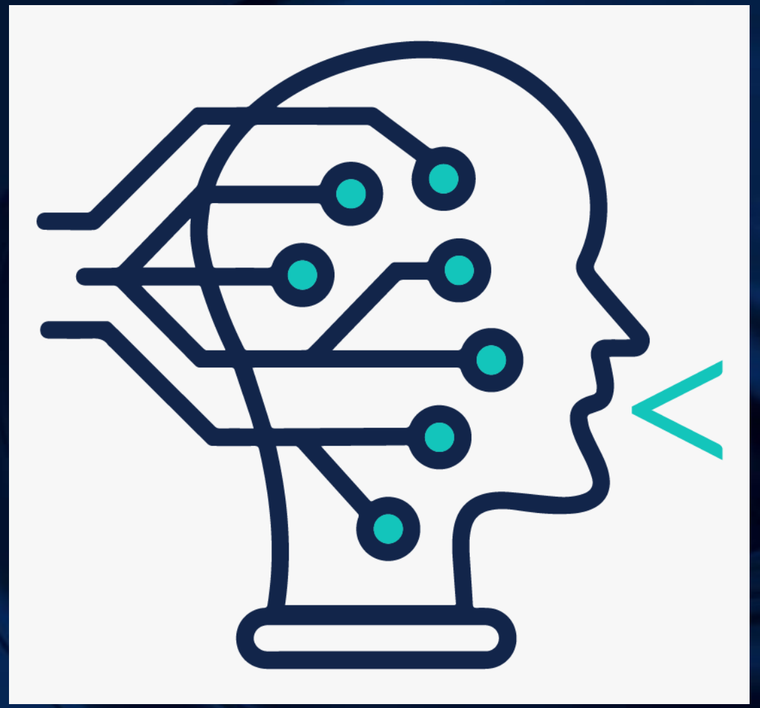
Как выглядит решение типовой задачи?

- Настройки k-means библиотеки scikit-learn для «игрушечного» (тестового) набора данных Ирисы
- Методы локтя и силуэта и интерпретация результата
- Как оценить качество, если есть значения кластеров (target)?
- Что если в данных есть категориальные переменные? Как поменяется задача?



Полезные ссылки

1. Модель кластеризации KMeans | К-средних | Метод локтя | KMeans часть 1 | МАШИННОЕ ОБУЧЕНИЕ. URL: https://www.youtube.com/watch?v=EHZJMz6zyFE
2. ОБУЧЕНИЕ С УЧИТЕЛЕМ, ОБУЧЕНИЕ БЕЗ УЧИТЕЛЯ, ОБУЧЕНИЕ С ПОДКРЕПЛЕНИЕМ | ЗАДАЧИ МАШИННОГО ОБУЧЕНИЯ. URL: https://www.youtube.com/watch?v=ku2oPMPht2I&t=326s
3. Кластерный анализ. URL: https://www.dmitrymakarov.ru/intro/clustering-16/
4. Евклидова, L1 и Чебышёва — 3 основные метрики, которые пригодятся в Data Science. URL: https://tproger.ru/translations/3-basic-distances-in-data-science
5. Задача поиска аномалий. URL: https://www.youtube.com/watch?v=aKYs4p8HcXA
6. Кластеризация множества объектов, алгоритм K-means++. URL: https://habr.com/ru/articles/829202/
7. Кластеризуем лучше, чем «метод локтя». URL: https://habr.com/ru/companies/jetinfosystems/articles/467745/



УНИВЕРСИТЕТ ИННОПОЛИС

ВОПРОСЫ И ОТВЕТЫ

Корнеева Елена
e.korneeva@innopolis.ru
https://t.me/Allyonzy