Факультет компьютерных наук | НИС Python | Москва 2025

Лекция 3
Кластерный анализ

Преподаватель: Меликян Алиса Валерьевна, amelikyan@hse.ru
кандидат наук, доцент Департамента программной инженерии

Кластерный анализ

Проводится с целью выделения однородных групп (кластеров) из исследуемой совокупности объектов (единиц анализа). Формируемые группы (кластеры) должны быть однородными (гомогенными) внутри и разнородными (гетерогенными) по отношению друг к другу по заданным характеристикам.

Наиболее часто при определении кластеров используется метод исследования расстояний между переменными в кластерах.

Кластерный и факторный анализы

Цель факторного анализа: сокращение числа переменных, участвующих в анализе.

Цель кластерного анализа: классификация единиц анализа (например, респондентов) на целевые группы на основании их характеристик (значений переменных).

Оба анализа могут использоваться в одном исследовании. Сперва проводится факторный анализ для снижения числа переменных, а затем кластерный анализ для формирования целевых сегментов.

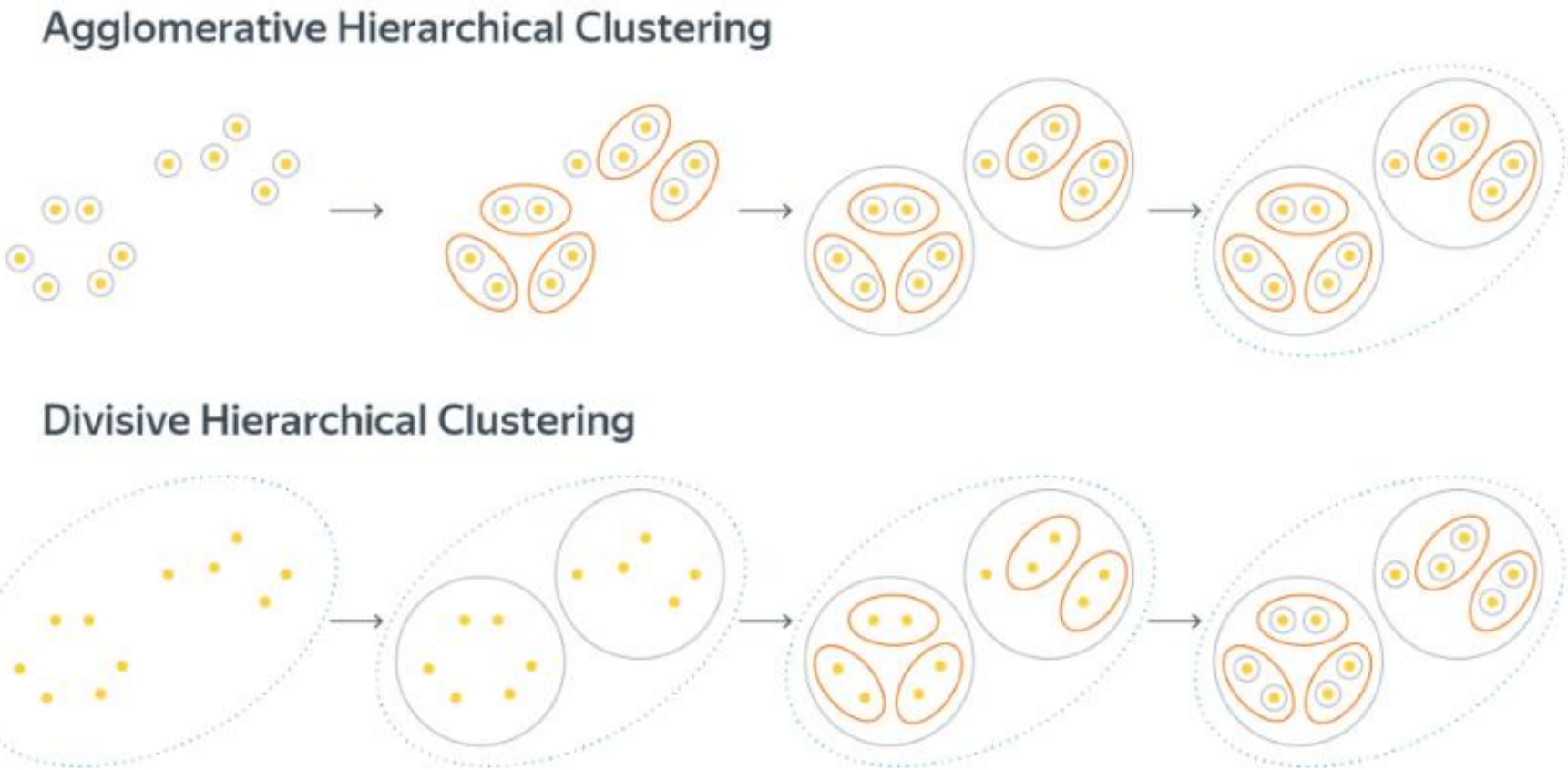
Порядок проведения кластерного анализа

- Шаг 1: отпределение алгоритма кластеризации и метода измерения расстояния между кластерами;
- Шаг 2: определение числа кластеров;
- Шаг 3: сохранение кластерного решения и описание кластеров;
- Шаг 4: оценка достоверности кластерного решения.

Особенности проведения кластерного анализа

- Анализ чувствителен к выбросам.
- Между переменными не должно быть высокой корреляции.
- Если переменные измерены в разных единицах, то их значения нужно стандартизировать.
- В кластере должно быть достаточное количество наблюдений.
- Не должно быть “влияющих наблюдений” и изменение алгоритма кластеризации не должно кардинально менять результат.

Иерархический кластерный анализ (агломеративный или дивизионный)



Иерархический кластерный анализ

Иерархический кластерный анализ предполагает многоступенчатый алгоритм проведения анализа.

1. Сначала каждое наблюдение представляет собой отдельный кластер (агломеративный метод).
2. На первом шаге два ближайших кластера объединяются в один.
3. Этот процесс продолжается пока все наблюдения не объединятся в один большой кластер.
4. Исследователю необходимо принять решение в какой момент прекратить объединение кластеров, чтобы кластеры были достаточно однородными и содержательно интерпретируемыми.

Определение оптимального числа кластеров

Оптимальным считается такое число кластеров, при котором сформированные кластеры:

- с одной стороны, объединяют в себе как можно больше объектов исследования;
- с другой стороны, являются как можно менее гетерогенными внутри.

Измерение расстояния между кластерами

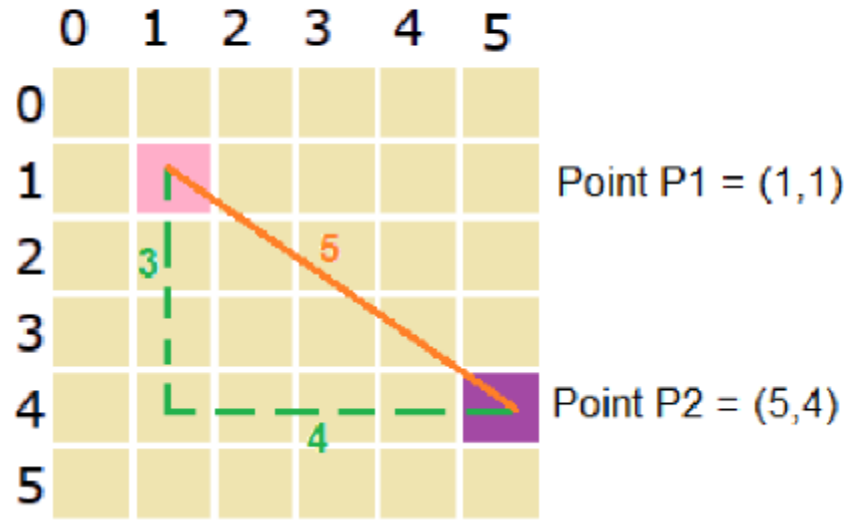
Евклидово расстояние чаще всего используется для измерения расстояния при проведении кластеризации на основе интервальных переменных. Иногда используется квадрат Евклидова расстояния для усиления эффекта больших расстояний. Рекомендуется предварительно стандартизировать шкалы переменных кластеризации, если они различаются.

$$D_{ij} = \sqrt{\sum_{k=1}^{n} \left(x_{ki} - x_{kj} \right)^2}$$

Измерение расстояния между кластерами

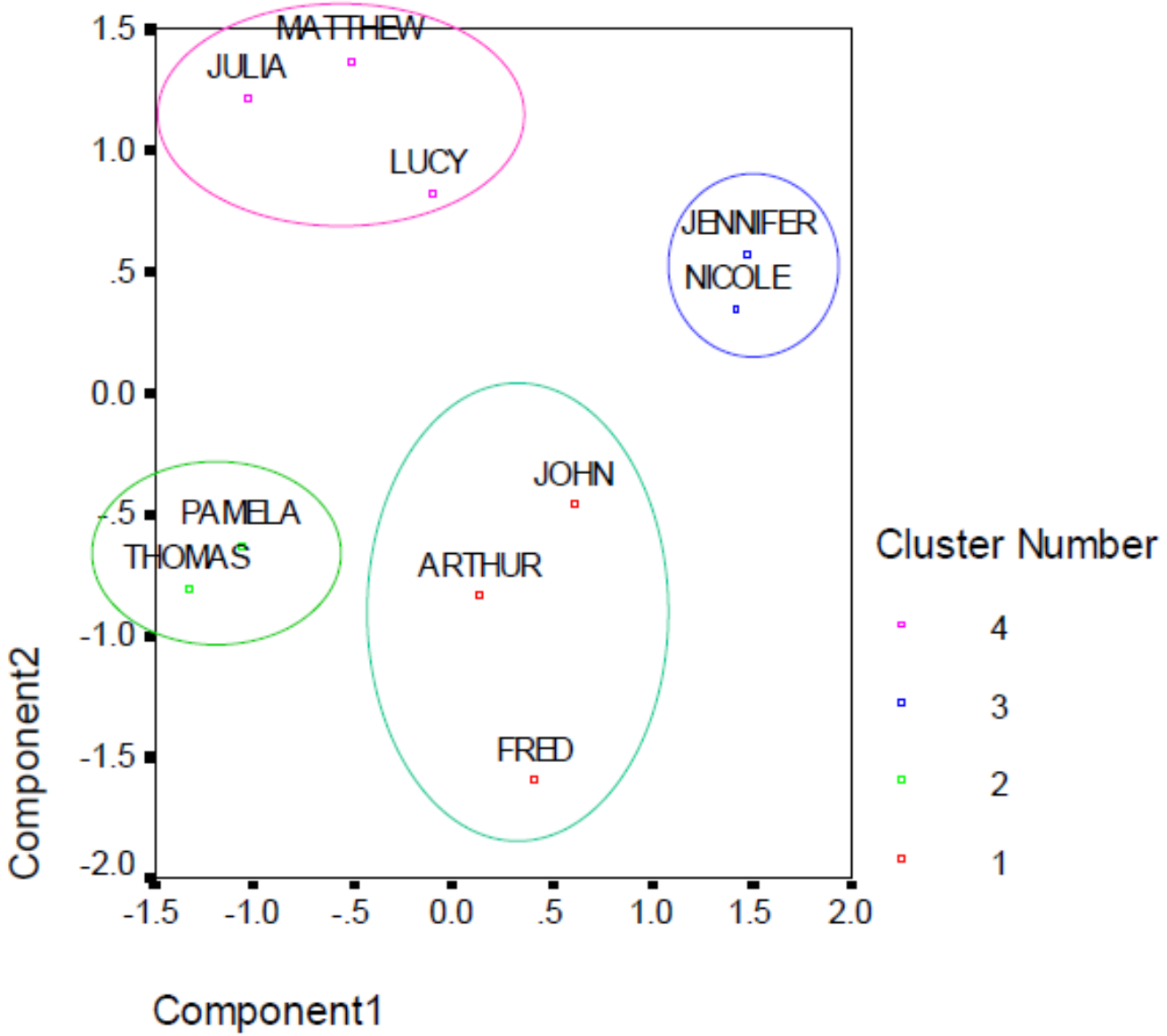
Расстояние городских кварталов (Манхэттенское расстояние) — расстояние между двумя точками равно сумме модулей разностей их координат.

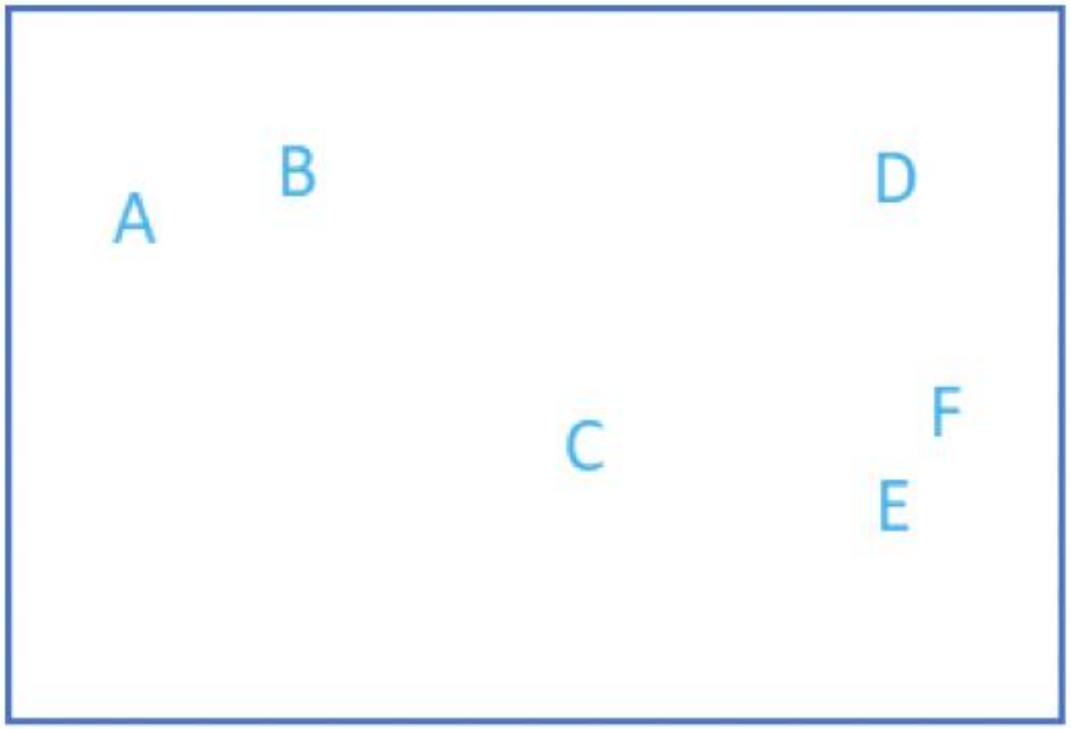
$$\rho(a,b) = |x_1 - x_2| + |y_1 - y_2|$$

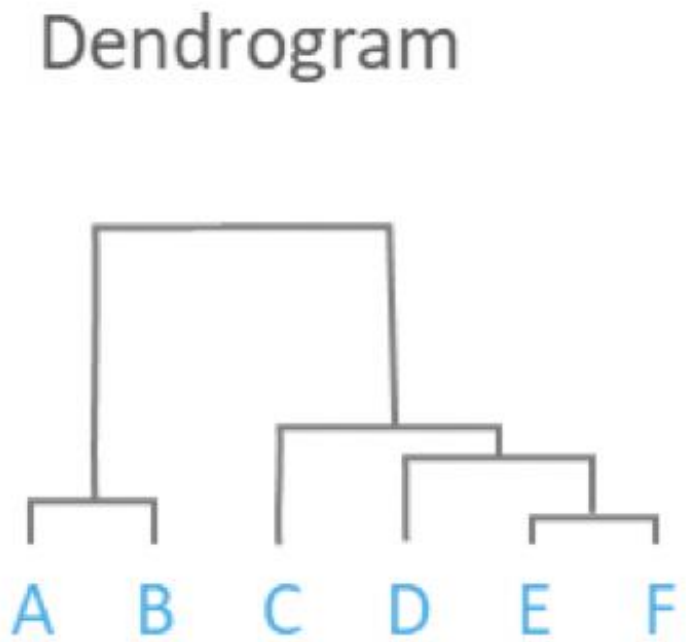


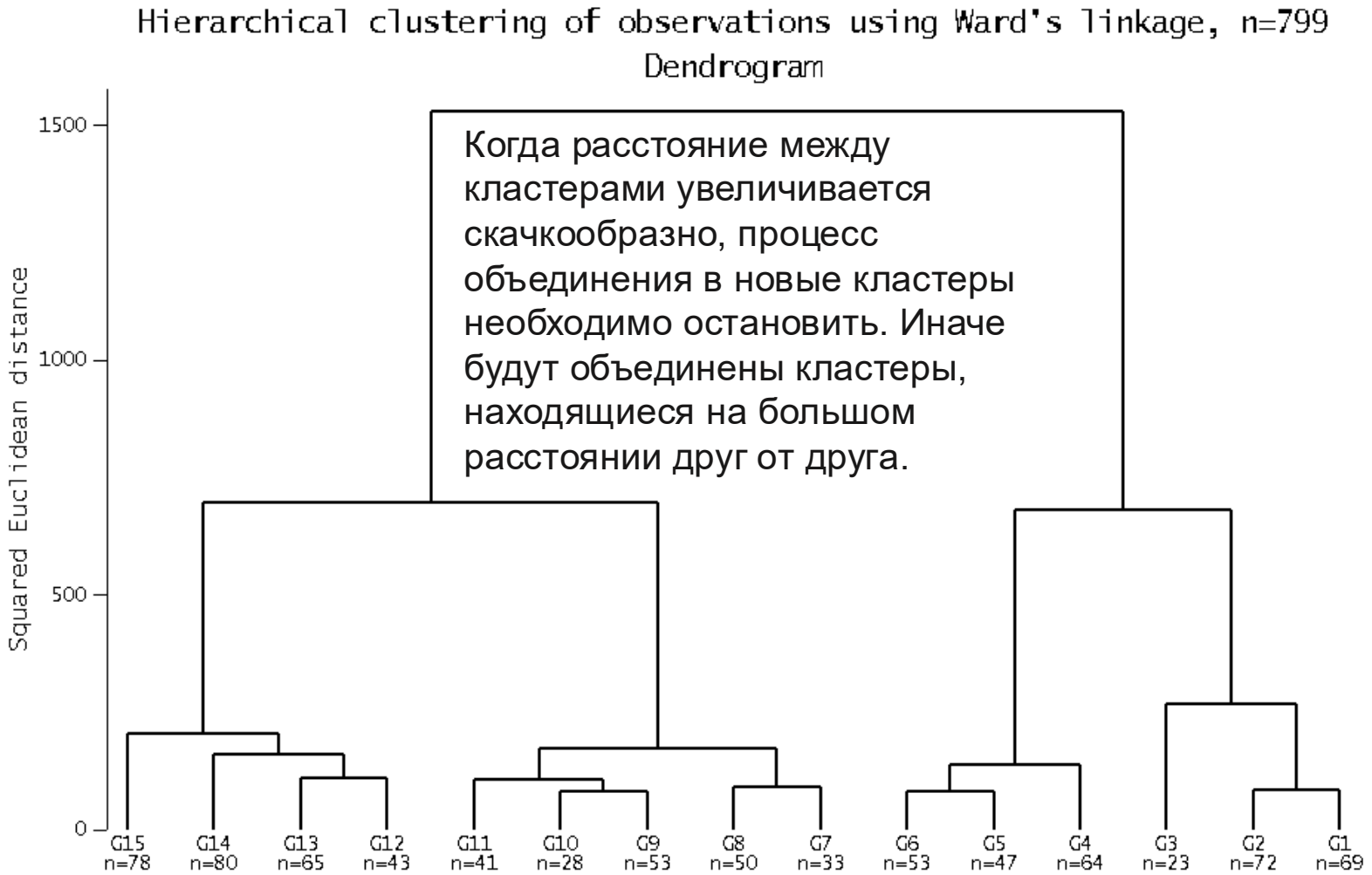
Euclidean distance = $\sqrt{(5-1)^2 + (4-1)^2} = 5$

Manhattan distance = $|5-1| + |4-1| = 7$

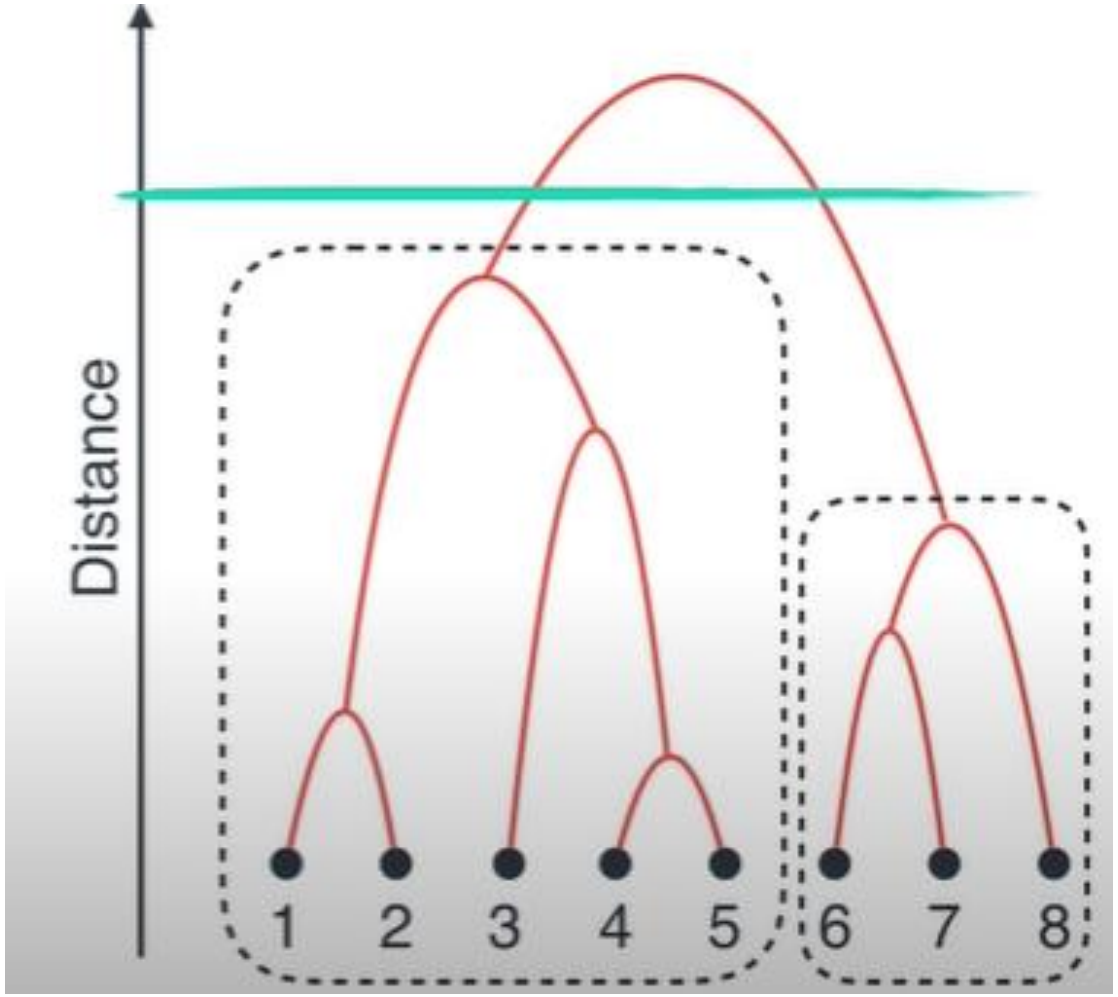

MATTHEW
JULIA
LUCY
JENNIFER
NICOLE
JOHN
PAMELA
THOMAS
ARTHUR
FRED
1.5
1.0
.5
0.0
-.5
-1.0
-1.5
-2.0
-1.5
-1.0
-.5
0.0
.5
1.0
1.5
2.0
Component2
Component1
Cluster Number
4
3
2
1

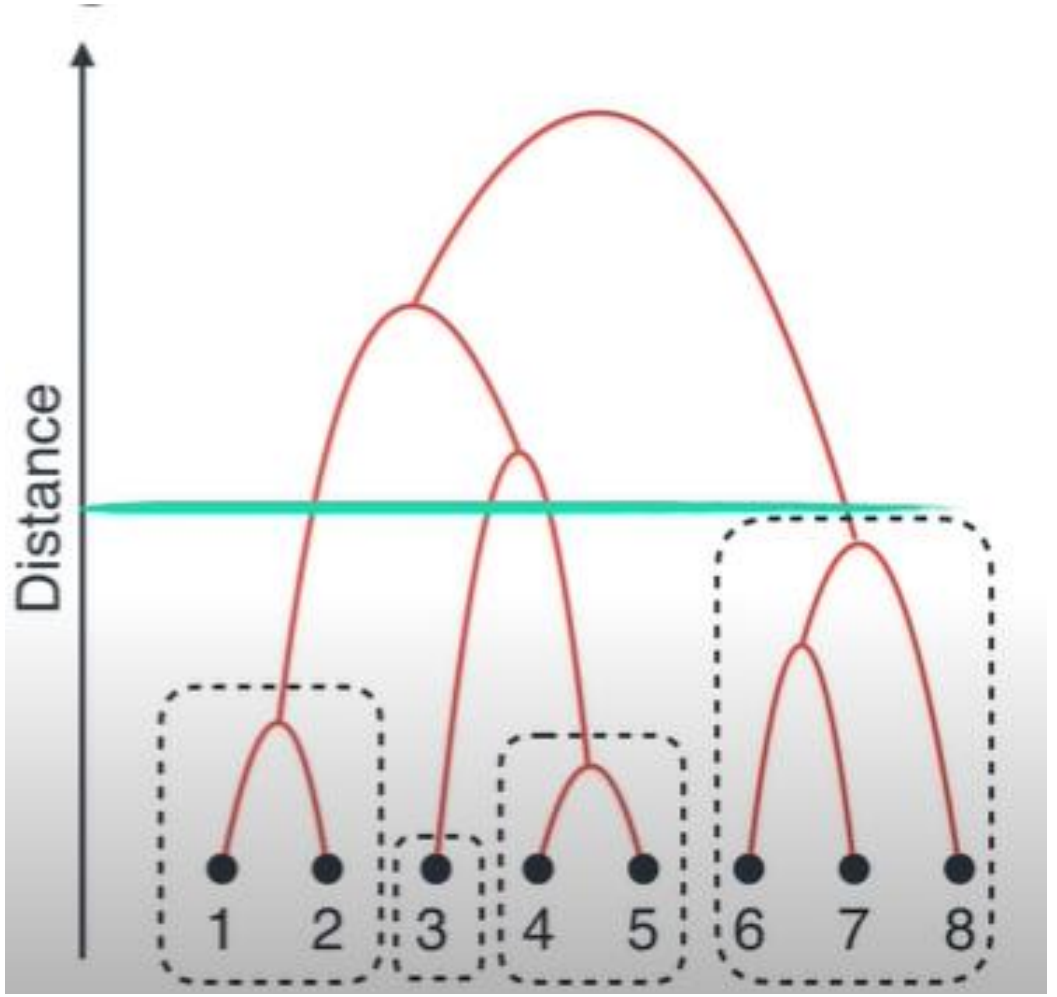

A
B
C
D
E
F


Dendrogram
A
B
C
D
E
F



Когда расстояние между кластерами увеличивается скачкообразно, процесс объединения в новые кластеры необходимо остановить. Иначе будут объединены кластеры, находящиеся на большом расстоянии друг от друга.


Distance
1
2
3
4
5
6
7
8


Distance
1
2
3
4
5
6
7
8

Алгоритмы иерархической кластеризации

Наиболее часто используемые алгоритмы для вычисления расстояний между кластерами:

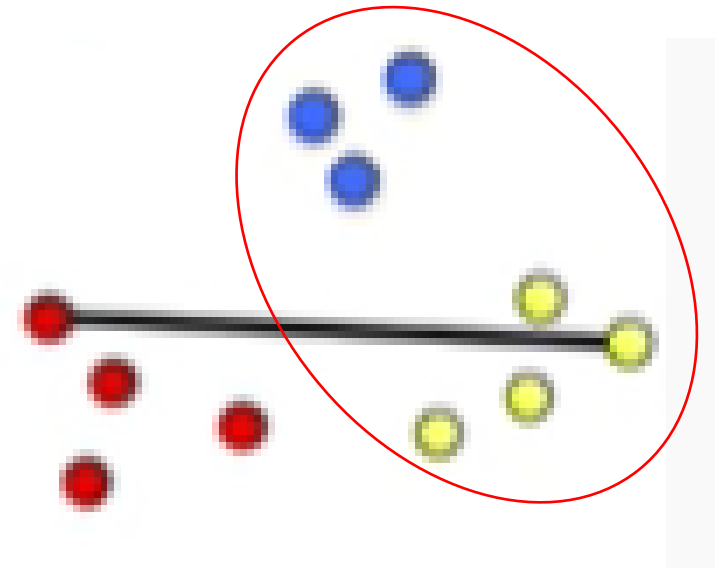
- Ближайший сосед (Nearest neighbor)
- Дальний сосед (Furthest neighbor)
- Межгрупповые связи (Between-groups linkage)
- Центроидная кластеризация (Centroid clustering)
- Метод Варда (Уорда)(Ward's method)

Близлежащий сосед (Nearest neighbor / Single-linkage)

Находится самое короткое расстояние между двумя наблюдениями и они объединяются в первый кластер. Находится следующее самое короткое расстояние и, либо создаётся второй кластер, либо наблюдение присоединяется к ранее созданному первому кластеру. Расстояние между двумя кластерами определяется расстоянием между двумя наиболее близкими объектами (ближайшими соседями) в различных кластерах. Может привести к цепочкам последовательностей.

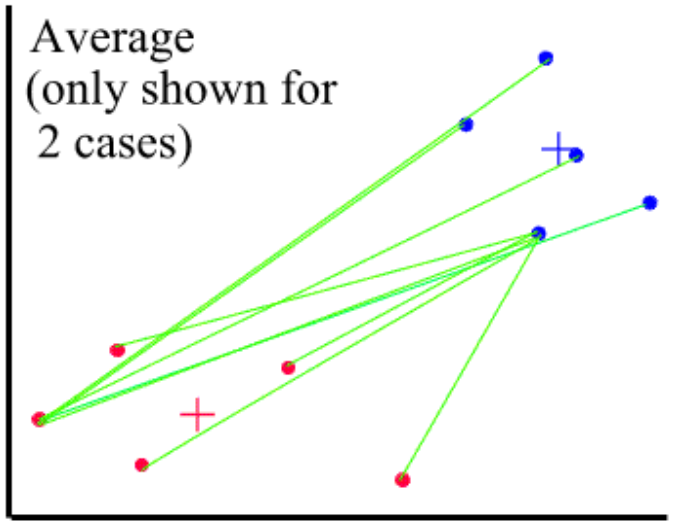
Дальний сосед (Furthest neighbor / Complete-linkage)

Расстояния между кластерами определяются наибольшим расстоянием между любыми двумя объектами в различных кластерах (т.е. наиболее удаленными соседями). Этот метод обычно работает очень хорошо, когда объекты происходят из отдельных групп. Старается минимизировать диаметр нового кластера. Больше шансов получить кластер, похожий на сферу.



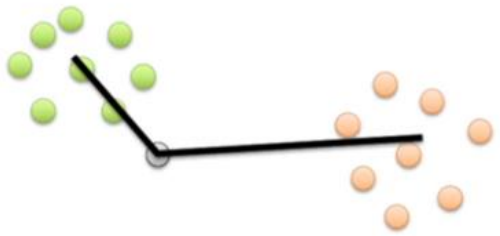
Межгрупповые связи (Average Linkage / Between Groups Linkage)

Дистанция между кластерами равна среднему значению дистанций между всеми возможными парами наблюдений, причём одно наблюдение берётся из одного кластера, а другое из другого. Информация, необходимая для расчёта дистанции, находится на основании всех теоретически возможных пар наблюдений. Снижается влияние выбросов на результат.

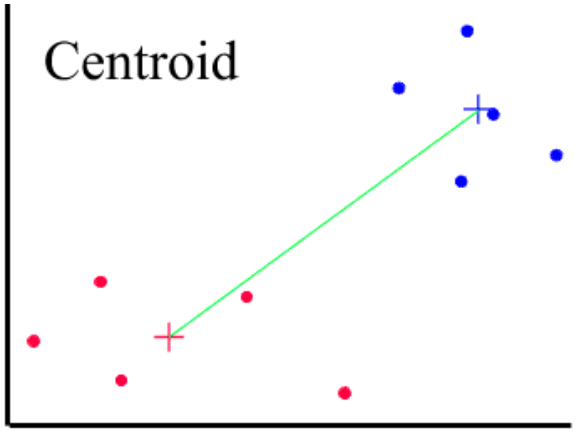


Центроидная кластеризация (Centroid Clustering)

Наблюдение будет отнесено к тому кластеру, центр которого ближе всего к нему расположен.

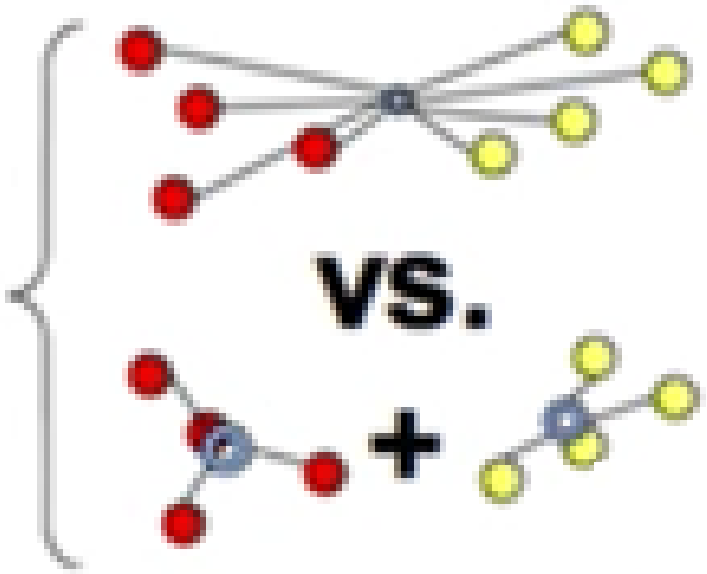


Расстояние между двумя кластерами определяется как расстояние между их центрами.



Метод Варда (Ward’s Method)

При попытке соединить два кластера оценивается изменение общей суммы дисперсий. Стараемся минимизировать отклонение значений наблюдений от центра. Измеряется ухудшение внутрикластерного разброса, которое произойдет в случае объединения двух кластеров.



Кластерный анализ по методу k-средних

Представляет собой итеративный метод группировки. Алгоритм стремится минимизировать суммарное квадратичное отклонение точек кластеров от центров этих кластеров.

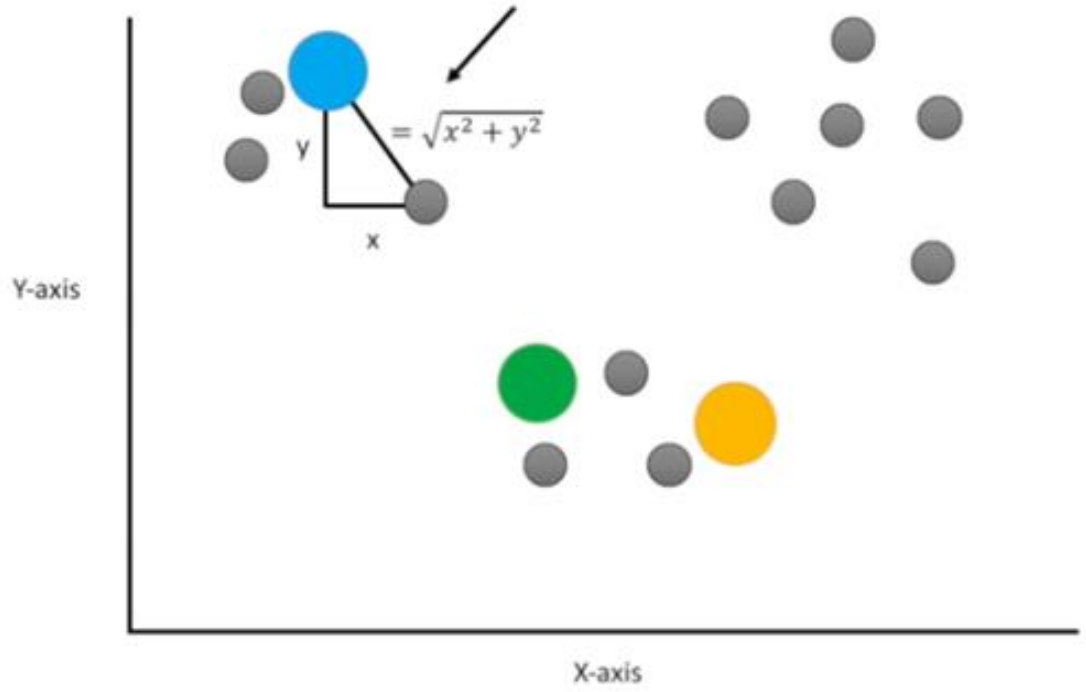
Рекомендован для выборок среднего и большого размера. Предполагается, что число кластеров известно заранее. Анализ чувствителен к выбросам.

Метод k-средних: шаг 1

Задать число кластеров. Если число кластеров заранее не известно, то можно сначала провести иерархический анализ на случайно отобранной подвыборке наблюдений, чтобы определиться с конкретным числом или допустимым диапазоном числа кластеров.

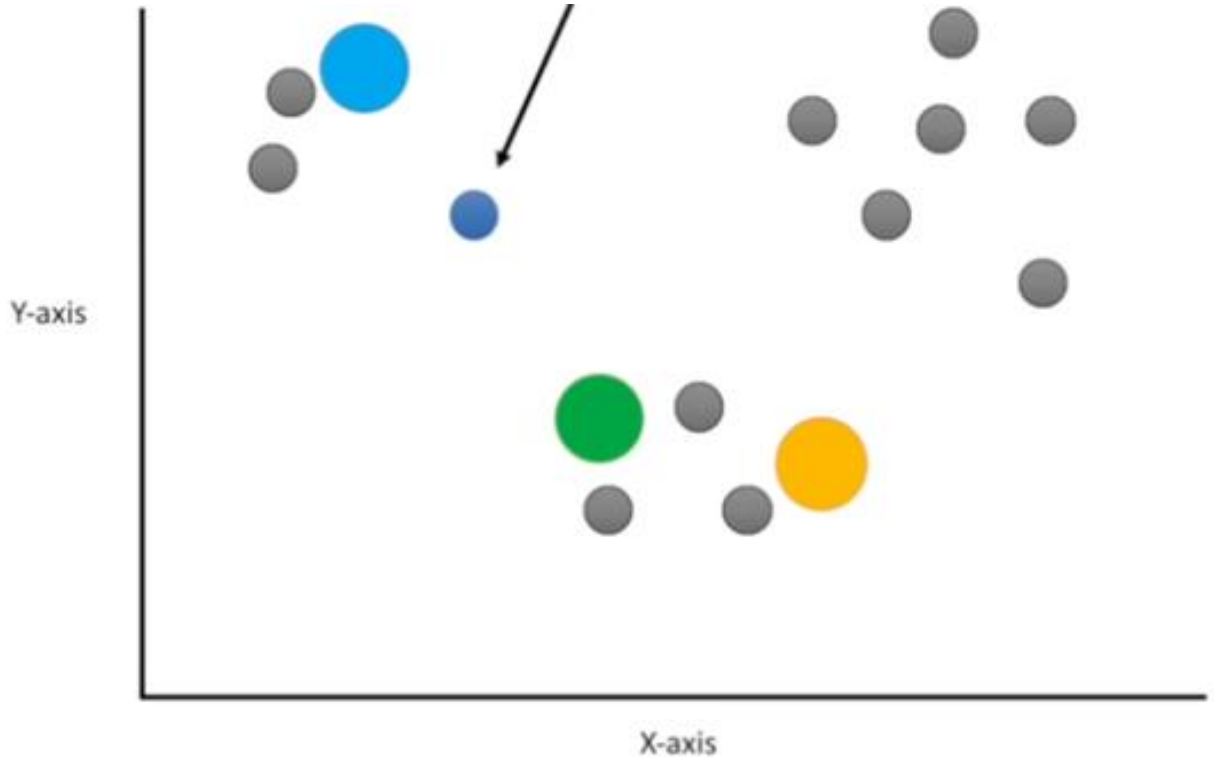
Метод k-средних: шаг 2

Выбрать случайные точки в соответствии с числом кластеров, заданных на предыдущем этапе.



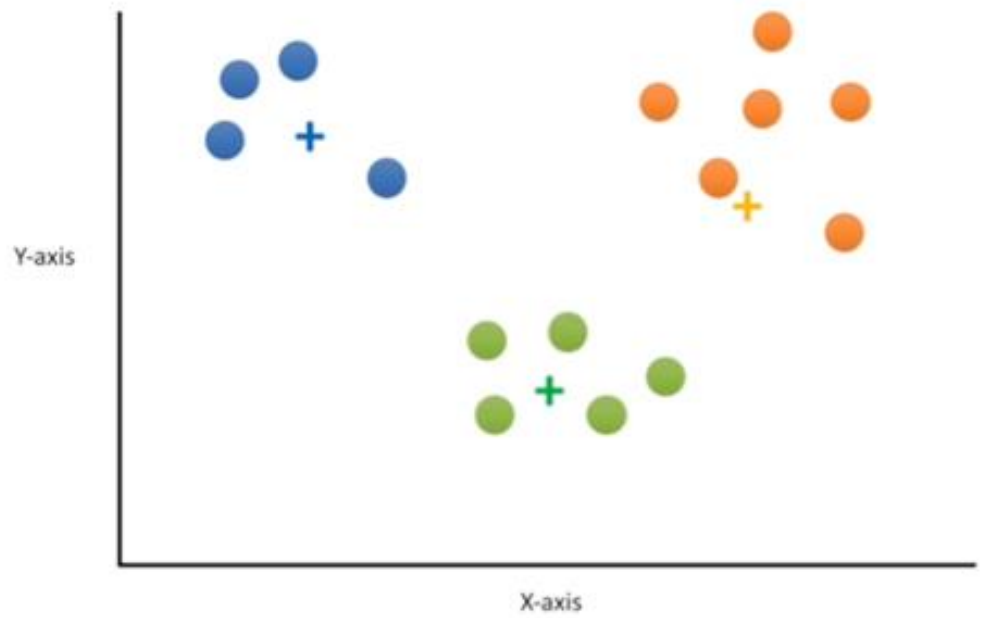
Метод k-средних: шаг 3

Присоединить наблюдения к ближайшим кластерам.



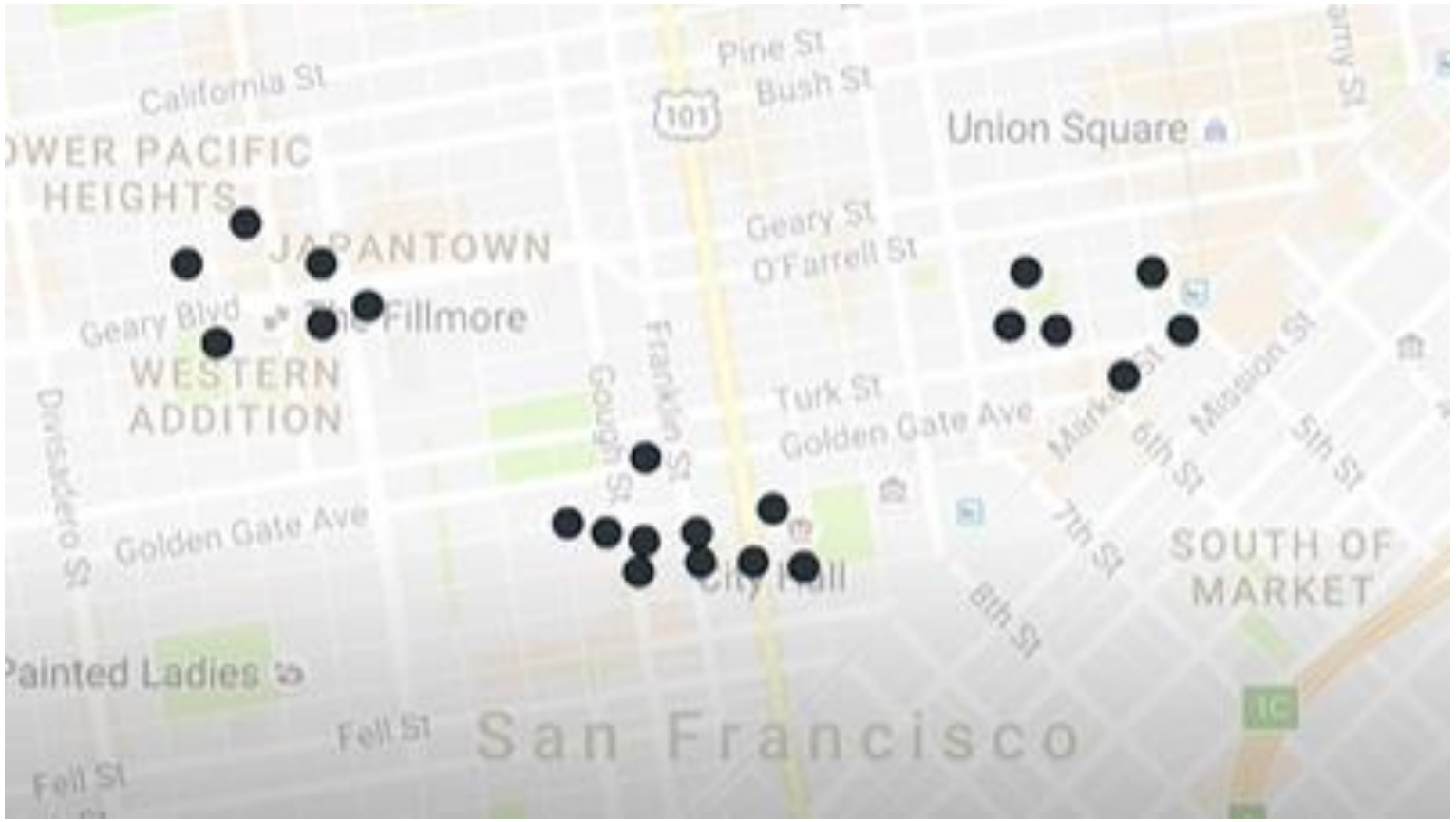
Метод k-средних: шаг 4

Рассчитать центры каждого получившегося кластера и повторить процесс кластеризации. Процесс кластеризации должен повториться пока возможны улучшение результатов кластеризации.



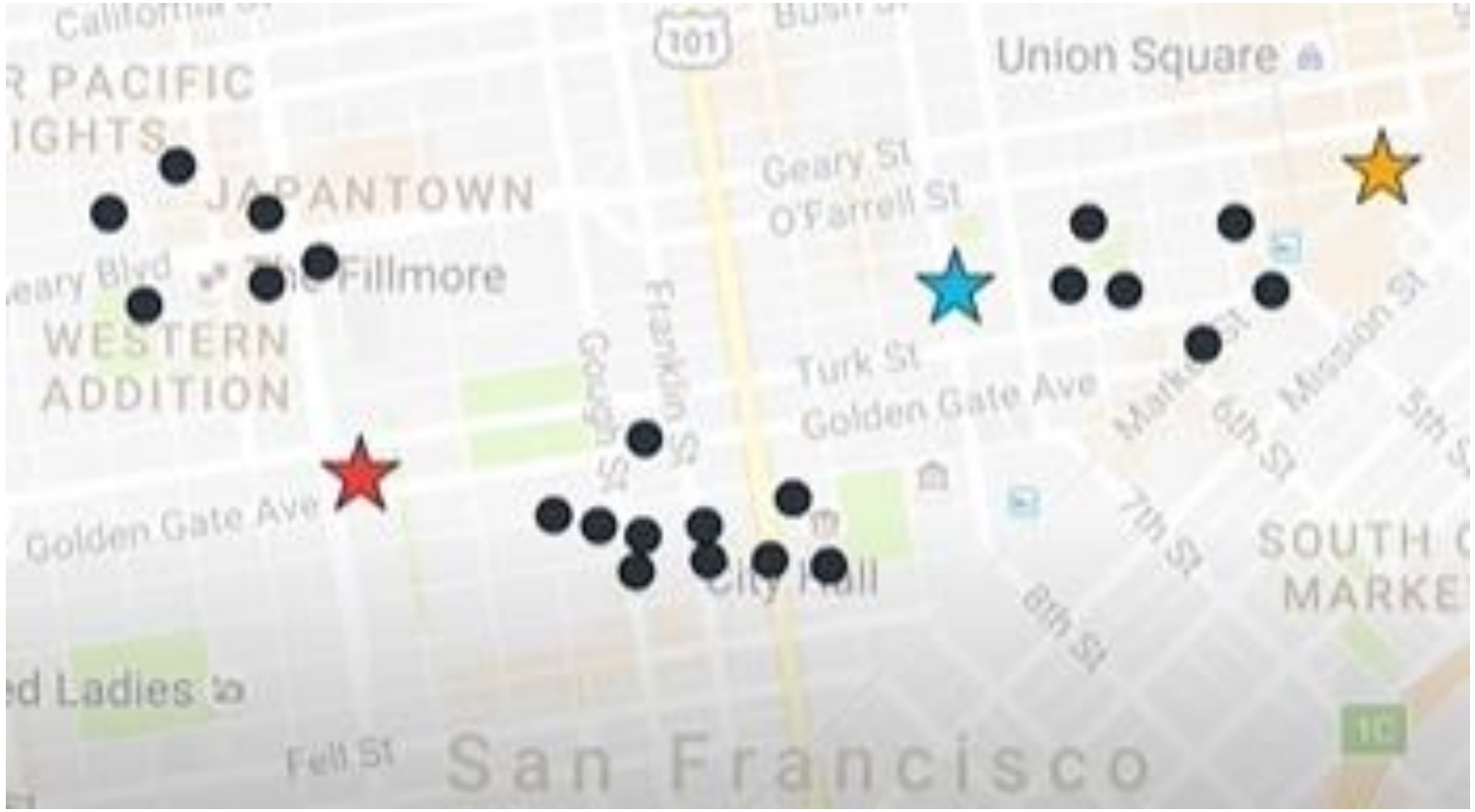
Пример: метод k-средних

Ищем лучшее расположение для 3-х кафе.



Пример: метод k-средних

Во-первых, помечаем 3 случайные точки (начальные центры).



Пример: метод k-средних

Во-вторых, распределяем точки по принципу близости к центру.

Пример: метод k-средних

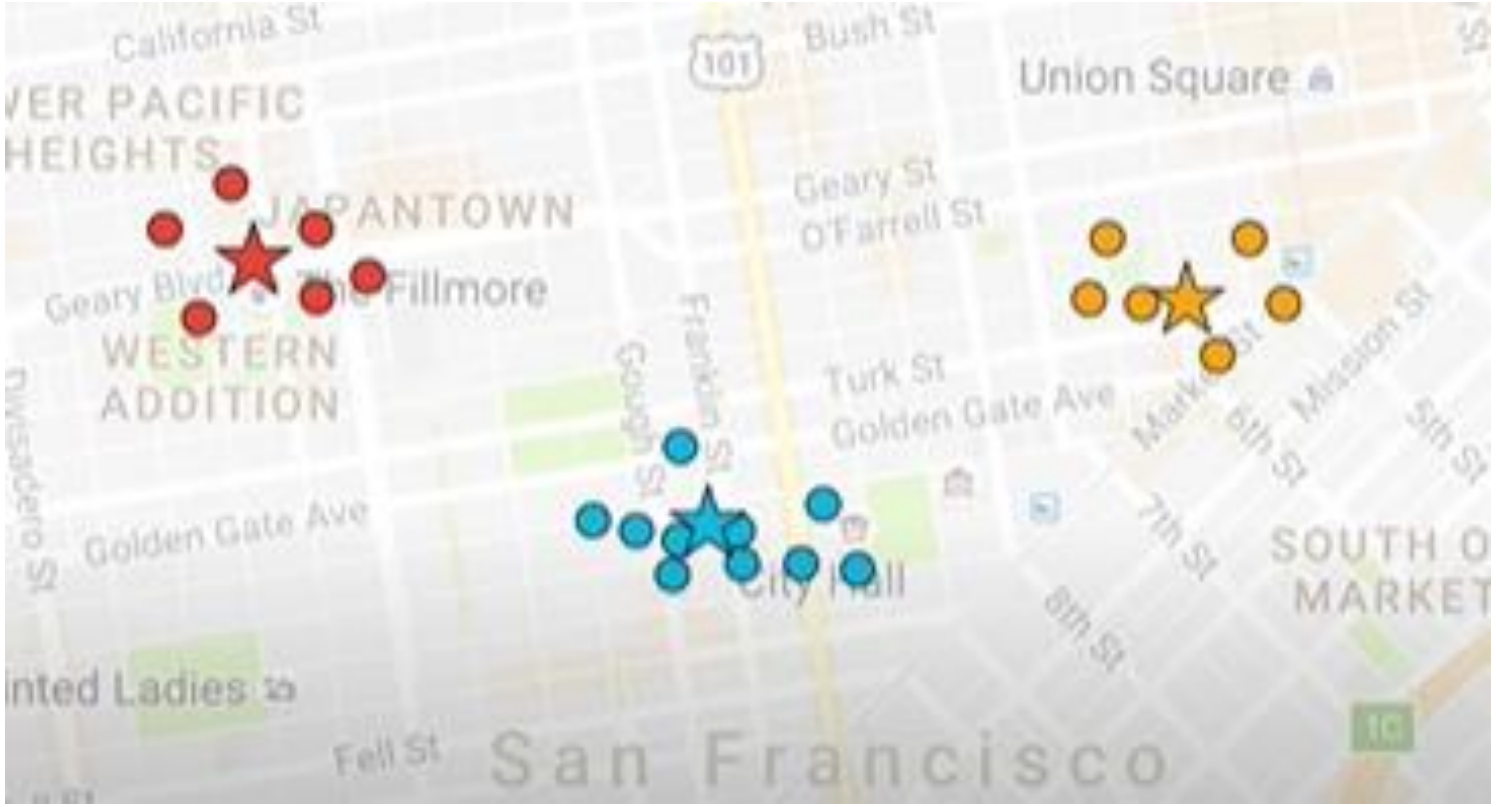
В-третьих, ставим центральные точки в центр домов, которые они обслуживают.

Пример: метод k-средних

В-четвёртых, снова пересчитываем центры, чтобы минимизировать расстояние между каждым зданием и кафе. Алгоритм повторяется пока возможны дальнейшие улучшения (сокращение расстояния между точками и центром).

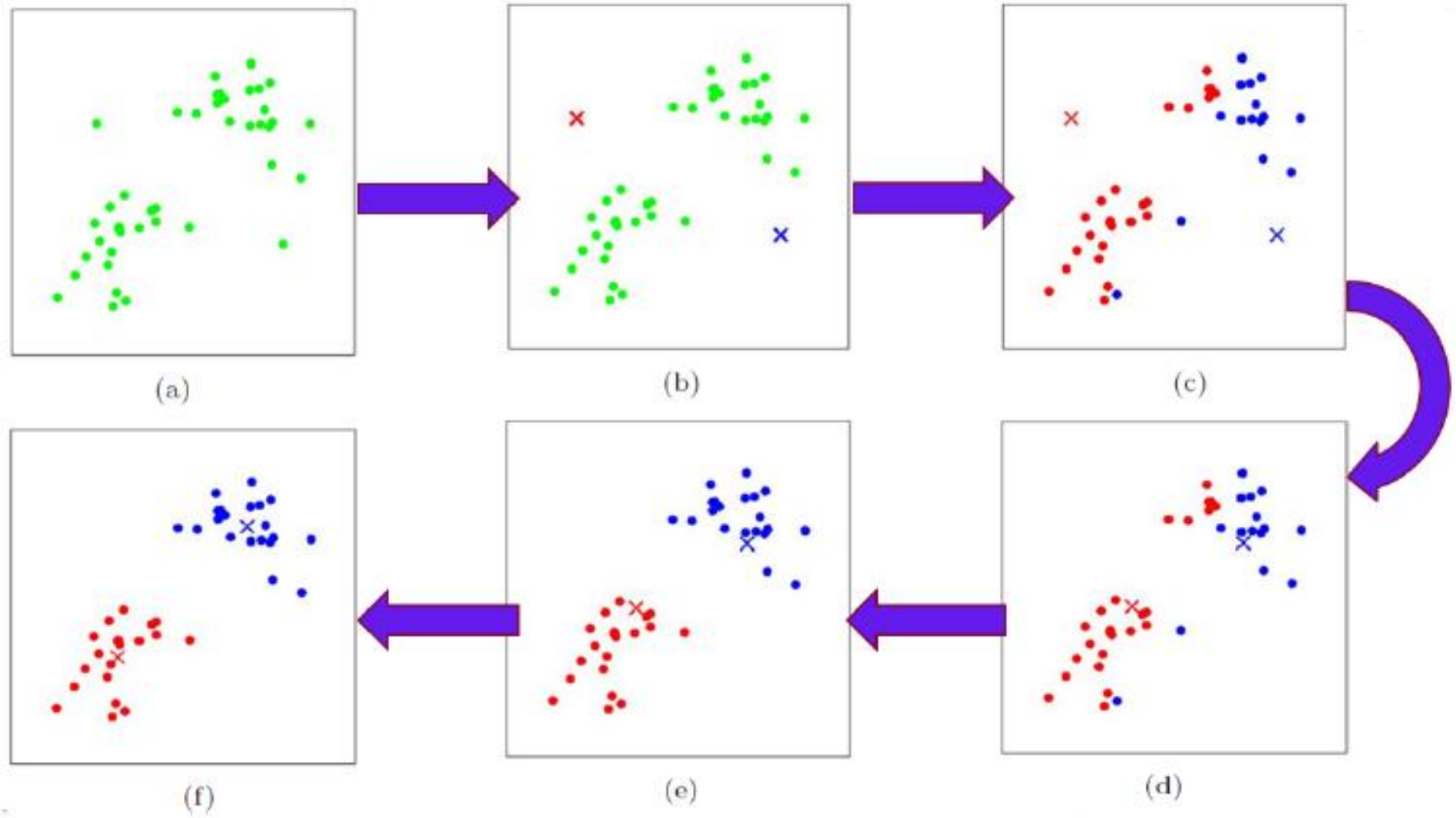
Пример: метод k-средних

Финальный результат.

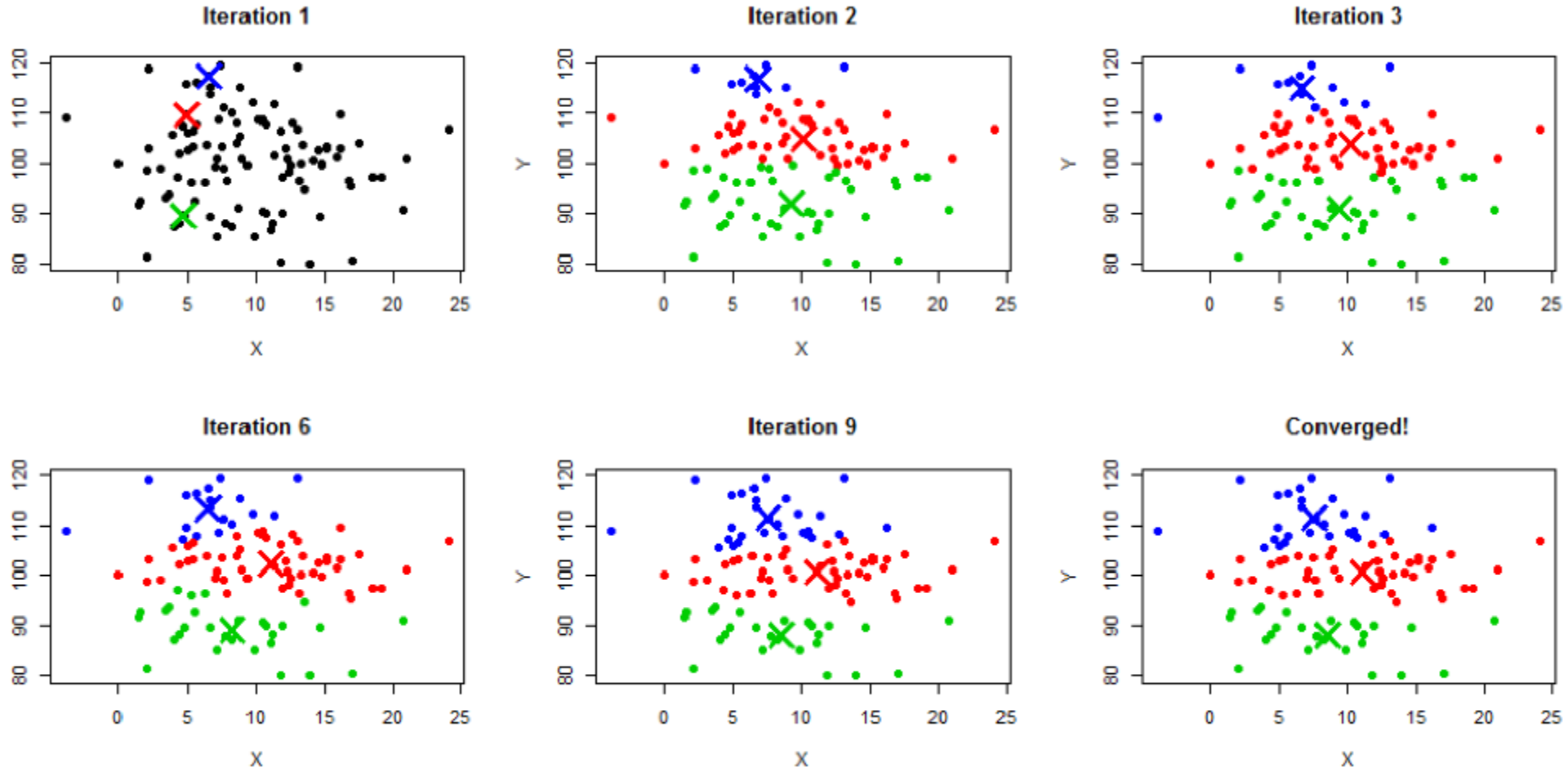


Источник: https://www.youtube.com/watch?v=QXOkPvFM6NU

Метод k-средних (2 кластера)



Метод k-средних (3 кластера)



k-medians

K-medians — это алгоритм кластеризации, похожий на K-means, но в качестве центра кластера он использует медиану. Благодаря этому K-medians более устойчив к выбросам. Он минимизирует манхэттенское расстояние.

K-medians использует в основе расчетов медиану, для вычисления которой требуется знать только порядок значений, а не их точные расстояния. Это делает его подходящим для порядковых данных, в отличие от K-means, который использует среднее значение и работает с метрическими переменными.

k-modes

K-modes специально разработан для работы с категориальными данными. В качестве центра кластера он использует моду (наиболее часто встречающуюся категорию). Для вычисления расстояния применяется мера несходства (количество несовпадений между категориями).

k-prototypes

K-Prototypes — это алгоритм кластеризации для наборов данных, содержащих как числовые, так и категориальные признаки. Он сочетает в себе подход K-Means (использует средние значения и евклидово расстояние для числовых данных) и K-Modes (использует моды и подсчёт несовпадений для категориальных данных). Каждый кластер представлен прототипом, состоящим из числовых средних значений и категориальных мод. Расстояние рассчитывается как взвешенная комбинация числового и категориального несходства. Такой подход делает алгоритм особенно полезным для реальных наборов данных со смешанными типами признаков.

k-means++

В алгоритме k-means начальные центры кластеров выбираются случайным образом. Если начальные точки выбраны неудачно, алгоритм может сойтись к плохому локальному минимуму, кластеры могут оказаться несбалансированными, а результаты — различаться между запусками.

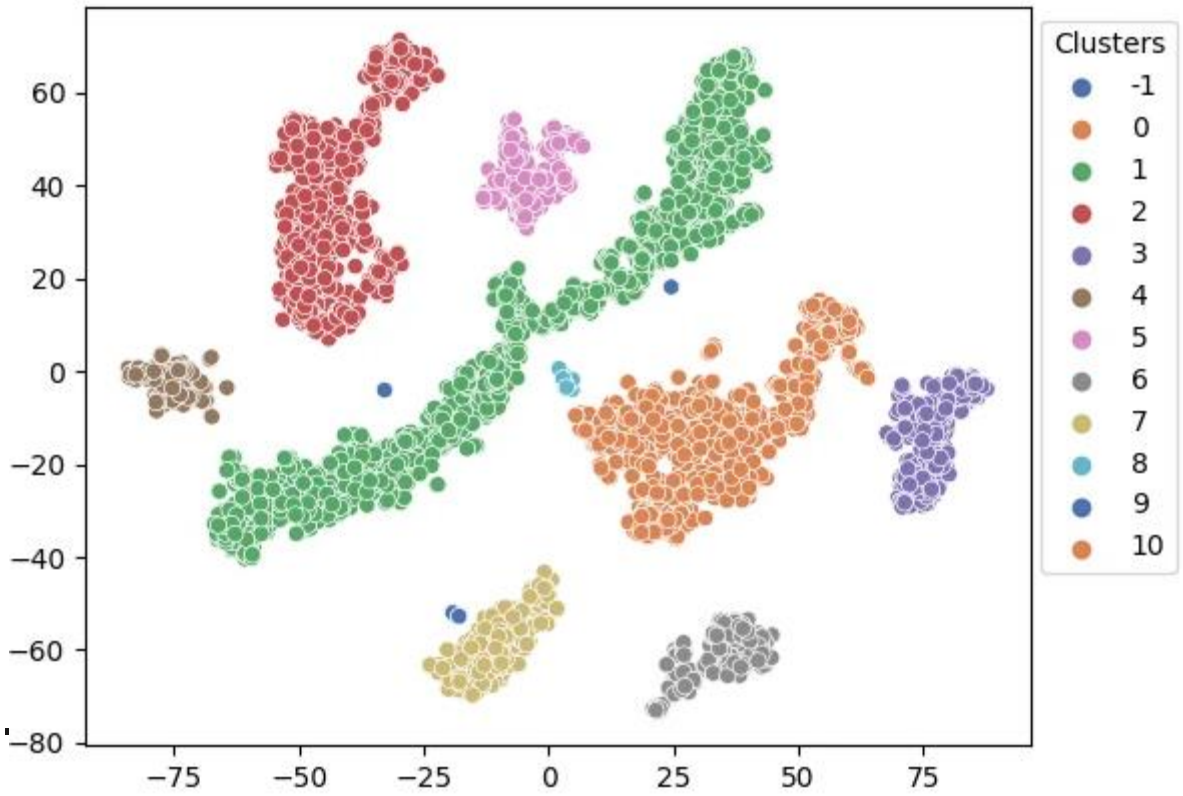
Алгоритм K-Means++ улучшает инициализацию, равномернее распределяя начальные центроиды перед запуском обычного K-Means. Это повышает вероятность нахождения хороших кластеров, снижает влияние случайности и ускоряет сходимость.

Плотностные алгоритмы кластеризации

Основной принцип работы плотностных алгоритмов – это выделение областей с высокой концентрацией точек. Точки, которые попали в области с низкой плотностью, считаются выбросами. Такие алгоритмы подходят для анализа пространственных данных. Например, для того, чтобы выявить области с наибольшим количеством ДТП. Плотностные алгоритмы учитывают выбросы. Если одно-два ДТП произошли вдали от города, эти объекты не будут отнесены к кластерам.

DBSCAN

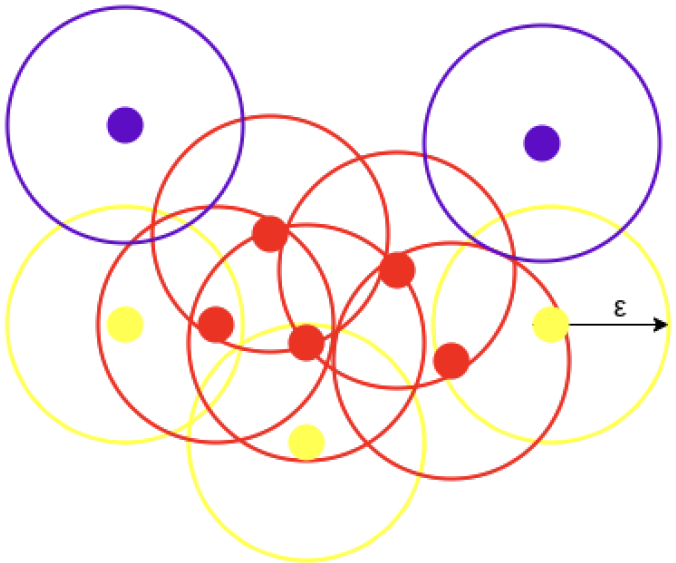
Алгоритм DBSCAN (Density Based Spatial Clustering of Applications with Noise) — плотностный алгоритм для кластеризации пространственных данных с присутствием шума. Способен распознать кластеры различной формы. При реализации алгоритма на вход подаются два параметра: радиус окрестности точек данных и минимальное число соседей.



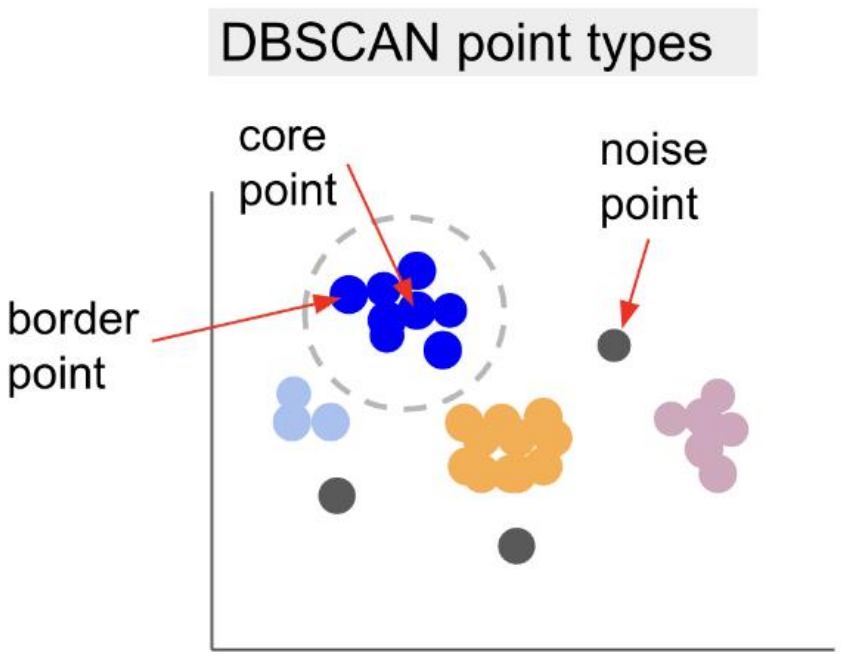
DBSCAN

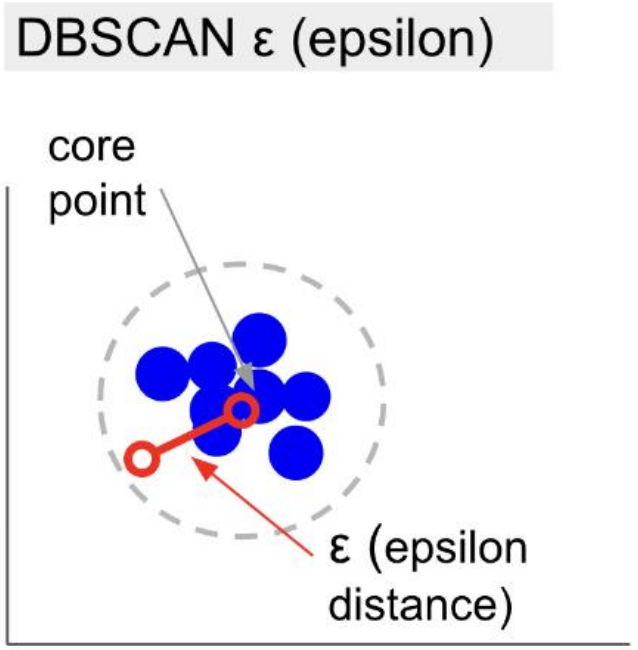
DBSCAN требует два параметра: epsilon (радиус круга, который должен быть создан вокруг каждой точки данных для проверки плотности) и minPoints (минимальное количество точек данных, необходимых внутри этого круга для того, чтобы эта точка данных была классифицирована как базовая).

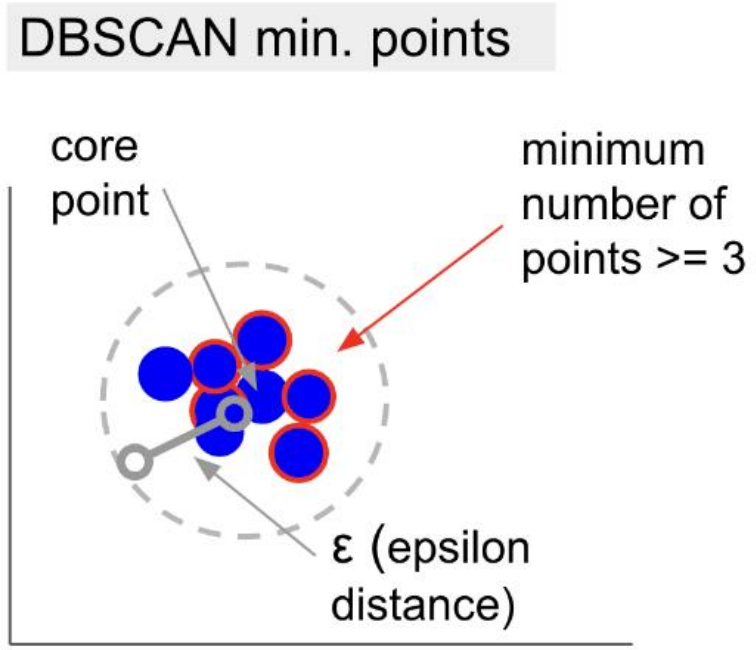
Точка данных является ***базовой***, если круг вокруг нее содержит не менее minPoints точек. Если количество точек меньше minPoints, то она классифицируется как ***граничная точка***, а если нет других точек в пределах эпсилон-радиуса, то точка рассматривается как ***шум***.



DBSCAN

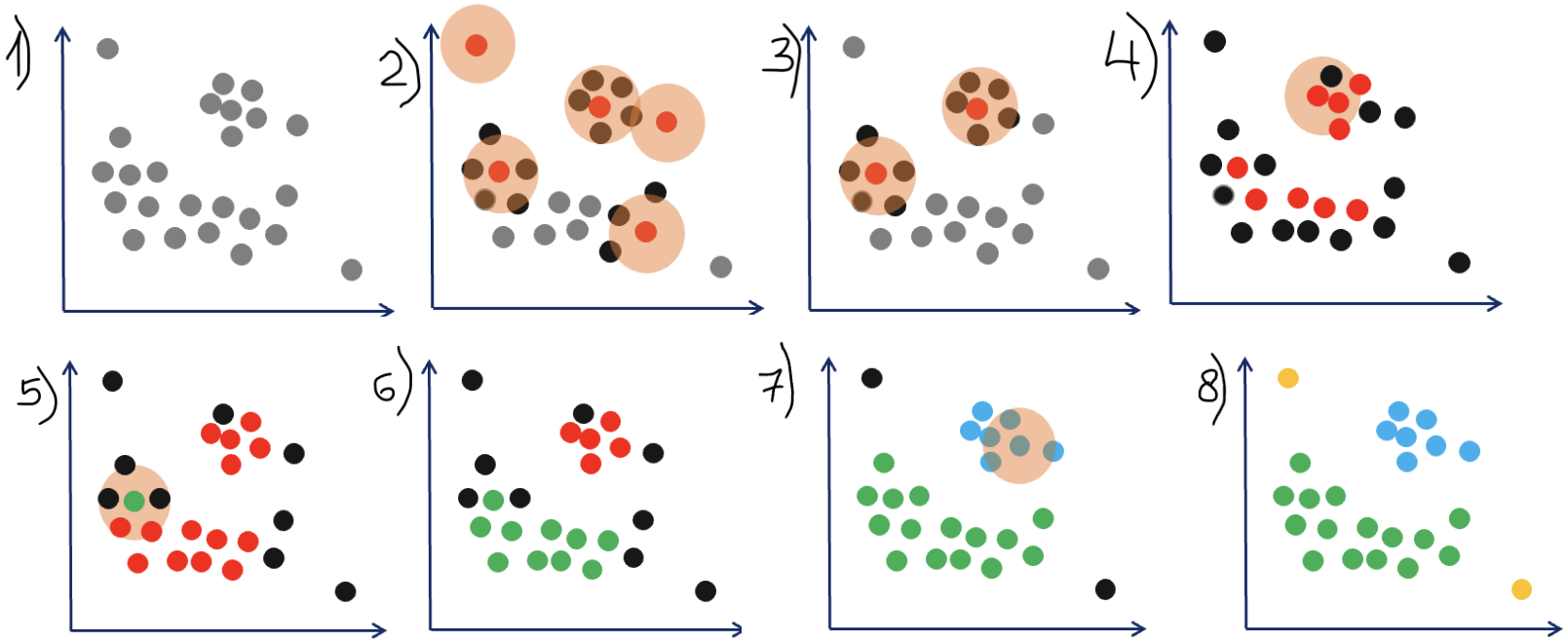

DBSCAN point types
core point
noise point
border point


DBSCAN ε (epsilon)
core point
ε (epsilon distance)


DBSCAN min. points
core point
minimum number of points >= 3
ε (epsilon distance)

DBSCAN

После того, как все точки классифицированы, алгоритм начинает формировать кластеры. Алгоритм создает кластер, соединяя все базовые точки и их граничные точки. Если две основные точки находятся достаточно близко друг к другу, они считаются частью одного кластера. Алгоритм продолжает формировать кластеры, пока все основные точки не будут присоединены к кластерам.



Преимущества и ограничения DBSCAN

Преимущества:
+ Находит кластеры различных форм и размеров;
+ Различает шумы и выбросы во входных данных;
+ Не требует предварительного указания количества кластеров;
+ Легко и быстро реализуется.

Ограничения:
- Параметрический, требует указания двух параметров (радиус и число соседей);
- Может быть чувствителен к выбору параметров и используемой метрике расстояния.

Выбор оптимального числа кластеров

Для принятия решения о числе кластеров необходимо учитывать специфику данных, результаты ранее проведённых исследований подобной тематики и интерпретируемость результатов. Возможно, стоит сохранить несколько кластерных решений и выбрать из них наилучшее с точки зрения однородности кластеров и содержательной интерпретации полученного решения.

Выбор количества кластеров может быть осуществлен на основе значений индекса Калински-Харабаша. Он рассчитывается как соотношение общего разброса объектов между кластерами и внутри их. Оптимальным считается число кластеров, при котором значение индекса максимально.

Индекс Калински-Харабаша

$$CH(K) = \frac{B(K)/(K-1)}{W(K)/(n-K)}$$

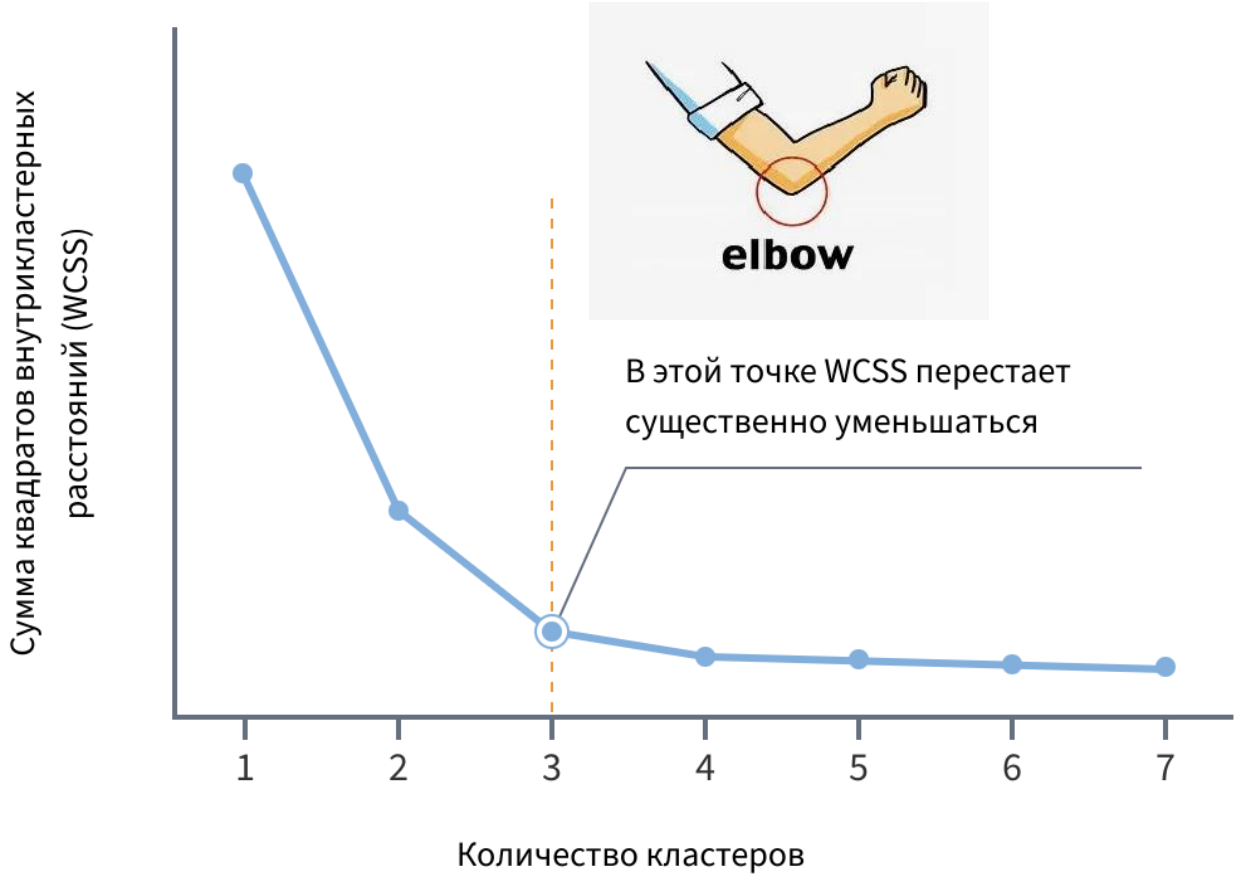
n – число наблюдений

K – число кластеров

B(K) – межкластерная вариация

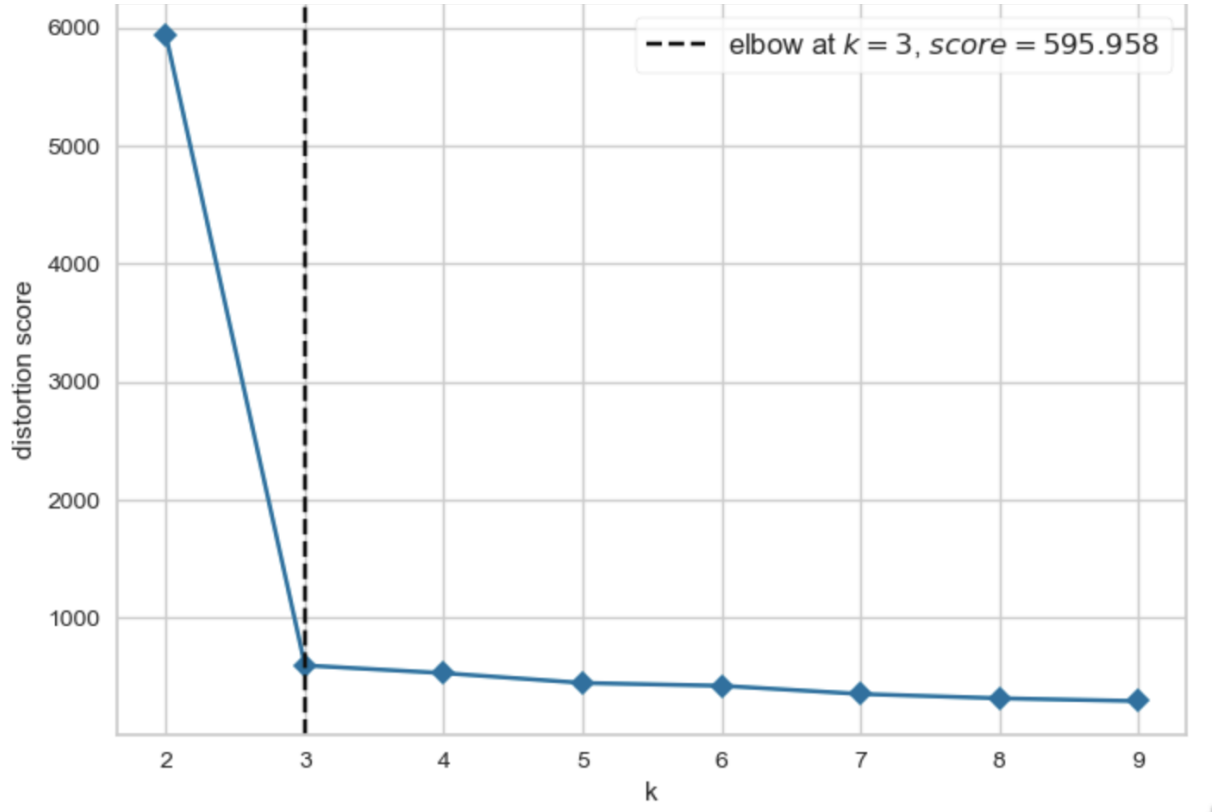
W(K) – внутрикластерная вариация

Метод локтя и инерция



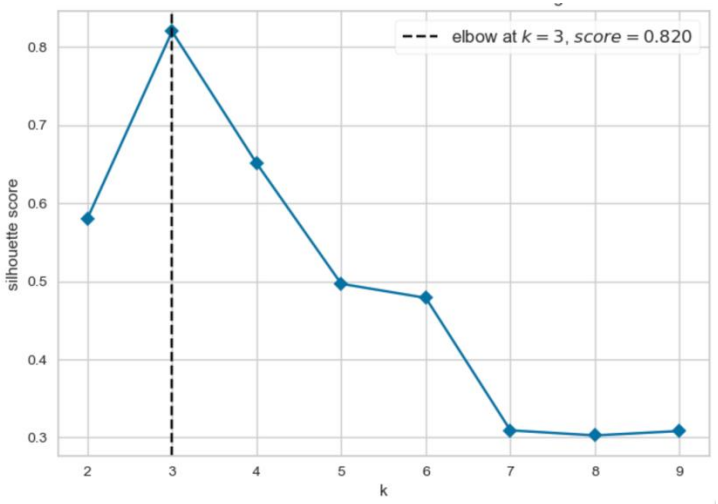
Инерция вычисляется путем измерения расстояния между каждой точкой данных и ее центроидом, возведения этого расстояния в квадрат и суммирования этих квадратов по каждому кластеру. Мы стремимся к получению модели с низкой инерцией.

Метод локтя и искажение



Искажение (distortion score) — это среднее значение квадрата евклидового расстояния от центроида до точек соответствующих кластеров.

Метод локтя и коэффициент силуэта



Коэффициент силуэта (silhouette score) показывает, насколько каждый объект «похож» на другие объекты в том кластере, в который он был распределен в процессе кластеризации, и «не похож» на объекты из других кластеров.

В основе идеи метода лежит вычисление коэффициентов, которые присваиваются каждому объекту в кластере и образуют так называемый силуэт кластера. Коэффициенты изменяются от -1 до 1. Значения, близкие к 1, указывают на то, что объект является похожим на другие объекты в кластере и не похожим на объекты из других кластеров. Если большинство объектов имеют значения коэффициентов близкими к 1, можно утверждать, что кластерная структура хорошо выражена, и количество кластеров соответствует естественной группировке данных.

Можно вычислить среднее значение силуэта по всем наблюдениям и использовать его как метрику для оценки количества кластеров.

Описание результатов и проверка их значимости

Для описания результатов кластеризации можно рассчитать описательную статистику по кластерам.

Для проверки статистической значимости полученных результатов можно провести тест ANOVA, чтобы убедиться, что средние значения анализируемых переменных значимо различаются по кластерам.

Для оценки качества кластеризации можно оценить дисперсию значений переменных по кластерам.

Описательная статистика по кластерам

Табл. 7. **Профили кластеров и распределение объектов**

Номер кластера	«ЕГЭ»	«Молодежь»	«УБФ»	«Размер»	«ИиР»	Количество объектов
1	59.81	0.24	82.75	5905.31	89.82	36
2	66.32	0.15	90.05	13495.43	86.92	37
3	57.42	0.06	95.82	5825.06	71.61	32
4	70.77	0.16	205.40	7947.39	70.72	18
5	59.53	0.21	85.15	7177.98	50.80	42
6	65.00	0.15	101.54	9005.85	13.41	54

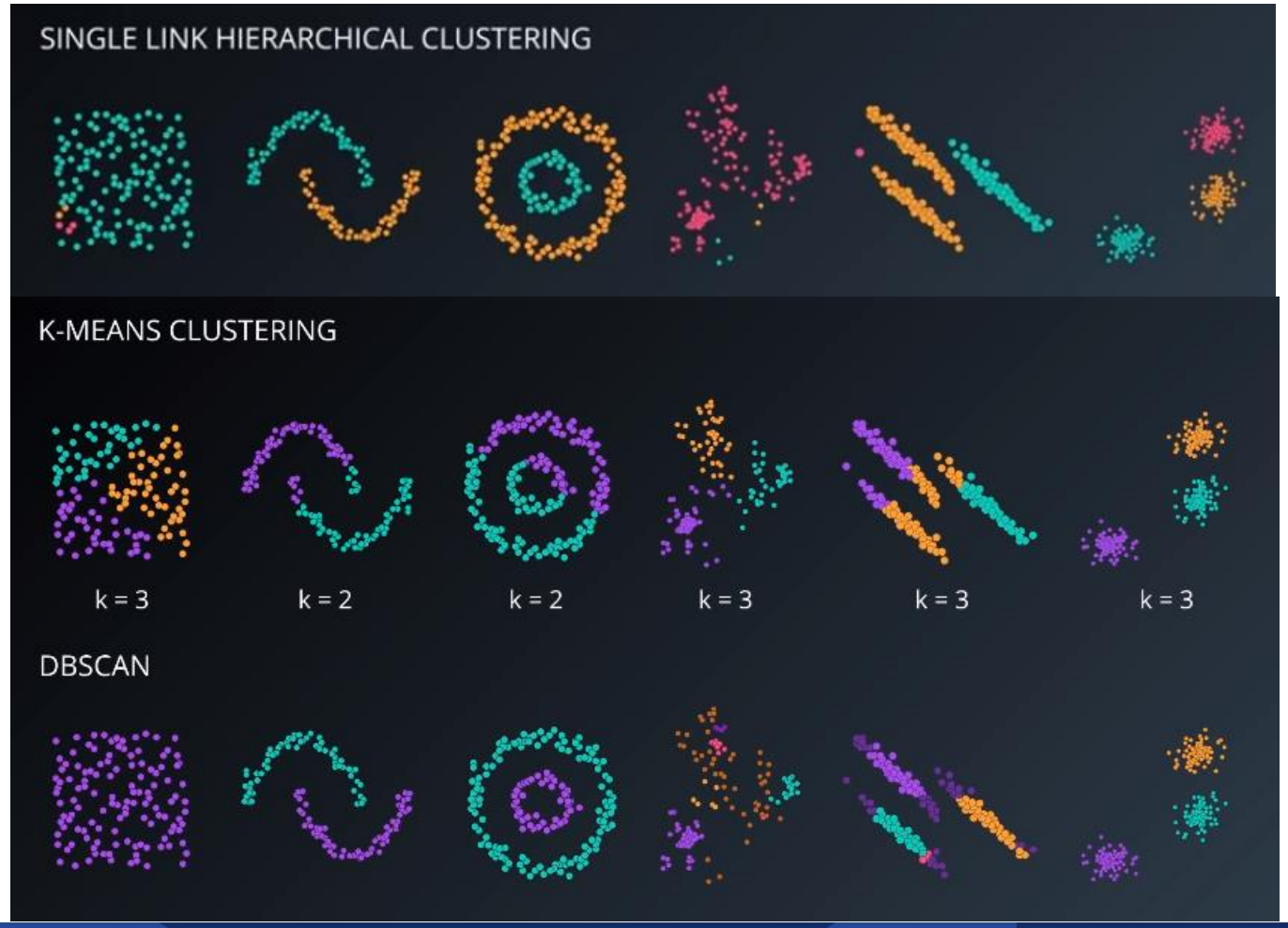
Источник: расчеты авторов.

Диагностика кластерной модели

Для подтверждения надёжности результатов кластеризации используются разные приёмы:

- Используются разные методы кластеризации и полученные результаты сопоставляются;
- Наблюдения разбиваются на две группы случайным образом, анализ выполняется отдельно для каждой группы, результаты сопоставляются;
- Анализ проводится по сокращённому набору переменных и результаты сравниваются с полноценным анализом.

Сравнение трех алгоритмов



Визуализация работы алгоритмов

- Visualizing K-Means Clustering (naftaliharris.com)
- Visualizing DBSCAN Clustering (naftaliharris.com)

Спасибо за внимание!