GeekBrains

Кластеризация и решение задачи группировки данных в машинном обучении.

Библиотеки Python для Data Science

Оглавление

Введение

Всем привет! В этой лекции мы поговорим о кластеризации и решении задачи группировки данных.

Кластеризация является важным аспектом машинного обучения, которые помогают организовать большие объемы информации и выявить скрытые закономерности. Кластеризация – это процесс разделения набора данных на группы или кластеры, состоящие из схожих объектов. Этот метод позволяет определить структуру данных, выделить характеристики, обнаружить аномалии и сделать выводы на основе полученных результатов.

В рамках лекции будут рассмотрены различные алгоритмы кластеризации, такие как k-средних, иерархическая кластеризация и DBSCAN. Мы узнаем, как эти алгоритмы работают и как выбрать подходящий для конкретной задачи. Также будет освещена проблема выбора оптимального числа кластеров и методы оценки качества кластеризации. Решение задачи группировки данных – это неотъемлемая часть машинного обучения, которая находит свое применение в различных сферах: от банковского дела до медицинских исследований. В данной лекции мы рассмотрим основные концепции и подходы к решению задачи группировки данных, а также изучим практические примеры их применения.
Давайте начнем!

На этой лекции вы найдете ответы на такие вопросы как / узнаете:

- Что такое кластеризация
- Понятие расстояния и меры сходства
- Алгоритмы кластеризации
- Методы оценки качества кластеризации

Термины, используемые в лекции

Кластеризация – это процесс разделения набора данных на группы или кластеры, состоящие из схожих объектов.

Кластер — это группа объектов или данных, которые похожи друг на друга больше, чем на объекты из других кластеров.

Дендрограмма — это графическое представление иерархии или структуры связей между объектами или группами объектов.

Метод K-средних (K-means) метод основан на итеративном повторении двух шагов, которые состоят в распределении объектов выборки по кластерам и пересчете центров кластеров.

Иерархическая кластеризация подразумевает создание иерархии кластеров.
DBSCAN (Density-based spatial clustering of applications with noise - пространственная кластеризация приложений с шумом на основе плотности) определяет кластеры путем анализа плотности точек.

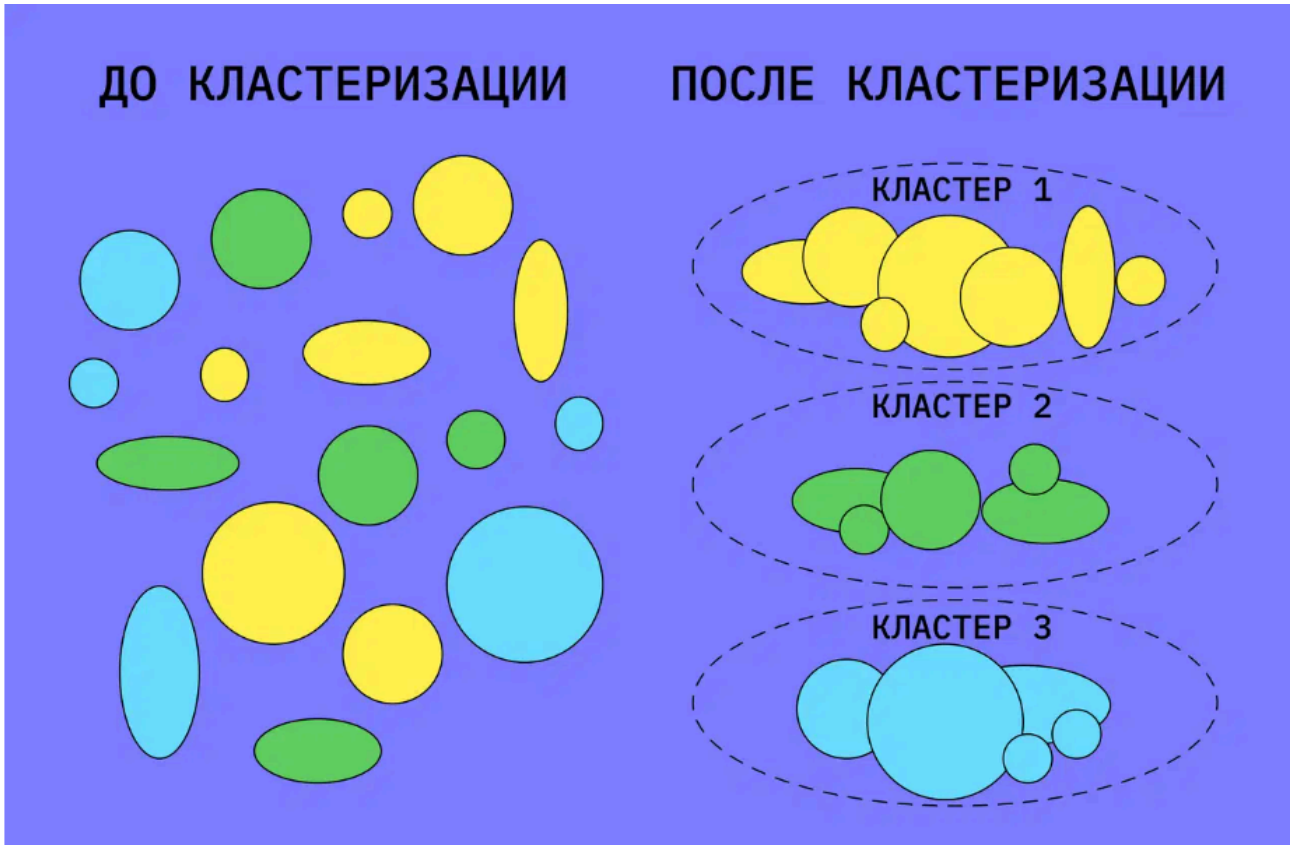
Введение в кластеризацию и группировку данных

В современном мире, охваченном цифровой революцией, накопление данных становится все более значимым и сложным процессом. Кластеризация представляет собой метод, позволяющий автоматически группировать данные на основе их сходства или различия, что помогает обнаруживать скрытые закономерности и структуры в информации.

В первой половине 20 века исследователи начали заниматься классификацией и сортировкой данных, открывая новую главу в истории кластерного анализа. Один из первых и наиболее известных методов, предложенный Гарри Чейном в 1957 году, был метод k-средних. Суть этого метода заключается в разделении объектов на кластеры путем вычисления среднего значения или центроида для каждого кластера. Это позволило исследователям классифицировать данные и распределить их по группам.

Преобразование необработанных данных в организованные структуры, выявление закономерностей и группировок — вот суть кластерного анализа в области анализа данных. Представьте себе обширную коллекцию точек данных, представляющих информацию о покупательском поведении клиентов или генетическую информацию различных организмов. В отдельности эти данные могут показаться беспорядочным набором цифр. Однако кластерный анализ приносит порядок в эту сложность и позволяет извлекать ценную информацию. Это дает нам преимущество.

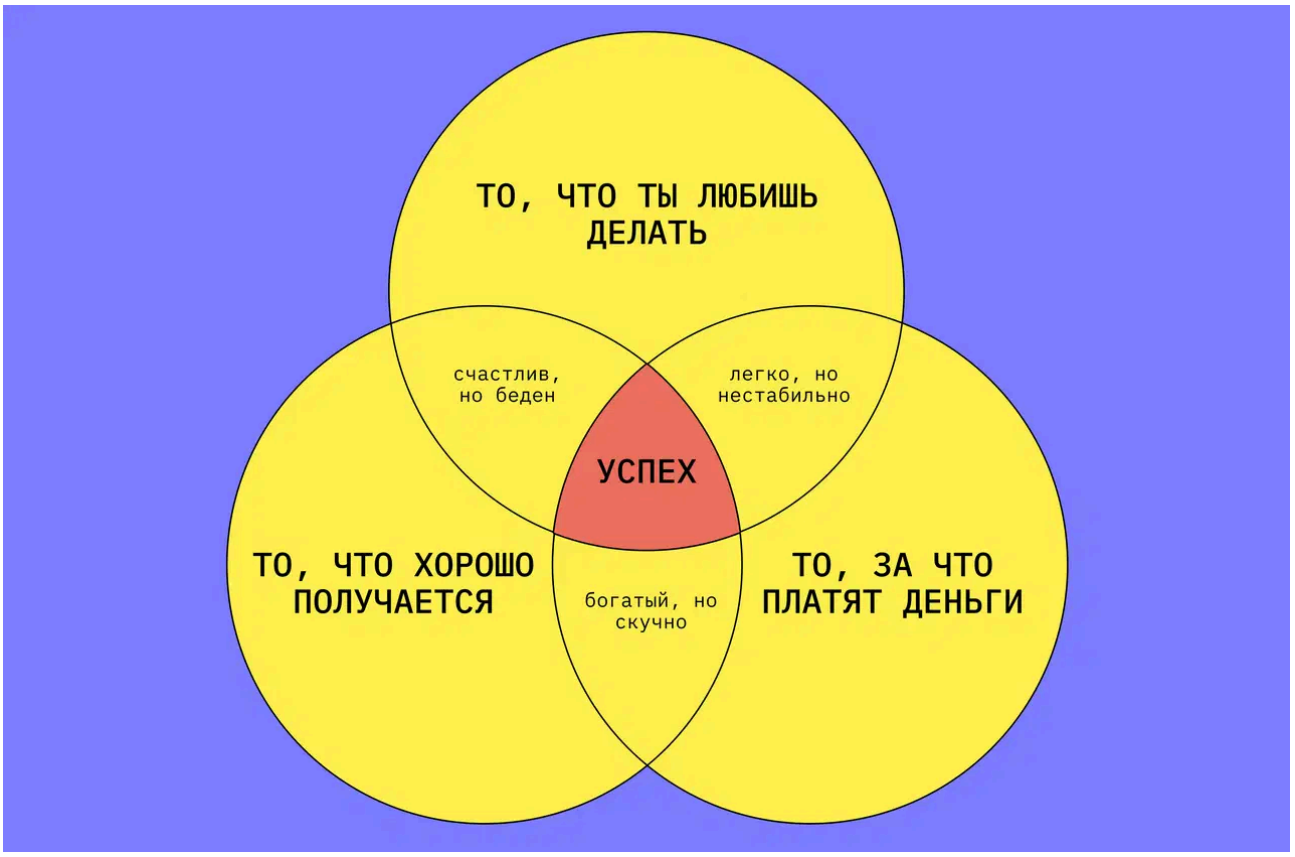
Кластерный анализ исходит из предположения, что объекты схожих атрибутов естественным образом группируются. Путем измерения сходства или несходства между точками данных, метод систематически организует их в кластеры, которые имеют общие характеристики. Этот процесс аналогичен группировке звезд на основе их спектральных характеристик или классификации животных на основе общих признаков.



Кластерный анализ данных предоставляет возможность проводить кластеризацию не только один раз, а множество раз. В результате таких повторных

кластеризаций можно выделить подкластеры, что способствует формированию иерархической структуры. В этой структуре каждому объекту может быть присвоено несколько кластеров, начиная от самых маленьких и заканчивая наибольшими.

Однако следует отметить, что кластерный анализ данных может представлять ситуации, где объекты пересекаются между различными кластерами. То есть, один и тот же объект может принадлежать к двум или более кластерам, если у него совпадают определенные критерии. Это свойство пересекающихся кластеров позволяет более точно описывать и классифицировать данные, учитывая их множественные свойства и особенности.



Рассмотрим конкретный пример, чтобы лучше понять важность и преимущества использования алгоритма кластерного анализа в онлайн-магазинах. Давайте представим, что у нас есть данные о посетителях определенного онлайн-магазина, в которых указан их возраст.

Существует возможность применить к этим данным алгоритм кластерного анализа, чтобы разделить посетителей на группы схожих возрастов. Например, мы можем создать кластеры по следующим возрастным интервалам: до 18 лет, от 18 до 25 лет, от 26 до 30 лет, от 31 до 40 лет, от 41 до 50 лет, от 51 года и старше.

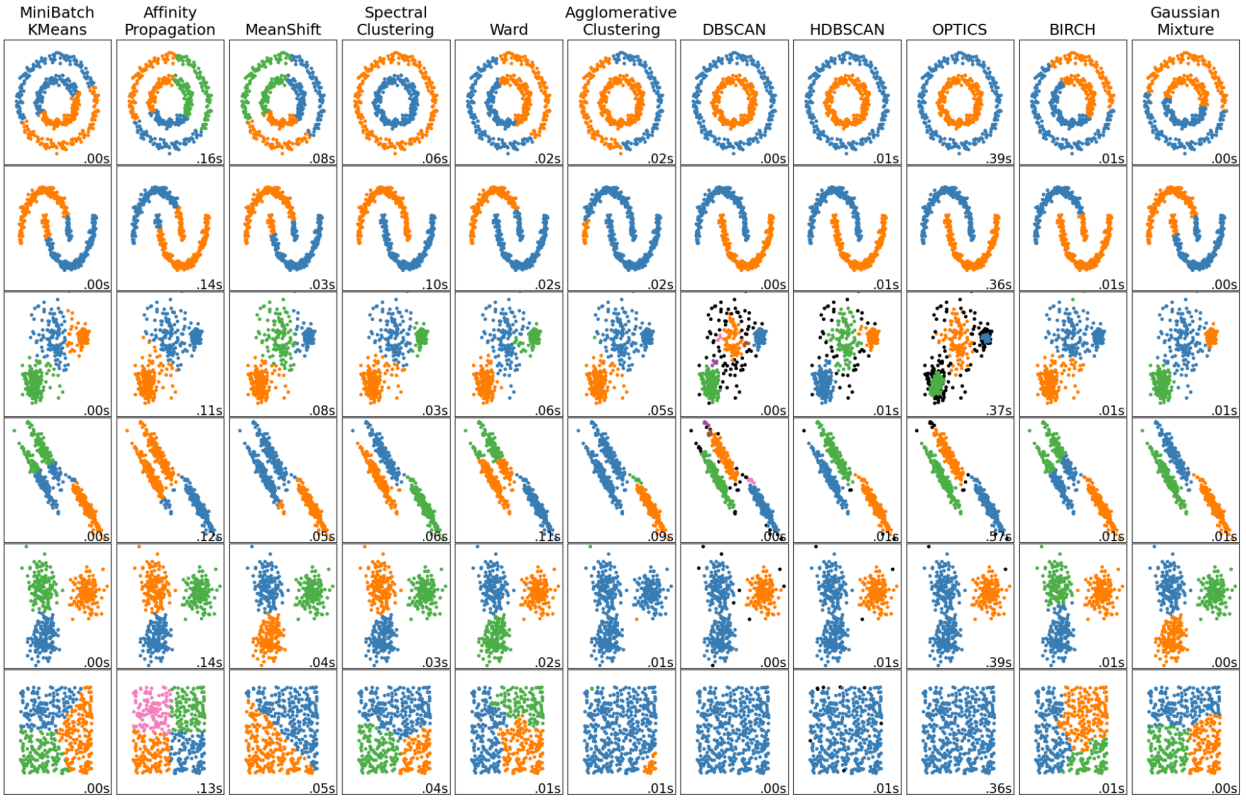
Такой подход позволит нам более детально и глубоко изучить поведение и предпочтения посетителей каждой возрастной группы. Мы сможем провести дополнительный анализ и выяснить, какие товары и услуги чаще всего покупают люди определенного возраста.

Кроме того, с помощью кластерного анализа мы сможем определить, сколько времени каждая возрастная группа проводит в онлайн-магазине. Это позволит нам понять, насколько активно каждая группа взаимодействует с платформой и какие дополнительные мероприятия можно предпринять для привлечения большего количества посетителей.

Также, благодаря анализу данных о покупках, мы сможем рассчитать общую сумму покупок каждой возрастной группы. Это поможет установить, какие группы клиентов наиболее ценны для бизнеса и на какие группы стоит ориентироваться при планировании маркетинговых акций и предложений.

Таким образом, использование алгоритма кластерного анализа на примере данных о посетителях онлайн-магазина позволяет получить ценную информацию о поведении и предпочтениях клиентов разных возрастных групп. Это поможет оптимизировать работу магазина, улучшить маркетинговые стратегии и повысить эффективность взаимодействия с клиентами.

Выбор подходящего показателя расстояния является начальным этапом процесса, который определяет количественное сходство между точками данных.



Алгоритм начинает создавать кластеры, объединяя или разделяя точки данных на основе их близости, как только матрица расстояний установлена. Евклидово расстояние, манхэттенское расстояние и косинусное сходство являются наиболее часто используемыми показателями. В результате формируются отдельные группы, каждая из которых представляет уникальное подмножество данных.

Кластерный анализ поддерживает различные методологии, которые адаптированы к конкретным типам данных и целям исследования.

Иерархическая агломеративная кластеризация создает дендрограмму (это графическое представление иерархии или структуры связей между объектами или группами объектов.), которая наглядно демонстрирует иерархические связи между кластерами. K-means кластеризация итеративно присваивает точки данных кластерам, минимизируя сумму квадратов расстояний внутри каждого кластера. DBSCAN и другие методы, основанные на плотности, определяют кластеры на основе плотности точек данных, эффективно выявляя кластеры необычной формы.

Кластерный анализ выходит за рамки простого организования данных — он предоставляет информацию, которая влияет на принятие решений.

Идентификация генов со сходными функциями в биологии путем кластеризации данных об экспрессии генов может пролить свет на биологические пути, аналогично тому, как понимание потребительских сегментов в бизнесе может привести к проведению целенаправленных маркетинговых кампаний, которые находят отклик у конкретной аудитории. Кроме того, в социальных науках кластеризация может помочь идентифицировать различные демографические группы, выявляя закономерности в ответах на опросы. Однако для достижения эффективности кластерного анализа необходимы вдумчивые соображения и осторожная интерпретация.

Для достижения оптимального количества кластеров, известного как "точка пересечения", необходимо найти баланс между простотой и детализацией. Ключевым фактором является выбор подходящих показателей расстояния и алгоритмов кластеризации, которые сыграют важную роль в достижении значимых результатов. Важно, чтобы кластеры были интерпретируемыми и соответствовали знаниям исследователя в конкретной области.

Кластерный подход предоставляет возможности для применения в различных сферах и областях. Он может быть использован не только для анализа данных, но и для исследования рынка, изучения статистики и анализа мнений. Важно отметить, что для успешной кластеризации данных необходимо наличие общих признаков.

Сферы применения кластерного подхода включают:

1. Анализ поведения клиентов: Кластеризация поможет выявить основные группы клиентов и понять их предпочтения и потребности. Это позволит улучшить стратегию маркетинга и предложить более персонализированный подход к каждой группе клиентов.

2. Исследование рынка: Кластерный анализ может использоваться для выявления основных конкурентов в определенной отрасли и проведения исследования их деятельности. Это поможет бизнесу разработать более эффективные стратегии для конкурентного преимущества.

3. Анализ статистики выздоровления: Кластеризация данных о заболеваниях и их статистике позволяет выявить общие закономерности и факторы, влияющие на выздоровление. Это может быть полезно для разработки более эффективных методов лечения и профилактики.

4. Анализ мнений и предпочтений: Кластеризация респондентов опроса позволяет разделить их на группы с похожими мнениями и предпочтениями. Это может быть полезно для понимания потребностей и ожиданий разных групп людей и адаптации маркетинговых стратегий под их запросы.

5. Формирование тематик страниц сайта: Кластерный анализ SEO-ключей помогает определить основные тематики и интересы пользователей, что позволяет оптимизировать контент и улучшить его релевантность для целевой аудитории.

Таким образом, кластерный подход имеет широкий спектр применения и может быть полезен во многих сферах. Он позволяет выявить общие закономерности и группы данных, что помогает принимать более обоснованные решения и разрабатывать эффективные стратегии.

1. Гораздо больше сфер применения открывается при использовании кластеризации для обработки различных файлов разных форматов.

2. Кластеризацию можно успешно применять не только к текстовым данным, но и к изображениям, аудиофайлам и видеофайлам.

3. Удобство обработки собранных файлов разных форматов становится особенно важным при работе с огромными объемами информации.

Собранные файлы разных форматов для их удобной обработки — это только начало. Кластеризация, как мощный инструмент, имеет гораздо больше сфер применения, чем мы могли бы подумать. Она может быть полезна в любой области, где требуется систематизация и структурирование данных.

Например, при работе с большими массивами клиентской информации, кластеризация поможет выявить группы схожих потребителей и оптимизировать маркетинговые стратегии для каждой группы. Кроме того, кластеризация может быть применена в медицине для классификации и анализа медицинских изображений, улучшения диагностики и разработки новых лечебных методов. Таким образом, кластеризация имеет огромный потенциал для оптимизации работы и достижения новых результатов. С обилием различных файлов разных форматов, кластеризация становится неотъемлемой частью работы с огромными объемами информации, обеспечивая удобство обработки и улучшенную организацию данных.

Понятие расстояния и меры сходства.

Итак, как же определять «сходство» объектов?

В процессе определения "сходства" объектов, первым шагом является составление вектора характеристик для каждого из них. Этот вектор может содержать различные числовые значения, например, рост и вес человека. Однако помимо таких числовых характеристик, существуют также категорийные характеристики, с которыми работают определенные алгоритмы.

Когда вектор характеристик готов, следующим шагом является нормализация. Это необходимо для того, чтобы все компоненты вектора вносили равный вклад при расчете "расстояния" между объектами. В процессе нормализации, все значения приводятся к определенному диапазону, например, [-1, 1] или [0, 1]. Это позволяет сравнивать объекты на основе их характеристик и учитывать их взаимное влияние на результат.

Основные метрики, используемые для измерения степени похожести между объектами, включают:

- Меры расстояния используются для измерения расстояния между двумя объектами или точками данных.

1. Евклидово расстояние: это наиболее распространенная метрика, которая измеряет физическое расстояние между двумя точками в n-мерном пространстве. Она основывается на теореме Пифагора и может быть применена в различных областях, включая геометрию, физику и машинное обучение.

$$\rho\left(x, x^{'}\right) = \sqrt{\sum_{i}^{n}\left(x_i - x_i^{'}\right)^2}$$

Пример:

Пусть есть две точки данных:

x = (1, 2, 3) y = (4, 6, 5)

Тогда евклидово расстояние между ними:

$$\rho(x, y) = \sqrt{(1-4)^2 + (2-6)^2 + (3-5)^2} = \sqrt{9 + 16 + 4} = \sqrt{29} = 5.29$$

Чем меньше евклидово расстояние между объектами, тем они ближе друг к другу в n-мерном пространстве признаков.

Это свойство используется в кластеризации - близкие по евклидову расстоянию объекты объединяются в один кластер. Например, в k-means кластеризации центры кластеров выбираются так, чтобы минимизировать суммарное евклидово расстояние от всех объектов кластера до его центра

2. Квадрата евклидова расстояния. Этот метод позволяет придать больший вес объектам, расположенным на большем расстоянии друг от друга. Расстояние вычисляется по следующей формуле:

$$\rho\left(x, x^{'}\right) = \sum_{i}^{n}\left(x_i - x_i^{'}\right)^2$$

В отличие от евклидова расстояния, здесь опускается операция извлечения квадратного корня. Квадрат евклидова расстояния сохраняет порядок близости объектов. При этом вычисления упрощаются, так как отпадает необходимость извлекать корень.

Это свойство часто используется в кластеризации для ускорения вычислений. Например, в k-means алгоритме при вычислении близости к центрам кластеров может применяться квадрат евклидова расстояния.

3. Манхэттенское расстояние: в отличие от евклидового расстояния, оно измеряет расстояние между двумя точками, перемещаясь только по перпендикулярным осям. Эта метрика особенно полезна для задач, связанных с пути и перемещением, и может быть использована, например, для определения кратчайшего пути в городе. Оно рассчитывается как среднее абсолютное значение разности координат двух точек. Эта мера расстояния обычно дает те же результаты, что и обычное евклидово расстояние, но с учетом влияния выбросов, которые не возводятся в квадрат.

$$\rho\left(x, x^{'}\right) = \sum_{i}^{n}\left|x_i - x_i^{'}\right|$$

В отличие от евклидова расстояния, здесь не вычисляется квадрат разности координат, а просто суммируются модули разностей по каждой координате. Интуитивно это расстояние можно представить как количество кварталов, которое нужно пройти в Манхэттене между двумя точками на карте с прямоугольной системой улиц.

Манхэттенское расстояние менее чувствительно к выбросам в отдельных признаках, чем евклидово. Это может быть полезно в некоторых задачах кластеризации, например, при наличии шумных данных.

4. Расстояние Чебышева представляет собой меру расстояния между двумя точками, которая учитывает только самую большую разницу по координатам. Это может быть полезно, когда нужно определить, насколько два объекта отличаются друг от друга по какой-то одной координате. Например, расстояние Чебышева может помочь определить, насколько две точки отличаются друг от друга по горизонтальной координате, не обращая внимание на их вертикальные координаты.

$$\rho\left(x, x^{'}\right) = max(\left|x_i - x_i^{'}\right|)$$

То есть это максимальная разность между соответствующими компонентами векторов.

Например:

x = (1, 3, 2) y = (2, 6, 4)

Расстояние Чебышева будет равно:

p(x, y) = max(|1 - 2|, |3 - 6|, |2 - 4|) = max(1, 3, 2) = 3

Основное свойство этой метрики в том, что она учитывает только максимальную разницу по одной из координат.

Это может быть полезно в некоторых задачах кластеризации, когда важно сгруппировать объекты с близкими максимальными значениями признаков, игнорируя другие различия. Однако в целом расстояние Чебышева используется реже классических метрик вроде евклидовой.

- Меры сходства наоборот измеряют степень схожести или близости между объектами.

1. Косинусное расстояние: это метрика, используемая для измерения угла между двумя векторами. Она основана на косинусе угла между векторами и обычно применяется в задачах, связанных с текстовым анализом, классификацией документов и рекомендательными системами.

$$\Delta cos(x, y) = 1 - cos(x, y)$$

Где cos(x, y) - косинус угла между векторами x и y.

Чем меньше косинусное расстояние, тем более похожи векторы.

Пример:

Пусть есть два документа x и y, представленные в виде векторов слов:

x = (3, 1, 0, 1, 2) y = (1, 0, 1, 1, 1)

Косинус угла между этими векторами равен:

$$cos(x, y) = \frac{(31 + 1 + 1 + 11 + 2 * 1)}{\left(\sqrt{3^2 + 2^1 + 0^2 + 1^2 + 2^2}\right) * \sqrt{(1^2 + 0^2 + 1^2 + 1^2 + 1^2)}} = 0.57$$

Следовательно, косинусное расстояние:

dcos(x, y) = 1 - 0.57 = 0.43

Чем меньше косинусное расстояние, тем более похожи документы по словарному составу. Это позволяет эффективно применять косинусную метрику в задачах кластеризации текстов и других объектов, представимых векторно.

5. Расстояние Хэмминга: данная метрика измеряет различия между двумя строками одинаковой длины. Она подсчитывает количество различных символов на соответствующих позициях в строках и используется в задачах, связанных с обработкой и сравнением данных.

$$d_{ij} = \sum_{k=1}^{p} |x_{ik} - x_{jk}|$$

1010001	25687914	пакет
1000101	24657934	багет
d=2	d=3	d=2

Чем меньше расстояние Хэмминга, тем более похожи два объекта.

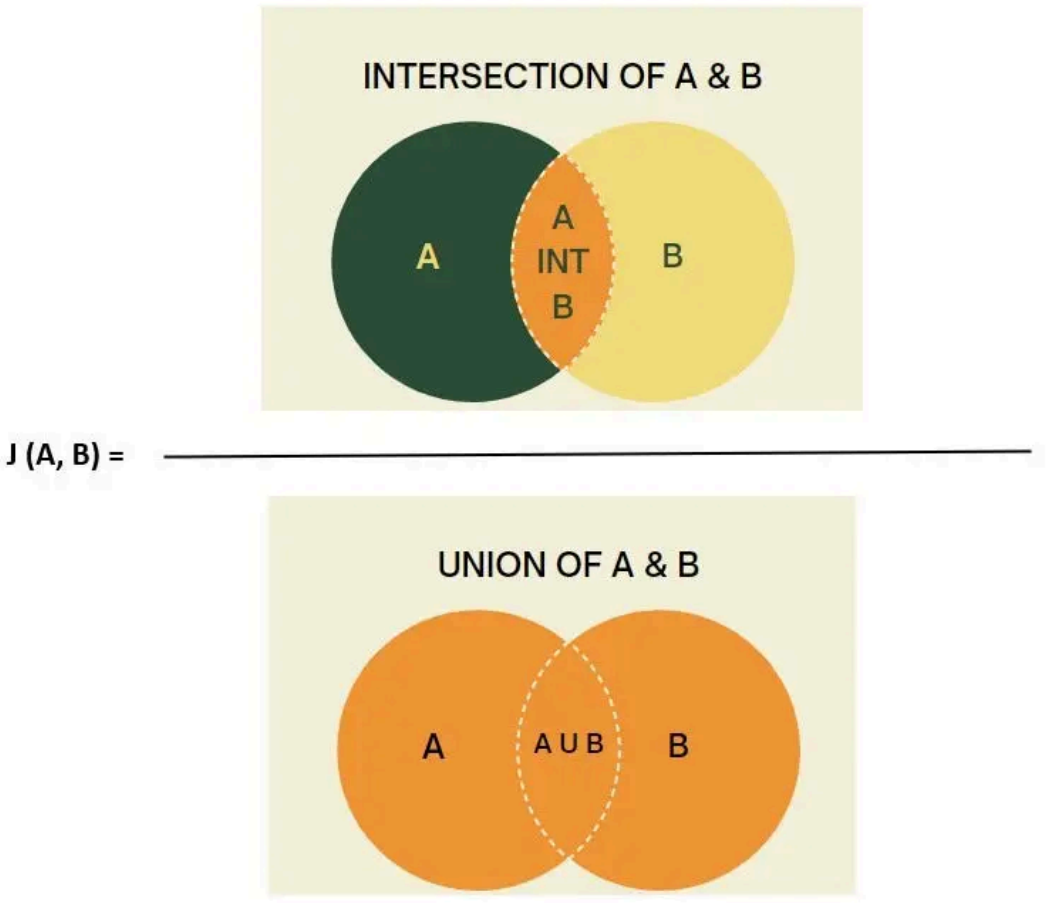
Преимущества использования расстояния Хэмминга:

- Простота вычисления для бинарных данных.
- Не требует нормализации данных, поскольку считает только различия в битах.
- Быстрое вычисление расстояний между большим количеством объектов.

Таким образом, расстояние Хэмминга хорошо подходит для кластеризации больших наборов категориальных или бинарных данных, где важна вычислительная эффективность. Оно позволяет эффективно группировать похожие объекты в кластеры.

6. Метрика Жаккара: она используется для измерения сходства между двумя множествами. Она основана на подсчете числа общих элементов в обоих множествах. Иными словами, это мера различия между двумя строками одинаковой длины, определяемая количеством позиций, в которых символы в этих строках различаются.

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|} = \frac{|A \cap B|}{|A| + |B| - |A \cap B|}.$$



Где:

A и B - два набора объектов |A ∩ B| - количество объектов, принадлежащих одновременно наборам A и B (пересечение) |A ∪ B| - количество объектов, принадлежащих хотя бы одному из наборов A или B (объединение)

Например, пусть:

A = {1, 2, 3} B = {2, 3, 4}

Тогда:

|A ∩ B| = 2 (общие элементы 2 и 3) |A ∪ B| = 4 (все различные элементы 1, 2, 3, 4)

J(A, B) = 2/4 = 0.5

Чем ближе значение метрики Жаккара к 1, тем похожее содержимое имеют сравниваемые наборы.

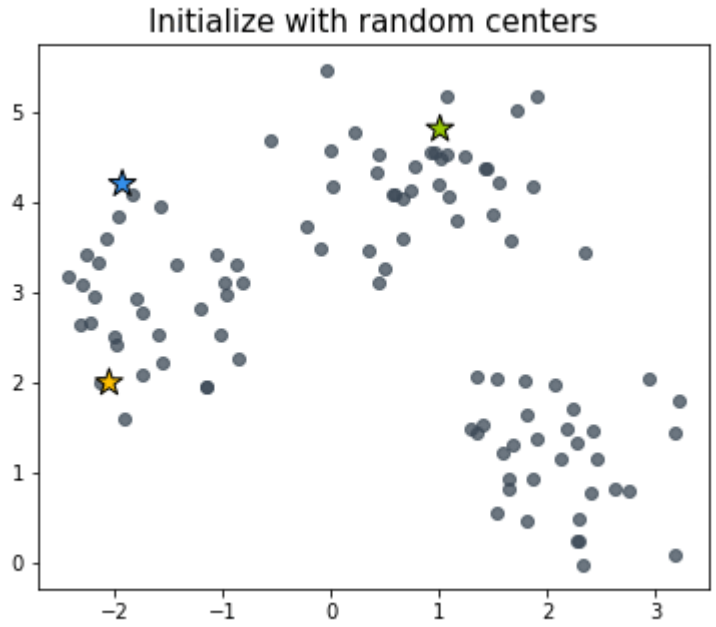
В кластеризации метрика Жаккара может применяться для определения схожести документов по наборам слов, пользователей по интересам, товаров по меткам и т.д. Она позволяет учитывать только наличие общих объектов, не обращая внимания на отсутствующие в одном из наборов.

Таким образом, выбор метрики зависит от конкретной задачи и типа данных, с которыми мы работаем. Каждая из этих метрик имеет свои преимущества и недостатки, и правильный выбор поможет нам получить более точные результаты и эффективно решить поставленную задачу.

Кластеризация на основе центроидов

Одним из наиболее распространенных методов кластеризации является кластеризация на основе центроидов. В этом методе каждый кластер представлен центроидом — точкой, которая является средним значением всех точек в данном кластере.

Кластеризация на основе центроидов имеет широкий спектр применений, начиная от анализа данных в контексте исследовательских работ до разработки систем рекомендаций компаниями. Этот метод может быть использован для выявления групп данных с общими характеристиками, таких как поведение клиентов или характеристики продуктов. Каждый полученный кластер может давать ценную информацию о связанных между собой данным или объектам.



Метод K-средних (K-means) является одним из самых распространенных методов кластеризации. Этот метод основан на итеративном повторении двух шагов, которые состоят в распределении объектов выборки по кластерам и пересчете центров кластеров.

Для начала работы алгоритма выбираются K случайных центров в пространстве признаков. Затем каждый объект выборки относится к кластеру, к центру которого объект оказался ближе. Далее производится пересчет центров кластеров путем вычисления среднего арифметического векторов признаков всех объектов, вошедших в данный кластер. Таким образом, получаем центр масс кластера. После обновления центров кластеров происходит повторное перераспределение объектов, а затем можно снова уточнить положение центров. Весь этот процесс повторяется до тех пор, пока центры кластеров не перестанут меняться.

Метод K-средних является итеративным алгоритмом, который позволяет найти наиболее оптимальные центры кластеров для данной выборки. Он широко применяется в различных областях, таких как машинное обучение, анализ данных и распознавание образов. Одно из его преимуществ заключается в том, что он прост в реализации и позволяет эффективно обрабатывать большие объемы данных.

Выбор начального приближения — это один из ключевых моментов при решении задачи кластеризации. Чтобы избежать попадания в область пространства признаков, где нет точек выборки, можно выбирать центры из некоторого случайного распределения. Однако такой подход может быть недостаточно эффективным.

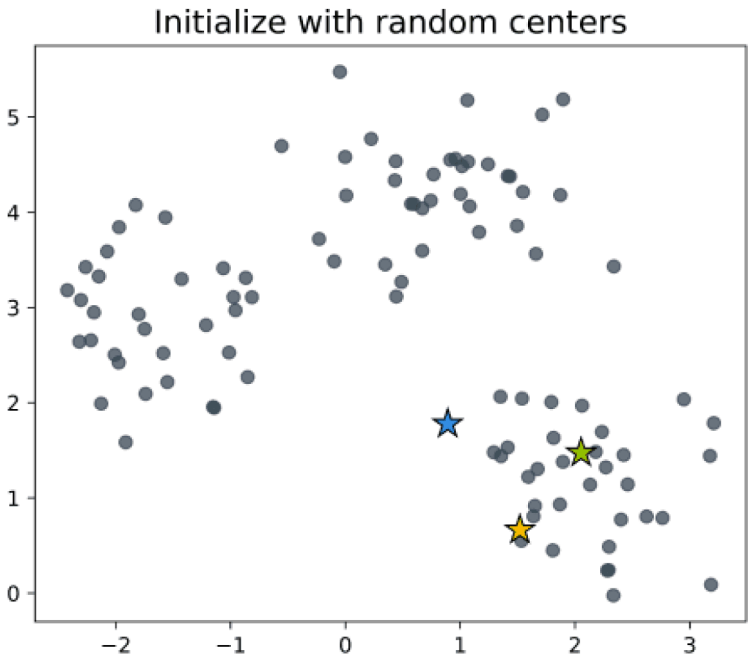
Для минимизации вероятности кучного размещения центров, необходимо учитывать их итоговое положение в кластерах. Например, если начальное

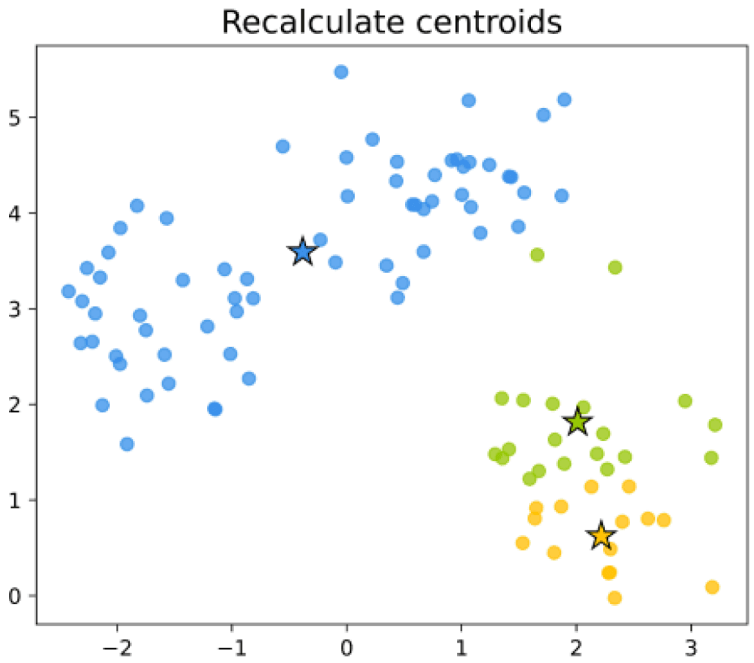
положение центров слишком далеко от финального положения, это может привести к неоптимальному результату.

Для решения этой проблемы можно использовать альтернативный подход. Например, можно выбрать в качестве начального положения центров какие-то из объектов выборки, которые наиболее репрезентативны для каждого кластера. Это позволит сократить вероятность кучного размещения центров и достичь более точных результатов.

Кроме того, для выбора начального приближения можно использовать методы оптимизации, которые учитывают как пространственные, так и структурные особенности данных. Например, можно использовать методы, основанные на графовых алгоритмах, с помощью которых можно учесть связи и зависимости между объектами выборки.

Таким образом, выбор начального приближения является важным шагом при решении задачи кластеризации. Кроме базового решения, существуют различные подходы, которые позволяют учесть возможные проблемы, связанные с выбором начального положения центров.





Метод борьбы с этим явлением заключается в стратегии выбора максимально удаленных друг от друга центров. Для этого на практике используется эвристика, называемая K-means++. Она включает в себя следующий алгоритм:

1. Вначале случайным образом выбирается первый центр из равномерного распределения на точках выборки.

2. Затем каждый следующий центр выбирается из случайного распределения на объектах выборки. При этом вероятность выбора объекта пропорциональна квадрату расстояния от него до ближайшего к нему центра кластера.

Таким образом, модифицированный K-means, использующий эту эвристику для выбора начальных приближений, позволяет более эффективно бороться с данной проблемой. Он позволяет получить оптимальное размещение центров в пространстве выборки, что повышает точность кластеризации и облегчает дальнейшую работу с данными. Это особенно важно при анализе больших объемов данных, где важно точно определить границы и структуру кластеров. Поэтому использование эвристики K-means++ является рекомендуемым подходом в задачах кластерного анализа.

Разберемся подробнее с выбором метрик для метода K-средних. Наш алгоритм состоит из двух шагов, которые повторяются до достижения сходимости. И здесь возникает вопрос, какие метрики использовать и как они влияют на каждый из этих шагов.

Первый шаг — отнесение объектов к ближайшим центрам кластеров, не зависит от выбора метрики. Он просто находит ближайший центр для каждой точки. Но второй шаг — пересчет центров кластеров — требует более тщательного рассмотрения.

При пересчете центров мы берем среднее арифметическое координат точек, входящих в каждый кластер. И здесь важно выбрать оптимальную метрику, которая будет работать наилучшим образом. Исследования показывают, что для получения оптимальных результатов в подсчете среднего арифметического наиболее эффективно использовать евклидову метрику.

Евклидова метрика является наиболее распространенной и широко используется в различных областях, включая кластерный анализ. Она измеряет расстояние между точками в пространстве и хорошо подходит для работы с числовыми данными. Однако стоит отметить, что в некоторых случаях выбор других метрик может быть более предпочтителен, в зависимости от специфики данных и требуемых результатов.

Таким образом, выбор метрик имеет важное значение для работы метода K-средних. Необходимо тщательно рассмотреть каждую метрику и ее влияние на два шага алгоритма, чтобы получить наилучшие результаты.

Важно отметить, что хотя на практике можно использовать метод без должного обоснования, это не гарантирует его успешной работы. Однако именно такая свобода позволяет нам экспериментировать с различными расстояниями, не ограничиваясь конкретными теоретическими предпосылками.

При анализе текстов наиболее распространенной альтернативой евклидовой метрике является косинусная мера близости векторов. Важно помнить, что эта мера является функцией близости, а не расстоянием. Таким образом, чем выше значение косинусной меры, тем ближе векторы друг к другу.

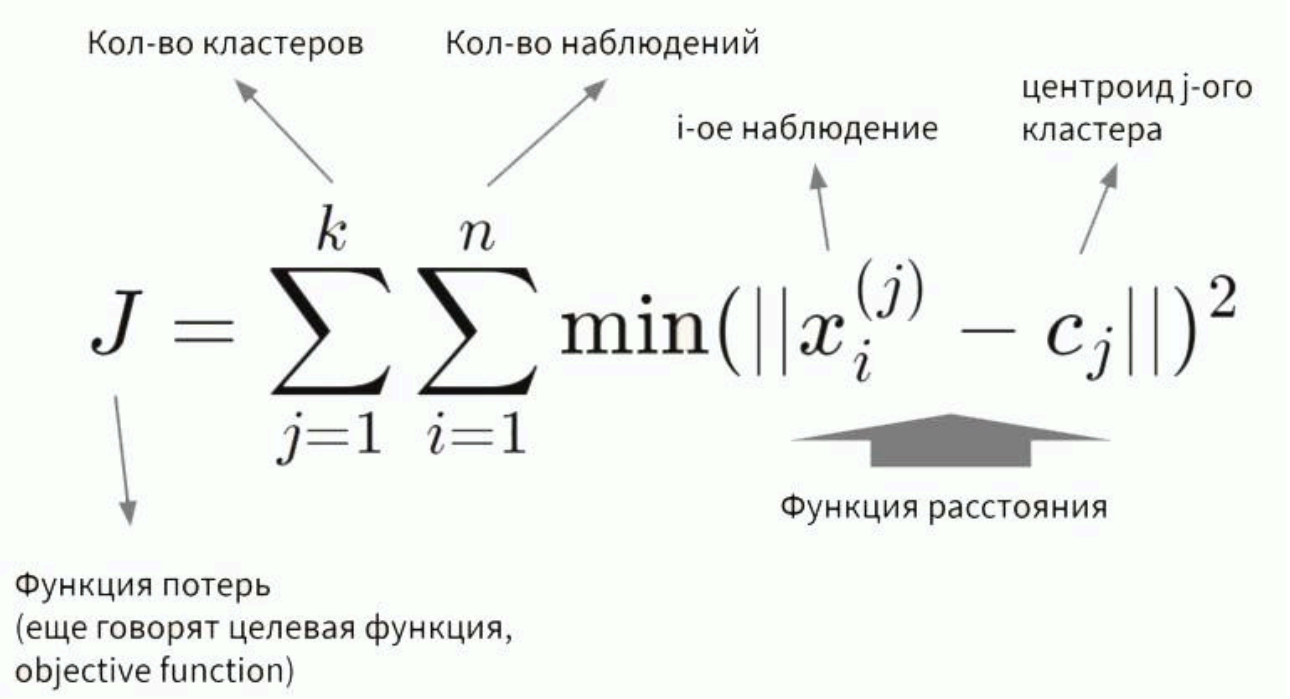
$$CosineSimilarity(\mu_k, x_i) = \frac{< \mu_k, x_i >}{|\mu_k|_2 \cdot |x_i|_2}$$

Одним из преимуществ использования косинусной меры является ее способность учитывать не только числовые значения, но и семантическую близость между векторами. Это особенно полезно при работе с текстовыми данными, где важно установить степень сходства между различными фрагментами текста.

В то же время, следует отметить, что косинусная мера также имеет свои ограничения. Например, она не всегда может точно оценить степень сходства между векторами, особенно если они имеют противоположные направления. Поэтому перед применением данной меры необходимо тщательно оценить ее применимость к конкретной задаче.

Центральная цель алгоритма заключается в минимизации суммы квадратов расстояний между точками внутри кластера до его центра. Это обозначается как within-cluster sum of squares (WCSS), которое является нашей функцией потерь. Однако важно заметить, что минимизация WCSS не является единственной целью алгоритма. Он также стремится максимизировать сумму квадратов расстояний

между центрами разных кластеров (between-cluster sum of squares, BCSS), чтобы обеспечить оптимальное разделение данных на кластеры.



Алгоритм построения кластеров имеет несколько этапов. Сначала происходит случайное выборка начальных центров кластеров. Затем каждая точка данных присваивается к ближайшему центру, исходя из минимального расстояния. После этого происходит пересчет центров кластеров на основе средних значений точек, принадлежащих кластеру. Эти шаги повторяются до тех пор, пока алгоритм не достигнет сходимости, то есть пока центры кластеров перестанут изменяться или изменения станут незначительными.

Таким образом, алгоритм кластеризации не только стремится минимизировать сумму квадратов внутрикластерных расстояний до центра кластера, но и учитывает максимизацию расстояний между центрами разных кластеров. Это позволяет достичь оптимального разделения данных на кластеры и повысить качество работы алгоритма.

Выбор количества кластеров в алгоритме кластеризации — это одна из важных задач, с которой сталкиваются исследователи и практики. Существует несколько подходов к решению этой проблемы.

Давайте рассмотрим первый способ — экспертный метод. Он основан на знании о предметной области, то есть эксперт может предположить, сколько кластеров может быть в данных. Например, при изучении поведения клиентов в интернет-магазине, эксперт может предложить, что существуют три типа потребителей: активные покупатели, случайные покупатели и новые клиенты. Экспертный метод особенно эффективен, когда есть достаточное количество знаний и опыта в данной области.

Другой способ — это метод локтя. Он основан на анализе внутрикластерных расстояний и суммы квадратов этих расстояний. Этот метод заключается в следующем: для разных вариантов количества кластеров мы обучаем модели алгоритмом кластеризации и измеряем сумму квадратов внутрикластерных расстояний. Затем мы строим график зависимости суммы квадратов от количества кластеров. Идея метода локтя заключается в том, чтобы найти точку на графике, где сумма квадратов перестает существенно уменьшаться. Эта точка называется "локтем", и она позволяет нам выбрать оптимальное количество кластеров.

Давайте рассмотрим пример. Представим, что мы анализируем данные о покупках в интернет-магазине. Мы обучаем модель кластеризации с разными вариантами количества кластеров, например, от 1 до 10. Затем мы измеряем сумму квадратов внутрикластерных расстояний для каждого варианта. Построив график зависимости суммы квадратов от количества кластеров, мы видим, что начиная с 3-го кластера, сумма квадратов перестает значительно уменьшаться. Это означает, что оптимальное количество кластеров в данном случае - 3. Этот метод позволяет нам объективно выбрать количество кластеров без предварительных знаний о предметной области.

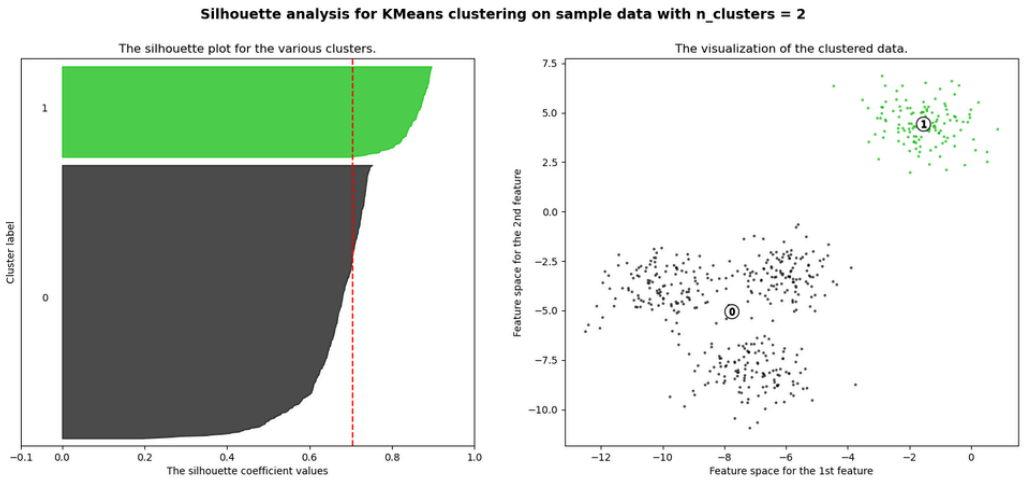
Таким образом, выбор количества кластеров может быть осуществлен как с помощью экспертного метода, основанного на знаниях о предметной области, так и с использованием метода локтя, который анализирует структуру данных. Каждый из этих способов имеет свои преимущества и недостатки, и выбор должен быть основан на конкретной задаче и условиях исследования.

Оптимальным значением в данном случае является три кластера. Это можно наблюдать, поскольку после достижения количества кластеров трех, сумма квадратов внутрикластерных расстояний перестает значительно уменьшаться. Однако стоит отметить, что определение оптимального значения количества кластеров не всегда так просто. В реальных задачах, иногда может быть необходимо проводить более детальный анализ, учитывая различные факторы и особенности данных. В таких случаях, использование дополнительных методов, например, оценки силуэта, может помочь в выборе оптимального числа кластеров. В конечном итоге выбор оптимального значения количества кластеров важен для достижения более точного и интерпретируемого результатов кластерного анализа.

Метод силуэта — это метод оценки качества кластеризации данных, позволяющий выбрать оптимальное число кластеров K.

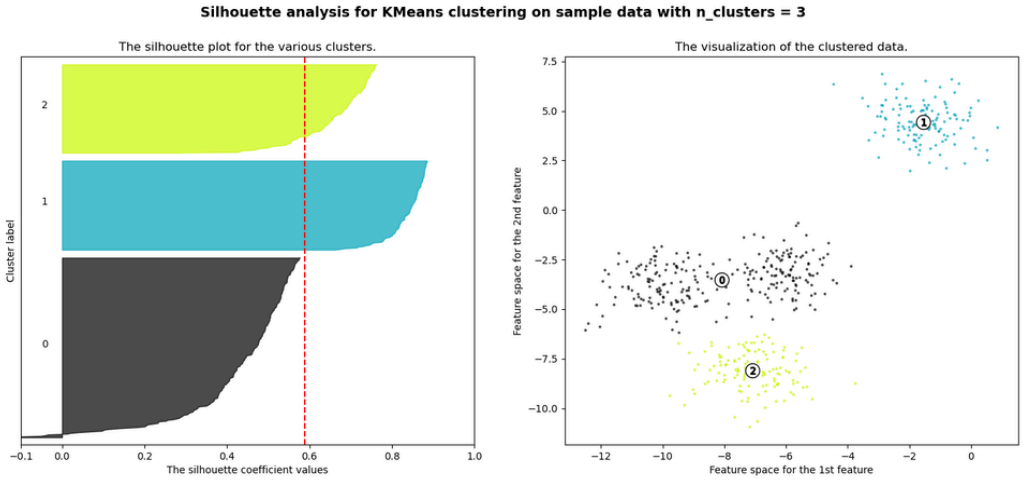
В его расчете задействованы среднее внутрикластерное расстояние (a), а также среднее расстояние до ближайшего кластера (b) для каждого образца. После этого, силуэт вычисляется как (b - a) / max(a, b). Однако что означает b? Это расстояние между точкой a и ближайшим кластером, в который она не входит. Очевидно, что мы можем вычислить среднее значение силуэта по всем образцам и использовать его в качестве метрики для оценки количества кластеров.

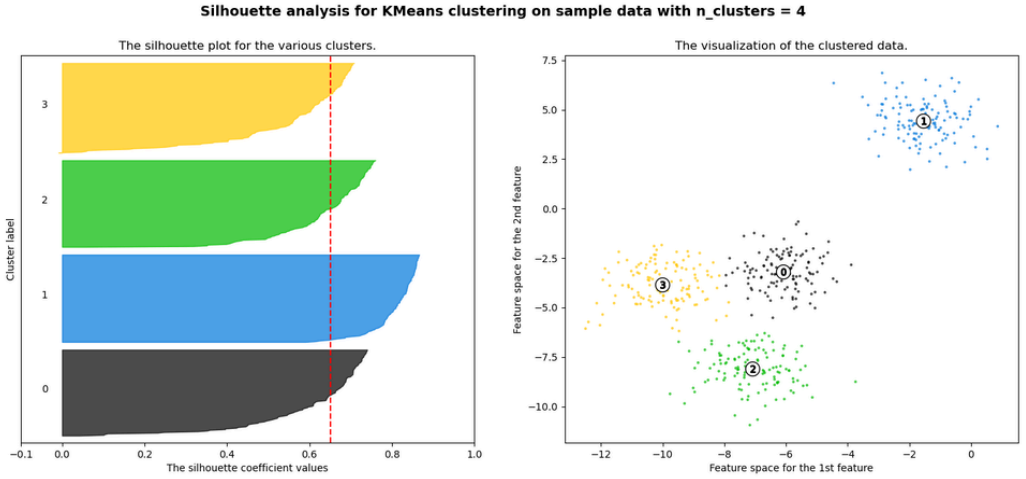
На практике, метрика "силуэта" представляет собой цифру от -1 до 1, где значение ближе к 1 указывает на хорошую кластеризацию, а значение ближе к -1 указывает на плохую кластеризацию. Однако следует отметить, что использование только этой метрики может быть недостаточно для полной оценки кластеризации. Важно всегда учитывать дополнительные факторы, такие как контекст задачи и предпочтения исследователя.

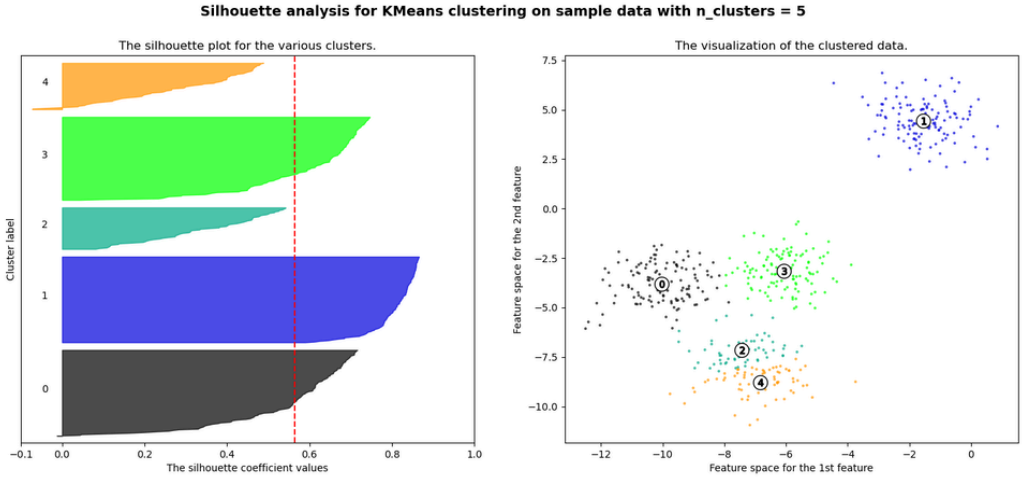


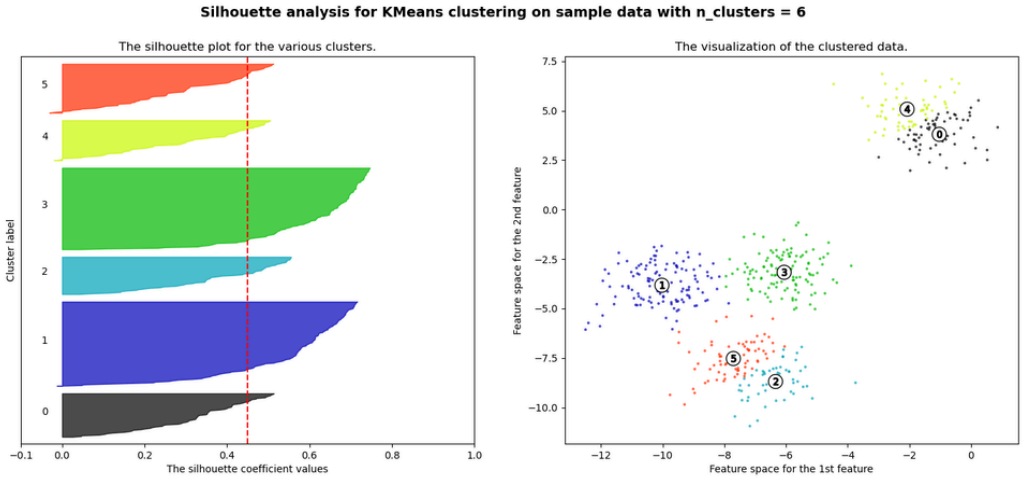
Оценки силуэта (как они называются) около +1 указывают на дальнее расположение образца от соседних кластеров. Значение, близкое к нулю, означает, что выборка находится на границе решения между двумя соседними кластерами или очень близко к этой границе, а отрицательные значения указывают на неправильное назначение выборки кластеру.

Анализ силуэта в этом примере используется для выбора наилучшего значения для числа кластеров (n_clusters). Графики ниже показывают, что значения n_clusters 3, 5 и 6 являются неоптимальным выбором для этих данных из-за наличия кластеров с оценками силуэта ниже среднего и значительных колебаний в размере участков силуэта. Анализ силуэта становится более неоднозначным при выборе между 2 и 4.









Размер кластера можно визуализировать по толщине силуэта. Например, график силуэта для кластера 0, при n_clusters равном 2, имеет больший размер из-за объединения 3 субкластеров в один большой кластер. Однако при n_clusters равном 4, все графики имеют примерно одинаковую толщину и, следовательно, имеют схожие размеры, что также отражено на помеченном графике разброса справа.

В заключение, коэффициент "силуэт" является более подходящей метрикой для оценки кластеризации, поскольку он учитывает как внутрикластерные, так и межкластерные расстояния. Однако его использование следует комбинировать с другими аналитическими методами для более полной и точной оценки результатов кластеризации.

Применение кластеризации на основе центроидов:

В медицине кластеризация на основе центроидов может быть использована для классификации пациентов по группам сходства при диагнозировании заболеваний. Например, исследователи могут использовать этот метод для выявления подгрупп пациентов с разным типом рака или других заболеваний, что поможет улучшить эффективность лечения и предложить индивидуализированный подход.

В бизнесе кластерный анализ на основе центроидов может быть полезным инструментом для сегментации клиентской базы данных. Это позволяет компаниям определить группы клиентов с общими характеристиками и потребностями. Такая информация может быть использована для создания персонализированных маркетинговых стратегий, улучшения уровня обслуживания и повышения удержания клиентов.

В социологии кластерный анализ на основе центроидов может помочь в изучении социальных групп и сетей. Исследователи могут использовать этот метод для определения подгрупп людей с общими интересами, поведением или демографическими характеристиками. Такие результаты позволяют лучше понять структуру общества и взаимодействия между его членами.

В экологии кластеризация на основе центроидов может быть полезной техникой для классификации видов и определения экосистем. Ученые могут использовать

этот метод для выявления групп организмов, которые имеют схожие характеристики и требуют похожих условий для выживания. Это позволяет лучше разбираться в сложности природных систем и принимать эффективные решения по их сохранению.

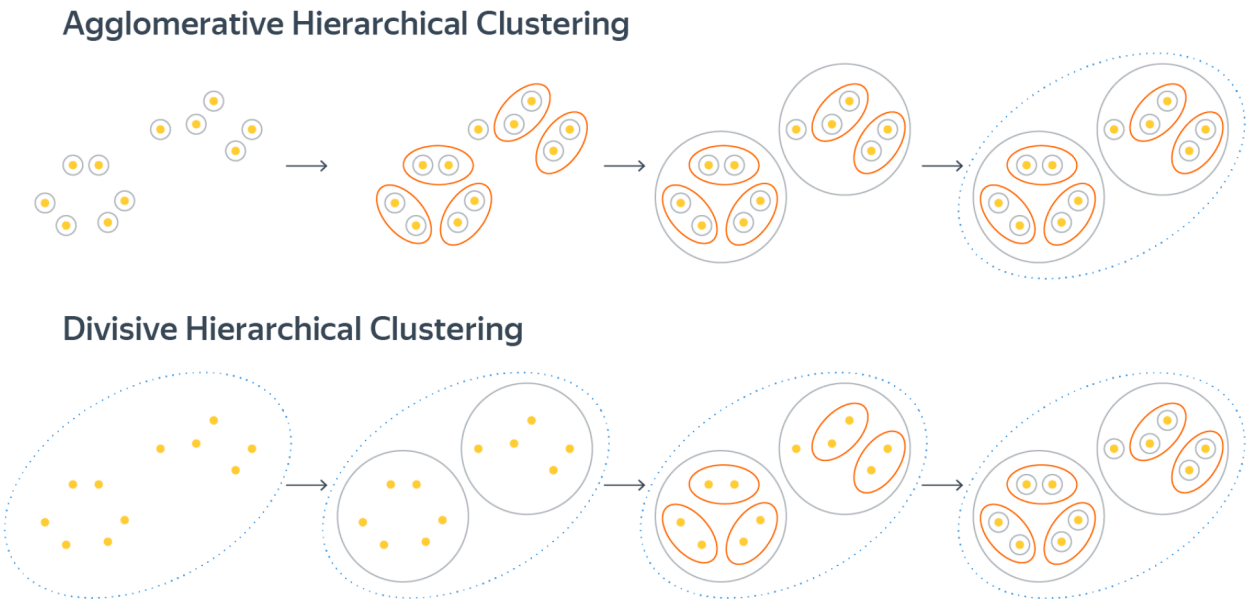
Иерархическая кластеризация

Иерархическая кластеризация, также известная как иерархическая агломеративная кластеризация, является ещё одним классическим методом кластеризации. Однако необходимо отметить, что название этого метода указывает на два важных момента.

Во-первых, алгоритмы кластеризации можно разделить на агломеративные и дивизивные. Агломеративные алгоритмы начинают с формирования небольших кластеров, которые обычно состоят из одного объекта, и постепенно объединяют их в крупные кластеры. С другой стороны, дивизивные алгоритмы начинают с одного большого кластера и постепенно разбивают его на более мелкие кластеры.

Во-вторых, иерархическая кластеризация подразумевает создание иерархии кластеров. Это значит, что кластеры могут быть организованы в виде дерева или иерархической структуры, где каждый узел соответствует кластеру. Такая иерархия позволяет анализировать данные на нескольких уровнях детализации и исследовать взаимосвязи между кластерами.

Иерархическая кластеризация является гибким методом, который может быть применен к различным типам данных. Он позволяет исследователям получать подробную информацию о структуре данных, а также проводить исследования на разных уровнях гранулярности. Кроме того, этот метод имеет широкий спектр применений, от финансового анализа до биологических исследований.



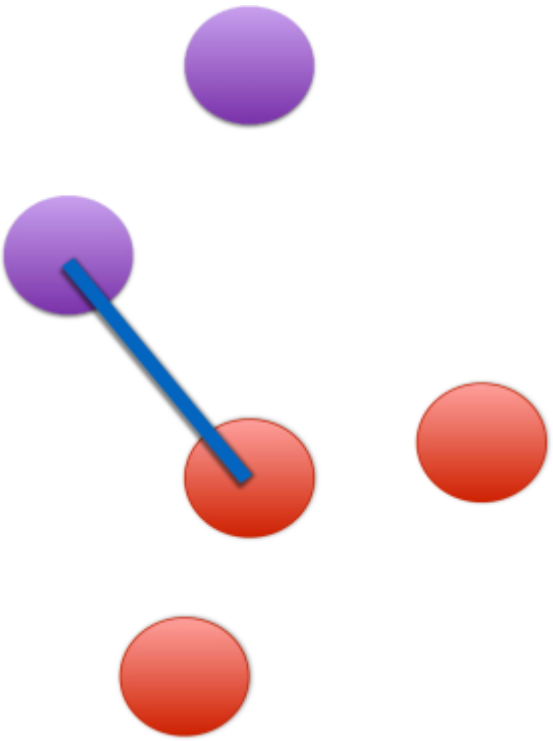
Во-первых, важно отметить, что кластеризация может быть реализована различными способами, включая плоскую и иерархическую структуры. Плоская кластеризация предполагает, что все кластеры равноправны и находятся на одном уровне. Однако иерархическая кластеризация предоставляет больше возможностей и гибкости.

В случае иерархической агломеративной кластеризации мы начинаем с кластеров, состоящих только из одного объекта, и постепенно объединяем их. Важно отметить, что последовательность этих объединений определяет структуру вложенности кластеров. Даже если нашей конечной целью является использование кластеров с одного уровня без углубления во вложенность, кластеризацию все равно называют иерархической, так как иерархия возникает естественным образом в процессе работы алгоритма. Такая иерархия может помочь нам лучше понять внутреннюю структуру данных и отношения между кластерами.

Начнем с того, что алгоритм иерархической кластеризации, несмотря на свою простоту, имеет множество вариаций и альтернатив. Например, можно начать с создания одного кластера, содержащего все объекты выборки, и затем последовательно разбивать его на более мелкие кластеры. Кроме того, кластеризация может осуществляться не только на основе евклидового расстояния, но и на основе других мер близости, таких как косинусное расстояние или корреляция.

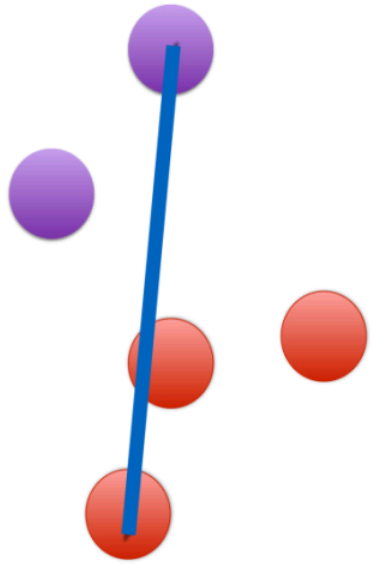
Объединение кластеров — важный этап в использовании иерархических алгоритмов. Но как именно происходит объединение и как вычисляются "расстояния" между кластерами? В этом вопросе приходится столкнуться с несколькими метриками, которые помогают определить наилучший способ объединения.

Одна из таких метрик — одиночная связь, которая использует расстояние между наиболее близкими объектами в разных кластерах. Это означает, что при объединении двух кластеров берется во внимание расстояние между их ближайшими соседями. Результатом такого объединения являются цепочки кластеров, которые имеют тенденцию объединяться в более крупные группы.



$$R^{б}(W, S) = \min_{w,s} \rho(w, s)$$

Еще одна метрика — полная связь, которая учитывает расстояние между наиболее удаленными соседями в разных кластерах. В этом случае при объединении кластеров рассматривается расстояние между самыми удаленными объектами. Этот подход может приводить к формированию более разрозненных кластеров, которые необязательно объединяются в цепочки.



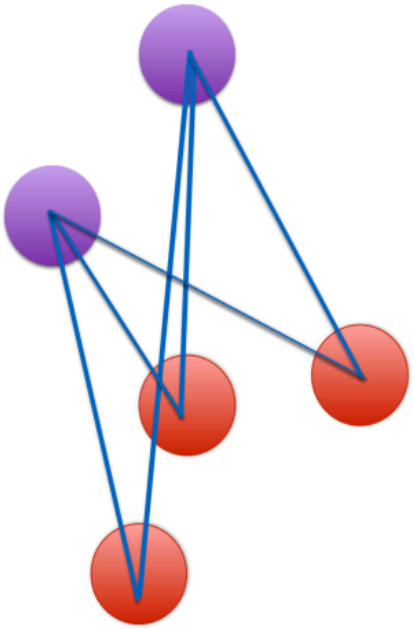
$$R^{Д}(W,S) = \max_{w,s} \rho(w,s)$$

Таким образом, выбор метрики для объединения кластеров важен и зависит от конкретной задачи и данных. Каждая метрика имеет свои особенности и может приводить к различным результатам. Поэтому важно тщательно анализировать данные и выбирать подходящую метрику для достижения наилучших результатов.

В методе невзвешенного попарного среднего расстояние между кластерами определяется как среднее расстояние между всеми парами объектов в них. Этот метод является эффективным, особенно когда объекты формируют различные группы. Однако он также работает хорошо в случаях, когда кластеры имеют удлиненную форму или представляют собой "цепочечный" тип.

Примерами таких кластеров могут быть группы генов, которые образуют связанные цепочки, или различные социальные группы, которые могут быть расположены на определенном пространственном расстоянии друг от друга.

Использование метода невзвешенного попарного среднего позволяет учесть все возможные комбинации объектов в кластерах, что может быть полезным при анализе данных. Этот метод также позволяет выявить наиболее характерные свойства и особенности каждого кластера.

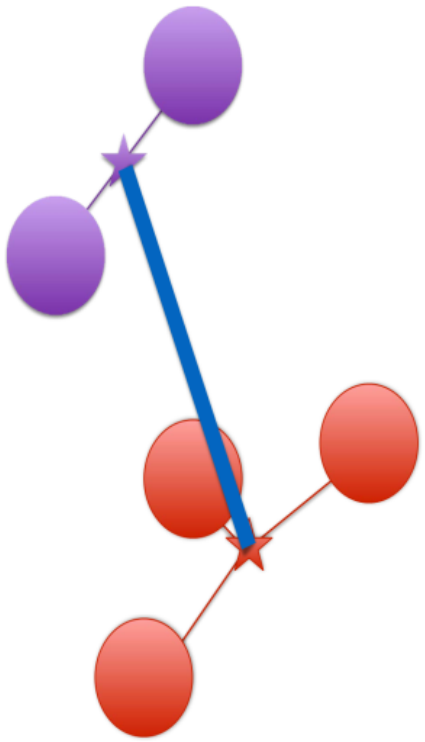


$$R^{\Gamma}(W,S) = \frac{1}{|W| * |S|} \sum_{w} \sum_{s} \rho(w,s)$$

Однако следует отметить, что этот метод может быть непригоден в случаях, когда объекты не образуют четко выраженных групп или когда кластеры имеют сложную форму. В таких случаях, возможно, будет более эффективно использовать другие методы для анализа и классификации данных.

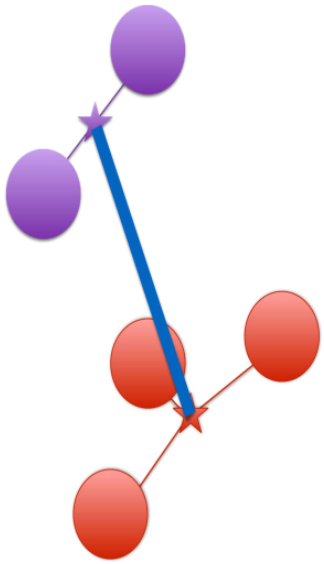
Взвешенное попарное среднее - это метод, который отличается от метода невзвешенного попарного среднего тем, что при вычислениях используется весовой коэффициент, равный размеру соответствующих кластеров. Этот метод рекомендуется использовать, когда предполагаются неравные размеры кластеров.

Невзвешенный центроидный метод определяет расстояние между двумя кластерами как расстояние между их центрами тяжести. Этот метод не учитывает размеры кластеров и рассматривает их центры как единое целое.



$$R^{Ц}(W,S) = \rho^2\left(\sum_w \frac{w}{|W|}, \sum_s \frac{s}{|S|}\right)$$

Взвешенный центроидный метод (медиана) (Расстояние Уорда) - это метод, который также учитывает размеры кластеров, но в отличие от взвешенного попарного среднего, определяет расстояние между кластерами как расстояние между их центроидами или медианами. Этот метод более точно учитывает геометрические особенности кластеров и может быть полезен при анализе данных с неравными размерами кластеров.



$$R^{y}(W,S) = \frac{|W| * |S|}{|W| + |S|}\rho^{2}\left(\sum_{w}\frac{w}{|W|}, \sum_{s}\frac{s}{|S|}\right)$$

При выборе метода объединения иерархической кластеризации (например, одиночная связь, полная связь, метод Уорда и т. д.) не существует универсального ответа, какой метод является лучшим, так как он зависит от особенностей данных и целей анализа. Поэтому важно понимать особенности каждого метода и выбирать соответствующий подход.

Ниже приведены некоторые аспекты и обоснования для выбора различных методов объединения:

1. Одиночная связь (single-linkage clustering): Этот метод вычисляет расстояние между двумя кластерами как минимальное расстояние между их элементами. Он полезен, когда кластеры имеют длинные вытянутые формы, так как он может обнаружить близкие элементы, находящиеся на разных частях кластера. Однако этот метод может быть чувствителен к шуму и выбросам, поэтому он может приводить к погрешностям при обработке таких данных.

2. Полная связь (complete-linkage clustering): Этот метод вычисляет расстояние между двумя кластерами как максимальное расстояние между их элементами. Он полезен, когда кластеры имеют компактную структуру, так как он учитывает удаленность наиболее удаленных элементов. Этот метод также более устойчив к шуму и выбросам, чем одиночная связь. Однако он также может создавать кластеры с более широкими границами.

3. Метод Уорда (Ward's method): Этот метод минимизирует внутрикластерную сумму квадратов отклонений (within-cluster sum of squares), что позволяет строить компактные и однородные кластеры. Метод Уорда обеспечивает низкое внутрикластерное отклонение в полученных кластерах. Он особенно полезен, когда важно обнаружить компактные кластеры с небольшой внутрикластерной дисперсией. Однако этот метод может быть чувствителен к шуму и выбросам.

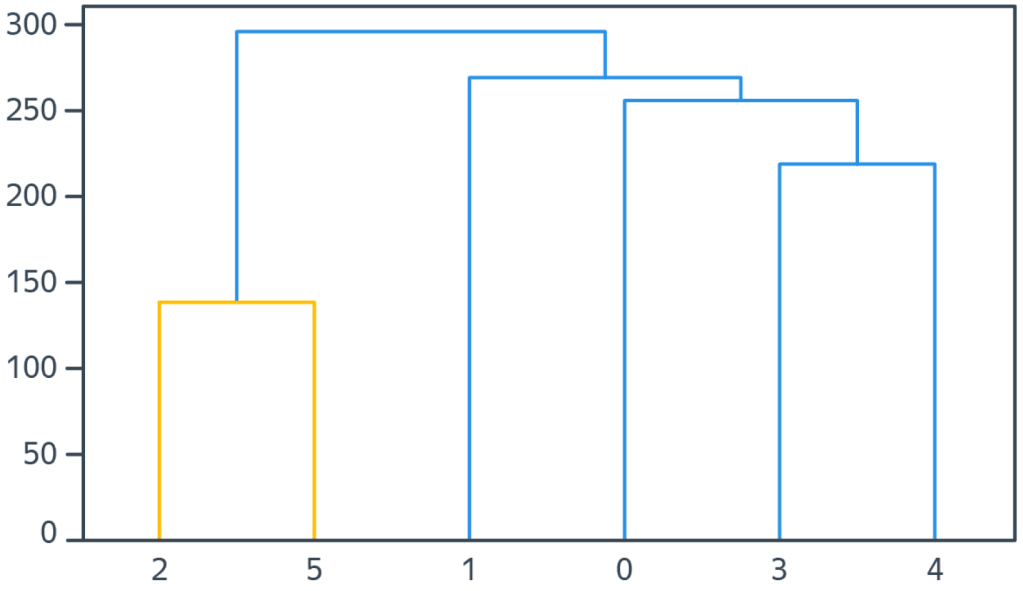
Важно провести эксперименты и проверить различные методы объединения на ваших данных, чтобы выбрать оптимальный подход к иерархической кластеризации. Также стоит отметить, что в некоторых случаях комбинированный подход с использованием нескольких методов может привести к лучшим результатам.

Размеры кластеров играют важную роль в анализе данных. Предлагается новый метод, который учитывает разницу в размерах кластеров, используя специальные веса при вычислениях. Этот метод позволяет более точно оценить влияние размеров кластеров на результаты анализа. Если имеются существенные различия в размерах кластеров, то этот метод будет предпочтительнее стандартного подхода.

Таким образом, новый метод поможет улучшить точность и качество анализа данных в случае, когда размеры кластеров имеют существенное значение.

Важным элементом алгоритма является выбор подходящего критерия останова. Это может быть достижение определенного числа кластеров или достижение определенного значения расстояния между кластерами. Кроме того, можно использовать статистические методы для определения оптимального числа кластеров.

Одним из преимуществ иерархической кластеризации является возможность визуализации результатов в виде дендрограммы. Дендрограмма представляет собой дерево, в котором объекты выборки располагаются на листьях, а каждый уровень дерева соответствует определенному числу кластеров. Это позволяет наглядно оценить степень схожести объектов и определить подходящее число кластеров.



Дендрограмма — это схема, которая отображает процесс объединения кластеров в алгоритме кластеризации. Каждой итерации алгоритма соответствует пара кластеров, которые объединяются, а также расстояние между ними в момент слияния. Важно отметить, что с ростом итерации расстояния между кластерами только увеличиваются.

Дендрограмма представляет собой графическое отображение этого процесса. На горизонтальной оси расположены объекты из кластеризуемой выборки, а под ней указаны номера этих объектов. Расположение объектов вдоль оси выбирается в соответствии с эстетическими соображениями, чтобы избежать пересечений дуг на дендрограмме.

Каждая дуга на дендрограмме представляет собой объединение двух кластеров на определенной итерации алгоритма. Чем выше на дендрограмме находится дуга,

тем больше расстояние между объединяемыми кластерами. Таким образом, дендрограмма позволяет наглядно представить процесс формирования кластеров и их иерархическую структуру.

Одним из способов визуализации результатов кластеризации является построение дендрограммы. В дендрограмме отображаются расстояния между кластерами на вертикальной оси, а каждая дуга представляет объединение двух кластеров. Однако, вместо указания конкретных объектов выборки, дуга указывает на дугу другого кластера, что делает визуализацию более наглядной.

Одним из интересных аспектов дендрограммы является возможность наблюдения за скачком расстояний между кластерами. Это может дать нам подсказку о "естественном" количестве кластеров в нашей задаче. Однако в практических задачах часто нет однозначного и "естественного" количества кластеров, так как кластеризация может быть выполнена на разных уровнях детализации и варьироваться в зависимости от целей и требований.

Таким образом, дендрограмма является полезным инструментом для визуализации процесса кластеризации, позволяя нам увидеть связь между объединением кластеров и изменением расстояний. Однако она не всегда может однозначно указать на оптимальное количество кластеров, оставляя это решение на усмотрение исследователя.

В заключение можно сказать, что иерархическая кластеризация является мощным инструментом в анализе данных, который позволяет группировать объекты по их схожести и выявлять скрытые закономерности. Однако необходимо учитывать, что алгоритм может быть требователен к вычислительным ресурсам и может потребовать оптимизации для работы с большими данными.

Кластеризация на основе плотности

В отличие от традиционных методов кластеризации, таких как иерархическая или k-средних, где каждый объект принадлежит только одному кластеру, кластеризация на основе плотности позволяет определять несколько типов структур в данных.

Основная идея метода заключается в том, чтобы определить области высокой плотности данных и использовать их для определения кластеров. Другими словами, объекты будут считаться членами одного кластера, если они находятся достаточно близко друг к другу и имеют достаточное количество соседей в заданном радиусе.

Давайте попробуем разобрать принцип работы алгоритма на основе интуиции:

Мы находимся в огромном зале, где происходит празднование дня рождения кого-то. Люди перемещаются по залу, некоторые в одиночку, но большинство в группах.

Наша цель — разделить всех людей в зале на разные группы.

Но как нам сделать это так, чтобы учесть разнообразную форму групп и не забыть о тех, кто находится в одиночестве? Мы решили использовать оценку плотности толпы вокруг каждого человека. Возможно, если плотность между двумя людьми превышает определенный порог, то можно предположить, что они принадлежат к одной группе. В конце концов, странно было бы, если люди, которые в данный момент общаются, были бы отнесены к разным группам, даже если плотность людей меняется в некоторых пределах.

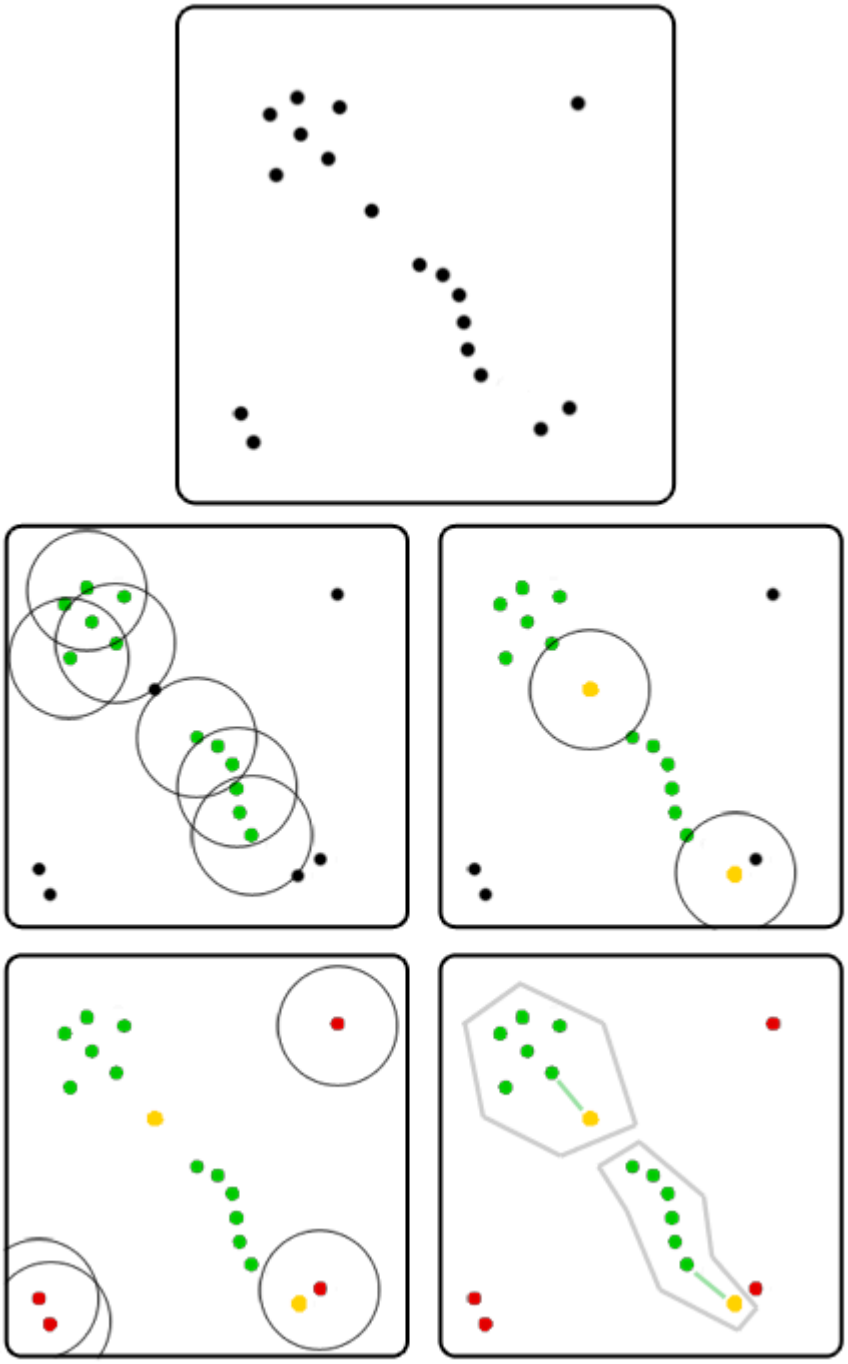
Давайте определим, что рядом с кем-то собралась толпа, если рядом с ним находится несколько других людей. Для этого нам нужно задать два параметра. Как мы определим параметр "близко"? Давайте примем некоторое интуитивно понятное расстояние. Если люди находятся на расстоянии около метра друг от друга, то они считаются близко расположенными. Теперь спросим, сколько таких людей считать "несколько других"? Пусть будет три человека.

Теперь давайте каждый человек подсчитает, сколько людей стоят в радиусе метра от него. Все, у кого есть хотя бы три соседа, возьмут в руки зеленые флажки. Таким образом, они станут основными элементами, именно они будут формировать группы.

Мы обратимся к тем, у кого меньше трех соседей. Мы выберем тех, у кого хотя бы один из соседей держит зеленый флаг, и мы вручим им желтые флаги. Мы скажем, что они находятся на границе групп.

Затем мы обратим внимание на одиночек, у которых не только нет трех соседей, но также ни один из них не держит зеленый флаг. Мы раздадим им красные флаги. Мы будем считать, что они не принадлежат ни одной группе.

Таким образом, если существует цепочка "зеленых" людей от одного человека до другого, то эти два человека принадлежат к одной группе. Очевидно, что все подобные группы разделены либо пустым пространством, либо людьми с желтыми флагами. Мы можем их пронумеровать: каждый человек в группе номер 1 может дотянуться рукой до каждого другого в той же группе номер 1, но не до кого в группе номер 2, номер 3 и так далее. То же самое относится к остальным группам.

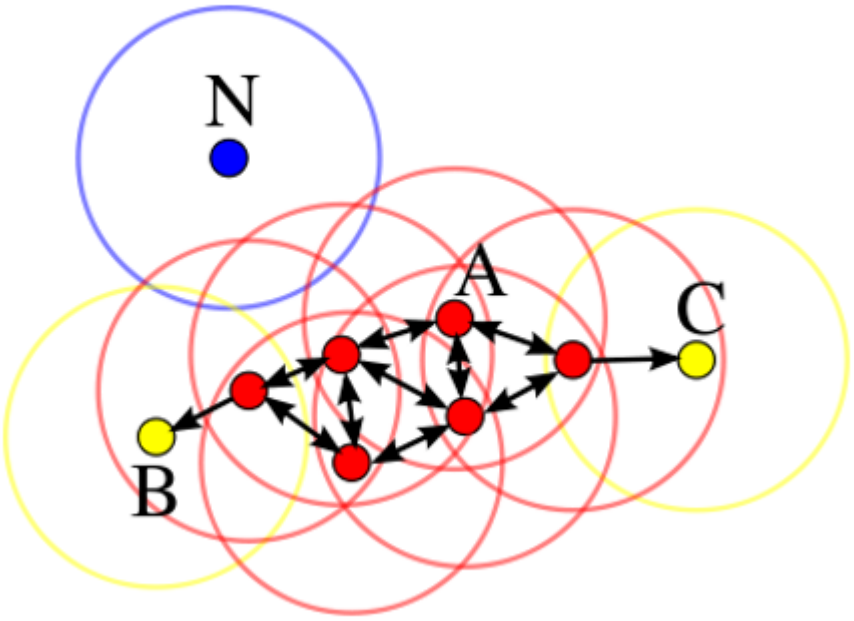


Выбор принадлежности человека к определенной группе, основываясь на наличии желтого флажка и наличии только одного зеленого соседа, можно осуществить. Однако, если таких соседей несколько и они принадлежат разным группам, то придется принимать решение. Для этого можно использовать различные методы, например, рассмотреть ближайшего соседа. Возможно придется учитывать краевые случаи, но этого не так страшно.

Методы кластеризации на основе плотности, разработанные для выявления кластеров произвольной формы, представляют собой алгоритмы кластеризации, основанные на плотности точек данных в пространстве признаков. Их принцип работы отличается от алгоритмов, основанных на расстоянии, которые опираются на меры близости, в то время как методы на основе плотности учитывают плотность

точек данных в окрестности для определения принадлежности кластерам. Эти методы особенно полезны при обработке сложных форм кластеров, различных уровней плотности и наличии шума в данных.

Один из наиболее популярных алгоритмов данной группы - алгоритм DBSCAN (Density-based spatial clustering of applications with noise - пространственная кластеризация приложений с шумом на основе плотности) определяет кластеры путем анализа плотности точек. Области с высокой плотностью указывают на наличие кластеров, в то время как области с низкой плотностью указывают на кластеры шума или выбросов. DBSCAN может эффективно работать с большими наборами данных, включая шум, и может идентифицировать кластеры различных размеров и форм. Основная концепция алгоритма DBSCAN заключается в том, что для каждой точки кластера необходимо, чтобы окружность определенного радиуса содержала как минимум определенное количество точек, то есть плотность в окружности должна превышать заданный порог.



Окрестность каждой точки данных расширяется до минимального количества точек, чтобы найти плотные области данных, разделенные более разреженными областями. Таким образом, DBSCAN классифицирует точки данных на основные, граничные и шумовые. Основные точки - это достаточно плотные регионы, граничные точки находятся рядом с основными, но имеют меньшую плотность, а шумовые точки являются изолированными и имеют недостаточную плотность. DBSCAN не требует предварительного указания количества кластеров и может работать с кластерами произвольной формы. Алгоритм использует три входных параметра.

DBSCAN - это алгоритм, который определяет группы объектов на основе параметров eps и min_samples. Параметр eps задает радиус окрестности точки, в которой ищутся соседние точки, а min_samples - минимальное количество точек, необходимых для образования кластера. DBSCAN использует различные метрики, такие как евклидова или манхэттенская, для измерения расстояния между точками.

Резюмируем работу алгоритма:

Процесс разделения точек данных в анализе кластеров включает их классификацию на разные категории:

1. Основные точки: Идентифицируются по наличию более N объектов в их радиусе.

2. Граничные точки: Хотя рядом есть основные точки, общее число соседей меньше N.

3. Шумовые точки: Не имеют основных точек поблизости и окружены менее чем N объектами.

В процедуре кластеризации наблюдается следующий этап: точки, классифицированные как шумовые, исключаются из дальнейшего анализа и не присоединяются к каким-либо кластерам.

Кластеризация данных с помощью метода DBSCAN вносит удобство благодаря его способности самостоятельно определить количество кластеров, с учётом некоторых настроек, таких как e и N. Процесс разделения данных на кластеры проходит в несколько шагов:

1. Определение рёбер: Точки с общими пространственными свойствами связываются вместе в графе.

2. Создание связностей: В сформированном графе выявляются отдельные связные группы.

3. Присвоение кластеров: Периферийные элементы объединяются в кластеры на основе близости к центральным точкам.

Примечательная особенность DBSCAN заключается в его гибкости: алгоритм не только справляется с обнаружением кластеров различных, в том числе и сложноструктурированных, форм, например напоминающих ленты или расположенных концентрически.

Одним из преимуществ DBSCAN является то, что не требуется заранее задавать количество кластеров. Также он хорошо отделяет кластеры высокой плотности от кластеров низкой плотности.

Однако у DBSCAN есть некоторые недостатки. Например, он не может корректно работать с данными разной плотности, так как параметры eps и min_samples нельзя выбирать индивидуально для каждого кластера.

Кластеризация на основе плотности представляет собой сложную задачу, требующую настройки параметров eps и min_samples. Ниже приведены некоторые популярные алгоритмы кластеризации на основе плотности, которые помогают в этом:

- OPTICS (Ordering Points To Identify the Clustering Structure):

Для каждой точки данных он вычисляет значение "достижимости" на основе расстояния до ближайшей соседней точки. Точки сортируются в определенном порядке на основе их достижимости. Этот порядок отражает кластерную структуру данных. OPTICS строит диаграмму достижимости, показывающую изменения плотности. Кластеры могут быть извлечены из этой диаграммы путем выбора соответствующего порогового значения. OPTICS устраняет необходимость предварительного задания радиуса поиска, как в DBSCAN. Он устойчив к шуму, может обрабатывать данные с разной плотностью кластеров. Недостатком является высокая вычислительная сложность.

- DENCLUE (DENsity CLUstEring кластеризация на основе плотности)

DENCLUE основан на модели плотности точки данных как функции влияния. Для каждой точки данных вычисляется функция влияния на основе её расстояния до соседних точек. Плотность в точке оценивается путем суммирования функций влияния соседних точек в этой точке. Точки с локально максимальной плотностью выбираются в качестве центров кластеров. Другие точки назначаются кластерам на основе максимального значения их плотности. DENCLUE не требует предварительного задания числа кластеров. Он может обнаруживать кластеры произвольной формы и размера. Недостатком является чувствительность к параметрам и высокая вычислительная сложность. DENCLUE эффективен при анализе данных со сложными нерегулярными кластерными структурами.

- HDBSCAN (Hierarchical Density-Based Spatial Clustering of Applications with Noise пространственная кластеризация приложений с шумом на основе иерархической плотности)

Является улучшенной версией DBSCAN, использует иерархический подход. Строит иерархию кластеров с использованием минимального расстояния между точками для создания дерева кластеров. Может найти кластеры с разной плотностью, устраняя параметр "радиус окрестности" из DBSCAN. Автоматически определяет оптимальное количество кластеров, анализируя иерархию. Робастен к шуму и выбросам, может их идентифицировать. Имеет лучшую вычислительную сложность по сравнению с OPTICS. Требует только один параметр - минимальный размер кластера. Эффективен на данных со сложными иерархическими кластерными структурами. Используется в различных областях, таких как биоинформатика, астрономия, машинное обучение.

- ST-DBSCAN (Spatial-Temporal Density-Based Clustering Кластеризация приложений на основе пространственно-временной плотности с шумом)

Это расширение DBSCAN для кластеризации пространственно-временных данных. Учитывает как пространственную, так и временную близость точек данных при определении плотных областей. Вводит понятия пространственного и временного радиусов окрестности. Точки считаются соседними, если они находятся близко как в пространстве, так и во времени. Позволяет обнаруживать кластеры произвольной формы и плотности в пространственно-временных данных. Может идентифицировать шумовые точки. Не требует предварительного задания количества кластеров. Параметры радиусов окрестности могут быть определены автоматически. Применяется в анализе движущихся объектов, временных рядов, видео и других пространственно-временных данных.

- SUBCLU (Density-Connected Subspace Clustering - кластеризация подпространств)

SUBCLU - это метод кластеризации подпространств на основе плотности для данных высокой размерности. Он разбивает пространство признаков на подпространства и ищет плотные области (кластеры) в каждом подпространстве. Для поиска плотных областей использует алгоритм DBSCAN. Подпространства определяются как максимальные множества коррелированных признаков. Разные кластеры могут иметь различные наборы релевантных признаков. Позволяет находить кластеры произвольной формы в данных с шумом и выбросами. Не требует указания числа кластеров заранее. Масштабируется на данные очень высокой размерности. Применяется в задачах анализа изображений, текстов, геномных данных.

- VDBSCAN (Varied Density-Based Spatial Clustering of Applications with Noise пространственная кластеризация приложений с шумом на основе вариационной плотности)

Является модификацией DBSCAN, которая позволяет находить кластеры с разной плотностью. Использует переменный радиус окрестности для каждой точки в зависимости от локальной плотности данных. Точки с высокой локальной плотностью получают больший радиус окрестности. Позволяет обнаруживать кластеры произвольной формы и плотности. Может идентифицировать шумовые точки и выбросы. Не требует предварительного задания числа кластеров. Имеет лучшую производительность, чем DBSCAN на данных с переменной плотностью. Недостатком является сложность определения порога локальной плотности. Применяется в задачах кластеризации пространственных, изображений, сетевых и других данных. Обеспечивает более точное выявление кластерной структуры по сравнению с DBSCAN.

Оценка качества кластеризации

Основные метрики качества кластеризации будут представлены в списке ниже. Также мы обсудим особенности использования данных метрик.

Среднее внутрикластерное расстояние (average intra-cluster distance) является одной из оценок качества кластеризации. Эта метрика позволяет измерить, насколько близко находятся объекты внутри каждого кластера.

Для вычисления среднего внутрикластерного расстояния сначала необходимо определить расстояние между всеми парами объектов внутри каждого кластера. Затем суммируются все эти расстояния и делятся на общее количество пар объектов. Таким образом, получается среднее значение расстояния внутри кластера.

$$F_0 = \frac{\sum_{i=1}^{n}\sum_{j=i}^{n}\rho(x_i, x_j)\mathbb{I}[a(x_i) = a(x_j)]}{\sum_{i=1}^{n}\sum_{j=i}^{n}\mathbb{I}[a(x_i) = a(x_j)]}$$

Чем меньше среднее внутрикластерное расстояние, тем более компактным и однородным является кластер. Идеальное значение среднего внутрикластерного расстояния будет равно 0, что означает, что все объекты внутри кластера находятся в одной точке.

Однако следует быть осторожным при использовании только этой метрики для оценки качества кластеризации. Она может быть подвержена некоторым проблемам:

1. Чувствительность к выбросам: Среднее внутрикластерное расстояние может быть недостаточно робустным к наличию выбросов в данных, так как даже небольшое количество выбросов может исказить результаты.
2. Чувствительность к форме кластеров: Метрика среднего внутрикластерного расстояния может слабо работать с кластерами разной формы. Например, если у вас есть кластеры в форме кольца или полукольца, среднее расстояние может быть маленьким, хотя кластеры не будут считаться хорошими кластерами.
3. Зависимость от выбора метода расстояния: Результаты среднего внутрикластерного расстояния могут зависеть от выбранной метрики расстояния.

Различные методы расстояния могут давать разные значения, что может повлиять на оценку качества кластеризации.

В целом, среднее внутрикластерное расстояние может быть полезной метрикой для измерения качества кластеризации. Однако рекомендуется использовать его вместе с другими метриками, чтобы получить более полное понимание результатов кластеризации.

Среднее межкластерное расстояние (average inter-cluster distance) - это еще одна метрика, которая может быть использована для оценки качества кластеризации. Эта метрика показывает среднее расстояние между центроидами (или средними значениями) каждой пары кластеров.
Формула для вычисления среднего межкластерного расстояния может быть представлена следующим образом:

$$F_1 = \frac{\sum_{i=1}^{n}\sum_{j=i}^{n}\rho(x_i, x_j)\mathbb{I}[a(x_i) \neq a(x_j)]}{\sum_{i=1}^{n}\sum_{j=i}^{n}\mathbb{I}[a(x_i) \neq a(x_j)]}$$

Для вычисления среднего межкластерного расстояния суммируются все расстояния между каждым центроидом одного кластера и каждым центроидом другого кластера. Затем эта сумма делится на общее количество пар кластеров.

Чем больше значение среднего межкластерного расстояния, тем лучше качество кластеризации. Это говорит, что центроиды кластеров находятся достаточно далеко друг от друга, что говорит о хорошей разделимости кластеров.

У этой метрики есть как плюсы так и минусы:
Плюсы среднего межкластерного расстояния (AICD) в качестве метрики оценки кластеризации:
1. Интерпретируемость: Среднее межкластерное расстояние легко понять и интерпретировать. Чем больше это расстояние, тем лучше разделяются кластеры и тем более однородными они являются.

2. Сопоставимость: AICD является относительной мерой, которая позволяет сравнивать качество кластеризации между разными наборами данных или алгоритмами кластеризации.
3. Простота расчета: AICD вычисляется путем нахождения среднего значения междуцентроидных расстояний всех пар кластеров. Это относительно простая операция и не требует особых вычислительных ресурсов.

Однако есть и некоторые минусы:
1. Зависимость от выбора алгоритма: Среднее межкластерное расстояние может быть варьировать в зависимости от выбранного алгоритма кластеризации. Разные методы могут иметь разные подходы к определению и обновлению центроидов, что может влиять на полученное значение AICD.
2. Чувствительность к выбросам: AICD недостаточно устойчив к наличию выбросов в данных. Даже небольшое количество выбросов может сильно влиять на расстояние между центроидами кластеров и искажать оценку качества кластеризации.
3. Недостаточность для некоторых типов данных: В некоторых случаях, особенно когда кластеры имеют нелинейную структуру или сложные формы, среднее межкластерное расстояние может не являться достаточно информативной метрикой и не учитывать другие аспекты качества кластеризации, такие как внутрикластерная схожесть или согласование с исходными метками (если они известны).

Гомогенность (homogeneity) является одной из метрик для оценки качества кластеризации. Она измеряет, насколько хорошо каждый кластер состоит из объектов одного и того же истинного класса.
Для вычисления гомогенности используется следующая формула:

$$Homogeneity = 1 - \frac{H_{class|clust}}{H_{class}}$$

где H(class, clust) - условная энтропия между разбиением кластеров clust и истинными классами class, а H(class) - энтропия истинных классов class.

Плюсы гомогенности в качестве метрики оценки кластеризации:
1. Интерпретируемость: Гомогенность позволяет оценить, насколько кластеры соответствуют истинным классам. Более высокое значение гомогенности указывает

на то, что кластеры состоят из объектов одного и того же класса, что облегчает их интерпретацию.
2. Простота расчета: Вычисление гомогенности относительно просто и не требует сложных вычислений или вычислительных ресурсов.
3. Сопоставимость: Гомогенность можно использовать для сравнения качества кластеризации между разными методами или алгоритмами.
Однако есть и некоторые минусы:
1. Ограничение на идентичность кластеров и классов: Гомогенность предполагает, что каждый кластер должен полностью соответствовать одному классу. Это может быть проблематично, если объекты одного класса распределены между разными кластерами.
2. Неучет сложностей внутри классов: Гомогенность не учитывает внутрикластерное распределение объектов. Это может быть проблематично, если кластеры содержат подгруппы или имеют сложную внутреннюю структуру.
3. Недостаточность для несбалансированных данных: Гомогенность может быть неинформативной в случае, когда классы имеют неравное количество объектов. Это связано с тем, что она основана на условной энтропии, учитывающей вероятности классов в разбиении. Несбалансированные данные могут привести к нерепрезентативным результатам.
Следующая метрика — **полнота**, которая задается аналогично гомогенности.

$$Completeness = 1 - \frac{H_{clust|class}}{H_{clust}}$$

Для достижения полноты равной единице, можно объединить все объекты класса в один кластер. Таким образом достигается максимальная полнота кластеризации.
Гомогенность и полнота кластеризации, аналогично точности и полноте классификации, представляют собой важные характеристики. В задаче кластеризации также существует аналог F-меры, известный как **V-мера**. Эта мера связана с гомогенностью и полнотой посредством использования той же формулы, что и для F-меры в отношении точности и полноты.

$$V_\beta = \frac{(1+\beta) \cdot Homogeneity \cdot Completeness}{\beta \cdot Homogeneity + Completeness}$$

Максимизация итоговой метрики V-меры не приводит к тривиальным решениям благодаря комбинированию гомогенности и полноты.

Плюсы полноты и V-меры в качестве метрик оценки кластеризации:

1. Включение информации о правильной классификации объектов: Полнота измеряет долю правильно классифицированных объектов, что позволяет учесть правильность кластеризации в контексте истинных классов.
2. Сбалансированность: V-мера сочетает в себе полноту и точность, обеспечивая сбалансированную метрику, которая учитывает истинно положительные и ложно отрицательные связи.
3. Интерпретируемость: Полнота и V-мера имеют простую интерпретацию. Высокие значения указывают на более точную классификацию и лучшее согласование с истинными метками.

Однако есть и некоторые минусы:
1. Зависимость от доступности истинных меток: Использование полноты и V-меры возможно только при наличии истинных меток для сравнения с кластеризацией. В отсутствие этих меток оценка становится трудной.
2. Ошибки при перекрестном сопоставлении: Полнота и V-мера могут быть восприимчивыми к ошибкам, связанным с неправильным перекрестным сопоставлением кластеров соответствующим классам.
Существует несколько разных сценариев использования метрик качества в задаче кластеризации. Итак, чтобы оптимизировать среднее внутрикластерное или среднее межкластерное расстояние, можно задать количество кластеров заранее. Если имеется разметка, то можно использовать метрики гомогенности и полноты. В целом, можно использовать V-меру для подбора количества кластеров, так как она сочетает в себе гомогенность и полноту.
Однако не всегда есть разметка, и задача кластеризации часто является субъективной.
- Задача кластеризации имеет множество возможных решений, что делает ее сложной.
- Формализация допустимых решений в практике кластеризации является сложной задачей.
- Задача кластеризации имеет более одного решения из-за некорректной постановки.
- Коэффициент силуэта является наиболее практичной метрикой в отсутствии разметки и фиксированного числа кластеров.
- Классификация изображений с использованием visual bag of words - исключение, где результат кластеризации используется для решения задачи обучения с учителем.
- Для достижения итоговой задачи рекомендуется выбирать алгоритм кластеризации и оптимальные гиперпараметры.
- Качество кластеризации можно проигнорировать в данном случае.
- Главное — найти такие параметры, которые смогут эффективно решить задач

Итоги:

На сегодняшней лекции мы рассмотрели основные концепции и алгоритмы кластеризации, которые позволяют представить разделить на группы неразмеченные данные.

1. Кластеризация является методом без учителя, который позволяет обнаруживать скрытые структуры и группировать данные на основе их сходства.

2. Кластеризация может иметь широкое практическое применение в различных областях, таких как маркетинг, биология, финансы и многие другие. Она может использоваться для сегментации клиентов, выявления групп схожих пациентов, анализа социальных сетей и многих других задач.

3. Существуют различные алгоритмы и подходы для кластеризации данных, такие как иерархическая кластеризация, k-средних, DBSCAN, агломеративная кластеризация и др. Каждый из этих алгоритмов имеет свои преимущества и недостатки, а выбор конкретного алгоритма зависит от характеристик данных и задачи.

- Наиболее часто используемый метод для кластеризации является K-means.
- Иерархическая кластеризация и DBSCAN занимают второе место в частоте использования.
- Иерархическая кластеризация проста в понимании и более известна, чем DBSCAN.
- Однако, иерархическая кластеризация часто работает некачественно из-за образования одного большого кластера и нескольких маленьких.
- DBSCAN менее известный метод, но его применение обычно приводит к лучшему качеству, чем K-means или иерархическая кластеризация.

4. При применении кластеризации следует учитывать следующие плюсы и минусы:

Плюсы:

- Обнаружение скрытых структур: Кластеризация позволяет выявить скрытые группы и структуры в данных, что может быть полезно для понимания особенностей набора данных и принятия решений в различных задачах.
- Без учителя: Кластеризация не требует наличия меток классов, что позволяет применять ее к неразмеченным данным.
- Масштабируемость: Некоторые алгоритмы кластеризации, такие как k-средних и DBSCAN, могут эффективно обрабатывать большие объемы данных.

Минусы:

- Чувствительность к выбору параметров: Некоторые алгоритмы кластеризации требуют настройки параметров, таких как количество кластеров, радиусы и т. д., что может быть сложно при отсутствии заранее известной информации о данных.

- Проблема с обработкой шума: Некоторые алгоритмы могут неэффективно работать с шумом и выбросами в данных.
- Субъективность интерпретации: Интерпретация результатов кластеризации может быть субъективной и зависеть от выбранного алгоритма и последовательности принятия решений.

В целом, кластеризация является мощным и гибким методом для группировки данных и обнаружения скрытых структур. Однако при применении кластеризации нужно учитывать особенности конкретной задачи и выбирать подходящий алгоритм с учетом плюсов и минусов каждого метода.

Что можно почитать еще?

1. DBSCAN
2. Метод силуэта
3. K-means

Используемая литература

1. https://questu.ru/articles/732117/
2. https://machinelearningmastery.ru/how-to-cluster-in-high-dimensions-4ef693bacc6/
3. https://gulweb.ru/articles/znachenie-mer-rasstoyaniya-i-podobiya-v-klasterizaczii.html