Титульный лист материалов по дисциплине

(заполняется по каждому виду учебного материала)

ДИСЦИПЛИНА	**Технологии извлечения знаний из больших данных** (полное наименование дисциплины без сокращений)
ИНСТИТУТ	ИКБ
КАФЕДРА	**Кафедра КБ-14 «Цифровые технологии обработки данных»** полное наименование кафедры)
ВИД УЧЕБНОГО МАТЕРИАЛА	**Лекция** (в соответствии с пп.1-11)
ПРЕПОДАВАТЕЛЬ	**Никонов В.В.** (фамилия, имя, отчество)
СЕМЕСТР	**3 семестр 2023/2024 уч. года** (указать семестр обучения, учебный год)

Продвинутая работа с табличными данными

Кластеризация и ее применения

Кластеризация — это деление объектов на группы похожих объектов, эти группы называют кластерами.

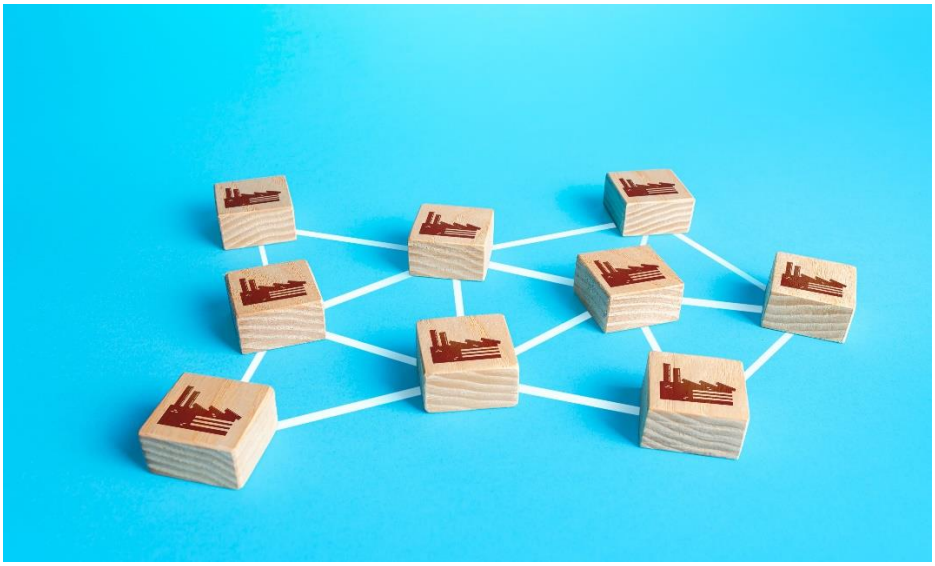
Пример кластеризации экспертным методом

Для примера рассмотрим клиентов кофейни (объект — клиент), признаки: сколько раз посещал(-а) кофейню утром, днем и вечером и сколько раз покупал позиции из меню (столько же признаков, сколько позиций: эспрессо, капучино, латте, маффины, сэндвичи и т. д.). В этом примере кластеризация задает деление клиентов на группы, соответственно, в каждой группе должны оказаться клиенты с похожими значениями признаков. Например, у нас могла бы получиться такая кластеризация клиентов кофейни на три группы:

- первая группа — работники близлежащего офисного центра, приходят в обеденное время, вместе с кофе обычно покупают салат или сэндвич;
- вторая группа — пассажиры автобусов, останавливающихся рядом с кофейней, приходят в разное время и покупают только кофе;
- третья группа — клиенты, приехавшие целенаправленно в данную кофейню после просмотра рекламы о ней, приходят в основном вечером и покупают фирменный напиток кофейни.

Если бы мы спросили бариста в кофейне, на какие группы он бы разделил клиентов своей кофейни, он бы наверняка смог назвать все перечисленные группы, потому что он работает с ними каждый день. Иными словами, пример хорошо иллюстрирует суть задачи кластеризации, но не является реальным кейсом применения, тут применять компьютерные методы не нужно. Однако если мы говорим о большой компании с миллионами клиентов, например банке, то вряд ли в компании найдется человек, который хорошо знает всех этих клиентов и сможет перечислить все группы клиентов. Конечно, руководитель отдела по работе с клиентами сможет перечислить

основные группы, например студенты, пенсионеры, работающие и VIP-клиенты, но он, скорее всего, не сможет выделить совершенно все имеющиеся группы клиентов, просто потому что он не может ознакомиться с информациях обо всех миллионах клиентов. А компьютер может — для этого и разрабатывают методы кластеризации данных.



Для чего же выполняют кластеризацию? В первую очередь для того, чтобы лучше понимать аудиторию, иметь структурированное представление о клиентских сегментах и таргетировать бизнес для отдельных групп клиентов. Например, в процессе кластеризации можно открыть новые группы клиентов, достаточно большие для того, чтобы разрабатывать для них отдельные продукты, но достаточно маленькие для того, чтобы раньше они были незаметны на фоне основной массы клиентов. Кластеризация может применяться для выделения типичных групп клиентов и разделения рынка для создания персонализированных предложений. Также кластеризацию используют для обнаружения особенных клиентов — тех, которые не похожи на других клиентов и не вошли ни в какой кластер.

Кластеризация клиентов игрового онлайн-сервиса

Число клиентов: 150 миллионов; признаки клиентов: среднее время, проведенное пользователем в играх за неделю, прогресс пользователя в разных играх, средний чек в месяц. Кластеризация была выполнена

алгоритмом CLARA, который похож на k-Means (который мы обсудим ниже), но лучше обрабатывает нестандартных клиентов и работает быстрее. Кластеризация позволила разделить клиентов на группы, для каждой группы был создан портрет характерного пользователя, который могут использовать разработчики и дизайнеры. Полезным оказалось также отслеживание доходов по каждой группе клиентов, может быть, пользователи, оказавшиеся в конкретной группе, перестают использовать сервис — тогда стоит проанализировать этих клиентов отдельно. Или после введения нового продукта некоторые пользователи переместились из более доходного кластера в менее доходный, тогда продукт стоит доработать.

Приложения кластеризации

Существует множество других приложений кластеризации за пределами бизнеса. Например, при анализе изображений кластеризацию можно использовать для сегментации изображений (выделения различных объектов на изображении и фона) — для этого в качестве объектов используют пиксели, в качестве признаков — цвет и расположение пикселей, и ищут группы похожих пикселей. Кластеризация текстов делит тексты по темам, что может упростить навигацию по большому набору текстов (пользователь читает только интересные ему/ей темы). В биоинформатике кластеризацию используют для группировки схожих геномных последовательностей в семейство генов, в медицине — для автоматического выделения различных типов тканей при диагностике.

Визуализация кластеризации

Еще один интересный пример применения кластеризации, который позволяет визуализировать результат, — это геокластеризация, то есть кластеризация точек на карте, например точек продаж. Такая кластеризация позволяет, к примеру, найти удобные расположения для складов (склады размещают в центрах кластеров), чтобы было удобнее развозить товары со складов в точки продаж. Для примера приведем кластеризацию городов России, цветами обозначены автоматически выделенные кластеры:

Для каждого города известна широта и долгота, алгоритм кластеризации объединяет в группы города, расположенные рядом — получаются цветные кластеры.

Стоит отметить, что чаще всего визуализировать результат кластеризации невозможно: на плоскости можно отрисовать только два признака (по двум осям, например, в данных с городами у нас есть всего два признака — широта и долгота), а в реальных данных признаков гораздо больше. Можно отрисовать два отдельных выбранных признака или воспользоваться методами понижения размерности, о которых мы поговорим в следующем блоке, но такие визуализации не показывают полную картину. Невозможность выполнения полной визуализации создает дополнительные сложности с оценкой результатов кластеризации, о чем мы поговорим далее.

Алгоритмы кластеризации

Мы обсудили постановку задачи и возможности применения кластеризации, давайте поговорим об особенностях алгоритмов кластеризации.

Представим, что мы имеем набор данных, которые хотим кластеризовать. В первую очередь нам необходимо вооружиться способом вычисления различия между объектами. Иными словами, нужно научиться для двух клиентов выдавать число, показывающее, насколько сильно они различаются. Чаще всего в кластеризации работают с табличными данными,

которые мы подробно обсудили в блоке 3: каждая строка таблицы задает объект, а столбцы описывают признаки этого объекта. В этом случае самый простой способ численно оценивать различия между клиентами — это суммировать разницу в значениях их признаков. Например, различие между клиентом 25 лет со стажем работы 7 лет и клиентом 28 лет со стажем работы 5 лет будет вычисляться как $|25 - 28| + |7 - 5| = 5$. Существует также множество других способов вычисления различности объектов для табличных данных.

Кластеризация текстовых данных

Кластеризацию можно использовать и для других видов данных, необязательно табличных. Например, чтобы кластеризовать текстовые последовательности, можно вычислять различие между двумя текстами как число вставок, удалений и замен символов, которые нужны, чтобы превратить один текст в другой. Такой способ может хорошо работать для несложных последовательностей вида названий продуктов или компаний, но для полноценных текстов он подходит плохо. Заказчик, хорошо понимающий аудиторию бизнеса, может подсказать, как можно оценивать схожесть и различие клиентов.

- Некоторые методы кластеризации (например k-Means, который мы рассмотрим ниже) требуют также задания числа кластеров. Дело в том, что задача кластеризации не является четко поставленной (можно выделять кластеры, состоящие из нескольких объектов, а можно — большие, в которые входят сотни или тысячи объектов), и указание числа кластеров в некотором смысле конкретизирует задачу.
- Другие методы кластеризации (например, DBSCAN) определяют число кластеров автоматически по данным в процессе обучения. Однако эти методы требуют задания некоторого другого значения, влияющего на кластеризацию, грубо говоря, вместо числа кластеров просят указать

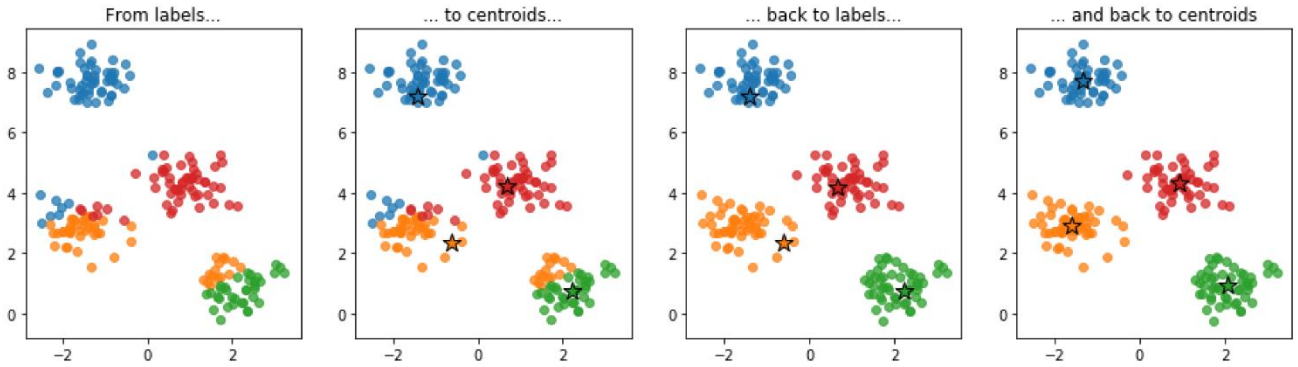
ожидаемый размер кластеров, то есть на самом деле эти методы просто требуют другой конкретизации задачи.

k-Means

Самый известный и часто используемый метод кластеризации называется k-Means (метод k средних). Число k в названии говорит о том, что алгоритм находит k кластеров по данным. Схема работы алгоритма довольно проста. Каждый кластер (группа клиентов, которую мы хотим выделить) задается центром — неким типичным клиентом этой группы.

Описанный алгоритм k-Means очень хорошо работает на несложных данных, например на данных, изображенных ниже. Если же данные сложнее, например требуется выделить кластеры с несколькими «центрами», то этот метод может не очень хорошо справиться с задачей. Кроме того, метод не очень хорошо работает, если число кластеров задано неправильно: тогда он будет разбивать кластеры на более мелкие или объединять несколько кластеров в один. Опишем последовательность работы:

1. В начале работы алгоритма мы назначаем эти центры случайно, например, выбираем произвольных клиентов в качестве центров.
2. Затем мы выполняем кластеризацию: каждого клиента из данных записываем в тот кластер, на центр которого он больше всего похож. Получается k групп (кластеров) клиентов.
3. Теперь, когда клиенты распределены по кластерам, мы находим новые центры кластеров: в каждой группе находим самого репрезентативного клиента.
4. И далее все повторяется снова: заново распределяем клиентов по кластерам, используя новые центры, заново ищем центры и т.д.

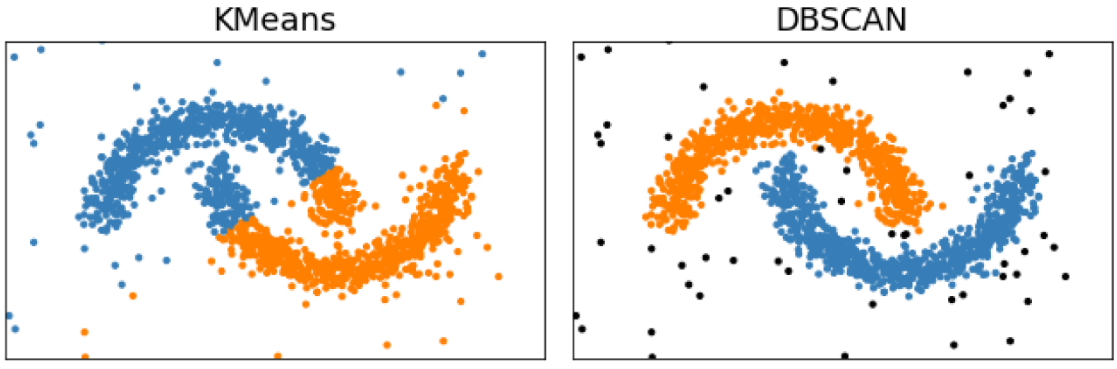


Применение k-Means для кластеризации точек на плоскости. Исходные кластеры (первое изображение) неправильные, но после выбора новых центров (черные звездочки, второе изображение) и перераспределения точек (третье) изображение кластеризация стала информативной.

Описанный алгоритм k-Means очень хорошо работает на несложных данных. Если же данные сложнее, например требуется выделить кластеры с несколькими «центрами», то этот метод может не очень хорошо справиться с задачей. Кроме того, метод не очень хорошо работает, если число кластеров задано неправильно: тогда он будет разбивать кластеры на более мелкие или объединять несколько кластеров в один.

DBSCAN

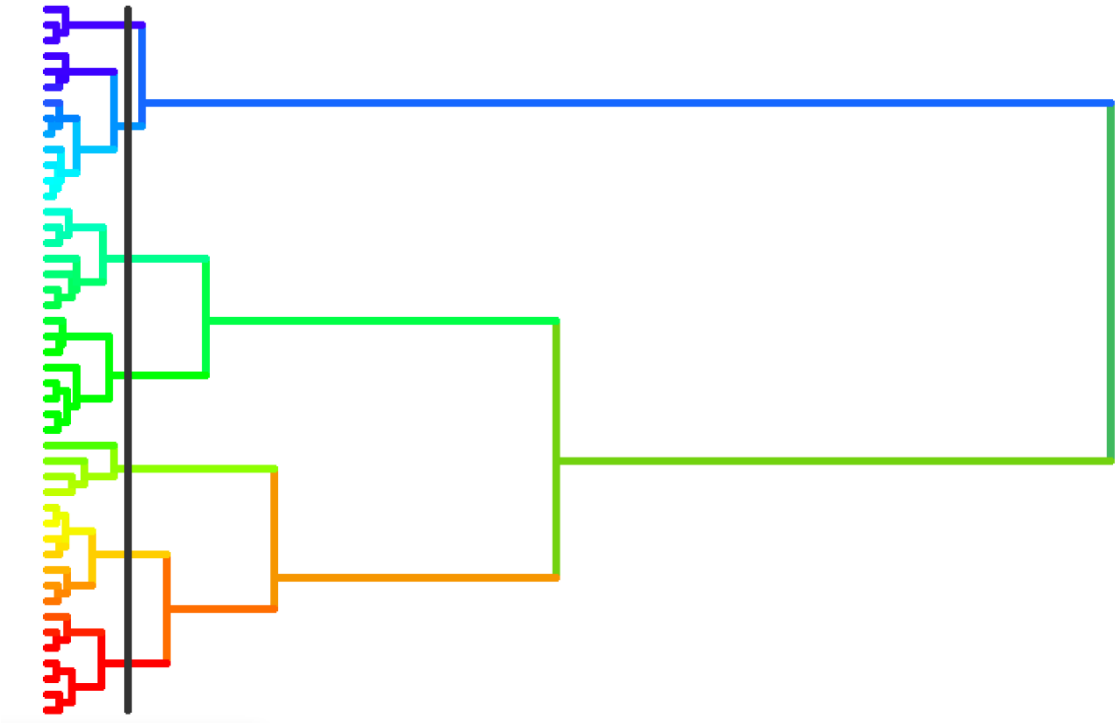
Другой известный метод — DBSCAN (Density-Based Spatial Clustering of Applications with Noise) — требует задания минимального числа объектов в кластере и минимальной схожести объектов в кластере и затем сам определяет число кластеров. Интересной особенностью DBSCAN является то, что некоторые объекты он называет шумовыми и не относит их ни к одному кластеру. Пример данных, для которых хорошо работает DBSCAN (шумовые точки отмечены черным) и плохо работает k-Means:



Кластеризация точек на два кластера (оранжевый и голубой). Слева метод KMeans, справа DBSCAN. Черный цвет отмечает шумовые точки.

Иерархическая кластеризация

Еще один метод кластеризации, иерархическая кластеризация (Agglomerative clustering), находит вложенные кластеры: например, в кластере «клиенты-студенты» могут быть выделены подкластеры «работающие студенты», «студенты с большими тратами в индустрии развлечений» и «иногородние студенты». Иерархическая кластеризация строит диаграммы следующего вида:



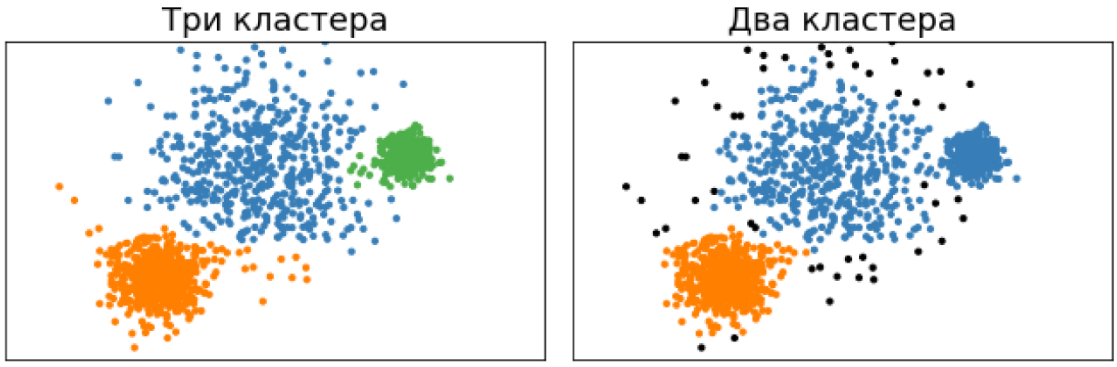
Здесь по вертикали отмечены объекты, и в самом начале работы алгоритма каждый кластер состоит из одного объекта. На каждом шаге алгоритм объединяет два кластера в один: это отмечается соединяющей скобкой. Самая последняя (самая большая) скобка означает объединение всех объектов в один кластер. Любое отсечение этой диаграммы задает одну кластеризацию, например черная линия пересекает семь отрезков и поэтому задает кластеризацию на семь кластеров.

Инструменты для выполнения кластеризации

Для выполнения кластеризации по конкретным данным чаще всего используется язык программирования Python и библиотека Scikit-learn — в ней реализовано множество различных алгоритмов кластеризаций, в частности все, описанные выше.

Измерение качества кластеризации

Как мы уже упомянули выше, задача кластеризации не является четко поставленной: как правило, в одних и тех же данных можно выделить разные кластеры, и не ясно, какая кластеризация лучше. Например, на изображении ниже можно выделить три кластера, а можно два (маленький кластер присоединится к большому).

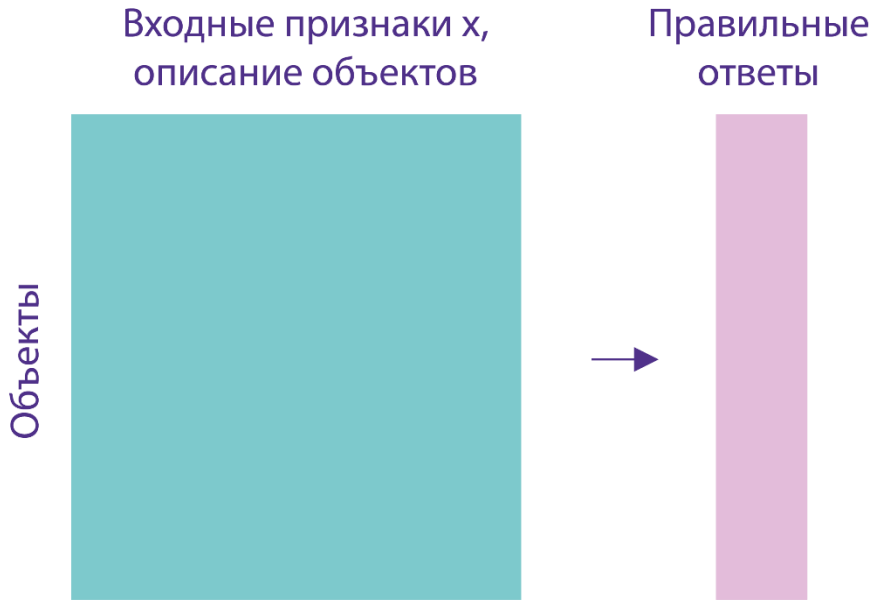


В задачах обучения с учителем, например, классификации или регрессии, можно измерить ошибку — насколько неточно алгоритм предсказывает целевые переменные. В задаче кластеризации целевые переменные не даны, и измерить качество гораздо сложнее — общепринятых метрик качества кластеризации не существует. Иногда рассматривают метрики наподобие внутрикластерого расстояния (насколько в среднем различны объекты внутри кластеров) или межкластерного расстояния (насколько в среднем различны объекты между кластерами), но они не всегда информативны и не отражают бизнес-цели. Зато кластеризацию всегда можно показать специалисту, он проанализирует выделенные группы клиентов и скажет, полезна ли построенная кластеризация для бизнеса.

Подведем итог: кластеризация — это разделение объектов на группы так, чтобы внутри одной группы объекты были похожи друг на друга. Кластеризация позволяет лучше понять аудиторию и таргетировать предложения для отдельных сегментов клиентов, а также используется в анализе текстов, изображений и других типов данных. Существует множество различных методов кластеризации, однако общепринятых метрик качества нет, поэтому обычно строят кластеризации разными методами и анализируют полученные кластеры.

Что можно делать с данными, если у них нет разметки и целевой переменной, а извлечь какую-то информацию очень нужно/

Парадигма обучения с учителем vs парадигма обучения без учителя



Данные в задаче **обучения с учителем** (или на размеченных данных) по структуре соответствуют картинке справа.
В таблице в части столбцов — значения признаков, которые описывают объект, в отдельном столбце — значения целевой переменной. Каждая строка соответствует отдельному объекту, паре *входное описание — значение целевой функции.*

Задача машинного обучения в этом случае понятна: научиться по описанию объекта предсказывать значение размеченной целевой функции. В задаче регрессии целевая функция принимает значения из некоторого

диапазона, в задаче классификации — из конечного набора возможных ответов, классов.

Так как у нас есть правильные ответы, то такую постановку задачи в машинном обучении называют *обучением с учителем*, который знает правильные ответы, или *обучением на размеченных данных.*

Данные в задаче **обучения без учителя** (или на неразмеченных данных) по структуре соответствуют другой картинке справа.

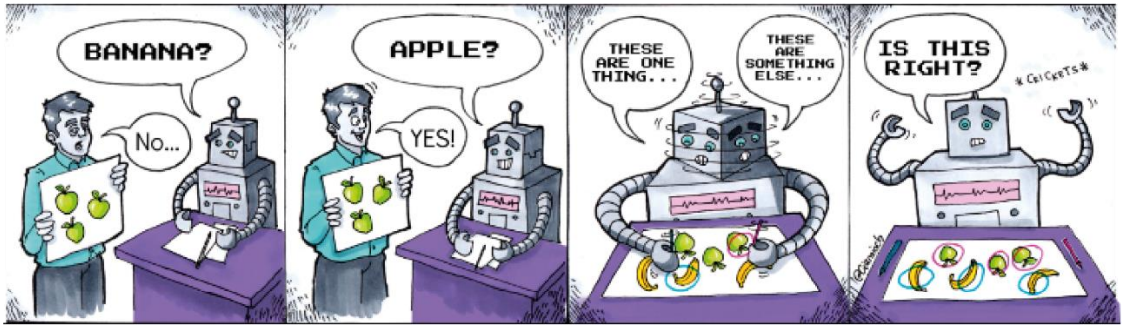
В таблице в части столбцов — значения признаков, которые описывают объект. Целевой переменной у нас в данных нет.

Про возможные постановки задач машинного обучения для таких данных мы и будем говорить в этом лонгриде. Они встречаются реже, чем

задачи обучения с учителем. Решать и понимать, насколько хорошее получилось решение, в этом случае сложнее.

Так как у нас нет правильных ответов, то такую постановку задачи в машинном обучении называют *обучением без учителя*, или *обучением на неразмеченных данных*.

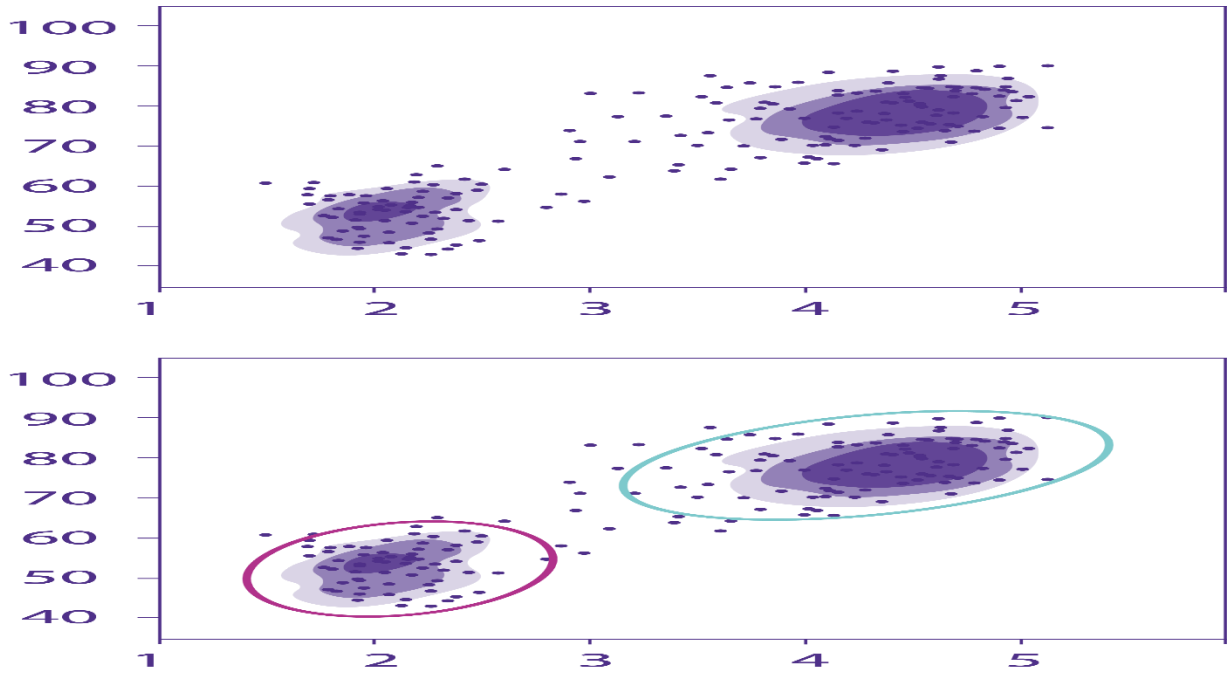
Кроме задачи обучения на размеченных и неразмеченных данных еще рассматривают задачи на частично размеченных данных, когда для части объектов ответы есть, а для части нет. Но на практике они встречаются не очень часто, поэтому в этом курсе мы их рассматривать не будем.



С учителем Без учителя

Восстановление распределения данных

Часто можно думать про задачу обучения без учителя как про задачу восстановления распределения данных: понимание того, как объекты распределены в пространстве признаков, какие наиболее характерные значения у них есть, где объектов мало, а где они лежат плотными кучками.



Пример полученного с помощью машинного обучения распределения данных можно увидеть на рисунке. Синие точки — значения двух признаков из обучающей выборки. Закрашенные синим разной насыщенности области — восстановленное распределение. Чем более интенсивный цвет, тем вероятнее, что в эту область попадет новая точка. Если у нас есть такая модель, то для новой точки мы можем сказать, насколько она вероятна в рамках существующего распределения данных. Если вероятность встретить точку в этом месте мала, то такая точка может быть *аномалией* или *выбросом*, и нужно аккуратно с ней работать.

Другой пример задачи машинного обучения на неразмеченных данных — кластеризация. Мы разбиваем все точки на кластеры похожих. На рисунке 2 мы выделили в данных два кластера, ограниченных зеленой и оранжевой линиями. Теперь для новой точки мы знаем, к какому кластеру она принадлежит и, следовательно, на какие точки из исходной выборки похожа.

Формальная постановка задачи кластеризации.

Пусть X - множество объектов, Y - — множество номеров (имен, меток) кластеров. Задана функция расстояния между объектами $\rho(x, x')$. Имеется конечная обучающая выборка объектов $X^m = \{x_1, \dots, x_m\} \subset X$.Требуется разбить выборку на непересекающиеся подмножества, называемые

кластерами, так, чтобы каждый кластер состоял из объектов, близких по метрике ρ, а объекты разных кластеров существенно отличались. При этом каждому объекту $x_i \in X^m$ приписывается номер кластера y_i.

Алгоритм кластеризации — это функция $\alpha: X \to Y$, которая любому объекту $x \in X$ ставит в соответствие номер кластера $y \in Y$. Множество Y в некоторых случаях известно заранее, однако чаще ставится задача определить оптимальное число кластеров, с точки зрения того или иного *критерия качества* кластеризации.

Зачем нужна кластеризация данных — разберем дальше.

Примеры задач кластеризации

Начнем с научной ***задачи кластеризации небесных тел:*** мы измеряем для небесных тел силу излучения на семи разных частотах. Проведя кластеризацию, можно выделить основные типы небесных тел. Когда на небе появится новый объект, мы сразу определим, что это звезда. Если какие-то объекты не попали ни в один из кластеров, нужно на них посмотреть внимательнее.

Ближе к приложениям — ***задача кластеризации документов*** или новостей. Каждый день в интернете публикуются тысячи новостей, причем часто про одно и то же рассказывают на разных сайтах. «Яндекс.Новости» кластеризует описания одного события с разных сайтов в сюжеты, чтобы пользователю было проще в них ориентироваться. Никакой разметки в этой задаче сделать не получится, потому что новые сюжеты появляются каждую минуту. Нужно использовать алгоритмы кластеризации, которые могут решать эту задачу без разметки.

Похожие задачи часто встречаются в бизнесе:

- Сегментация пользователей: пользователи в каждом кластере должны быть похожи друг на друга, а в соседних — отличаться. Тогда, посмотрев на характеристики пользователей в каждом кластере, мы сможем выбрать наиболее подходящий кластер целевой аудитории для новой маркетинговой кампании

- Детектирование аномалий: ищем те транзакций, которые не попадают ни в один из кластеров. Такие аномальные транзакции могут быть мошенническими
- Подготовить данные для машинного обучения (разбить на группы): если пользователи в группах более однородны, для каждой группы можно обучить свою модель машинного обучения, которая будет более интерпретируемой и точной
- Аналитика доступных данных: проведя кластеризацию, мы поймем, с какими клиентами мы работаем, что общее для каждого кластера пользователей, а чем они отличаются

Постановка задачи кластеризации

Можно описать задачу кластеризации как разбиение объектов на однородные группы. В результате обучения модель кластеризации может для нового объекта определить номер кластера, к которому этот объект с ее точки зрения принадлежит.

Пример работы алгоритма кластеризации — в таблице ниже. Меток кластера в обучающей выборке не было, алгоритм сам придумал правило для разбиения объектов на кластеры.

Входные признаки				Результат кластеризации
x_1 возраст	x_2 пол	x_3 доход за последний месяц	x_4 количество членов семьи	Метка кластера
25	1	70000	0	1
32	0	63000	1	2
29	0	81000	0	1
44	0	54000	2	2

У нас должен выполняться *внутренний критерий качества кластеризации:* объекты внутри одного кластера похожи друг на друга.

Кроме того, должен выполняться *внешний критерий качества кластеризации:* объекты из разных кластеров отличаются друг от друга.

Метрики качества кластеризации — обычно некоторая комбинация этих двух критериев, и их пытаются формализовать тем или иным образом.

Однако так как точного критерия близости у нас нет, с точки зрения и математики, и бизнеса, задача поставлена не очень хорошо. Поэтому трудно понять, насколько хорошую кластеризацию мы получили. Можно сделать это косвенно, посмотрев на качество решения бизнес-задачи, которая использует полученную кластеризацию.

Другие постановки задачи кластеризации

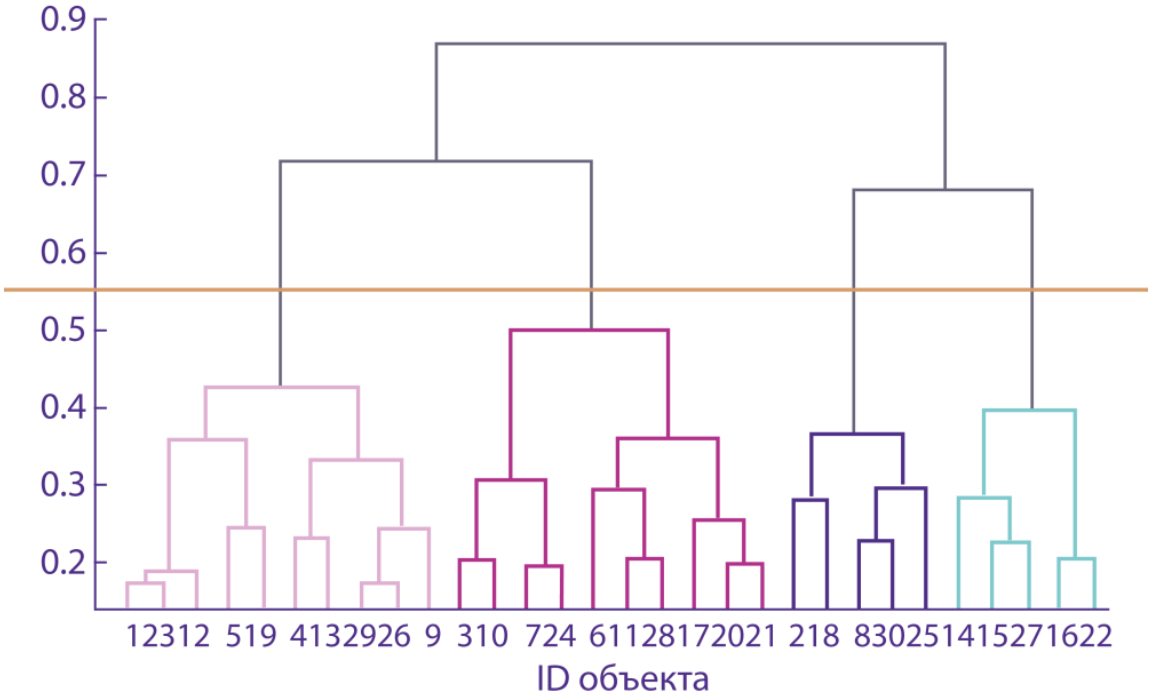
Выделяют другие постановки задачи кластеризации: мягкую кластеризацию и иерархическую кластеризацию.

Кластеризация может быть не жесткой (hard), а **мягкой** (soft). В этом случае результатом работы модели кластеризации будет не жесткая метка кластера для каждого объекта, а набор вероятностей принадлежности к каждому из классов. Как и для задачи классификации, легко перейти от мягкой к жесткой кластеризации, взяв в качестве метки кластера индекс кластера с максимальной вероятностью. Пример работы алгоритма мягкой кластеризации приведен в таблице ниже.

Входные признаки				Мягкая кластеризация		Жесткая кластеризация
x_1 возраст	x_2 пол	x_3 доход за последний месяц	x_4 количество членов семьи	вероятность принадлежности к первому кластеру	вероятность принадлежности ко второму кластеру	жесткая метка кластера
25	1	70000	0	0,82	0,18	1
32	0	63000	1	0,24	0,76	2
29	0	81000	0	0,87	0,13	1
44	0	54000	2	0,11	0,89	2

Другой способ создать более гибкую модель кластеризации — иерархическая кластеризация. В этом случае результатом работы алгоритма кластеризации будет дерево, похожее на изображенное ниже. Для того чтобы получить из такого дерева кластеризацию, мы выбираем уровень, изображенный на рисунке оранжевой линией. После этого идем от

пересечений оранжевой линии с черной вниз, получая в данном примере четыре кластера. Если мы сделаем разбиение выше или ниже, получим больше или меньше кластеров, более релевантных решаемой задаче бизнеса.



Методы кластеризации: метод k-средних, метод гауссовских смесей, DBSCAN

https://colab.research.google.com/drive/1-2w2dXPtAvv81Rk5Ow8hiQe4M9jryZYu?usp=sharing#scrollTo=cdM2Hjt0TXEG

Внешние и внутренние метрики качества: коэффициент Джаккарда, коэффициент силуэта

Метрики качества в задаче кластеризации

Для классификации и обучения с учителем имеется множество различных метрик качества модели: точность, полнота, доля правильных классификаций, ROC AUC. Так как у нас есть правильные ответы, на тестовой выборке мы сравниваем полученные моделью ответы с правильными и получаем оценку качества как меру несовпадения правильных и предсказанных моделью ответов.

Для кластеризации возникает вопрос: как оценить ее качество, если у нас нет разметки или если имеющаяся разметка недостаточно полно отражает желаемое разбиение данных?

Необходимы метрики для сравнения алгоритмов кластеризации и двух полученных множеств кластеров. Различают *внутренние и внешние метрики качества кластеризации* (по смыслу внешние и внутренние метрики отличаются от внешних и внутренних критериев).

Внешняя метрика сравнивает полученные метки кластеров с метками, которые мы получили из какого-то другого источника.

В качестве примера внешней метрики будем использовать **индекс Джаккарда.**

Пусть у нас есть истинные метки классов и разбиение на кластеры, полученное с помощью алгоритма кластеризации. Тогда для пары точек из выборки наблюдается одна из четырех ситуаций:

1. Две точки лежат в одном кластере, но метки классов у них разные
2. Две точки лежат в одном кластере, и метки классов у них одинаковые
3. Две точки лежат в разных кластерах, и метки классов у них разные
4. Две точки лежат в разных кластерах, но метки классов у них одинаковые

Ситуации 1 и 4 соответствуют неадекватной кластеризации с точки истинной разметки, ситуации 2 и 3 говорят о соответствии полученной кластеризации и истинной разметки.

Пусть количество пар для каждой ситуации соответственно A1, A2, A3, A4. Тогда индекс Джаккарда равен $J = A2 / (A2 + A4 + A1)$.

В идеальном мире A1 и A4 равны нулю, и индекс Джаккарда равен своему максимальному значению — 1. В худшем случае A2 равно нулю, и индекс Джаккарда равен своему минимальному значению — 0. В реальности мы наблюдаем промежуточную ситуацию: чем больше индекс Джаккарда, тем лучше соответствие между заданной «истинной» разметкой и полученной кластеризацией.

Читателю может показаться, что мы забыли про A3, и он прав. Но количество A3 про то, что объекты из разных кластеров должны сильно отличаться, а в индексе Джаккарда мы смотрим на то, насколько похожи точки внутри одного кластера. Поэтому A3 и не входит в формулу.

У индекса Джаккарда есть и другая интерпретация: насколько сильно пересекаются множества одинаковых пар с точки зрения истинной разметки и с точки зрения алгоритма кластеризации.

Обычно истинной разметки у нас нет. Поэтому используют внутренние метрики.

Внутренняя метрика качества кластеризации не использует никакой дополнительной информации.

Посмотрим на типичную внешнюю метрику — **коэффициент силуэта.**

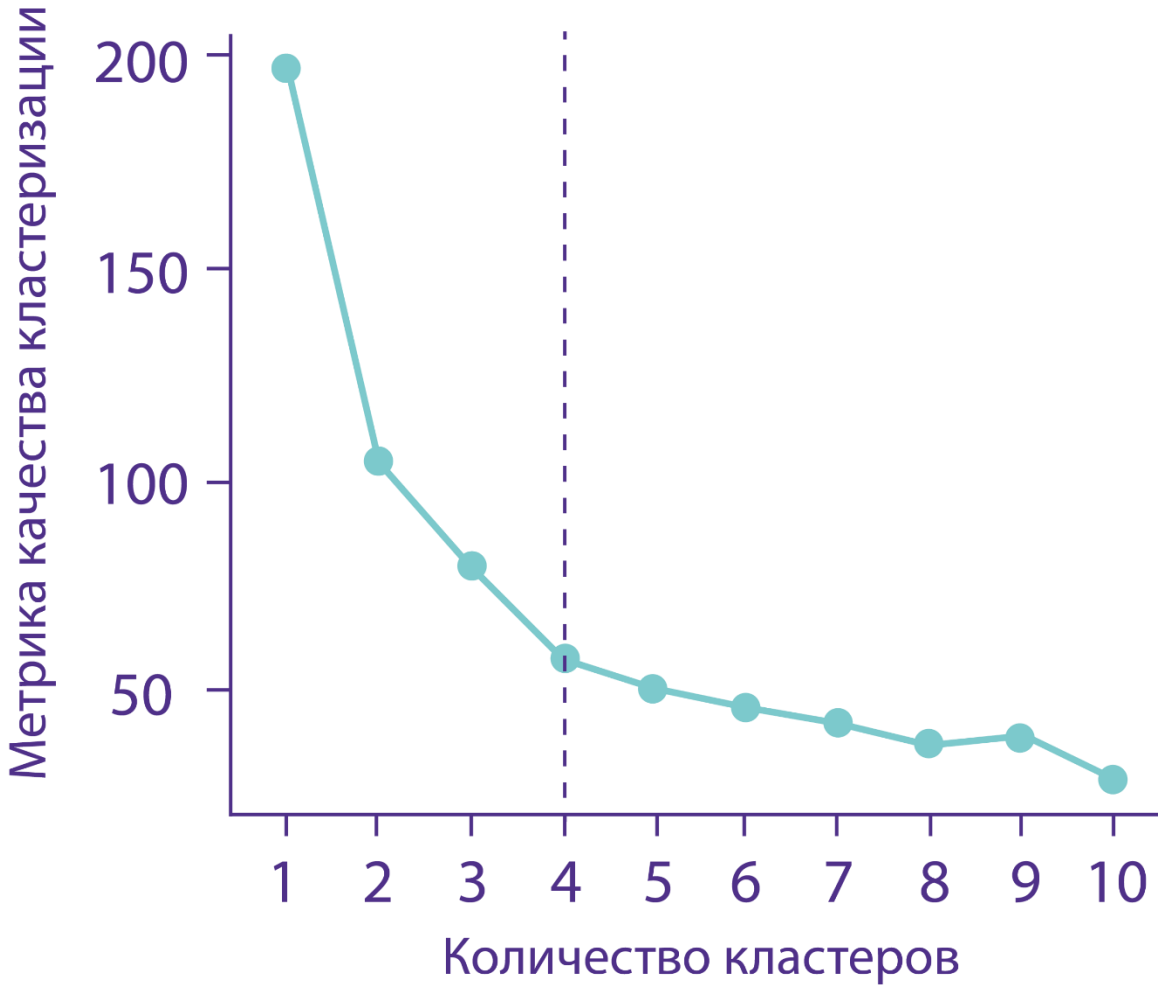
Сначала научимся считать коэффициент силуэта для одной точки *i*. Посчитаем среднее расстояние от нее до всех точек этого же кластера *a(i)*, посчитаем среднее расстояние от этой же точки до точек из других кластеров *b(i)*. Чем меньше *a(i)*, тем лучше; чем больше *b(i)*, тем лучше. Поэтому коэффициент силуэта считается по следующей формуле:

$$s(i) = (b(i) - a(i)) / \max(a(i), b(i)).$$

Если *s(i)* больше нуля, то в среднем точки того же кластера ближе, чем точки из других кластеров, потому что $b(i) > a(i)$. Если *s(i)* меньше нуля, то ситуация обратная, и эта точка кластеризовалась не очень удачно. Для нормировки мы делим разность между *b(i)* и *a(i)* на максимум из этой пары и получаем нормированную разность. Если она близка к 1, все хорошо; если близка к –1, все плохо.

Усреднив коэффициент силуэта по всем точкам обучающей выборки, получим общую характеристику качества кластеризации. Так как каждое из значений от –1 до 1, то и их среднее будет из этого диапазона. Коэффициент силуэта близкий к 1 соответствует высокому качеству кластеризации, коэффициент силуэта близкий к –1 — низкому.

Как и многие другие метрики, связанные с расстояниями между объектами, коэффициент силуэта работает, если расстояния отражают похожесть объектов друг с другом. Поэтому в больших размерностях, когда входных признаков мало, в силу проклятия размерности расстояния слабо отражают похожесть объектов. Следовательно, и коэффициент силуэта плохо применим.



Кроме выбора алгоритма кластеризации значения метрики качества могут использоваться для *выбора числа кластеров*. Если с какого-то количества кластеров метрика перестает падать, значит в данных именно такое истинное их число. Так как форма кривой обычно похожа на две прямые с таким изгибом (как рука), а метрика перестает уменьшаться в ее сгибе, такой критерий называют *критерием локтя* для выбора числа кластеров.

Снижение размерности: PCA, t-SNE

Демонстрация в google colab

https://colab.research.google.com/drive/1_tK9IMMrI1clH-0dOANo9IRyacxBK0OB?usp=sharing