Литература

1. Коэльо Л.П., Ричарт В. Построение систем машинного обучения на языке Python. М.: ДМК Пресс, 2016. - 302 с. https://e.lanbook.com/reader/book/82818/#4

2. Шарден Б., Массарон Л., Боскетти А. Крупномасштабное машинное обучение вместе с Python. М.: ДМК Пресс, 2018. - 358 с. https://e.lanbook.com/reader/book/105836/#4

3. Пролубников А.В. Математические методы распознавания образов: учебное пособие. Омск: Изд-во Ом. гос. ун-та, 2020. - 110 с. https://e.lanbook.com/reader/book/142454/#2

4. Хейдт М., Груздев А.В. Изучаем pandas. М.: ДМК Пресс, 2019. - 682 с. https://e.lanbook.com/reader/book/131693/#4

5. Шалев-Шварц Ш., Бен-Давид Ш. Идеи машинного обучения. От теории к алгоритмам. М.: ДМК Пресс, 2019. - 436 с. https://e.lanbook.com/reader/book/131686/#4

Литература

6. Ростовцев В.С. Искусственные нейронные сети: учебник. Санкт-Петербург: Лань, 2019. – 216 с. https://e.lanbook.com/reader/book/122180/#2

7. Плас Дж. Вандер. Python для сложных задач: наука о данных и машинное обучение. - СПб.: Питер, 2018. - 576 с.

8. Элбон Крис. Машинное обучение с использованием Python. Сборник рецептов: Пер. с англ. - СПб.: БХВ-Петербург, 2019. - 384 с.

9. Бенгфорт Бенджамин, Билбро Ребенка, Охеда Тони. Прикладной анализ текстовых данных на Python. Машинное обучение и создание приложений. - СПб.: Питер, 2019. - 368 с.

Алгоритмы анализа данных и машинное обучение

Мы живём в век 4-ой промышленной революции. **Первая** промышленная революция пришла с изобретением ткацкого станка, **вторая** – с открытием электричества и автоматизацией станков, **третья** происходила в прошлом веке, когда люди научились делать конвейерную сборку и освоили массовое производство. На наших глазах происходит **четвертая** промышленная революция, когда новая экономика будет построена на знаниях, которые получают люди. В связи с этим приобретает огромную актуальность профессия **инженера по знаниям (Data Scientist)**. Ожидается, что в ближайшем будущем будет требоваться огромное количество выпускников с навыками анализа данных.

Введение в машинное обучение

Машинное обучение – это область искусственного интеллекта, связанная с созданием алгоритмов, способных самостоятельно строить модели, обучаться на большом количестве данных без жестко запрограммированных правил. Данная область ИТ в настоящее время очень популярна и **связана с цифровизацией экономики**. По оцифрованным данным можно проводить машинное обучение. Обычно решение принимает человек, а в случае машинного обучения эту функцию выполняет компьютер. Машинное обучение к тому же строит модели, которые могут предсказывать результаты каких-то будущих ситуаций и явлений.

Применяется машинное обучение (МО) в самых различных областях. Например, в безопасности для распознавания лиц и обнаружения вторжений в компьютерные сети; в медицине и в фармацевтике – начиная с автоматического выставления диагноза до получения информации какие вещества могут быть потенциально перспективными для создания новых лекарств от каких-то серьезных болезней; для предсказания продаж, создания рекламы, принятия различных решений, например, о выдаче кредита и т.д. Одна из основных областей применения МО – **обработка текстов**, как печатных, так и рукописных, а также распознавание голоса и перевод его в текст.

Введение в машинное обучение

ПРИМЕНЕНИЯ МАШИННОГО ОБУЧЕНИЯ

БАНКИ И ФИНАНСЫ

Решение о выдаче кредита
Торговля на бирже
Выявление мошенничества

ИНТЕРНЕТ

Контекстная реклама
Рекомендательные системы в онлайн-магазинах и социальных сетях

БИЗНЕС

Голосовые помощники
Классификация почты
Управление колл-центрами
Прогнозирование оттока клиентов

МЕТЕОРОЛОГИЯ

Прогнозирование стихийных бедствий
Прогноз погоды

БЕЗОПАСНОСТЬ

Распознавание лиц
Контроль безопасности дорожного движения

МЕДИЦИНА И ФАРМАЦЕВТИКА

Диагностика заболеваний
Выбор стратегии лечения
Выявление побочных эффектов лекарственных средств

ОБРАБОТКА ТЕКСТОВ

Распознавание печатного и рукописного текста
Проверка грамматики и орфографии
Анализ отзывов клиентов
Спам-фильтры

РИТЕЙЛ

Прогнозирование продаж
Предотвращение потерь

Этапы создания модели машинного обучения

1. **Сбор данных** (самый трудоемкий этап) может быть ручным, автоматическим или полуавтоматическим. При анализе текста разметка текста делается вручную.
2. **Выделение признаков (предикторов)** – подготовка данных для машинного обучения (делается часто вручную). **Краудсорсинг** – это метод сбора данных, когда разметка текста производится большим количеством людей, например, когда в браузере всплывает картинка и вас спрашивают что-то, чтобы проверить, что вы не робот (где пешеходный переход) и вы делаете разметку текста, которая в дальнейшем будет где-то использована.
3. **Выбор алгоритма обучения –** определяет качество и скорость обучения.
4. **Обучение модели** (в процессе применения решений множества сходных задач).

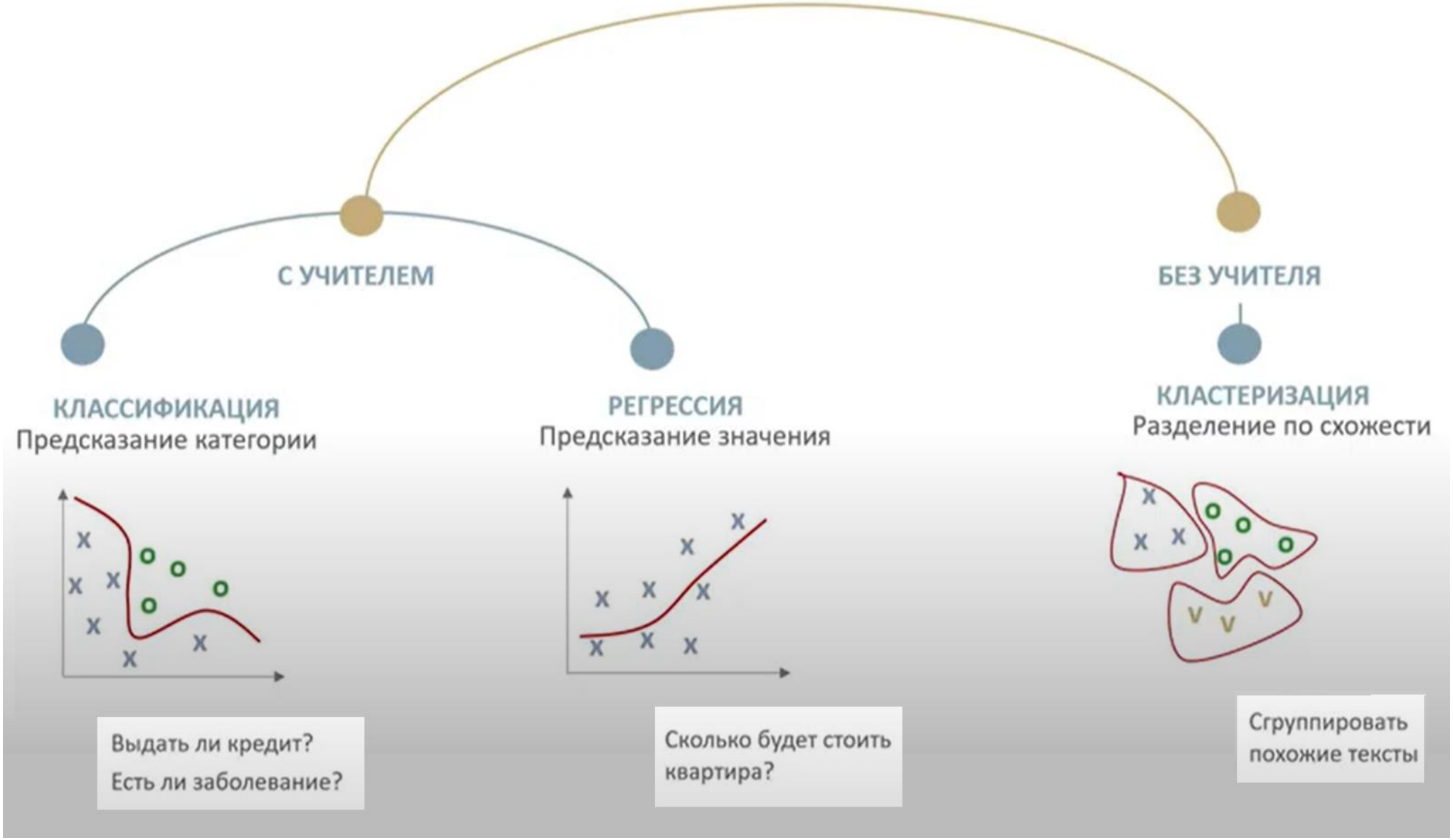
Обзор алгоритмов МО

Различают 2 типа алгоритмов МО:

1. **Классическое МО** – основано на статистических алгоритмах, развивается с 1950-х годов.
2. **Нейронные сети и глубокое обучение** – используют модель, вдохновленную устройством мозга, в 1980-х годах потерпели неудачу, так как не было достаточных вычислительных мощностей, активно развивается с 2012 года.

Эти алгоритмы решают одни и те же задачи, но классический подход требует извлечения признаков, а глубокое обучение получает признаки из текста автоматически. Глубокое МО – очень популярная на сегодняшний день тема, но **большинство задач** по-прежнему решаются методами **классических** алгоритмов МО, так как они имеют следующие **преимущества**: высокая скорость обучения, нужен меньший объем данных при обучении, больше стабильность алгоритмов при изменении данных и они легче интерпретируются человеком.

Алгоритмы классического МО



Алгоритмы классического МО

Обучение с учителем производится, если есть целевая переменная и мы знаем значение этой переменной на исторических данных. Для текстов также решаются **задачи классификации** (группировка текста на основе заранее определенного пользователем списка классов, например, спорт, политика, наука и др.) и **задачи группировки** (близких по содержанию текстов без предварительной информации). Чаще всего на текстах не решаются задачи предсказания численного решения.

Одним из самых популярных и самых старых методов классического МО является **метод деревьев решений**. Деревья решений симулируют процесс, который раньше производили люди. Например, сотрудник банка встречается с клиентом, который захотел получить займ и начинает задавать вопросы, исходя из своего опыта – есть ли у клиента недвижимость, какой у него доход, есть ли семья и т.д. На основе полученных ответов он принимает решение о надежности клиента и можно ли ему выдавать займ.

Классические алгоритмы МО

ДЕРЕВЬЯ РЕШЕНИЙ

Классические алгоритмы МО

Дерево решений автоматизирует процесс принятия решения. Система получает на исторические данные, изучает когда был выдан или не выдан кредит, получает параметры человека, например, наличие постоянной работы, среднегодовой доход, семейное положение, наличие недвижимости и т.д. Система находит по каким признакам можно разделить данные на 2 класса, причем количество “да” и “нет” в получившихся подмножествах данных должно быть максимизированно. Деревья решений могут использовать разные функции для расщепления (например, есть деревья решений, построенные на максимизации энтропии).

Решение системы, построенное на дереве решений, выдается с некоторой вероятностью, чем выше вероятность, тем лучше полученное решение. Таким образом можно решать задачи как бинарной классификации, например, выдать ли кредит, так и задачи множественной классификации, когда есть много классов и нужно определить к какой тематике относится текст, например, про экономику, политику, спорт и т.д. К тому же есть способы применения деревьев решений для предсказания численных значений.

Классические алгоритмы МО

Модели, построенные на основе алгоритма **дерева решений,** популярны до сих пор не смотря на то, что это один из самых старых алгоритмов МО, потому что они имеют очень серьезные **преимущества**:

1. Простота интерпретации и наглядность результатов для человека, здесь можно сформулировать четкие логические правила. Даже если результаты будут чуть менее точны, чем результаты, полученные с помощью глубокого обучения с применением нейронных сетей, здесь можно легко объяснить причины выбора решения системой.

2. Возможность работы как с категориями, так и с количественными значениями.

3. Высокая производительность при классификации по построенному дереву, т.е. если построена модель, то классификация по дереву решений производится очень быстро для новых данных.

Классические алгоритмы МО

Модели, построенные на основе **дерева решений,** имеют **недостатки**:

1. **Структура часто бывает нестабильна** и некоторые изменения в данных и переобучение может привести к тому, что дерево решений перестроится полностью. Например, для принятия решения о выдаче займа клиенту первым делом мы будем смотреть не на наличие недвижимости, а на его семейное положение. Поэтому надзорным органам придется объяснять почему так сильно изменилась система принятия решения.

2. **Сложно контролировать размер дерева решений** и это часто приводит к появлению лишних веток. Но есть специальные статистические алгоритмы, которые позволяют избежать этого.

3. Использование наилучшего бинарного вопроса для разбивки данных **не всегда ведет к точным предсказаниям**. Иногда для лучшего прогнозирования нужны менее эффективные первоначальные разбиения.

Классические алгоритмы МО

Чтобы обойти эти ограничения, можно избежать ориентации на лучшую разбивку данных и использовать различные варианты деревьев решений совместно. То есть можно получить более точные и постоянные результаты путем комбинирования прогнозов, полученных от различных деревьев решений.

Есть два способа сделать это:

1. Сначала различные комбинации бинарных вопросов для создания деревьев выбирают случайным образом, а затем полученные предсказания суммируются. Этот метод известен как построение **случайного леса**.

2. Бинарные вопросы выбираются стратегически, вследствие чего точность прогнозирования последовательно улучшается. Результатом становится взвешенное среднее значение, полученное при помощи всех деревьев решений. Этот метод называется **градиентным бустингом (gradient boosting)**.

Классические алгоритмы МО

Хотя случайные леса и градиентный бустинг позволяют делать более точные прогнозы, их сложность мешает визуализации, в связи с этим их прозвали **черными ящиками**. Это объясняет, почему популярным инструментом продолжают оставаться обычные деревья решений. Их наглядность упрощает оценку предикторов (признаков) и их взаимодействия.

Для того, чтобы **проверить результаты работы дерева решений**, данные разбивают на **обучающие** и **тестовые**. Обучающих берется порядка 70-80%, строится модель и эта модель применяется на оставшихся 20-30% данных, которые не использовались в процессе обучения. Если процент ошибок, которые делает построенная модель, не сильно отличается от процента ошибок, полученных на обучающей выборке, то данная модель **устойчива** и может применяться на практике. Если же процент ошибок получается гораздо больше на тестовых данных по сравнению с тем, что получали на обучающей выборке, то это сигнализирует о том, что произошло **переобучение**, т.е. мы использовали слишком много конкретных признаков и эта модель не имеет предсказательной силы.

Классические алгоритмы МО

Например, если мы хотим предсказывать купил или не купил товар такой-то покупатель и если мы в качестве одного из признаков возьмем паспортные данные покупателей, то мы сможем однозначно предсказать на обучающих данных, что люди с такими-то паспортными данными покупают наш товар. Но эта модель в будущем будет абсолютно бессмысленна, потому что в следующем потоке покупателей будут люди с другими паспортными данными и наша модель не сможет предсказать ровно ничего. Это пример того, как происходит **переобучение системы**, когда система использует слишком специфические данные. Проверка модели на тестовой выборке как раз исключает подобные переобучения.

Часто методы, которые исключают переобучение модели встроены в саму систему для построения моделей. Эти методы называются **кросс-валидацией**, когда система сама внутри себя выкладывает алгоритм, какое-то количество данных, сама учится на обучающей выборке и потом проверяет их на отложенных данных. Как правило, этот процесс происходит много раз и каждый раз откладываются другие непересекающиеся подмножества данных.

Классические алгоритмы МО

Еще один из очень популярных и довольно старых классических методов МО – это **Байесовский классификатор**. Он построен на информации из теории вероятности. **Теорема Байеса** – это одна из фундаментальных теорем теории вероятности, которая позволяет установить вероятность наступления сложного события А, если имеет место возникновение другого несвязанного с ним события В.

Классические алгоритмы МО

На основе Байесовского классификатора строили спам-фильтры, когда, например, мы пытаемся определить будет ли связано со словом “скидка” то, что данное сообщение будет отнесено к спаму или нет. Можно решить и другую задачу, если мы встретили в сообщении какое-то слово, например, слово “скидка”, то данное сообщение будет относиться к спаму. Спамеры научились **обходить этот алгоритм**, они стали добавлять в конце спам-сообщения несколько слов, которые являются легитимными и обманывают Байесовский классификатор, в результате чего спам проходит через подобные фильтры. В настоящее время придуманы гораздо более продвинутые фильтры, но Байесовский классификатор остался как памятник одному из самых успешных первоначальных применений МО.

Классические алгоритмы МО работают не с исходным “сырым” текстом, а с набором признаков, которые каким-то образом извлекли. Признак (или предиктор) – это информация, полученная из исходных данных, с которой работает МО, например, это может быть заболевание, которое есть у человека, возраст или цена товара, наличие семьи и др.

Классические алгоритмы МО

Малозначительные признаки в тексте часто нужно отсекать, чтобы облегчить работу и сузить размерность задачи. Например, человек звонит и сообщает о поломке, то алгоритм МО должен выбрать слова о том, что сломалось, каким образом сломалось, …

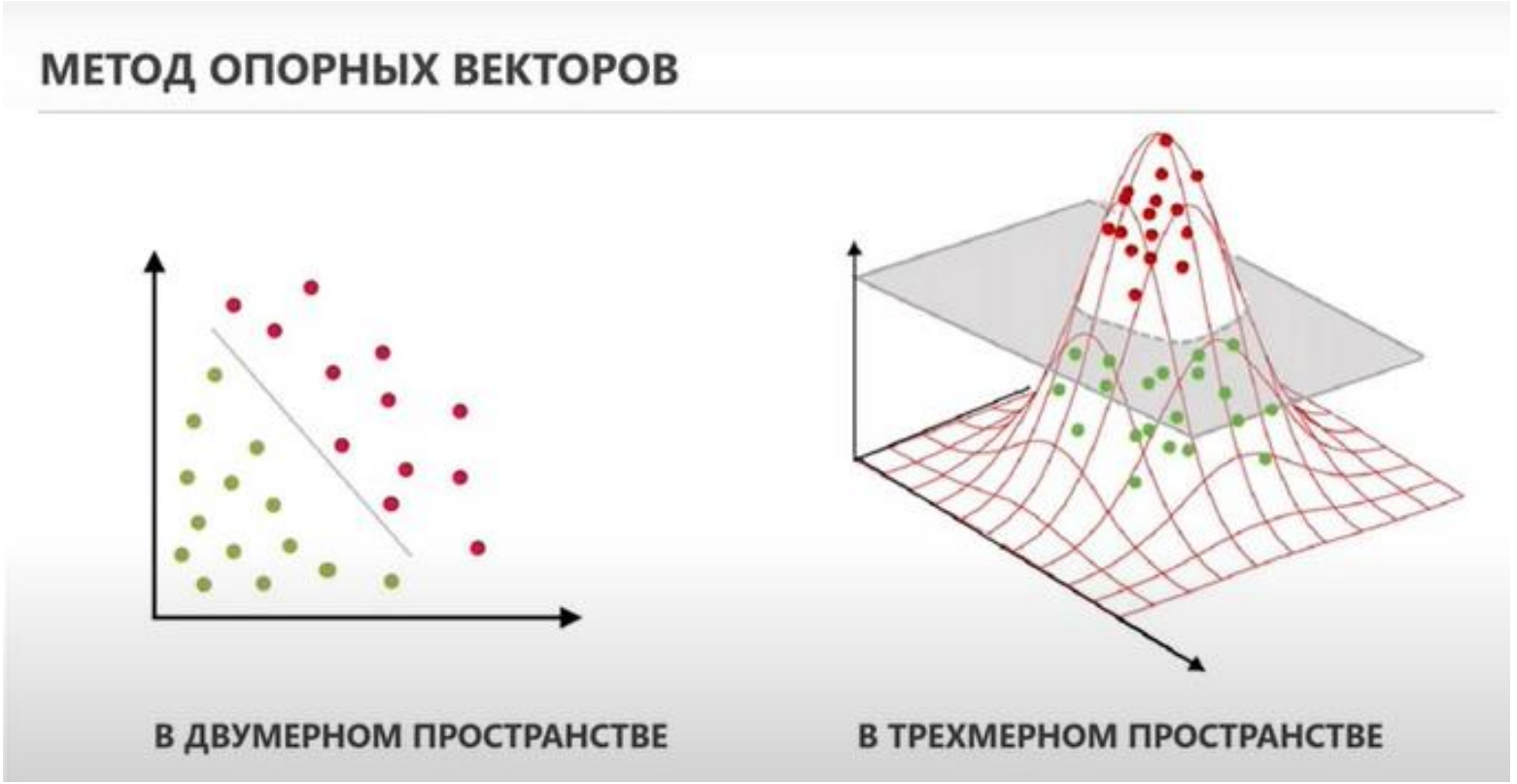
В случае **анализа текстов в МО** часто приходится делать **предобработку данных**, чтобы получить хорошие результаты с помощью МО. Часто одно и тоже слово встречается в разных падежах и для того, чтобы машина поняла, что это одно и то же слово, нужна некоторая обработка.

Есть разные типы предобработки текста:

- текст нужно привести к единому регистру;
- нужно удалить стоп-слова, т.е. слова не несущие нагрузки, например, предлоги, союзы, …
- нужно привести все слова к начальной форме, т.е. в именительный падеж.

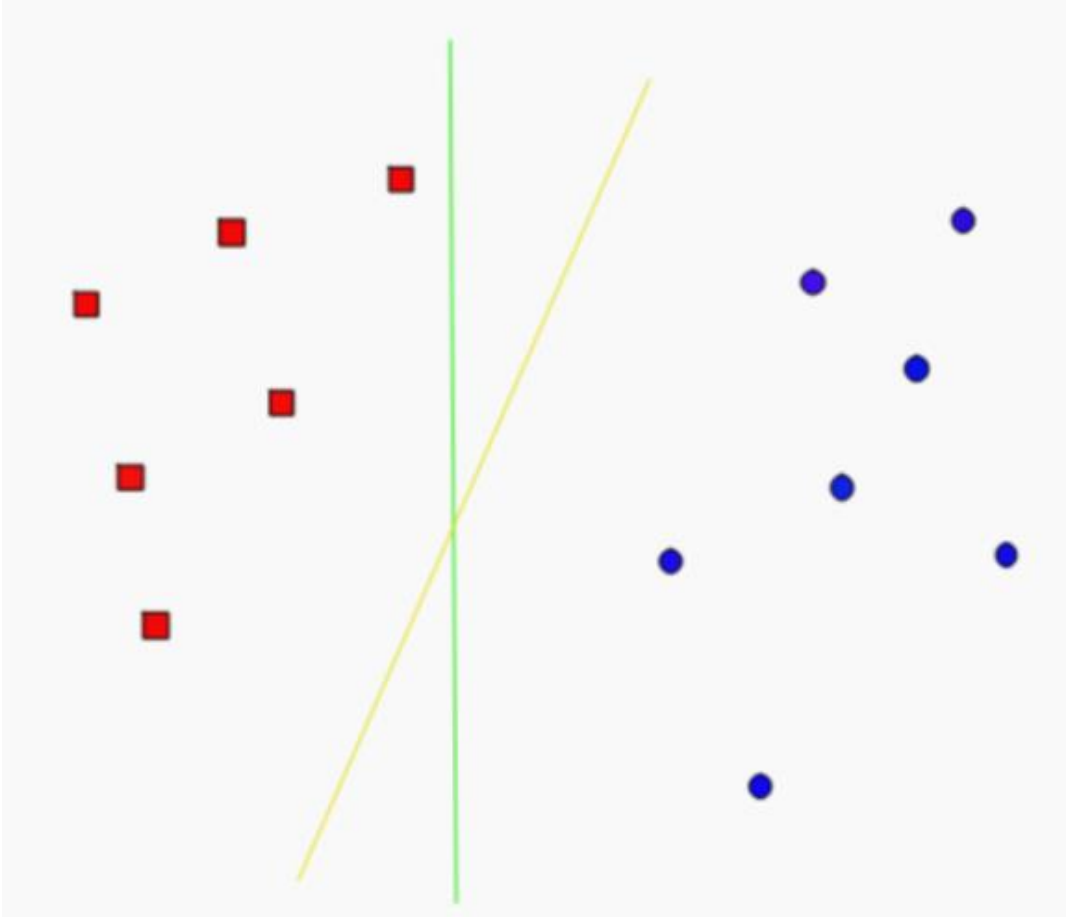
Классические алгоритмы МО

Основная идея **метода опорных векторов** состоит в увеличении размерности задачи, на которую мы смотрим, для того, чтобы точки, которые относятся к разным группам, могли быть разделены гиперплоскостью наиболее оптимальным образом. **Гиперплоскость** – это объект, у которого размерность на единицу меньше, чем размерность пространства, в котором мы смотрим. Например, для 3-мерного пространства гиперплоскость – это двумерная плоскость, а для двумерной плоскости – это просто прямая.



Классические алгоритмы МО

Метод опорных векторов или **SVM** (Support Vector Machines) – это алгоритм МО, широко используемый на практике для решения задач **классификации**, он выявляет оптимальную границу для разделения данных на две группы. Это не так просто, как кажется, поскольку возможных вариантов очень много.



Какая из двух линий (желтая или зеленая) лучше разделяет два класса и является более оптимальной?

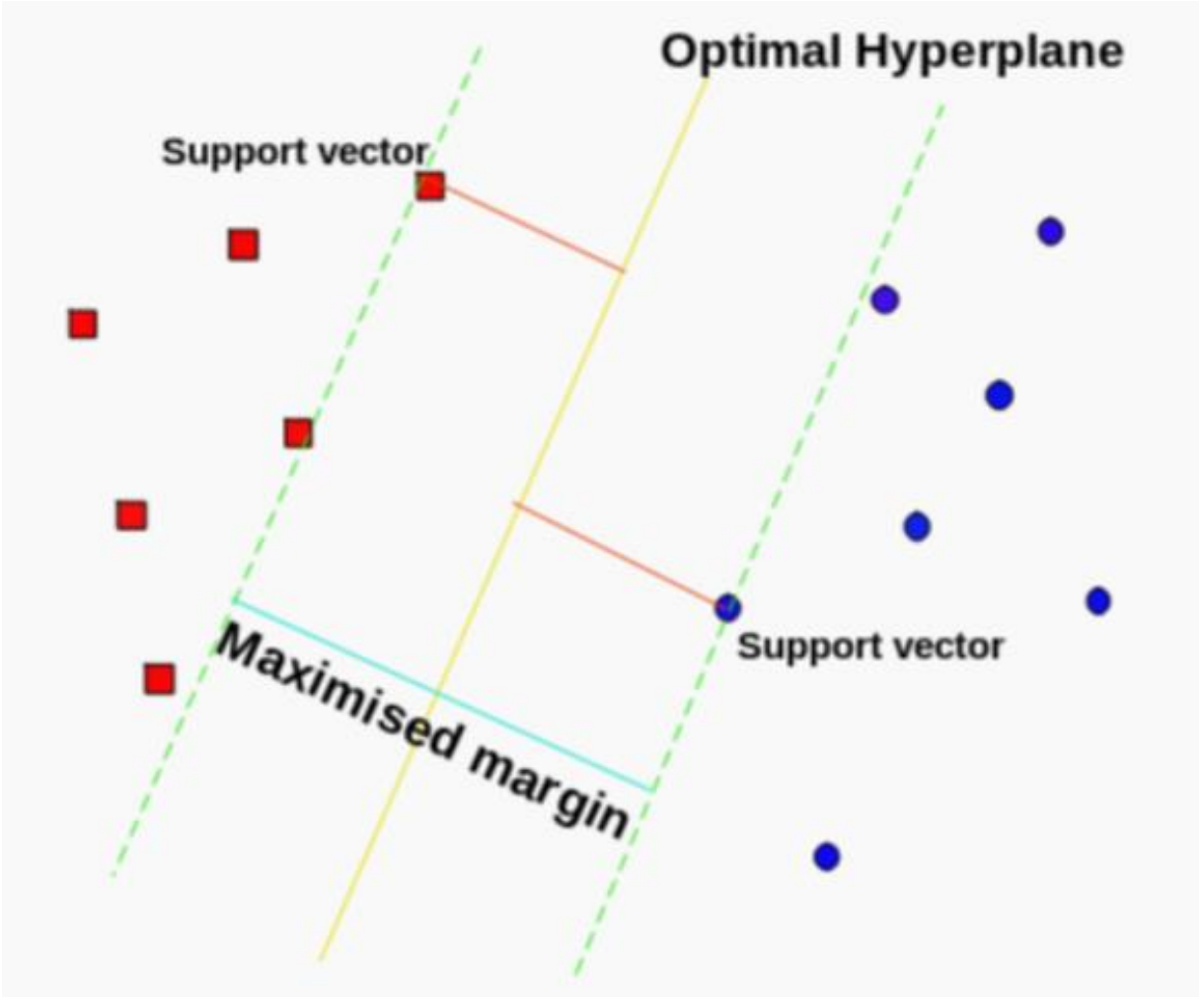
Метод опорных векторов (SVM)

Чтобы найти оптимальную линию разграничения, нужно сначала найти периферийные элементы данных, которые находятся ближе всего к противоположной группе. Оптимальная граница проводится посередине между такими периферийными элементами данных в обеих группах. Поскольку эти элементы данных помогают обнаружить оптимальную линию разграничения, то их называют **опорными векторами**.

Алгоритм SVM вычисляет **зазор** – расстояние между опорными векторами и разделяющей плоскостью. **Лучшей гиперплоскостью** считается такая гиперплоскость, для которой этот зазор является максимально большим. Таким образом **основная цель** алгоритма SVM – максимизировать расстояние зазора.

В рассмотренном ранее примере алгоритм SVM выберет желтую линию, поскольку она создает больший зазор, чем зеленая линия.

Метод опорных векторов (SVM)



В рассмотренном примере мы можем и интуитивно понять, что более оптимальной границей разделения данных на две группы является желтая линия.

Метод опорных векторов (SVM)

Одно из **преимуществ** метода опорных векторов – это высокая **скорость** вычисления, поскольку линия разграничения определяется только по периферийным элементам данных. Тем не менее эта манера опираться на отдельные элементы данных имеет и оборотную сторону. Разделительная граница оказывается чувствительной к положению опорных векторов, а значит, слишком зависит от используемого набора данных. Более того, элементы данных редко делятся так ровно, как было показано на рисунке, в реальности они часто перекрываются.

Еще одно существенное **достоинство** метода состоит в его способности **обнаруживать** в данных **криволинейные паттерны** (регулярности, закономерности). Вместо того чтобы сразу прочерчивать границу на плоскости данных, метод опорных векторов сначала проецирует их на дополнительное измерение, которое может быть определено прямой линией. Эти прямые линии легче как вычислять, так и преобразовывать в кривые при возврате к изначальной размерности.

Метод опорных векторов (SVM)

Хотя **метод SVM** является **адаптивным** и **быстрым** инструментом, но он **может не подходить** в следующих случаях:

1. **Малые наборы данных** (поскольку для определения границ метод опирается на опорные векторы, то небольшой набор данных сокращает их число и отрицательно влияет на точность расчета).
2. **Множество групп** (метод способен классифицировать данные только на две группы за раз, а если групп три и более, то необходимо применять итеративно для выявления каждой отдельной группы метод, который называется *многоклассовая классификация – multi-class SVM*).
3. **Большое перекрытие данных** (метод классифицирует элементы данных исходя из того, с какой стороны границы разграничения они оказались, но если элементы данных сильно перекрываются обеими группами, то те из них, которые находятся ближе к границе, могут быть классифицированы ошибочно; более того, метод не дает информации о вероятности ошибочной классификации для отдельного элемента данных, можно лишь ориентироваться на расстояние от него до границы раздела).

Метод опорных векторов (SVM)

Метод опорных векторов (Support Vector Machine — SVM) – это очень мощная и универсальная модель машинного обучения, способная выполнять **линейную** или **нелинейную классификацию, регрессию и даже выявление выбросов**, она является одной из самых популярных моделей в МО. Особым свойством метода опорных векторов является непрерывное уменьшение эмпирической ошибки классификации и увеличение зазора, поэтому метод также известен как **метод классификатора с максимальным зазором**. Методы SVM особенно хорошо подходят для классификации сложных, но не очень больших наборов данных.

Способность метода SVM работать с несколькими измерениями обеспечивает его популярность в анализе наборов данных со множеством переменных. Его нередко **применяют для** решения таких сложных задач машинного обучения, как расшифровка генетической информации, анализ тональности текста, определение пола человека по фотографии, вывод рекламных баннеров на сайты и др.

Метод K-ближайших соседей

Метод K-ближайшего соседа (англ.: *k-nearest neighbors method*, *k-NN*) – один из методов решения задачи классификации.

Предполагается, что уже имеется какое-то количество объектов с точной классификацией (т.е. для каждого них точно известно, какому классу он принадлежит). Нужно выработать *правило, позволяющее отнести новый объект к одному из возможных классов* (т.е. сами классы известны заранее).

В основе k-NN лежит следующее правило: *объект считается принадлежащим тому классу, к которому относится большинство его ближайших соседей.* Под «соседями» здесь понимаются объекты, близкие к исследуемому в том или ином смысле.

Заметим, что здесь необходимо уметь определять, насколько объекты близки друг к другу, т.е. уметь измерять «расстояние» между объектами. Это не обязательно евклидово расстояние. Это может быть мера близости объектов, например, по цвету, форме, вкусу, запаху, интересам (если речь идёт о формировании групп людей), особенностям поведения и т.д. Следовательно, для применения метода k-NN в пространстве признаков объектов должна быть введена некоторая *метрика* (т.е. *функция расстояния*).

Метод K-ближайших соседей

Предполагается, что объекты с близкими значениями одних признаков будут близки и по другим признакам (т.е. относиться к одному и тому же классу).

Рассмотрим работу метода k-NN на простом примере.

Продукт	Сладость	Хруст	Класс
яблоко	9	8	фрукт
бекон	1	4	протеин
банан	10	1	фрукт
…	…	…	…

Здесь качества продуктов (сладость и хруст) оцениваются по 10-балльной шкале. Эти значения можно рассматривать как координаты точек (продуктов) в 2-мерном пространстве.

По оси будем откладывать степень сладости продукта, по оси ординат – степень хруста.

Получим график, изображённый на Рис.

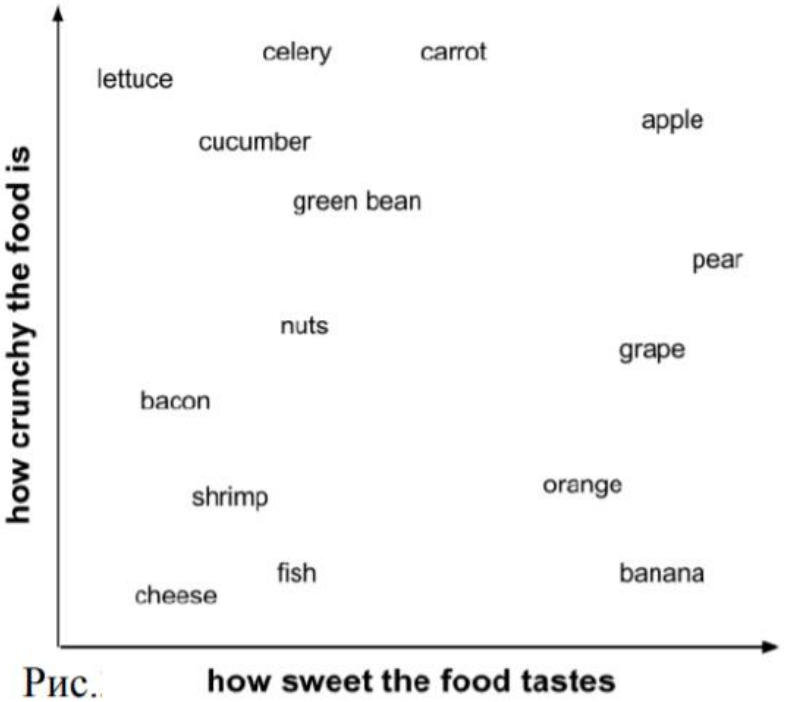


Рис.

Для каждого из названных продуктов мы точно знаем тип (класс) – см. последний столбец таблицы.

Можно заметить, что точки (продукты) на графике можно разбить на классы:

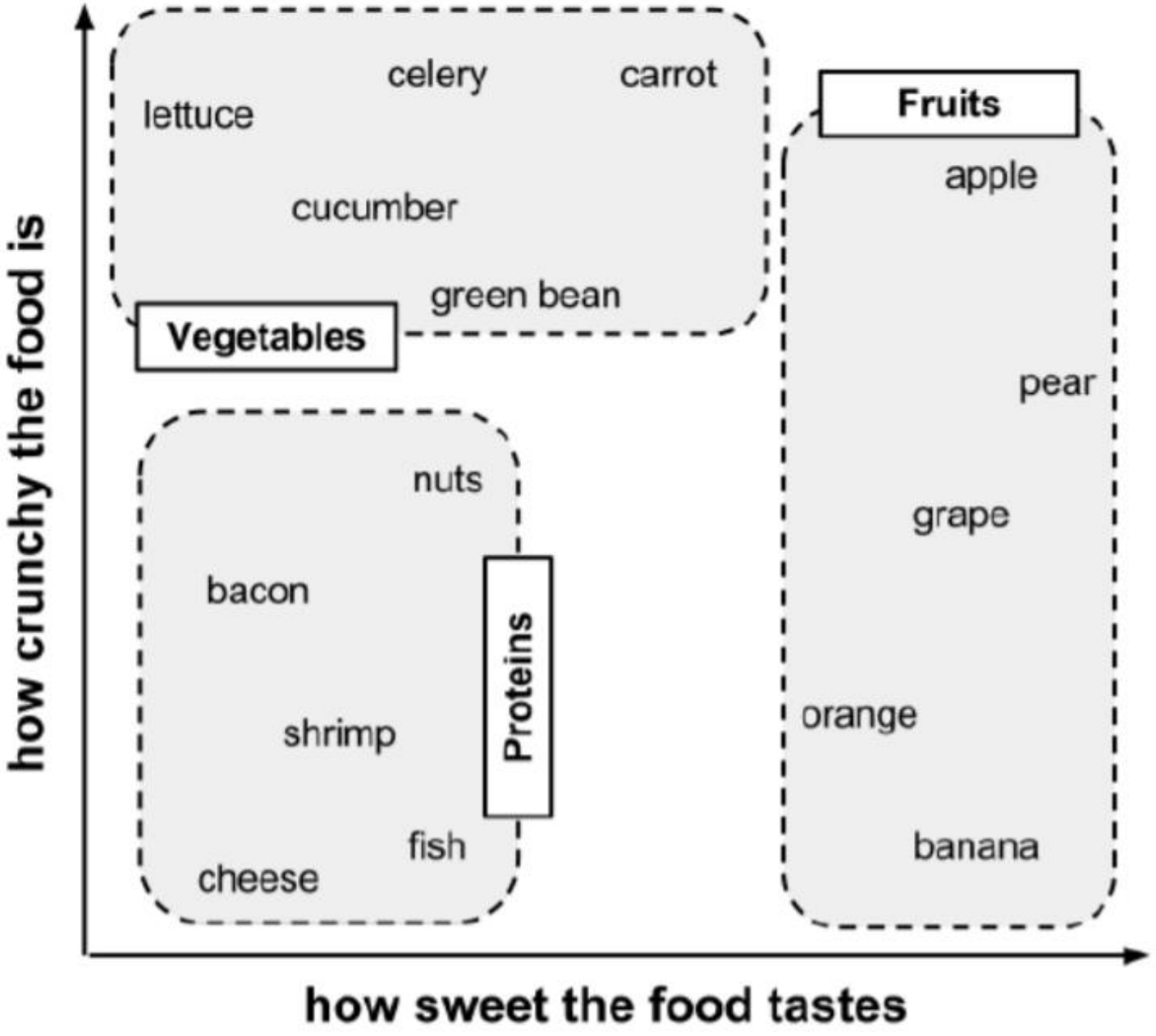
– в левом верхнем углу «группируются» овощи (огурец, морковь, салат-латук, сельдерей) – они хрустящие и несладкие,

– в левом нижнем – продукты, богатые протеином (бекон, креветки, сыр, рыба, орехи) – они нехрустящие и несладкие,

– справа «выстроились» фрукты (яблоко, груша, виноград, апельсин, банан) – они сладкие по сравнению с другими классами, но неоднородны в отношении хруста.

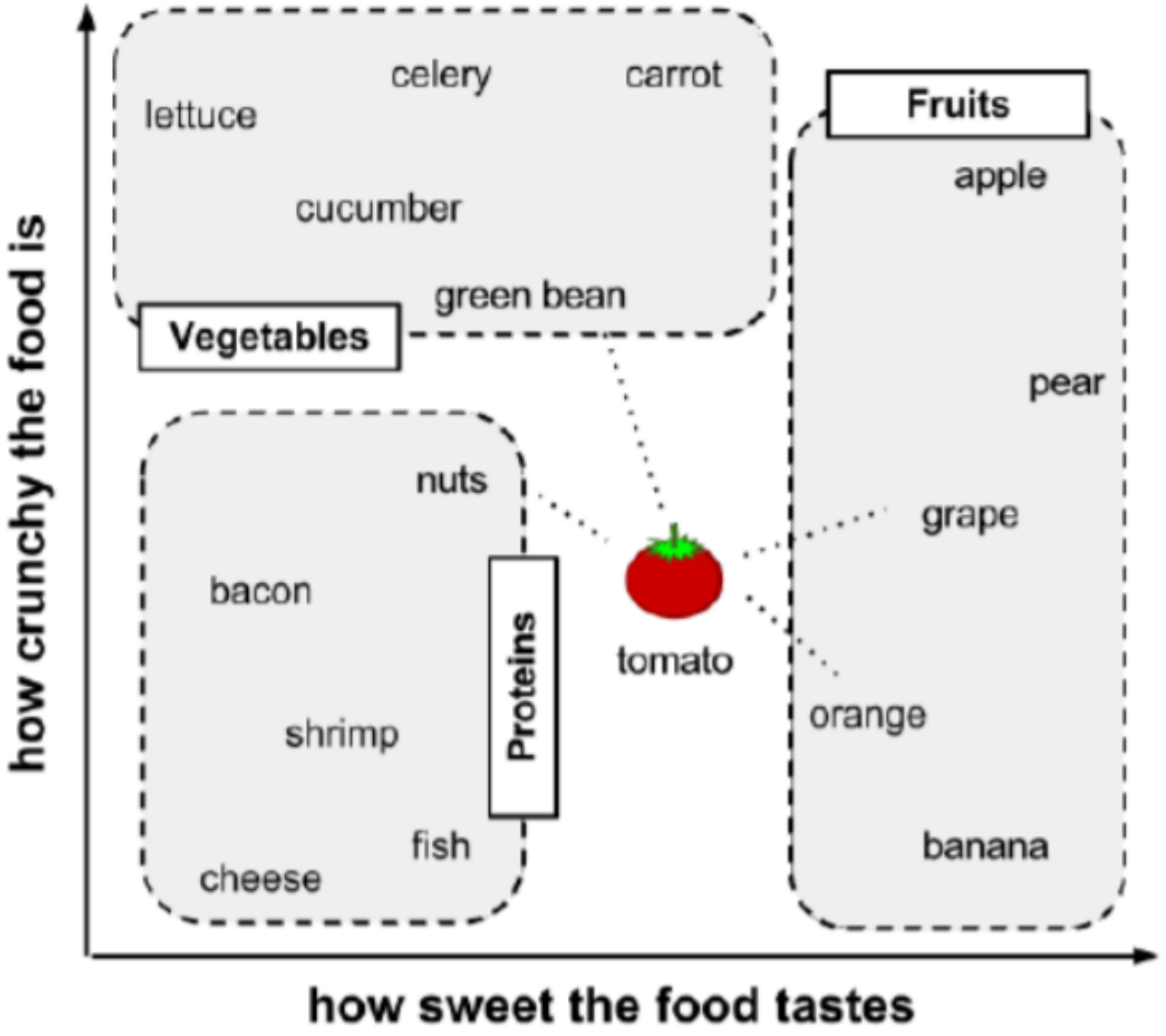
Метод K-ближайших соседей

Разбиение на классы показано на Рис.



Метод K-ближайших соседей

Предположим теперь, что нам предложен новый продукт, и мы должны определить, к какому классу он относится – см. Рис.



Метод K-ближайших соседей

Согласно методу k-NN мы отнесём его к тому классу, к которому принадлежит большинство из k его ближайших соседей. Расстояние между объектами будем понимать в смысле *Евклидовой нормы*, т.е. расстояние между объектами с координатами (x_1, y_1) и (x_2, y_2) равно $\sqrt{((x_1 - y_1)^2 + (x_2 - y_2)^2}$.

Так, если у томата показатель сладости равен 3, а показатель хруста равен 7, то его расстояния до яблока, бекона и банана равны примерно 6,1; 3,6, 9,2, соответственно. Приведём расстояния от томата до всех остальных продуктов, упорядочив их по возрастанию:

№	Продукт	Класс	Расстояние до томата
1	апельсин	фрукт	1,4
2	виноград	фрукт	2,2
3	креветка	протеин	3,5
4	бекон	протеин	3,6
5	орехи	протеин	3,6
6	сыр	протеин	4,0
7	бобы	протеин	4,2

№	Продукт	Класс	Расстояние до томата
8	огурец	овощ	5,9
9	яблоко	фрукт	6,1
10	морковь	овощ	6,8
11	сельдерей	овощ	7,0
12	салат-латук	овощ	7,2
13	банан	фрукт	9,2

Метод K-ближайших соседей

Теперь нужно выбрать число k и определить, к какому классу принадлежат большинство из k ближайших соседей томата. Так, если k=1, то ближайший сосед – один, и это апельсин, он – фрукт. При k=2 это апельсин и виноград, оба фрукты. При k=3 мы имеем 2 фрукта (апельсин и виноград) и креветку (протеин). Значит, метод k-NN опять даст ответ: «фрукт». При k=4 ответом сбудет «фрукт или протеин с равной вероятностью». При k=5, k=6, k=7, k=8 «побеждает» протеин. Этот процесс можно продолжать и далее, увеличивая значение k. Мы видим, что *результат, получаемый методом k-NN, сильно зависит от выбора параметра k.*

Если мы выберем значение *k* слишком малым, то есть опасность, что единственным ближайшим объектом окажется «выброс», т.е. объект с неправильно определённым классом, и он даст неверное решение. Казалось бы, увеличивая значение параметра *k*, мы снижаем вероятность случайного попадания на такие «выбросы» в качестве ближайших соседей исследуемого объекта. Но здесь возникает другая опасность. Чтобы понять в чём она заключается, рассмотрим случай, когда *k* равно общему числу объектов *N*. Понятно, что тогда «победит» самый популярный (модальный) класс, и расстояние до исследуемого объекта не будет играть вообще никакой роли. Проблему выбора оптимального значения параметра k называют «*bias-variance tradeoff*», т.е. «компромисс между «выбросами» и дисперсией». На практике чаще всего полагают $k=[\sqrt{N}]$. Т.е. в нашем примере k=3 и результатом классификации будет то, что *помидор – фрукт*.

Метод K-ближайших соседей

В том случае, если мы уверены в «чистоте» выборки, мы можем выбирать k меньшим. Существует также приём под названием «*weighted voting*» (т.е. буквально «взвешенное голосование»), при котором более близкие соседи исследуемого объекта имеют больший вес, чем более дальние.

Рассмотрим ещё один аспект применения метода k-NN – предварительную подготовку данных.

Подготовка данных для применения метода k-NN

Заметим, что в рассмотренном примере оба признака (уровень сладости и хруста продуктов) измерялись в одной шкале – принимали значения от 0 до 10. На практике различные признаки могут иметь разные единицы измерения и разные шкалы, что может существенно искажать реальное расстояние между объектами. Для решения этой проблемы перед применением метода k-NN производят так называемую *нормализацию* (или *масштабирование*) данных (англ.: *scaling*).

Существуют различные способы нормализации. Приведём некоторые наиболее часто используемые: $$x_i \equiv \frac{x_i - x_{min}}{x_{max} - x_{min}}. \quad (1)$$

Формула (1) означает переход от абсолютных значений признаков к относительным. Преимущество новых переменных состоит в том, что они принимают значения от 0 до 1 (или, если перейти к процентному выражению, то о 0 до 100).

Метод K-ближайших соседей

Второй способ масштабирования имеет вид:

$$x_i \equiv \frac{x_i - \bar{x}}{s}, \qquad (2)$$

где $\bar{x}$ – выборочное среднее (т.е. $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$), s – выборочное среднеквадратическое отклонение (т.е. s =. $\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2}$.

Как известно, если с.в. ξ имеет нормальное распределение с параметрами μ и σ, то с.в. $\eta \equiv \frac{\xi - \mu}{\sigma}$ также является нормально распределённой, но параметры её распределения равны 0 и 1, соответственно (такие с.в. называются *стандартными* гауссовыми).

Не все признаки имеют количественное выражение. В этом случае прибегают к так называемому *dummy coding*. Например, значение признака «пол» можно обозначить 1 для мужчин и 0 для женщин.

Кластеризация методом K-средних

Кластеризация методом *k*-средних – это способ сгруппировать вместе похожие элементы данных по *k* кластерам. Действие алгоритма таково, что он стремится минимизировать суммарное квадратичное отклонение точек кластеров от центров этих кластеров. Для группировки элементов данных сначала каждый из них соотносится с кластером, а потом обновляются позиции кластерных центров. Эти два шага повторяются до тех пор, пока изменения не будут исчерпаны. Кластеризация методом *k*-средних лучше работает для сферичных, непересекающихся кластеров.

Определив общие предпочтения или характеристики, этим методом можно разделить клиентов на группы, которые затем можно использовать для таргетированной рекламы. Однако определение таких групп – хитроумная задача. Мы изначально можем не знать, как следует группировать клиентов и сколько групп существует.

Чтобы определить кластеры клиентов с помощью кластеризации методом *k*-средних, нам потребуется **информация о клиентах**, которую можно соизмерять. Общая переменная – это доход. Группы с высоким доходом более склонны приобретать продукцию известных брендов, чем с низким. В итоге магазины смогут использовать эту информацию, чтобы адресовать рекламу дорогих товаров группам с высоким уровнем дохода.

Кластеризация методом K-средних

Особенности личности тоже хороший способ группировки клиентов, который лучше показать на примере пользователей Facebook.

Пользователей пригласили пройти опрос, чтобы распределить их, исходя из четырех свойств: *экстраверсии* (насколько им в радость социальные взаимодействия), *добросовестности* (насколько они трудолюбивы), *эмоциональности* (как часто они испытывают стресс) и *открытости* (насколько они восприимчивы к новому).

Первичный анализ показал наличие связи между этими личностными особенностями. Добросовестные люди обычно более экстравертны. Кроме того, хотя это проявляется в меньшей степени, но высокоэмоциональные люди имеют тенденцию быть более открытыми. Поэтому для лучшей визуализации этих свойств мы их объединили (добросовестность с экстраверсией, эмоциональность с открытостью) путем сложения очков для каждой пары. После этого мы получили двумерный график.

Кластеризация методом K-средних

Рис. Персональные профили кинозрителей

Кластеризация методом K-средних

Суммарные очки черт характера были соотнесены с информацией о связанных с фильмами страницах, которые пользователь лайкнул на Facebook. Это дало нам возможность соотнести группы кинозрителей с профилями. На рис. мы видим два больших кластера.

- **Светлый:** добросовестные экстраверты, которым нравятся боевики и романтические фильмы.
- **Темный**: эмоциональные и открытые люди, которым нравится авангардное кино и фэнтези.

Фильмы посередине, по-видимому, фавориты семейного просмотра.

Обладая такой информацией, можно разработать таргетированную рекламу. Если зрителю нравится *«Славные парни»*, то владелец магазина может порекомендовать ему другой фильм этого жанра или даже продавать такие фильмы вместе, предложив скидку.

При определении кластеров нам нужно ответить на два вопроса:

1. Сколько кластеров существует?
2. Что включают в себя кластеры?

Кластеризация методом K-средних

Сколько же кластеров существует? Это субъективно, но по мере возрастания численности кластеров члены каждого из них становятся больше похожи друг на друга, и соседние кластеры при этом становятся менее различимы. Если довести это до крайности, то каждый элемент данных окажется в отдельном кластере, что не даст нам никакой полезной информации. Поэтому нужен компромисс. **Число кластеров** должно быть достаточно велико, чтобы позволить нам выявить важные закономерности, но не слишком, чтобы кластеры сохраняли отчетливые различия.

Одним из способов определить **оптимальное количество кластеров** является использование так называемого **графика каменистой осыпи** или **графика Каттела** (**scree plot**).

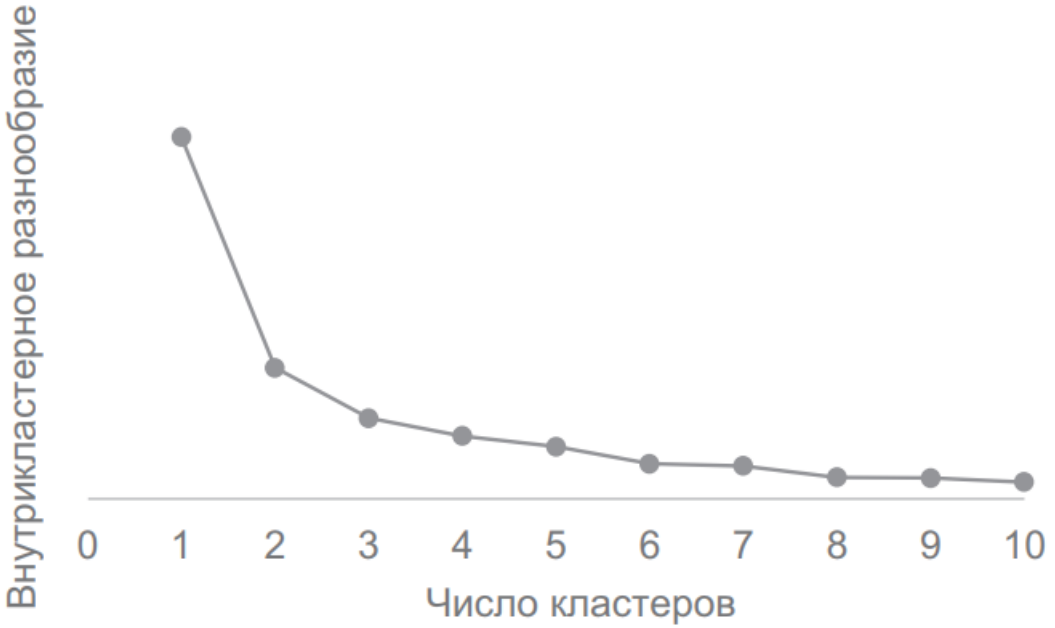


Рис. График осыпи показывает изломы, из которых следует, что оптимальное число кластеров от 2 до 3

Кластеризация методом K-средних

График осыпи показывает, насколько снижается разнообразие внутри кластеров при увеличении их числа. Если все члены отнесены к единственному кластеру, то разнообразие максимально. Но по мере увеличения числа кластеров сами они становятся плотнее, а их члены однороднее.

Излом – это острый изгиб на графике осыпи, который предлагает **оптимальное число** кластеров, исходя из разумной степени внутрикластерного разнообразия. На рис. мы видим излом на двойке, которая соответствует двум кластерам. Другой излом, поменьше, находится на тройке, говоря о том, что для данного случая мы можем ввести третий кластер. А вот введение еще большего их числа для данного примера уже даст слишком малые кластеры, слабо отличающиеся друг от друга.

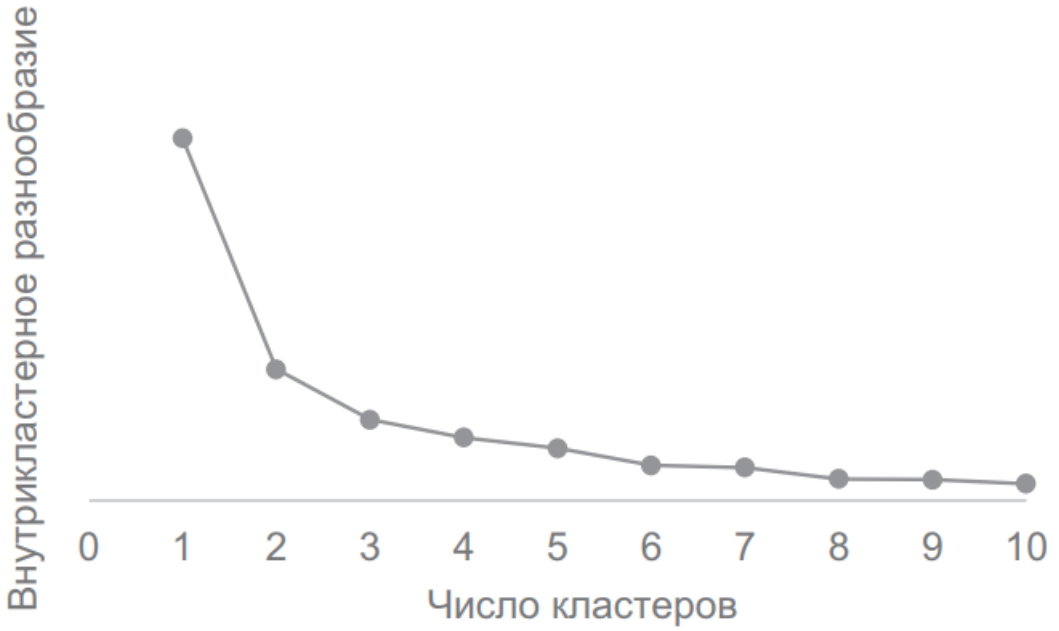
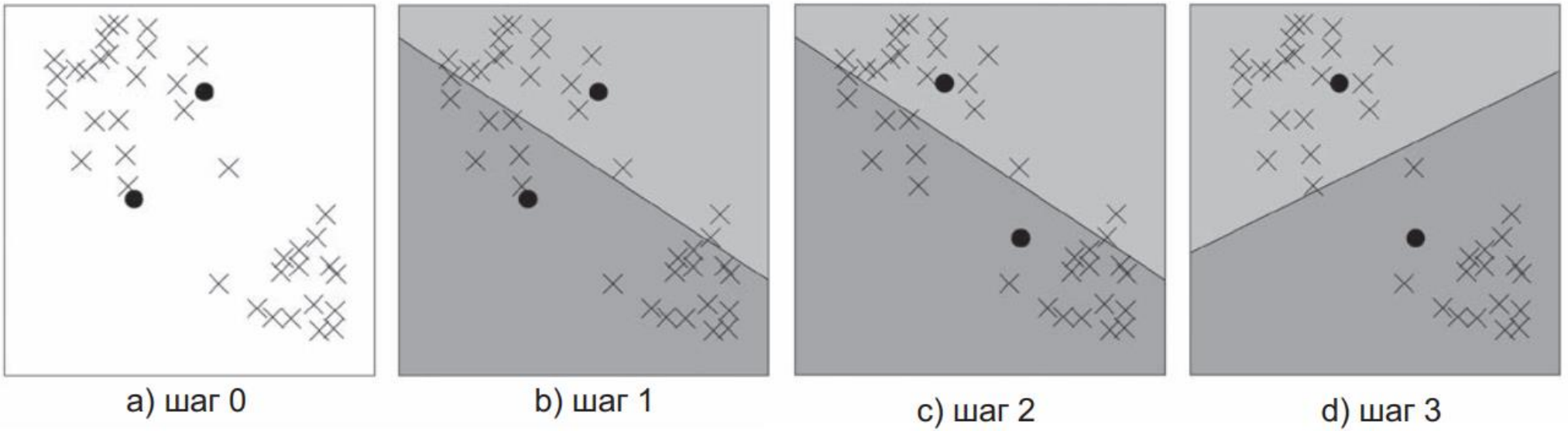


Рис. График осыпи показывает изломы, из которых следует, что оптимальное число кластеров от 2 до 3

Кластеризация методом K-средних

Что включают в себя кластеры? Данные распределяются по кластерам в итеративном процессе, показанном для двухкластерного примера на рисунке ниже.



Поскольку хороший кластер содержит похожие элементы данных, мы можем оценить его по тому, как далеко его члены находятся от центра. Но изначально позиции кластерных центров неизвестны, поэтому они берутся приблизительно. Затем элементы данных связывают с ближайшим к ним кластерным центром. После этого кластерный центр снова вычисляется для своих членов, а для элементов данных процедура повторяется, и если элемент данных окажется ближе к центру другого кластера, его членство будет изменено.

Кластеризация методом K-средних

Следующие шаги точно описывают процесс определения членства в кластере и могут использоваться при любом количестве кластеров.

Шаг 0: начать с предположения о том, где находятся центры кластеров. Условно можно назвать их псевдоцентрами, поскольку мы пока не знаем, соответствуют ли они в действительности центральному положению.

Шаг 1: связать каждый элемент данных с ближайшим псевдоцентром. Сделав это, мы получаем два кластера.

Шаг 2: вычислить новое положение псевдоцентров, ориентируясь на центр отнесенных к кластеру членов.

Шаг 3: повторять переназначение членов кластера (шаг 1) и его репозиционирование (шаг 2) до тех пор, пока все изменения в членстве не прекратятся.

Хотя здесь мы рассмотрели шаги для двумерного анализа, группирование в кластеры может быть также выполнено для трех и более измерений.

Кластеризация методом K-средних

Хотя кластеризация методом *k*-средних очень полезна, у нее есть **ограничения**:

- **Каждый элемент данных может быть связан только с одним кластером**. Иногда элемент данных находится ровно посередине между двух центров, отчего его включение в эти кластеры равновероятно.
- **Предполагается, что кластеры сферичны.** Итеративный процесс поиска ближайшего кластерного центра для элементов данных ограничен его радиусом, поэтому итоговый кластер похож на плотную сферу. Это может стать проблемой, если фактическая форма кластера, например, эллипс. Тогда кластер может быть усечен, а некоторые его члены отнесены к другому.
- **Кластеры предполагаются цельными.** Метод *k*-средних не допускает того, чтобы они пересекались или были вложены друг в друга.

Вместо принудительного назначения каждого элемента данных в единственный кластер можно воспользоваться более гибкими методами группировки, которые вычисляют то, с какой вероятностью каждый элемент данных может принадлежать другим кластерам, помогая нам находить несферические или пересекающиеся кластеры.

Метод главных компонент

Представьте, что вы диетолог. Как лучше всего дифференцировать пищевые продукты? По содержанию витаминов? Или белка? Или, может, по тому и другому?

Знание о переменных, которые лучше всего дифференцируют ваши данные, может иметь несколько **применений**:

- **Визуализация**. Отображение элементов на графике с подходящей шкалой может дать их лучшее понимание.
- **Обнаружение кластеров**. При хорошей визуализации могут быть обнаружены скрытые категории или кластеры. Например, если говорить о пищевых продуктах, мы можем выявить такие широкие категории, как мясо и овощи, а также подкатегории различных видов овощей.

Вопрос в том, как нам получить переменные, которые дифференцируют наши данные наилучшим образом.

Рис. Обычная пирамида питания

Метод главных компонент

Метод главных компонент (Principal Component Analysis, МГК) – это способ нахождения основополагающих переменных (известных как главные компоненты), которые дифференцируют ваши элементы данных оптимальным образом.

Эти главные компоненты дают наибольший разброс данных (см. рис.).

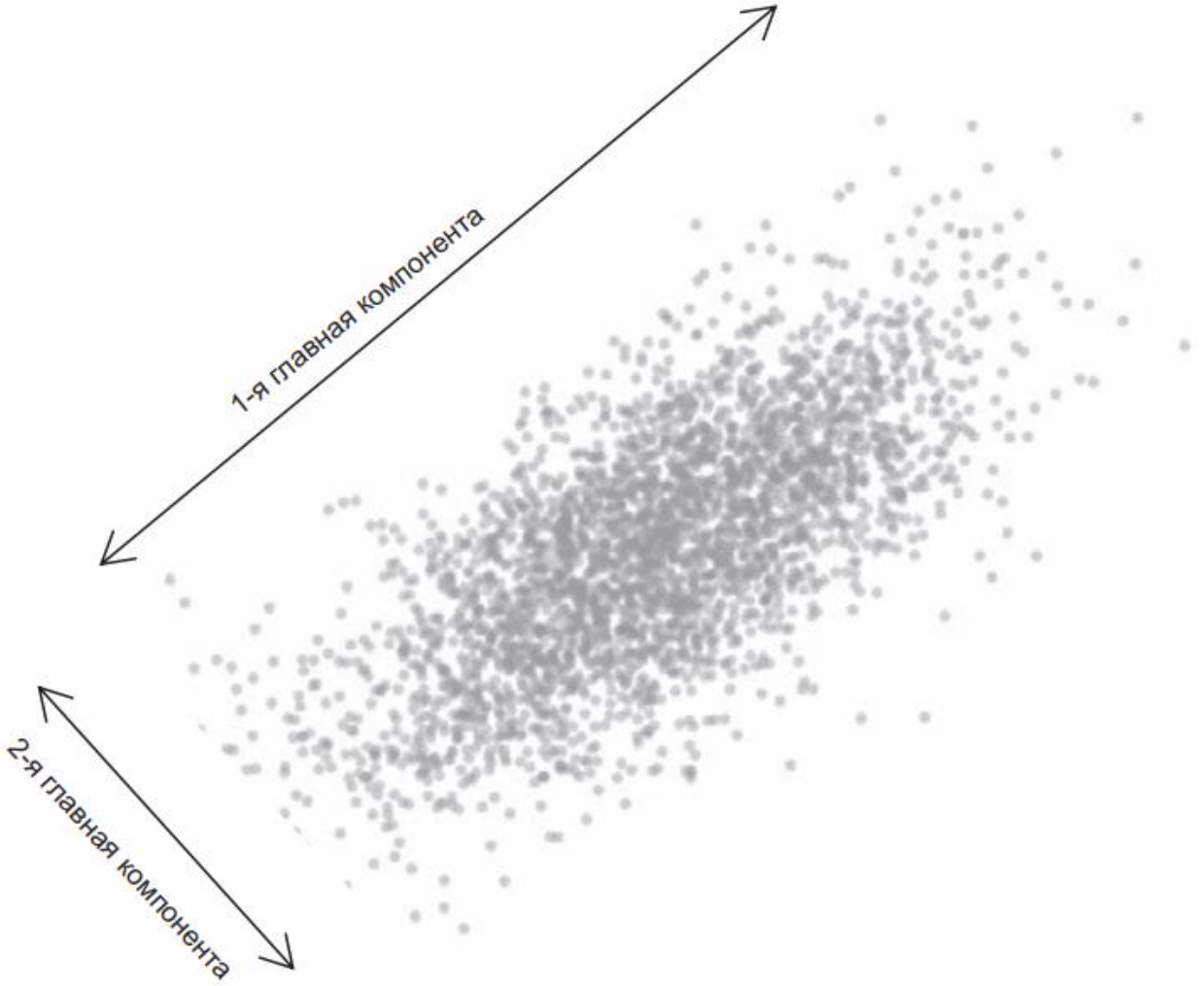


Рис. Визуальное представление главных компонент

Метод главных компонент

Главная компонента может выражать одну или несколько переменных.

Например, мы можем воспользоваться единственной переменной «Витамин С».

Поскольку витамин С есть в овощах, но отсутствует в мясе, итоговый график (левая колонка на рис.) распределит овощи, но все мясо окажется в одной куче.

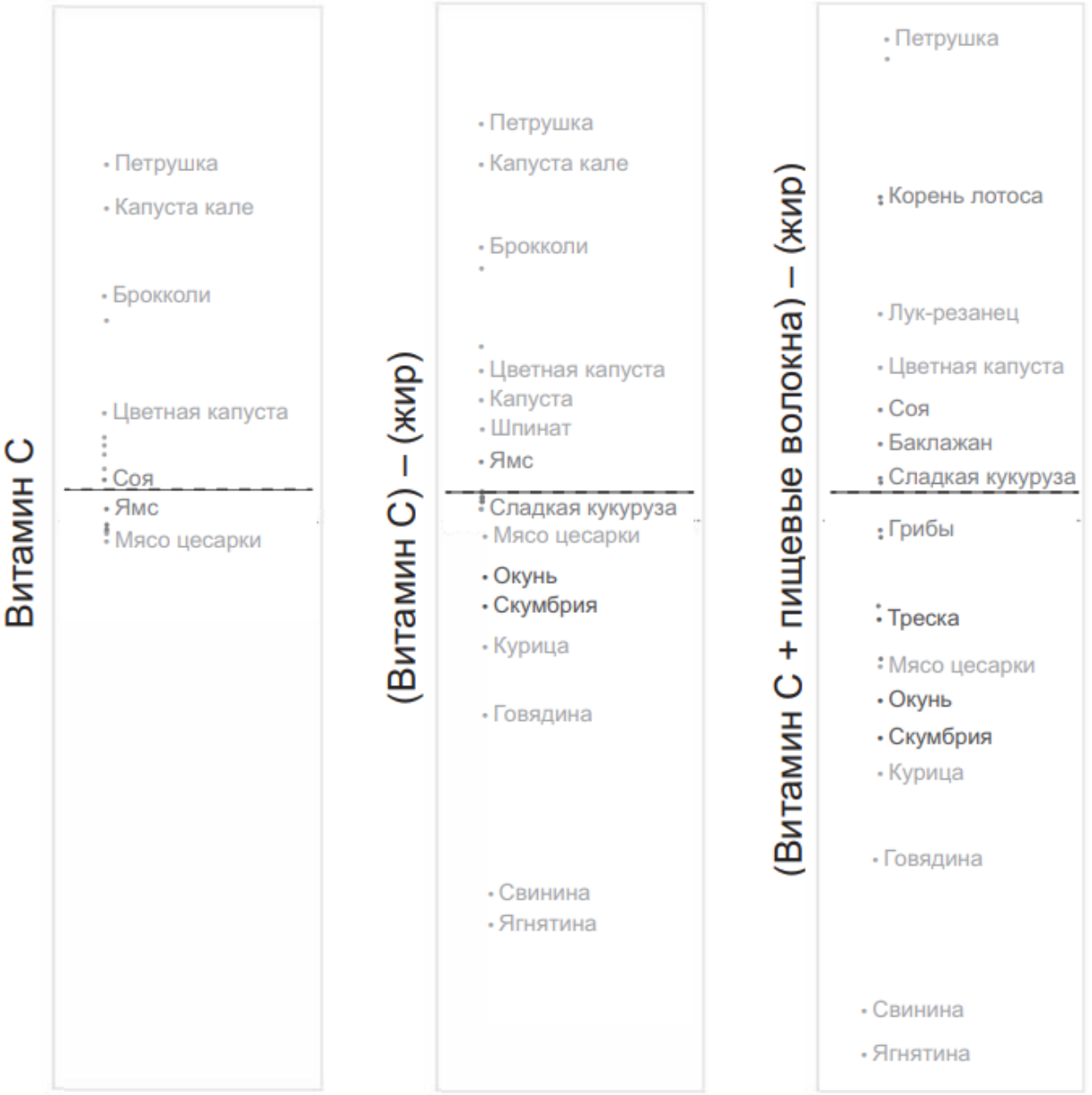


Рис. Пищевые продукты, распределенные разными комбинациями переменных

Метод главных компонент

Для распределения мясных продуктов мы можем использовать в качестве второй переменной жир, поскольку он присутствует в мясе, но его почти нет в овощах. Тем не менее, поскольку жир и витамин С измеряются в разных единицах, то прежде чем их скомбинировать, мы должны стандартизировать их.

Стандартизация – это выражение каждой переменной в процентилях, которые преобразуют эти переменные в единую шкалу, позволяя нам комбинировать их для вычисления новой переменной:

витамин С – жир

Поскольку витамин С уже распределил овощи вверх, то жир мы вычитаем, чтобы распределить мясо вниз. Комбинирование этих двух переменных поможет нам распределить как овощи, так и мясные продукты (столбец посередине).

Мы можем улучшить разброс, приняв во внимание пищевые волокна, содержание которых в овощах различается:

(Витамин С + пищевые волокна) – жир

Метод главных компонент

Мы можем улучшить разброс, приняв во внимание пищевые волокна, содержание которых в овощах различается:

(Витамин С + пищевые волокна) – жир

Эта новая переменная дает нам оптимальный разброс данных (см. правый столбец).

В то время как мы получили главные компоненты в этом примере методом проб и ошибок, МГК может делать это на системной основе. Мы увидим, как это работает, на следующем примере.

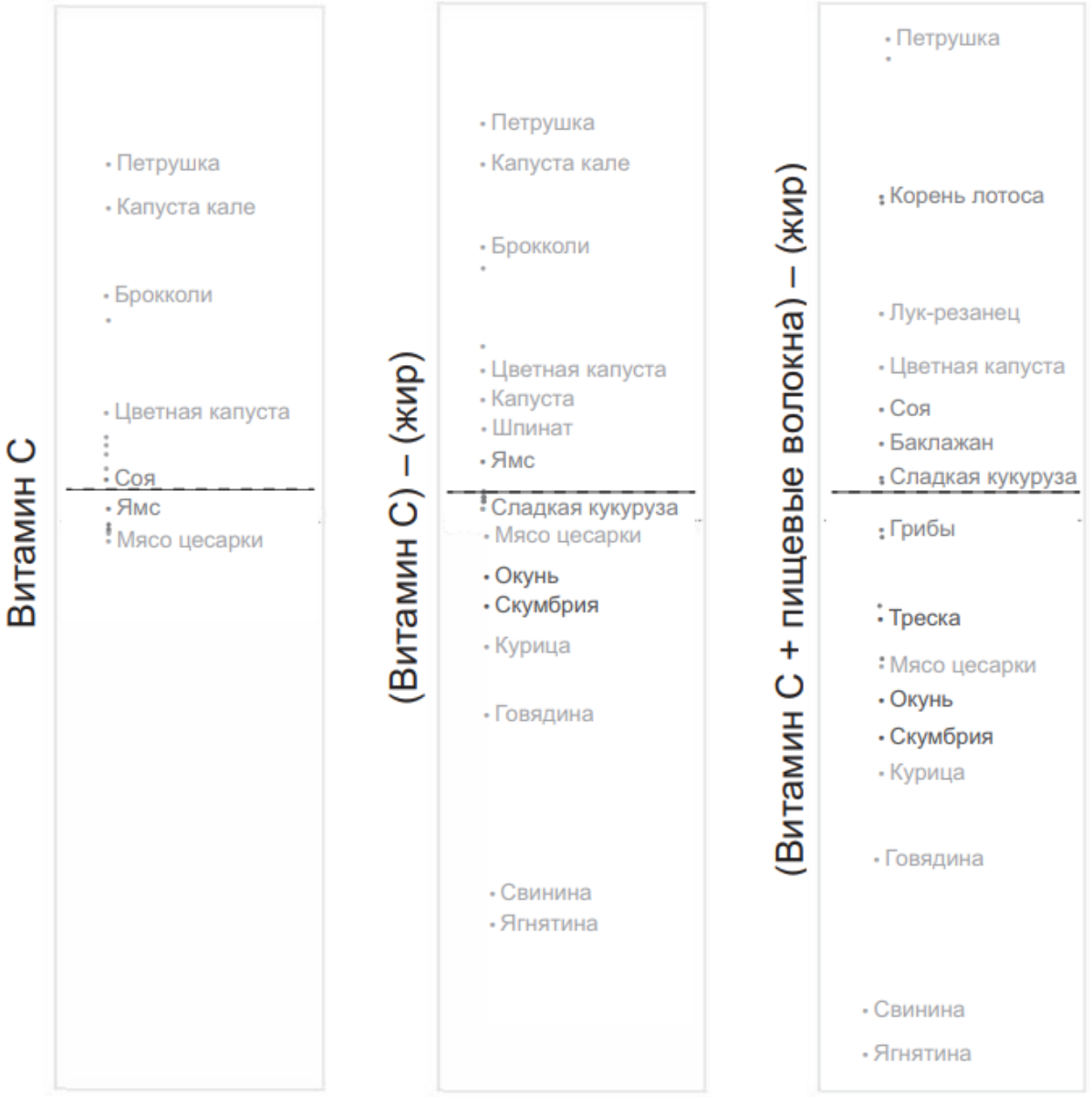


Рис. Пищевые продукты, распределенные разными комбинациями переменных

Метод главных компонент

Пример:

анализ пищевых групп

Используя данные Министерства сельского хозяйства США, были проанализированы питательные свойства случайного набора продуктов, рассмотрев четыре пищевых переменных: жиры, белки, пищевые волокна и витамин С. Как видно на рис., определенные питательные вещества часто встречаются в продуктах вместе.

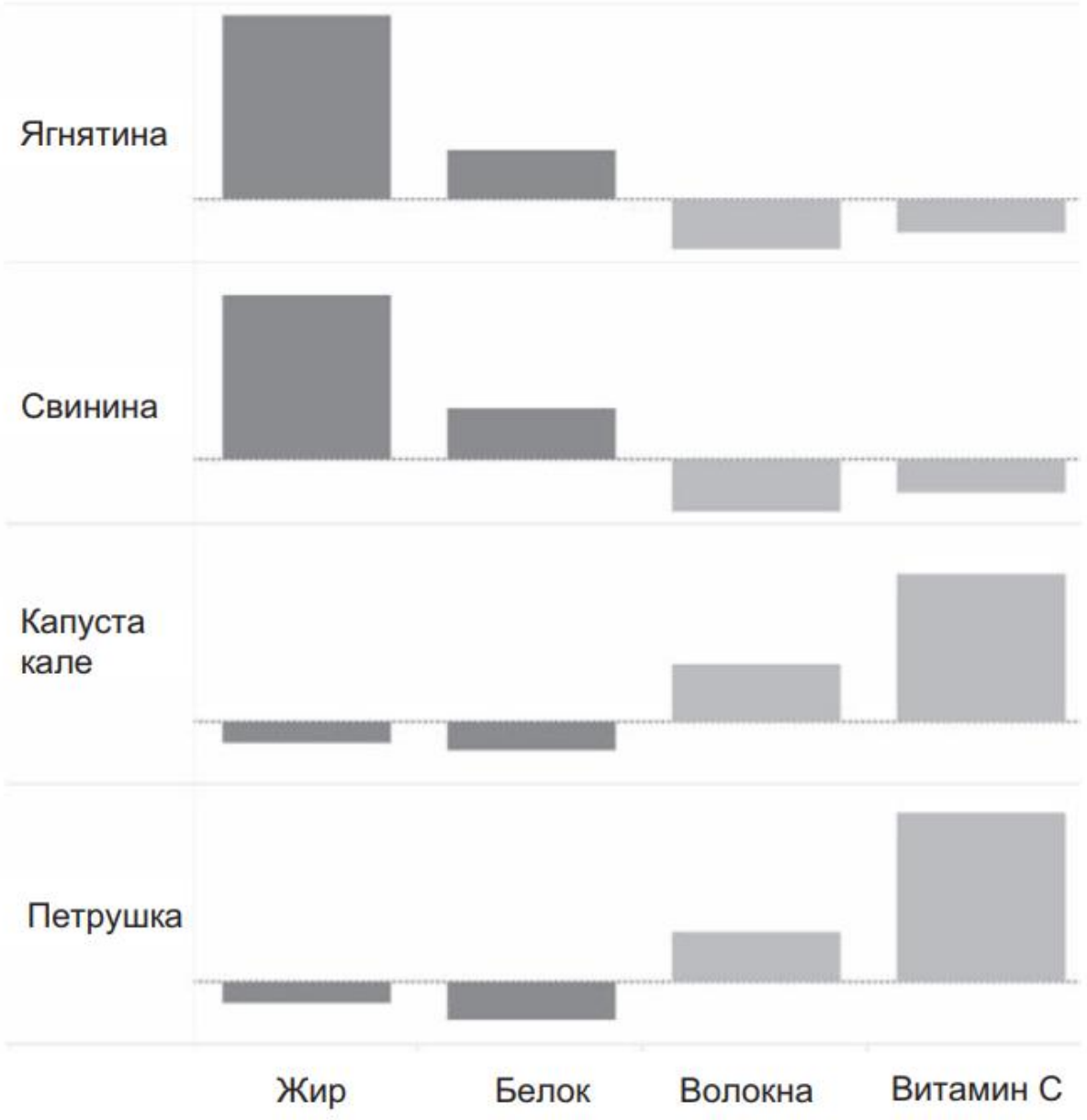


Рис. Сравнение пищевой ценности различных продуктов

Метод главных компонент

В частности, уровни содержания жиров и белков возрастают в одном направлении, противоположном тому, в котором растут уровни пищевых волокон и витамина С.

Мы можем подтвердить наши предположения, проверив, какие переменные коррелируют. И действительно, мы находим значимую положительную корреляцию как между уровнями белков и жиров (r = 0,56), так и между уровнями пищевых волокон и витамина С (r = 0,57).

Таким образом, вместо анализа четырех пищевых переменных по отдельности мы можем скомбинировать высококоррелирующие из них, получив для рассмотрения всего две. Поэтому метод главных компонент относят к техникам **уменьшения размерности**.

Применив его к нашему пищевому набору данных, мы получим главные компоненты, изображенные на следующем рисунке.

Метод главных компонент

	PC1	PC2	PC3	PC4
Жир	−0,45	0,66	0,58	0,18
Белок	0,55	0,21	−0,46	−0,67
Волокна	0,55	0,19	0,43	−0,69
Витамин С	0,44	0,70	−0,52	0,22

Рис. Главные компоненты — это комбинации пищевых переменных. Светло-серые ячейки у одной и той же компоненты представляют собой связанные переменные

Каждая **главная компонента** — это комбинация пищевых переменных, значение которой может быть положительным, отрицательным или близким к нулю. Например, чтобы получить компоненту 1 для отдельного продукта, мы можем вычислить следующее:

0.55(пищевые волокна) + 0.44(Витамин С) – .045(жир) – 0.55(белок)

Метод главных компонент

То есть вместо комбинирования переменных методом проб и ошибок, как мы делали раньше, метод главных компонент сам вычисляет точные формулы, при помощи которых можно дифференцировать наши позиции.

Основная для нас главная компонента 1 (PC1) сразу объединяет жиры с белками, а пищевые волокна с витамином С, и эти пары обратно пропорциональны. В то время как PC1 дифференцирует мясо от овощей, компонента 2 (PC2) подробнее идентифицирует внутренние подкатегории мяса (исходя из жирности) и овощей (по содержанию витамина С). Лучший разброс данных мы получим, используя для графика обе компоненты (см. рис.).

Рис. График продуктов при использовании двух главных компонент

Метод главных компонент

У мясных товаров низкие значения компоненты 1, поэтому они сосредоточены в левой части графика, в противоположной стороне от овощных. Видно также, что среди не овощных товаров низкое содержание жиров у морепродуктов, поэтому значение компоненты 2 для них меньше, и сами они тяготеют к нижней части графика. Схожим образом у тех овощей, которые не являются зеленью, низкие значения компоненты 2, что видно в нижней части графика справа.

Выбор количества компонент. В этом примере созданы четыре главных компоненты по числу изначальных переменных в наборе данных. Поскольку главные компоненты создаются на основе обычных переменных, информация для распределения элементов данных ограничивается их первоначальным набором.

Вместе с тем для сохранения простоты и масштабируемости результатов нам следует выбирать для анализа и визуализации только несколько первых главных компонент. Главные компоненты отличаются по эффективности распределения элементов данных, и первый из них делает это в максимальной степени. Число главных компонент для рассмотрения определяют с помощью графика осыпи.

Метод главных компонент

График показывает снижающуюся эффективность последующих главных компонент в дифференцировании элементов данных. Как правило, используется такое количество главных компонент, которое соответствует положению острого излома на графике осыпи. На рис. излом расположен на отметке в две компоненты. Это означает, что хотя три и более главных компонент могли бы лучше дифференцировать элементы данных, эта дополнительная информация может не оправдать сложности итогового решения.

Как видно из графика осыпи, две первые главные компоненты уже дают 70 %-ный разброс. Использование небольшого числа главных компонент для анализа данных дает гарантию того, что схема подойдет и для будущей информации.

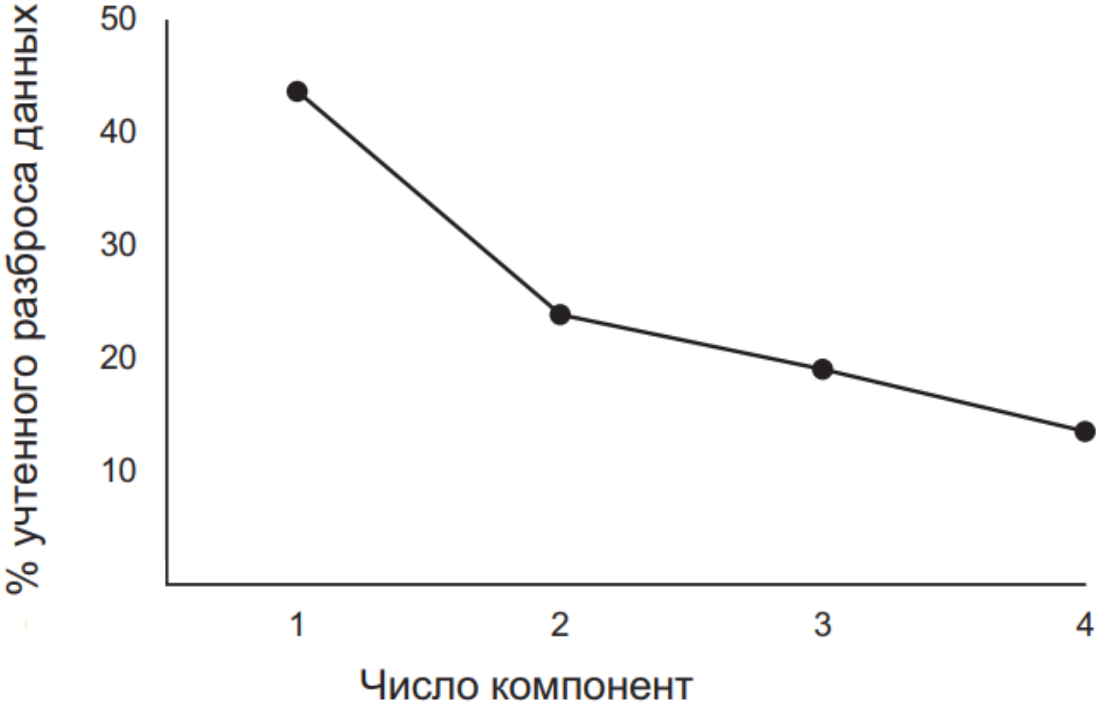


Рис. На графике осыпи виден излом, обозначающий, что оптимальное число главных компонент — две

Метод главных компонент

Метод главных компонент – это полезный способ анализа наборов данных с несколькими переменными. Однако у него есть и **недостатки**.

- **Максимизация распределения**. МГК исходит из важного допущения того, что наиболее полезны те измерения, которые дают наибольший разброс. Однако это не всегда так. Известным контрпримером является задача с подсчетом блинчиков в стопке.

Чтобы сосчитать блинчики, мы отделяем один от другого по вертикальной оси (то есть по высоте стопки). Однако если стопка невелика, МГК ошибочно решит, что лучшей главной компонентой будет горизонтальная ось (диаметр блинчиков), из-за того что в этом измерении можно найти больший разброс значений.

Рис. Аналогия с подсчетом блинчиков

Метод главных компонент

- **Интерпретация компонент.** Главная трудность с МГК состоит в том, что необходима интерпретация сгенерированных компонент, и иногда нужно очень постараться, чтобы объяснить, почему переменные должны быть скомбинированы именно выбранным способом.

Тем не менее нас могут выручить предварительные общие сведения. В нашем примере с продуктами скомбинировать пищевые переменные для главных компонент нам помогают именно предварительные знания об их категориях.

- **Ортогональные компоненты**. МГК всегда формирует *ортогональные* главные компоненты, то есть такие, которые размещаются в отношении друг друга под углом 90°. Однако это допущение может оказаться излишним при работе с неортогональными информационными измерениями. Для решения этой проблемы мы можем воспользоваться альтернативным методом, известным как **анализ независимых компонент (АНК).**

Метод главных компонент

АНК (**анализ независимых компонент**) допускает неортогональность компонент, но запрещает ситуации взаимного перекрытия содержащейся информации (см. рис.). Каждая из выделенных им главных компонент будет содержать уникальную информацию о наборе данных. Помимо обхода ортогонального ограничения АНК в поисках главных компонент принимает во внимание не один только разброс данных и поэтому менее подвержен «блинчиковой ошибке».

Хотя АНК может показаться совершеннее, самым популярным способом уменьшения размерности остается МГК, и понимание того, как он работает, весьма полезно. В случае сомнений имеет смысл всегда запускать АНК в дополнение к МГК для получения более общей картины.

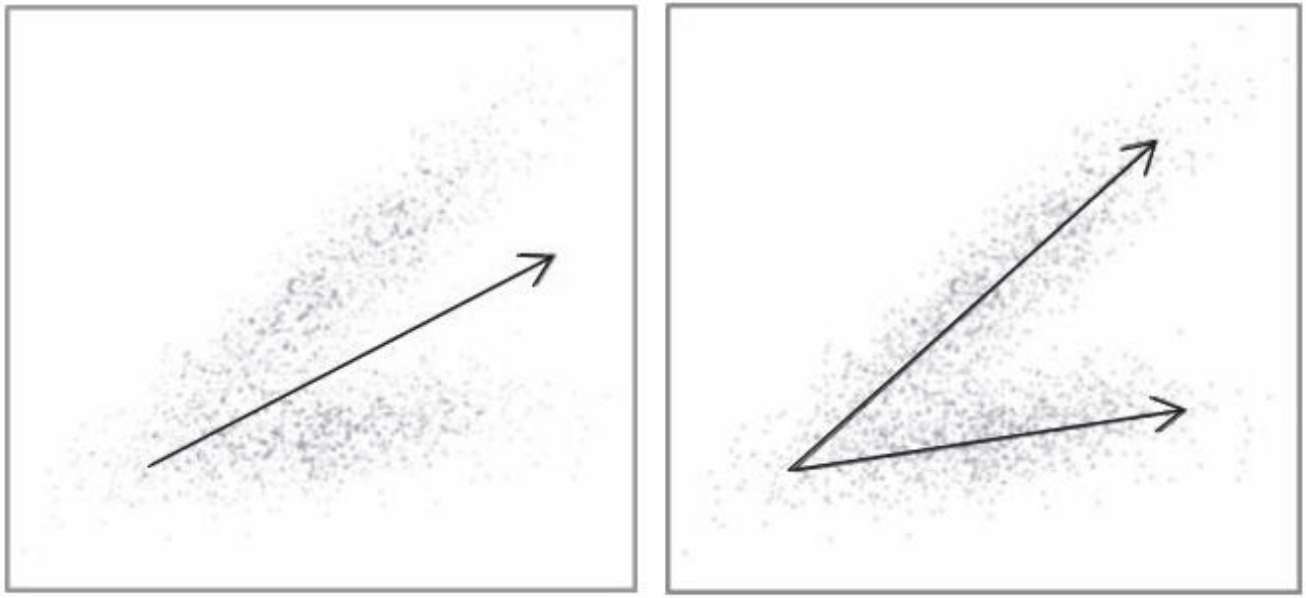


Рис. Сравнение того, как МГК и АНК определяют главные компоненты

Ассоциативные правила

Отправляясь в продуктовый магазин, вы наверняка возьмете с собой список покупок, исходя их ваших потребностей и предпочтений. Домохозяйка, возможно, купит полезные продукты для семейного ужина, а холостяк, скорее всего, возьмет пива и чипсов. Понимание таких закономерностей поможет увеличить продажи сразу несколькими способами. Например, если пара товаров X и Y часто покупается вместе, то:

- реклама товара X может быть направлена на покупателей товара Y;
- товары X и Y могут быть размещены на одной и той же полке, чтобы побудить покупателей одного товара к приобретению второго;
- товары X и Y могут быть скомбинированы в некий новый продукт, такой как X со вкусом Y.

Узнать, как именно товары связаны друг с другом, нам помогут **ассоциативные правила**. Кроме увеличения продаж ассоциативные правила могут быть также использованы в других областях. В медицинской диагностике, например, понимание сопутствующих симптомов может улучшить заботу о пациентах.

Ассоциативные правила

Существуют **три основные меры** для определения ассоциаций. **Мера 1: поддержка**. Поддержка показывает *то, как часто данный товарный набор появляется*, что измеряется долей покупок, в которых он присутствует. В табл. 1 {яблоко} появляется в четырех из восьми покупок, значит, его поддержка 50%. Товарные наборы могут содержать и несколько элементов. Например, поддержка набора {яблоко, пиво, рис} – два из восьми, то есть 25%.

Для определения часто встречающихся товарных наборов может быть установлен **порог поддержки**.

Товарные наборы, встречаемость которых выше заданного числа, будут считаться **частотными**.

$$\text{Поддержка } \{\text{🍎}\} = \frac{4}{8}$$

Рис. Мера «поддержка»

Таблица 1. Пример покупок

Покупка 1	🍎 🥛 🍚 🍗
Покупка 2	🍎 🥛 🍚
Покупка 3	🍎 🥛
Покупка 4	🍎 🍐
Покупка 5	🍾 🥛 🍚 🍗
Покупка 6	🍾 🥛 🍚
Покупка 7	🍾 🥛
Покупка 8	🍾 🍐

Ассоциативные правила

Мера 2: достоверность. Достоверность показывает, *как часто товар Y появляется вместе с товаром X*, что выражается как {X->Y}. Это измеряется долей их одновременных появлений. Согласно табл. 1, достоверность {яблоко->пиво} соответствует трем из четырех, то есть 75 %.

Рис. Мера «достоверность»

Одним из недостатков этой меры является то, что она может исказить степень важности предложенной ассоциации. Пример на рис. принимает во внимание только то, как часто покупают яблоки, но не то, как часто покупают пиво. Если пиво тоже довольно популярно, что и видно из табл. 1, то неудивительно, что покупки, включающие яблоки, нередко содержат и пиво, таким образом увеличивая меру достоверности. Тем не менее мы можем принять во внимание частоту обоих товаров, используя третью меру.

Ассоциативные правила

Мера 3: лифт. Лифт отражает *то, как часто товары X и Y появляются вместе, одновременно учитывая, с какой частотой появляется каждый из них*. Таким образом, лифт {яблоко->пиво} равен достоверности {яблоко->пиво}, деленной на частоту {пива}.

$$\text{Лифт}\{\text{🍎} \rightarrow \text{🍺}\} = \frac{\text{Поддержка}\{\text{🍎}, \text{🍺}\}}{\text{Поддержка}\{\text{🍎}\} \times \text{Поддержка}\{\text{🍺}\}}$$

Рис. Мера «лифт»

Согласно табл. 1, лифт для {яблоко->пиво} равен единице, что означает отсутствие связи между товарными позициями. Значения лифта больше единицы означают, что товар Y *вероятно* купят вместе с товаром X, а значение меньше единицы – что их совместная покупка *маловероятна*.

Ассоциативные правила

Пример:

ведение продуктовых продаж

Чтобы продемонстрировать использование мер ассоциации, были проанализированы данные одного продуктового магазина за 30 дней. Рисунок показывает **ассоциации между товарными парами**, в которых достоверность выше 0,9 %, а лифт – 2,3. **Большие круги** означают высокую поддержку, а **темные** – больший лиф.

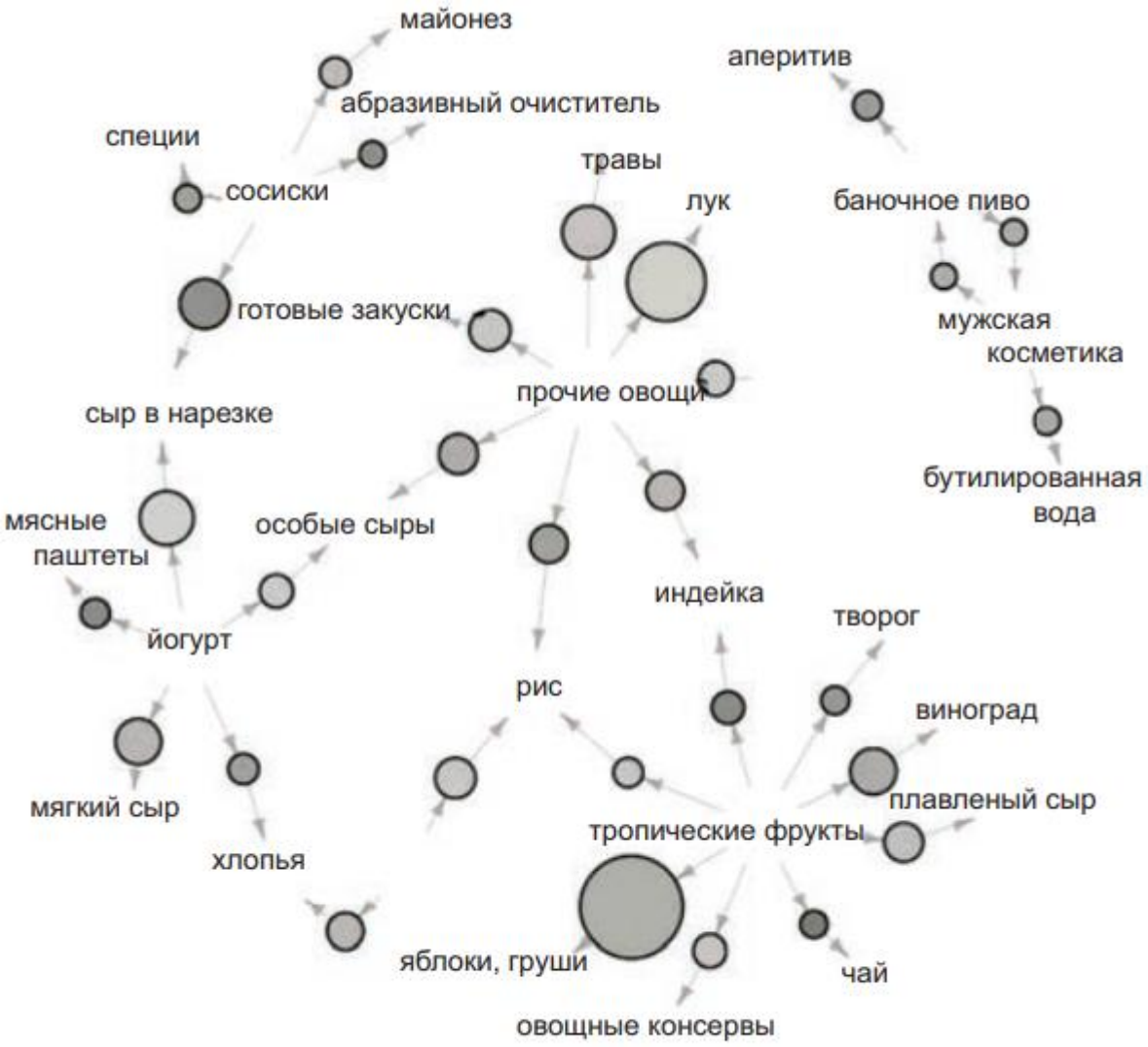


Рис. Граф ассоциаций между товарными позициями

Ассоциативные правила

Мы можем наблюдать такие закономерности в покупках:

- чаще всего покупают яблоки и тропические фрукты;
- другая частая покупка: лук и овощи;
- если кто-то покупает сыр в нарезке, он, скорее всего, возьмет и сосиски;
- если кто-то покупает чай, то он, вероятно, возьмет и тропические фрукты.

Вспомним, что одним из недостатков меры «достоверность» является то, что она может создавать искаженное впечатление о значимости ассоциации. Чтобы показать это, рассмотрим три ассоциативных правила, содержащих пиво.

Таблица 2. Ассоциативные метрики для трех правил, связанных с пивом

Покупка	Поддержка	Достоверность	Лифт
Пиво → Газировка	1,38 %	17,8 %	1,0
Пиво → Ягоды	0,08 %	1,0 %	0,3
Пиво → Мужская косметика	0,09 %	1,2 %	2,6

Ассоциативные правила

Правило {пиво->газировка} имеет высокую достоверность — 17,8 %. Однако и пиво, и газировка вообще часто появляются среди покупок (табл. 3), поэтому их ассоциация может оказаться простым совпадением. Это подтверждается значением лифта, указывающим на отсутствие связи между газировкой и пивом.

Таблица 3. Значение поддержки для отдельных товаров в правилах, связанных с пивом

Покупка	Поддержка
Пиво	7,77 %
Газировка	17,44 %
Ягоды	3,32 %
Мужская косметика	0,46 %

Ассоциативные правила

С другой стороны, правило {пиво->мужская косметика} имеет низкую достоверность из-за того, что мужскую косметику вообще реже покупают. Тем не менее если кто-то покупает ее, он, вероятно, купит также и пиво, на что указывает высокое значение лифта в 2,6. Для пары {пиво->ягоды} верно обратное. Видя лифт меньше единицы, мы заключаем, что если кто-то покупает пиво, то он, скорее всего, не возьмет ягод.

Хотя несложно определить частотность отдельных товарных наборов, владелец бизнеса обычно заинтересован в получении полного списка часто покупаемых товарных наборов. Для этого потребуется вычислить значения поддержки для каждого возможного товарного набора, после чего выбрать те, поддержка которых выше заданного порога.

В магазине со всего десятью товарами суммарное число возможных конфигураций для анализа составит 1023 (то есть 2^10 – 1), и это число экспоненциально возрастает для магазина с сотнями товарных позиций. Ясно, что нам потребуется более эффективное решение.

Ассоциативные правила

Одним из способов снизить количество конфигураций рассматриваемых товарных наборов является использование **принципа Apriori.** Если вкратце, то принцип Apriori утверждает, что если какой-то товарный набор редкий, то и большие наборы, которые его включают, тоже должны быть редки. Это значит, что если редким является, скажем, {пиво}, то редким должно быть и сочетание {пиво, пицца}. Таким образом, составляя список частотных товарных наборов, мы уже не будем рассматривать ни пару {пиво, пицца}, ни какую-либо другую с содержанием пива.

С применением принципа Apriori **мы можем получить список частотных** товарных наборов, используя следующие шаги.

Шаг 0: начать с товарных наборов, содержащих всего один элемент, таких как {яблоки} или {груши}.

Шаг 1: вычислить поддержку для каждого товарного набора. Оставить наборы, удовлетворяющие порогу, и отбросить остальные.

Шаг 2: увеличить размер анализируемого товарного набора на единицу и сгенерировать все возможные конфигурации, используя товарные наборы из предыдущего шага.

Шаг 3: повторять шаги 1 и 2, вычисляя поддержку для возрастающих товарных наборов до тех пор, пока они не закончатся.

Ассоциативные правила

На рисунке показано, как число рассматриваемых товарных наборов может значительно **сократиться** при использовании принципа Apriori. Если у элемента {яблоки} низкая поддержка, то он будет удален из списка анализируемых товарных наборов вместе со всем, что его содержит, тем самым это сократит число наборов для анализа более чем вдвое.

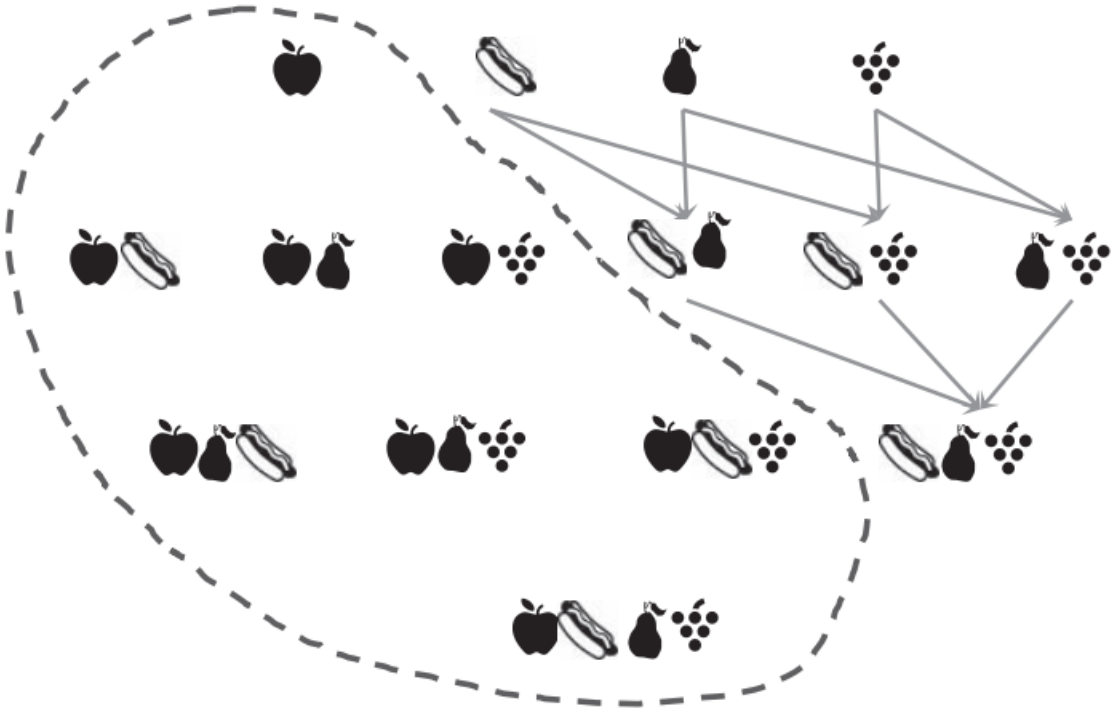


Рис. Товарные наборы в пределах пунктирной линии будут отброшены

Ассоциативные правила

Поиск товарных правил с высокой достоверностью или лифтом

Кроме определения товарных наборов с высокой поддержкой, принцип Apriori также может помочь найти товарные ассоциации с высокой достоверностью или лифтом. Поиск этих ассоциаций требует меньше вычислений, поскольку если товарные наборы с высокой поддержкой известны, то достоверность и лифт вычисляются уже с использованием значения поддержки.

Возьмем для примера задачу поиска правил с высокой достоверностью. Если правило {пиво,чипсы->яблоки} имеет низкую достоверность, то и все другие правила с теми же образующими элементами и яблоком с правой стороны будут тоже иметь низкую достоверность, включая {пиво->яблоки,чипсы} и {чипсы->яблоки,пиво}. Как и прежде, эти правила могут быть отброшены благодаря принципу Apriori, тем самым снижая число потенциально рассматриваемых правил.

Ассоциативные правила

Ограничения

1. **Требует долгих вычислений.** Хотя принцип Apriori и снижает число потенциальных товарных наборов для рассмотрения, оно все еще может быть достаточно значительным, если список товаров большой или указан низкий порог поддержки. В качестве альтернативного решения можно сократить число сравнений, используя расширенные структуры данных, чтобы отобрать потенциальные товарные наборы с большей эффективностью.

2. Ложные ассоциации. В больших наборах данных ассоциации могут быть чистой случайностью. Чтобы убедиться, что обнаруженные ассоциации масштабируемы, их нужно оценить. Несмотря на эти ограничения, ассоциативные правила остаются интуитивно-понятным методом обнаружения закономерностей в наборах данных с управляемым размером.

Ассоциативные правила

- Ассоциативные правила выявляют то, как часто элементы появляются вообще и в связи с другими.
- Есть три основных способа оценки ассоциации:

  1. *Поддержка* {X} показывает, как часто X появляется.

  2. *Достоверность* {X->Y} показывает, как часто Y появляется в присутствии X.

  3. *Лифт* {X->Y} показывает, как часто элементы X и Y появляются вместе по сравнению с тем, как часто они появляются по отдельности.

- *Принцип Apriori* ускоряет поиск часто встречающихся товарных наборов, отбрасывая значительную долю редких.

Анализ социальных сетей

Большинство из нас имеет множество кругов общения, включающих такие категории людей, как родственники, коллеги или одноклассники. Чтобы выяснить, как устроены отношения всех этих людей, определив, например, активных персон и то, как они влияют на групповую динамику, мы можем воспользоваться методом под названием **анализ социальных сетей** (**Social Network Analysis**).

Этот метод можно **применять** в вирусном маркетинге, моделировании эпидемий и даже для стратегий в командных играх. Тем не менее он больше известен своим использованием для анализа отношений в социальных сетях, что и дало ему название. На рисунке пример того, как анализ социальных сетей показывает отношения.

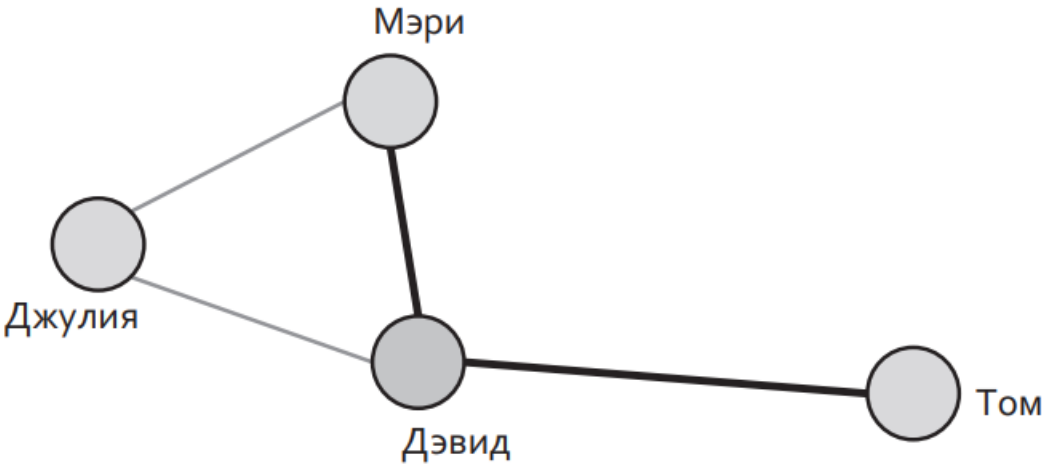


Рис. Простая сеть друзей. Более близкие отношения показаны утолщенными линиями

Анализ социальных сетей

Рисунок показывает сеть из четырех индивидов, также известную как **граф**, в котором каждый из этих персон представлен **узлом** (node). Отношения между узлами представлены линиями, называемыми **ребрами** (edges). Каждое ребро может иметь **вес**, показывающий силу отношений.

Из рисунка мы можем заключить следующее:
- Дэвид имеет больше всех связей, будучи знакомым с остальными тремя персонами;
- Том не знает никого, кроме Дэвида, с которым они близкие друзья;
- Джулия знает Мэри и Дэвида, но не близка с ними.

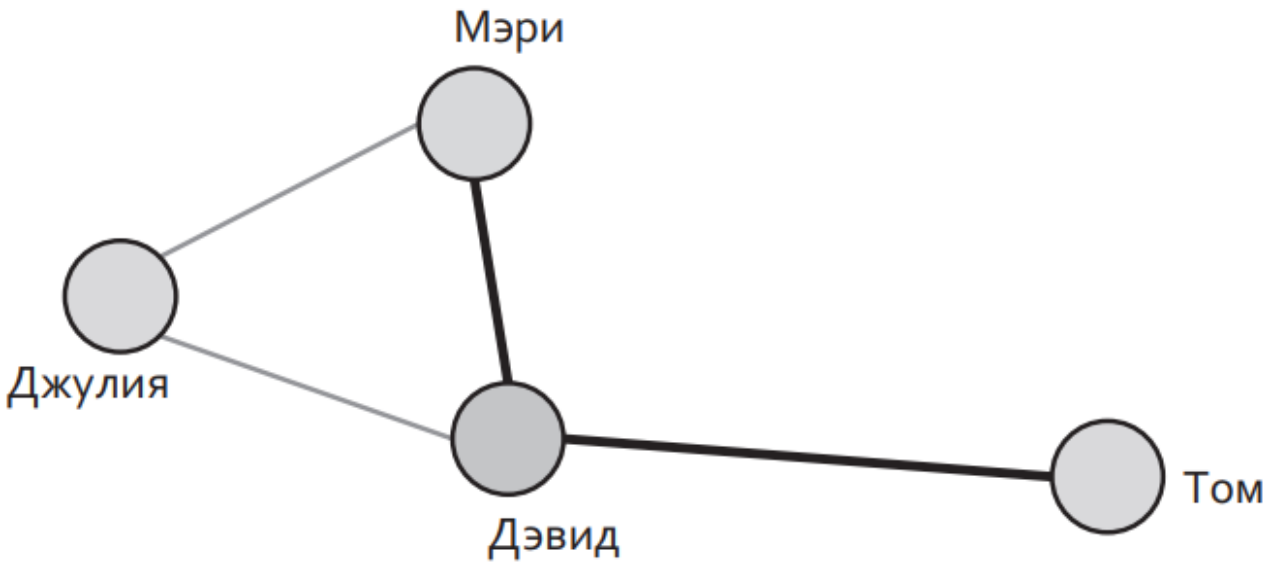


Рис. Простая сеть друзей. Более близкие отношения показаны утолщенными линиями

Анализ социальных сетей

Кроме отношений анализ социальных сетей может строить схемы и для других сущностей, при условии, что между ними есть связи. Воспользуемся им для анализа международной сети торговли оружием, чтобы выявить доминирующие силы и их сферы влияния.

Пример: геополитика в торговле оружием

Были получены данные о двусторонних трансферах основных видов обычных вооружений из Стокгольмского международного института по исследованию проблем мира. Военные поставки были выбраны в качестве косвенного показателя двусторонних отношений, поскольку должны свидетельствовать о тесной связи стран на международной арене.

В этом анализе были стандартизированы стоимость оружия на уровне цен 1990 года в долларах США, после чего приняли в расчет только сделки, сумма которых превысила 100 млн долларов. Чтобы учесть флуктуации в торговле оружием, обусловленные производственными циклами новых технологий, был рассмотрен 10-летний период, с 2006 по 2015 год, построив сеть из 91 узла и 295 ребер.

Анализ социальных сетей

Для визуализации сети использовался **силовой алгоритм (force-directed algorithm)**: узлы без связей отталкиваются друг от друга, а связанные узлы, наоборот, притягиваются с той степенью близости, которая отражает силу их связи (см. рис.). Например, максимальный объем торговли зафиксирован между Россией и Индией ($ 22,3 млрд), поэтому эти государства соединены толстой линией и близко расположены.

Рис. Сеть стран, исходя из военных поставок

Анализ социальных сетей

После анализа получившейся сети **лувенским методом (Louvain Method)** геополитические альянсы были сгруппированы в **три кластера:**

- **Светло-серый:** это крупнейший кластер, в котором доминируют США и который включает их союзников, таких как Великобритания и Израиль.
- **Светлый:** в нем лидирует Германия, и он включает в основном европейские страны, а также тесно связан со светло-серым кластером.
- **Темный:** в этом кластере доминируют Россия и Китай, он дистанцирован от двух других и включает в основном азиатские и африканские государства.

Кластеры отражают геополитические реалии XXI столетия, такие как долгосрочные альянсы между западными нациями, поляризацию между демократическими и коммунистическими странами и возрастающую роль противостояния между США и Китаем.

Кроме группировки в кластеры также проранжировали отдельные страны по уровню их влияния, воспользовавшись **алгоритмом PageRank** (описывается дальше). На следующем рисунке представлены 15 самых влиятельных государств, которые также отмечены на рисунке более крупными узлами и подписями.

Согласно нашему анализу, в пятерку самых влиятельных стран входят США, Россия, Германия, Франция и Китай. Эти результаты подтверждаются тем обстоятельством, что четыре из пяти этих государств имеют влияние еще и как члены Совета Безопасности ООН.

Анализ социальных сетей

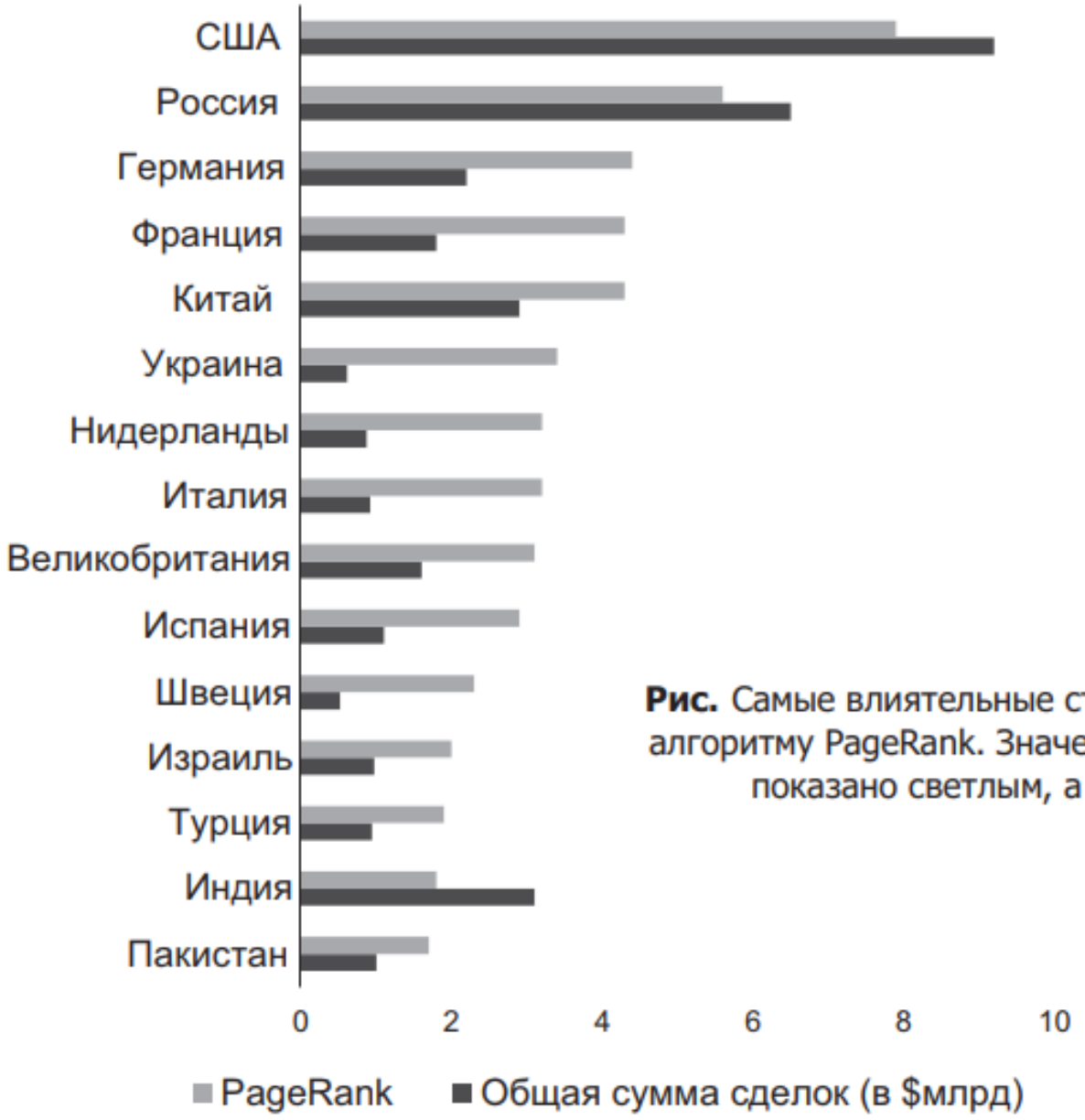


Рис. Самые влиятельные страны в торговле оружием, согласно алгоритму PageRank. Значение PageRank для каждой страны показано светлым, а торговый объем — темным

Анализ социальных сетей

По рисунку сети стран были найдены кластеры сети путем группировки узлов. Изучение этих кластеров поможет понять, чем различаются части сети и как они соединены.

Лувенский метод – это один из способов определения кластеров сети. Он подбирает различные кластерные конфигурации, чтобы:

1) максимизировать число и силу связей между узлами в одном кластере;
2) минимизировать при этом связи между узлами различных кластеров.

Степень удовлетворения этим двум условиям известна как **модулярность** (**modularity**), и более высокая модулярность – это признак более оптимальных кластеров.

Анализ социальных сетей

Чтобы получить оптимальную конфигурацию кластеров, лувенский метод итеративно проходит следующие **стадии**.

Стадия 0: рассматривает каждый узел в качестве кластера, то есть начинает с числа кластеров, равного числу узлов.

Стадия 1: меняет кластерное членство узла, если это приводит к улучшению модулярности. Если модулярность больше нельзя улучшить, узел остается на месте. Это повторяется для каждого узла до тех пор, пока изменения кластерного членства не будут исчерпаны.

Стадия 2: строит грубую версию сети, в которой каждый кластер, найденный на стадии 1, представлен отдельным узлом, и объединяет бывшие межкластерные соединения в утолщенные ребра этих новых узлов в соответствии с их весом.

Стадия 3: повторяет стадии 1 и 2 до тех пор, пока не закончатся дальнейшие изменения членства и размера связей.

Таким образом, **лувенский метод помогает выявить более значимые кластеры**, начав с обнаружения малых из них, а затем при необходимости соединяя их.

Анализ социальных сетей

Простота и эффективность делают **лувенский метод** популярным решением для кластеризации сети. Однако он имеет свои **ограничения**.

1. Важные, но малые кластеры могут быть поглощены. Итеративный процесс слияния кластеров может привести к тому, что значимые, но небольшие кластеры будут обойдены вниманием. Чтобы избежать этого, мы можем при необходимости проверять идентифицированные кластеры на промежуточных фазах итераций.

2. Множество возможных кластерных конфигураций. Для сетей, содержащих перекрывающиеся или вложенные кластеры, определить оптимальное кластерное решение может оказаться трудным. Тем не менее, когда имеются несколько решений с высокой модулярностью, мы можем сверить кластеры с другими информационными источниками, что мы и проделали на рисунке с сетью стран, приняв во внимание географическое местоположение и политическую идеологию.

Анализ социальных сетей

Алгоритм PageRank

Поскольку кластеры выявляют области высокой концентрации взаимодействий, эти взаимодействия могут управляться ведущими узлами, вокруг которых эти кластеры и сформированы. Для определения этих доминирующих узлов можно использовать их ранжирование.

Алгоритм PageRank, названный по имени сооснователя Google Ларри Пейджа, стал одним из первых алгоритмов Google для ранжирования веб-сайтов. Он может быть использован для того, чтобы классифицировать узлы любого типа.

Значение PageRank для веб-сайта определяется **тремя факторами**.

1. Число ссылок. Если на один веб-сайт ссылаются другие, то он, скорее всего, привлекает больше пользователей.

2. Сила ссылок. Чем чаще переходят по этим ссылкам, тем больше трафик сайта.

3. Источник ссылок. Ранг веб-сайта повышается и оттого, что на него ссылаются другие высокоранговые сайты.

Анализ социальных сетей

Чтобы увидеть **работу PageRank**, посмотрим на пример сети на рисунке, где узлы представляют **веб-сайты**, а ребра – **гиперссылки**.

Входящая гипер-ссылка с большим весом означает больший объем трафика для сайта. На рисунке видно, что посетитель сайта *M* с вдвое большей вероятностью посетит сайт *D*, чем *J*, и может никогда не посетить сайт *T*.

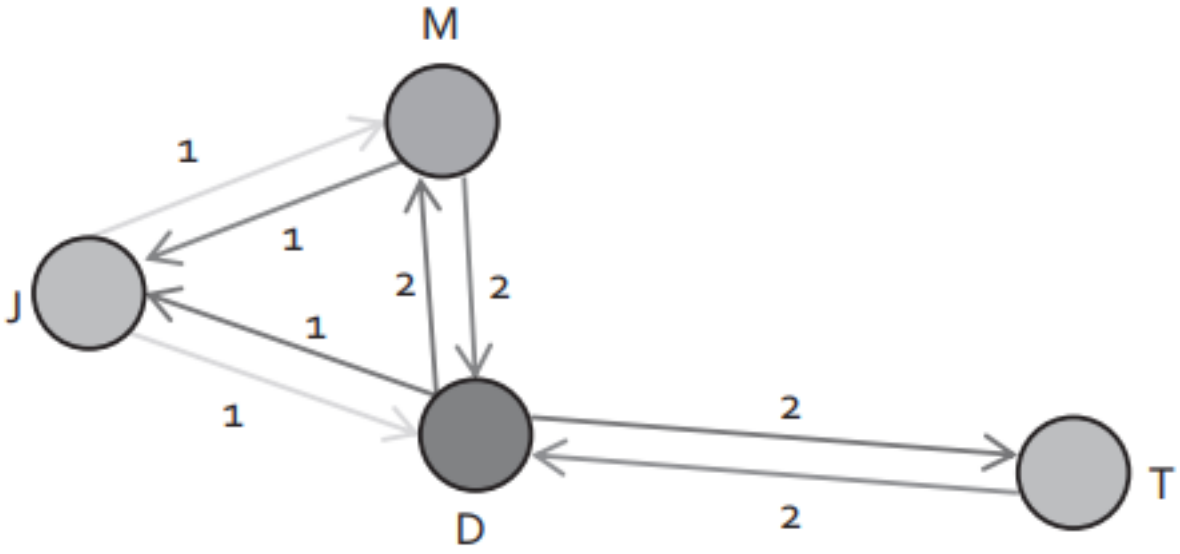


Рис. Сеть, в которой узлы — это веб-сайты, а ребра — гиперссылки

Чтобы понять, какой сайт привлекает больше пользователей, можно смоделировать поведение сайта из рисунка для 100 пользователей и посмотреть, на какой сайт они в итоге попадут.

Анализ социальных сетей

Сначала равно распределим 100 пользователей по четырем веб-сайтам, как показано на рисунке справа.

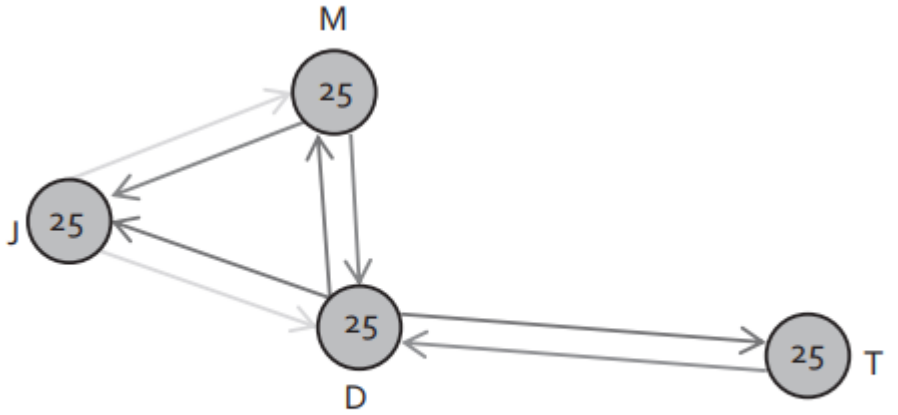


Рис. Начальное положение, в котором 100 пользователей распределены по четырем веб-сайтам

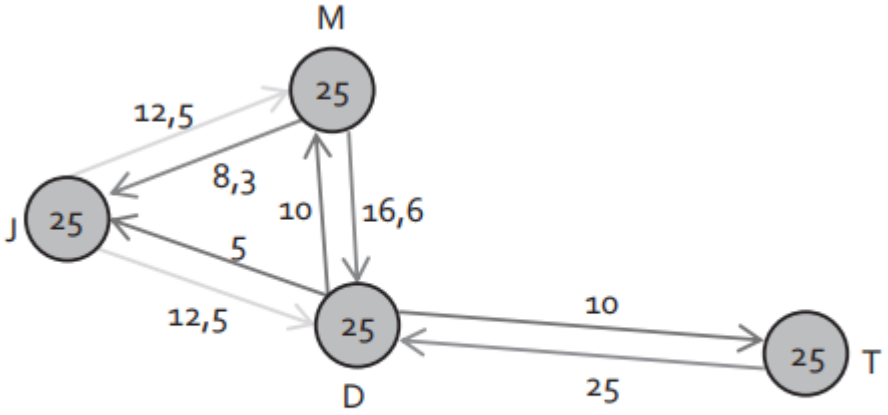


Рис. Перераспределение пользователей, основанное на силе исходящих ссылок

Затем перераспределим пользователей каждого сайта в соответствии с его исходящими ссылками. Например, две трети пользователей сайта M отправятся на сайт D, в то время как оставшаяся треть посетит сайт J. Ребра на рисунке внизу показывают число приходящих и уходящих пользователей для каждого сайта.

Анализ социальных сетей

После перераспределения всех пользователей на сайте M оказалось около 23 пользователей, из которых 10 пришли с сайта D и 13 с сайта J. Следующий рисунок показывает результаты распределения пользователей по каждому сайту, округленные до целого.

Чтобы получить **значение PageRank для каждого сайта**, нужно **повторять** этот процесс до тех пор, пока численность пользователей сайта не перестанет меняться. Итоговое число пользователей для каждого веб-сайта будет соответствовать его значению PageRank: чем больше пользователей он привлечет, тем выше его **ранг**.

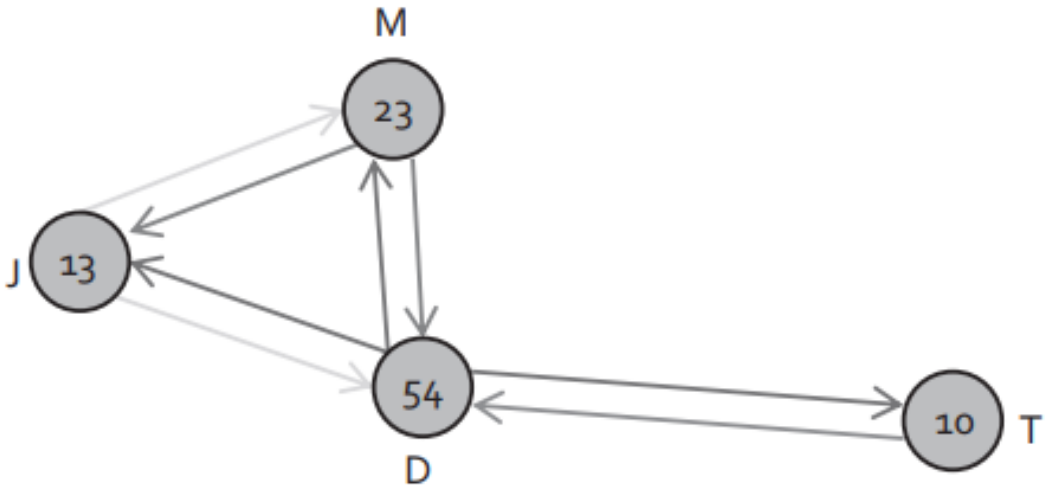


Рис. Число пользователей на каждом веб-сайте после распределения

Тем же способом с помощью PageRank можно измерить и влиятельность государства. В сети, иллюстрирующей торговлю оружием, страной с высоким значением PageRank будет та, которая участвует во многих значительных торговых сделках с другими высокоранговыми странами, что делает ее влиятельным игроком в мировых военных поставках.

Анализ социальных сетей

Несмотря на простоту использования, у алгоритма PageRank есть **недостаток: необъективность в отношении старых узлов**. Например, хотя новая веб-страница и может содержать отличный контент, ее относительная безвестность в момент появления даст ей низкое значение PageRank, что потенциально может привести к исключению из перечней рекомендуемых сайтов. Чтобы избежать этого, значения PageRank могут регулярно **обновляться**, давая новым сайтам возможность поднимать свои ранги по мере зарабатывания репутации.

Тем не менее такое смещение не всегда критично, особенно при моделировании доминирования за долгие периоды времени, например, когда мы ранжируем страны, исходя из степени их влияния. Это показывает то, как ограничения алгоритмов могут не быть их недостатками, в зависимости от целей исследования.

Хотя методы кластеризации и ранжирования позволяют нам получить очень интересные результаты, интерпретировать их нужно с большой осторожностью.

Анализ социальных сетей

Возьмем, к примеру, наше использование данных по поставкам оружия для оценки влиятельности государств. У такой упрощенной оценки есть несколько подводных камней.

- **Игнорирование дипломатических отношений при отсутствии покупок вооружения.** Большинство ребер проведены между экспортерами и импортерами оружия. Таким образом, дружественные отношения между странами, обе из которых являются импортерами (либо экспортерами), не отражены.
- **Игнорирование других соображений**. Возможно, нужно учесть сложившиеся системы отношений, ограничивающие потенциальных покупателей. Кроме того, страны-экспортеры при принятии решений о продаже оружия могут предпочесть двусторонним отношениям внутренние сделки (например, из экономических соображений). Это могло бы объяснить, почему Украина, значительный экспортер оружия, получила шестой ранг, вопреки отсутствию репутации влиятельной страны.

Поскольку обоснованность наших выводов зависит от того, насколько качественное построение для анализа дают данные, используемые для генерации сети, то они должны выбираться с особой тщательностью. Чтобы убедиться, что наши исходные данные и методы анализа достаточно надежны, нужно **проверять** полученные результаты по другим источникам информации.

Регрессионный анализ

Выведение линии тренда

Линии тренда – популярный инструмент для прогнозирования, поскольку они просты как для вычисления, так и для понимания. Достаточно открыть любую ежедневную газету, чтобы увидеть графики трендов в самых различных областях: от цен на акции до прогноза погоды.

Общие тренды обычно применяют единственный предиктор для предсказания результата, используя, например, время (предиктор) для прогнозирования цен на акции компании (результат). Однако можно улучшить предсказание цен на акции, добавив другие предикторы, такие как уровень продаж.

Это становится возможным с **регрессионным анализом**, позволяющим не только улучшать прогнозирование путем учета множества предикторов, но и сравнивать эти предикторы между собой по степени влияния. Рассмотрим пример с предсказанием цен на дома.

Регрессионный анализ

Пример 1: предсказание цен на дома

Здесь были использованы данные за 1970-е годы о ценах на дома в Бостоне. Предварительный анализ показывает, что двумя сильнейшими предикторами цен на дома являются число комнат в доме и доля соседей с низким доходом.

На рис. 1 видно, что у дорогих домов обычно больше комнат. Для предсказания цены дома можно построить **линию тренда**, известную также как линия наилучшего соответствия. Она проходит близко к наибольшему числу элементов данных на графике. Например, если у дома восемь комнат, его цена составит приблизительно $ 38 150.

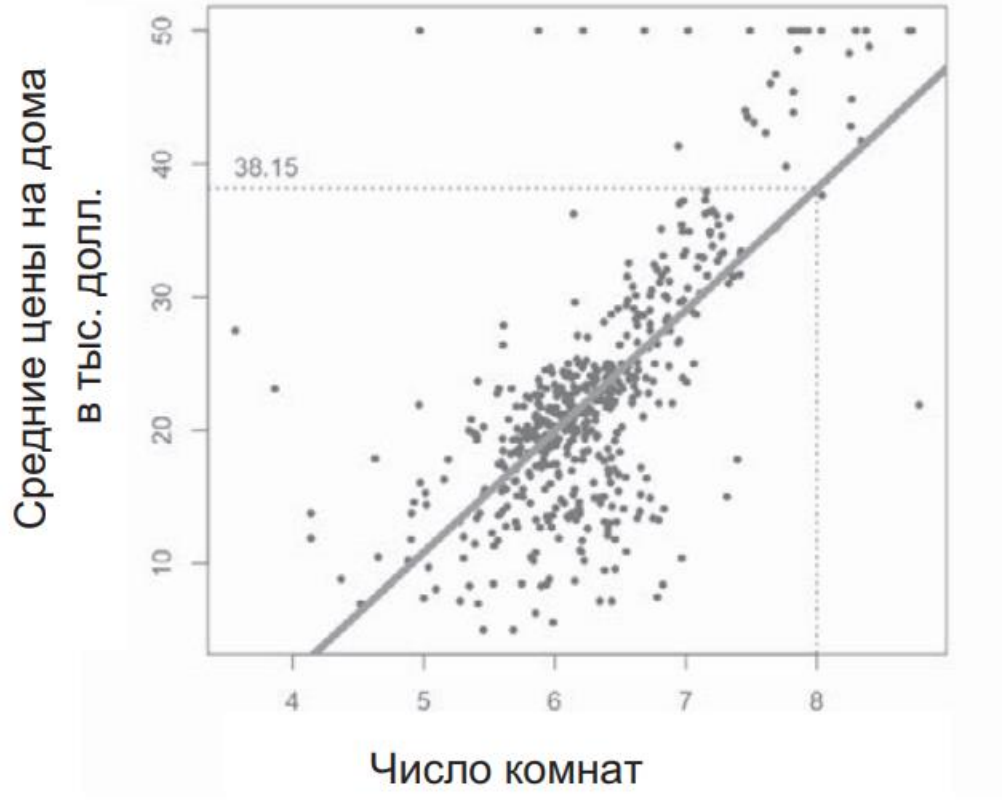


Рис. 1. Цены на дома в сравнении с числом комнат

Регрессионный анализ

Пример 1: предсказание цен на дома

Кроме числа комнат на цену дома также влияло его окружение. Дома оказались дешевле там, где была выше пропорция соседей с низким доходом (рис. 2). Поскольку тренд получался немного изогнутым (рис. 2, a), то применили к предикторам взятие логарифма. Благодаря этому через элементы данных проще провести прямую линию тренда (рис. 2, b).

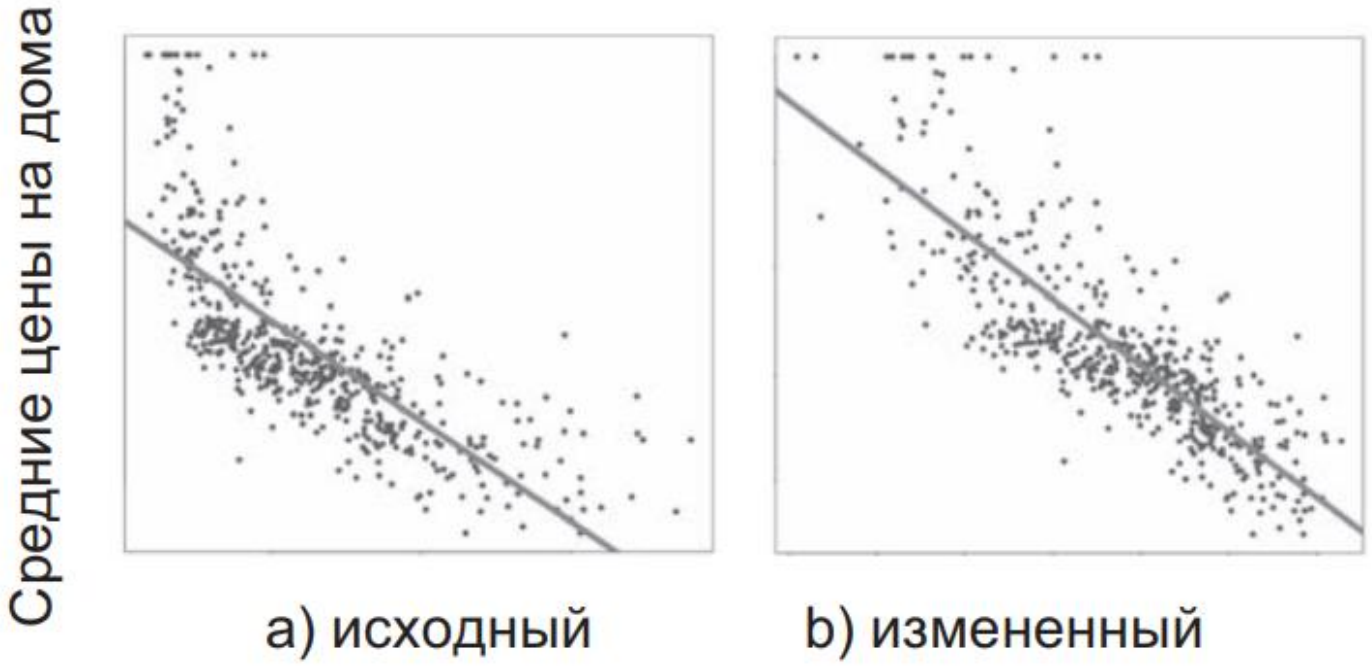


Рис. 2. Цены на дома в сравнении с долей соседей с низким доходом

Регрессионный анализ

Пример 1: предсказание цен на дома

Можно заметить, что элементы данных на рис. 2, b плотнее прилегают к линии тренда, чем на рис. 1. Это означает, что фактор соседства оказался более точным предиктором цены дома, чем число комнат.

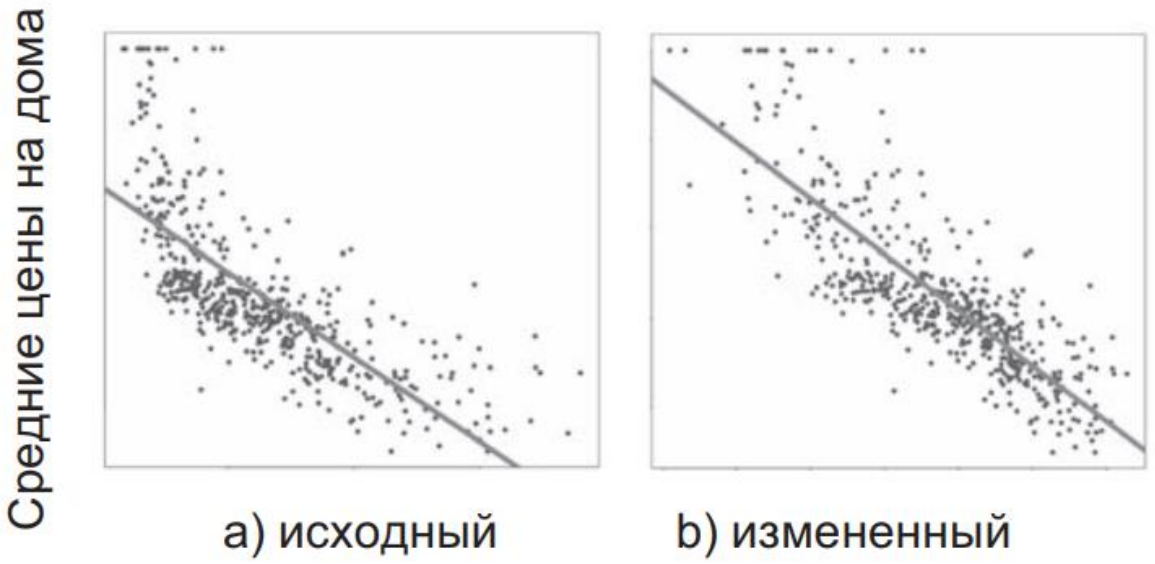


Рис. 2. Цены на дома в сравнении с долей соседей с низким доходом

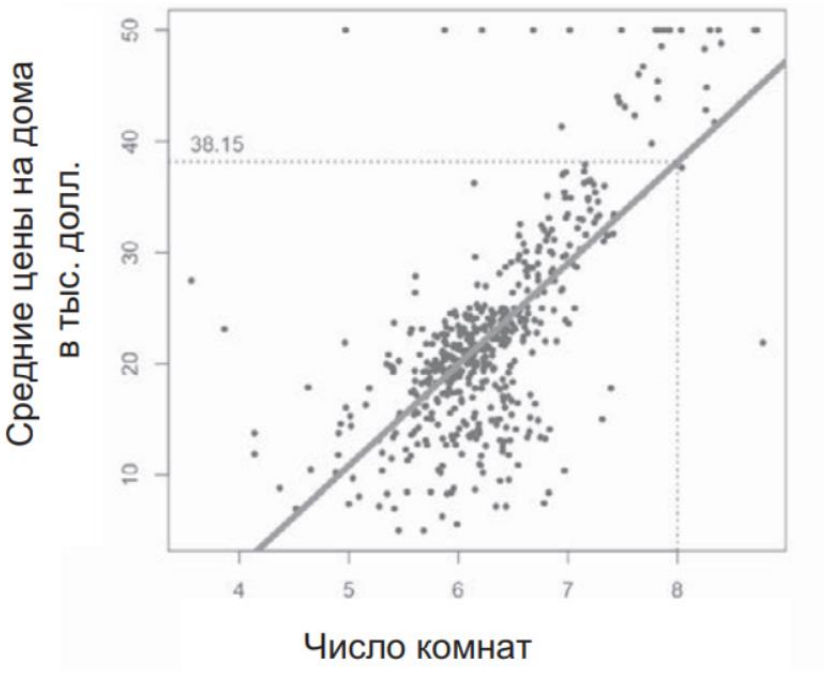


Рис. 1. Цены на дома в сравнении с числом комнат

Регрессионный анализ

Пример 1: предсказание цен на дома

Для улучшения этих расчетов цен на дома можно учесть и число комнат, и влияние соседства. Но поскольку выяснилось, что влияние соседства лучше предсказывает цену дома, простое сложение этих двух предикторов не станет идеальным решением. Вместо этого предиктору соседства нужно задать больший вес.

Рис. 3 показывает график цен на дома согласно оптимальной комбинации двух предикторов. Здесь элементы данных располагаются еще ближе к итоговой линии тренда, чем раньше, поэтому прогноз с использованием такой линии тренда должен оказаться точнее. Чтобы проверить это, можно сравнить погрешность трех линий тренда (табл. 1).

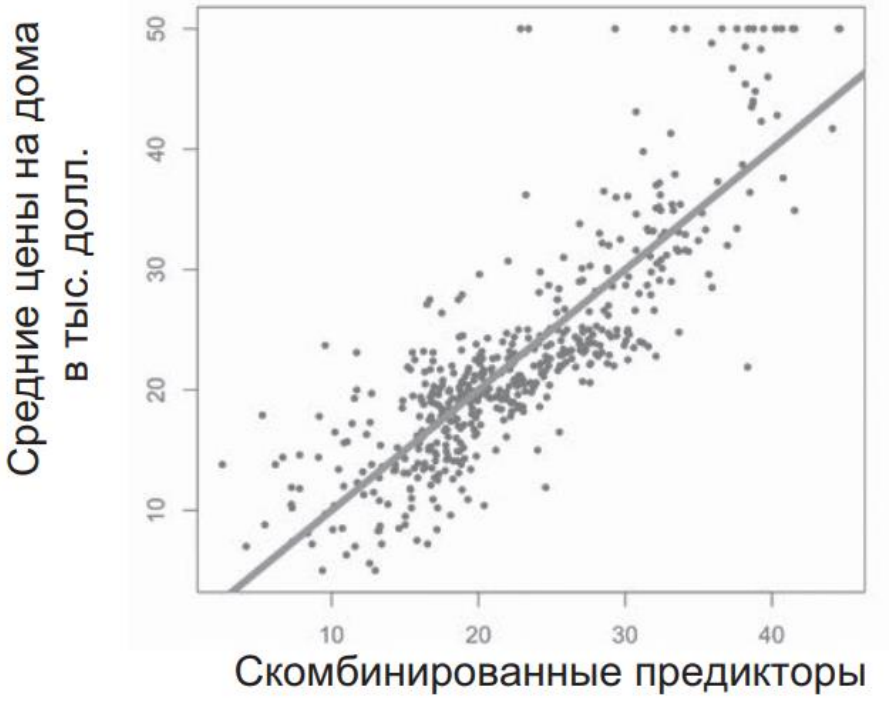


Рис. 3. Цены на дома в сравнении со скомбинированным предиктором из числа комнат и доли соседей с низким доходом

Регрессионный анализ

Пример 1: предсказание цен на дома

Таблица 1. Средняя прогностическая ошибка при использовании трех разных линий тренда

	Погрешность прогнозирования (в тыс. долл.)
Число комнат	4,4
Влияние окружения	3,9
Число комнат и влияние окружения	3,7

Хотя очевидно, что уравновешенная комбинация предикторов ведет к более точным предсказаниям, возникают **два вопроса**:

1) как вычислить оптимальный вес предикторов;

2) как следует их проинтерпретировать.

Регрессионный анализ

Градиентный спуск

Вес предиктора – главный параметр регрессионного анализа, и оптимальный вес обычно вычисляется путем решения уравнений. Тем не менее, поскольку регрессионный анализ прост и годится для визуализации, воспользуемся им для демонстрации альтернативного способа оптимизации параметров. Этот метод называется **градиентным спуском (gradient descent)** и используется в случаях, когда параметры нельзя получить напрямую.

Вкратце: алгоритм градиентного спуска делает первоначальное предположение о наборе весовых составляющих, после чего начинается итеративный процесс их применения к каждому элементу данных для прогнозирования, а затем они перенастраиваются для снижения общей ошибки прогнозирования.

Регрессионный анализ

Градиентный спуск

Этот процесс можно сравнивать с пошаговым спуском в овраг в поисках дна. На каждом этапе алгоритм определяет, какое направление даст наиболее крутой спуск, и пересчитывает весовые составляющие. В конечном итоге мы достигнем самой нижней позиции, которая представляет собой точку, в которой погрешность прогнозирования минимальна.

Рисунок 4 показывает, как оптимальная линия тренда регрессии соответствует нижней точки градиента.

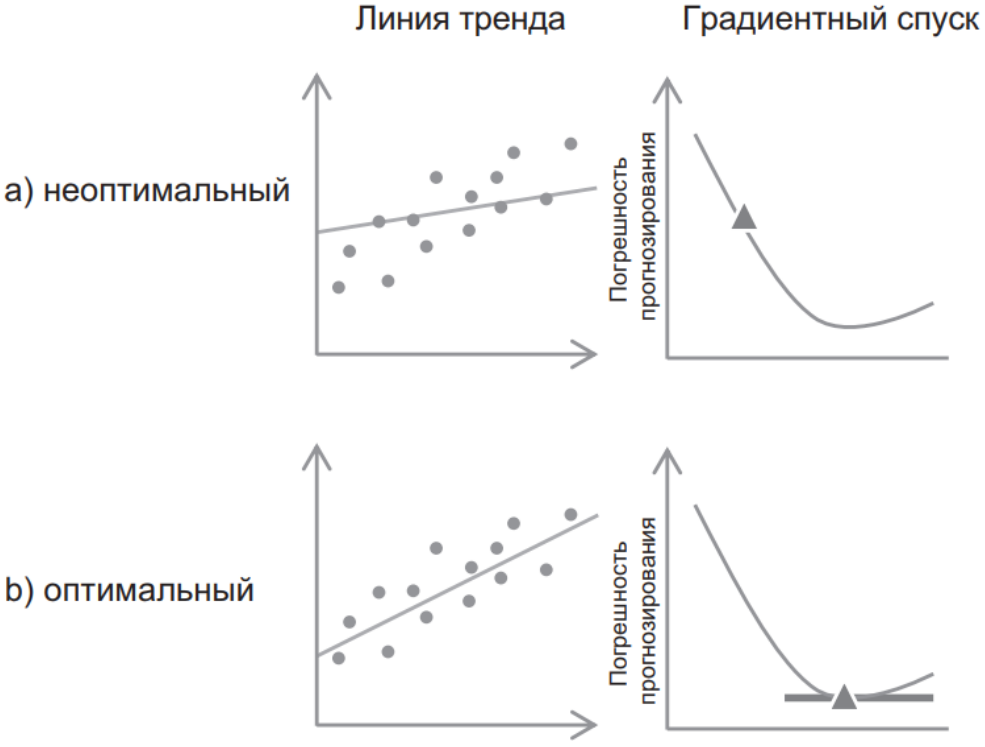


Рис. 4. Как линия тренда достигает оптимальности благодаря градиентному спуску

Регрессионный анализ

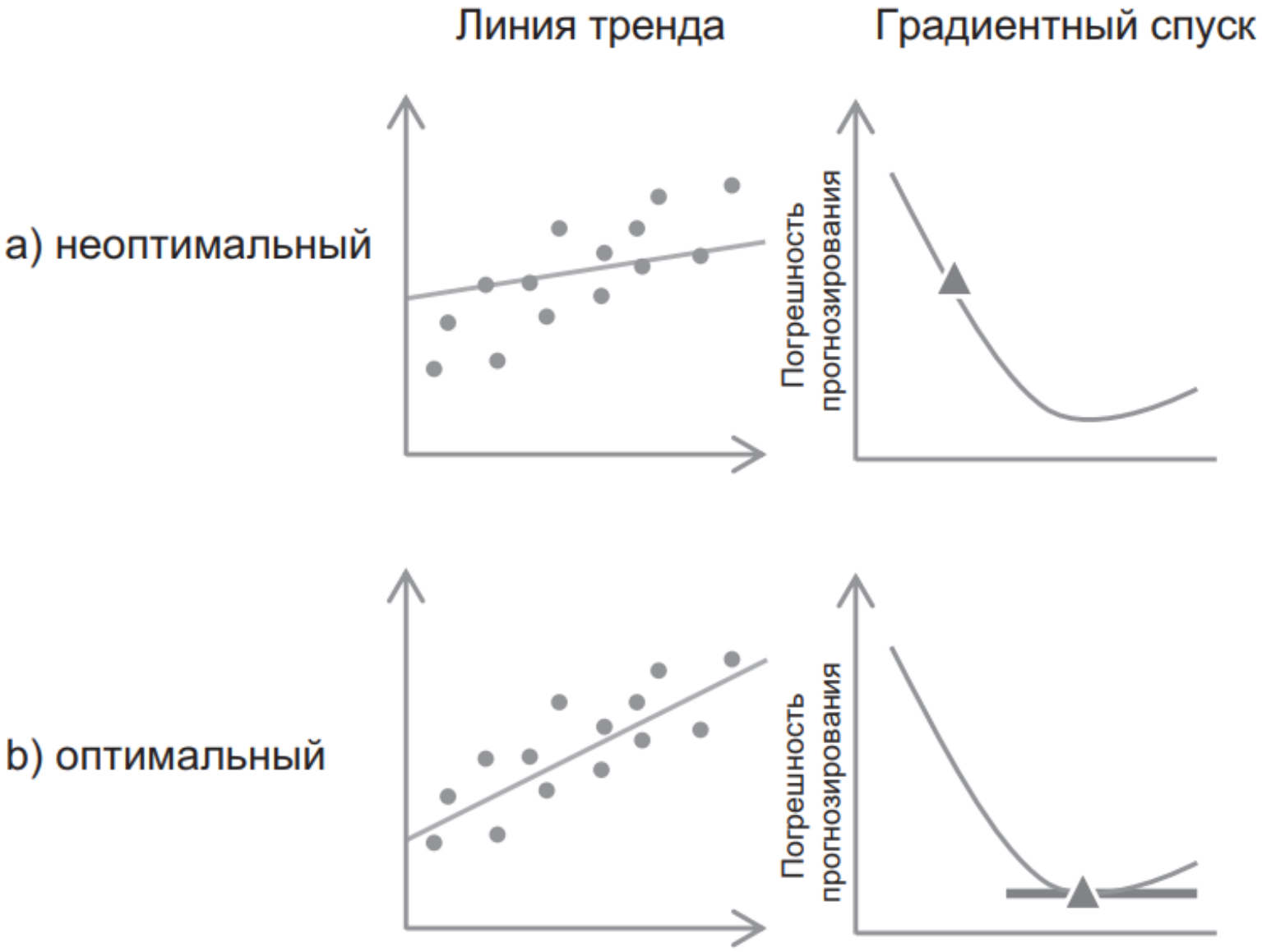


Рис. 4. Как линия тренда достигает оптимальности благодаря градиентному спуску

Регрессионный анализ

Градиентный спуск

Кроме регрессии градиентный спуск может также использоваться для оптимизации параметров в других моделях, таких как метод опорных векторов или в нейронных сетях. Однако в этих более сложных моделях результаты градиентного спуска могут зависеть от стартовой позиции в овраге (то есть изначальных значений параметра).

Например, если случится начать в небольшой яме, алгоритм градиентного спуска может ошибочно принять это за оптимальную точку (рис. 5).

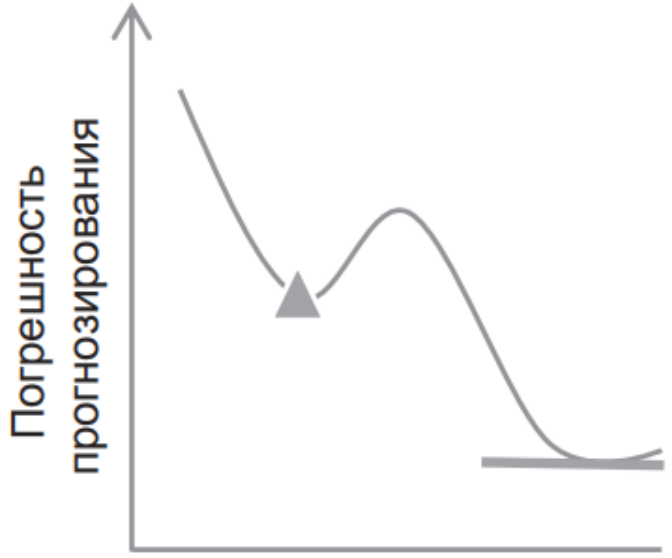


Рис. 5. Как ближайшая яма может быть ошибочно принята за оптимальную точку (треугольник), хотя истинная оптимальная точка находится ниже ее (черта)

Регрессионный анализ

Градиентный спуск

Чтобы снизить риск попадания в такую яму, можно воспользоваться **стохастическим градиентным спуском**, при котором вместо использования всех элементов данных для регулировки параметров при каждой итерации берется только один. Это привносит вариативность, позволяя алгоритму избегать ям. Хотя итоговые значения параметров после работы стохастического процесса могут оказаться не оптимальными, они, как правило, обеспечивают достаточно высокую точность.

Тем не менее этот «недостаток» относится только к более сложным моделям, поэтому не стоит об этом беспокоиться, когда используем регрессионный анализ.

Регрессионный анализ

Коэффициенты регрессии

После получения оптимального набора регрессионных предикторов их нужно интерпретировать.

Вес регрессионных предикторов называется **коэффициентом регрессии**. Коэффициент регрессии показывает то, насколько силен предиктор при совместном использовании с другими. Иными словами, это значение, добавляемое к предиктору, а не его собственная предсказательная способность.

Например, если кроме числа комнат для предсказания цены дома использовать его общую площадь, то значимость числа комнат может показаться незначительной. Поскольку и число комнат, и общая площадь дома связаны с его размером, это добавляет к предсказательной силе не так уж и много.

Регрессионный анализ

Коэффициенты регрессии

Толковой интерпретации регрессионных коэффициентов мешает также различие в единицах измерения. Например, если предиктор измеряется в сантиметрах, его вес будет в 100 раз отличаться по весу от предиктора, берущегося в метрах. Чтобы избежать такого, нужно **стандартизировать** единицы измерения предикторных переменных перед тем, как проводить регрессионный анализ. Стандартизация – это выражение переменных в процентилях. Когда предикторы стандартизированы, то коэффициент, который называется **бета-весом**, может быть использован для более точных сравнений.

В примере с ценами на дома два предиктора (первый – число комнат, второй – соседи с низким доходом) были стандартизированы в соотношении 2,7 к 6,3. Это означает, что доля жильцов с низким доходом является более мощным предиктором цены на дом, чем количество комнат.

Регрессионный анализ

Коэффициенты регрессии

Уравнение регрессии будет выглядеть примерно так:

цена = 2,7 (количество комнат) – 6,3(% соседей с низким доходом).

В этом уравнении доля жильцов с низким доходом имеет отрицательный вес, что выражено знаком «минус».

Дело в том, что предиктор имеет обратную корреляцию с ценами на дома, как показано на устремленной вниз линии тренда на рис. 2.

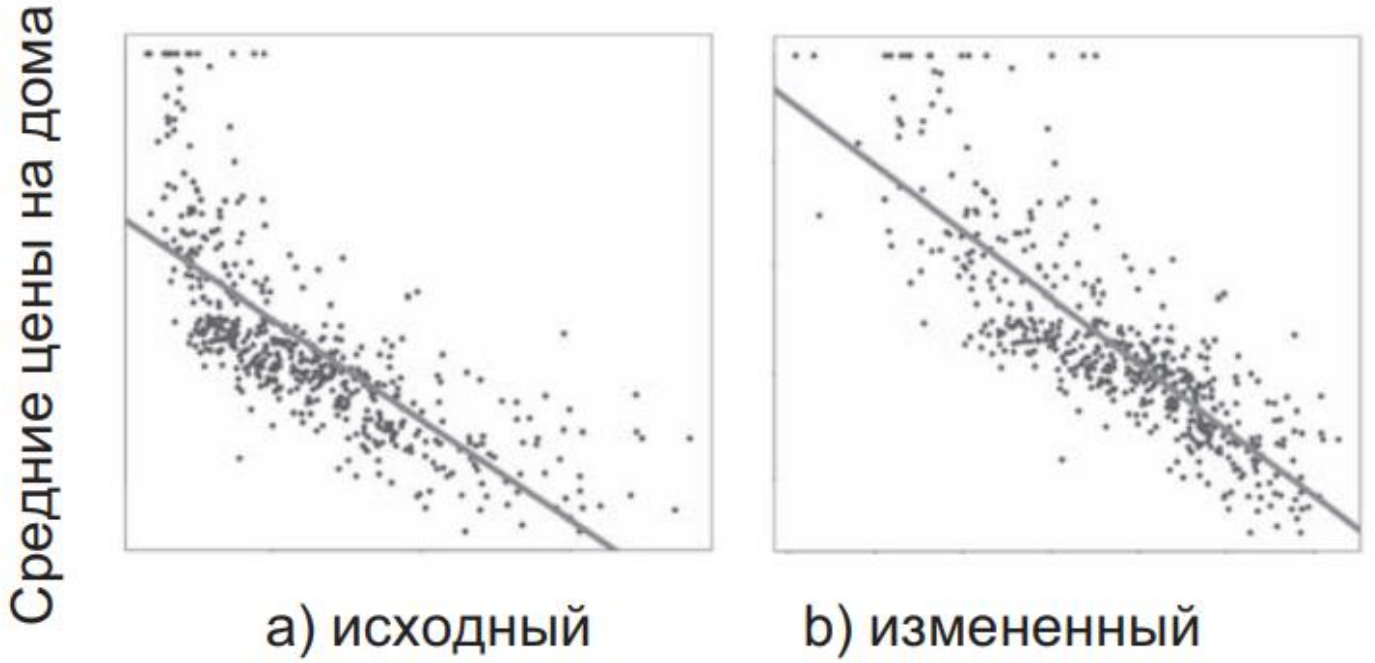


Рис. 2. Цены на дома в сравнении с долей соседей с низким доходом

Регрессионный анализ

Коэффициенты корреляции

Если предиктор только один, бета-вес такого предиктора называется **коэффициентом корреляции** и обозначается как **r**. Коэффициенты корреляции варьируются **от –1 до 1** и несут две единицы информации.

Направление. При положительных коэффициентах предиктор стремится в том же направлении, что и результат. При отрицательных — в обратном направлении. Цены домов положительно коррелируют с числом комнат, но отрицательно коррелируют с долей жильцов с низким доходом по соседству.

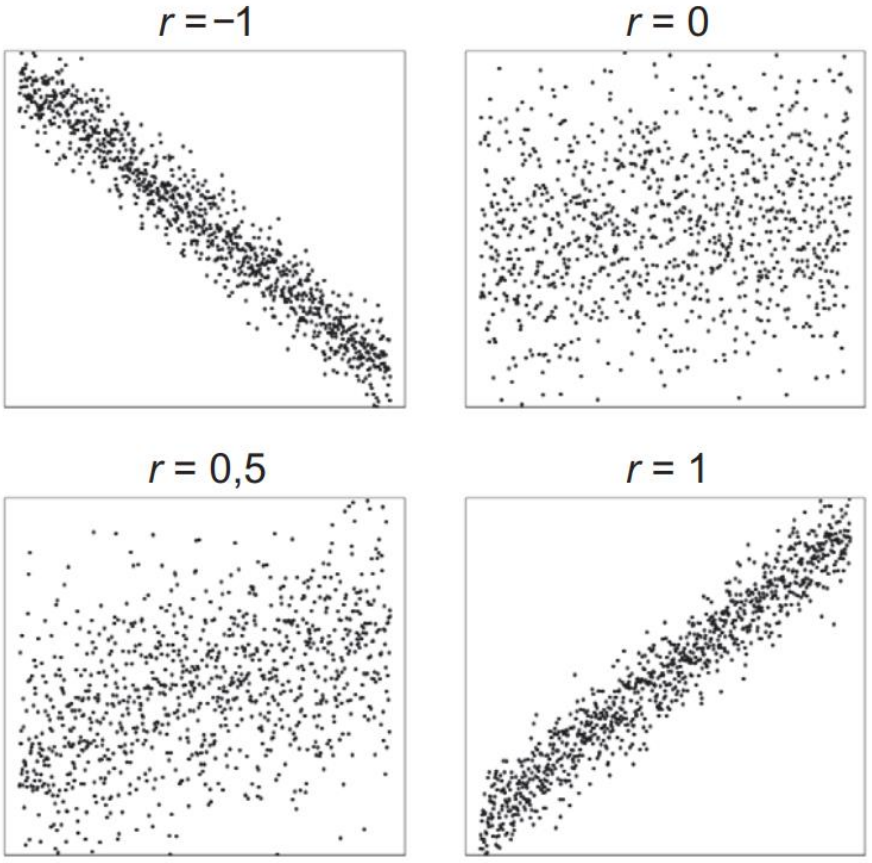


Рис. 6. Пример распределения данных в соответствии с различными коэффициентами корреляции

Регрессионный анализ

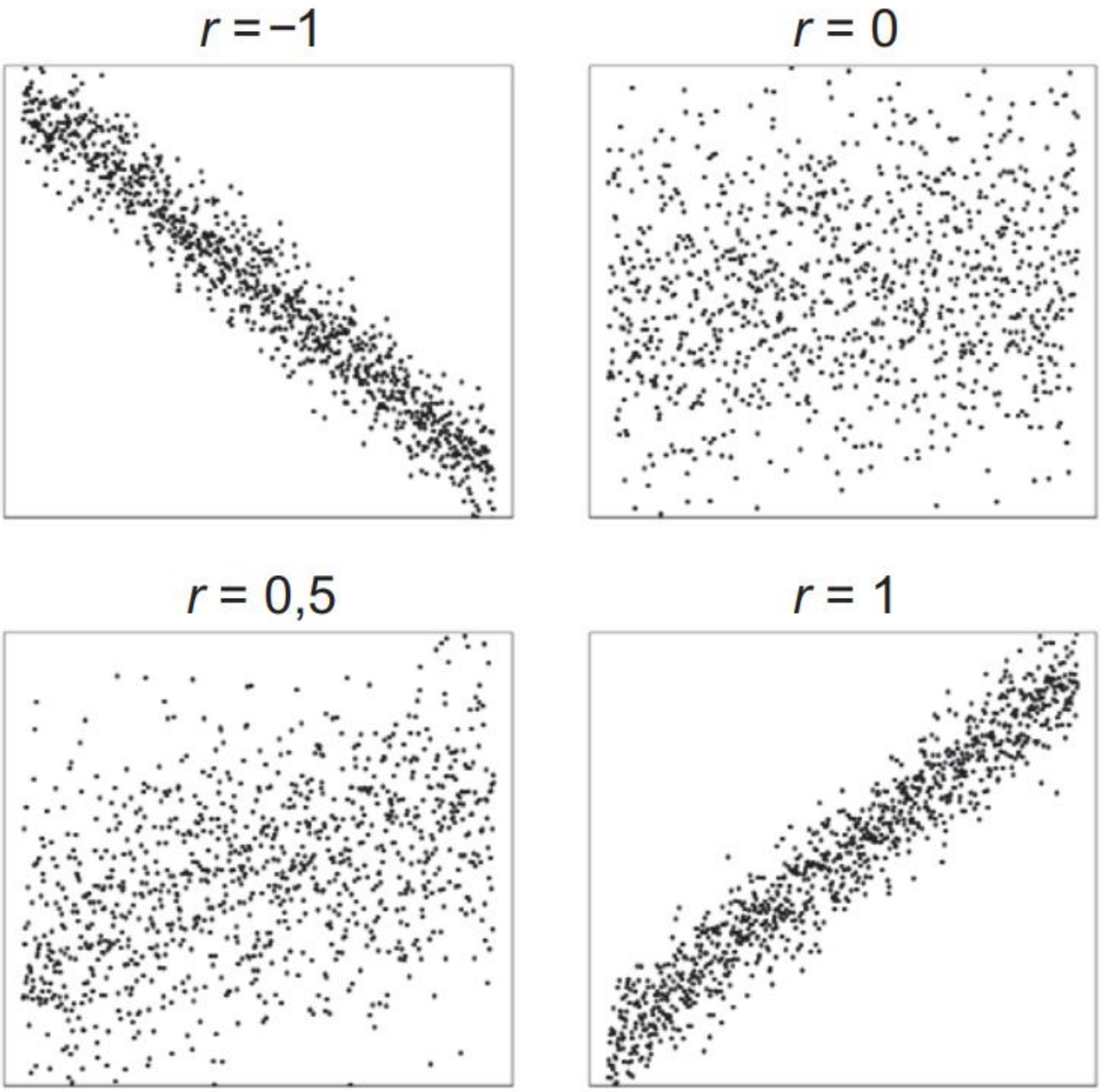


Рис. 6. Пример распределения данных в соответствии с различными коэффициентами корреляции

Регрессионный анализ

Коэффициенты корреляции

Величина. Чем ближе коэффициент к −1 или 1, тем сильнее предиктор. Например, коэффициент корреляции, показанный линией тренда на рис. 1, равен 0,7, в то время как на рис. 2, b это –0,8. Это означает, что достаток соседей — более достоверный предиктор цен на дома, чем число комнат. Нулевая корреляция означала бы отсутствие связи между предиктором и результатом.

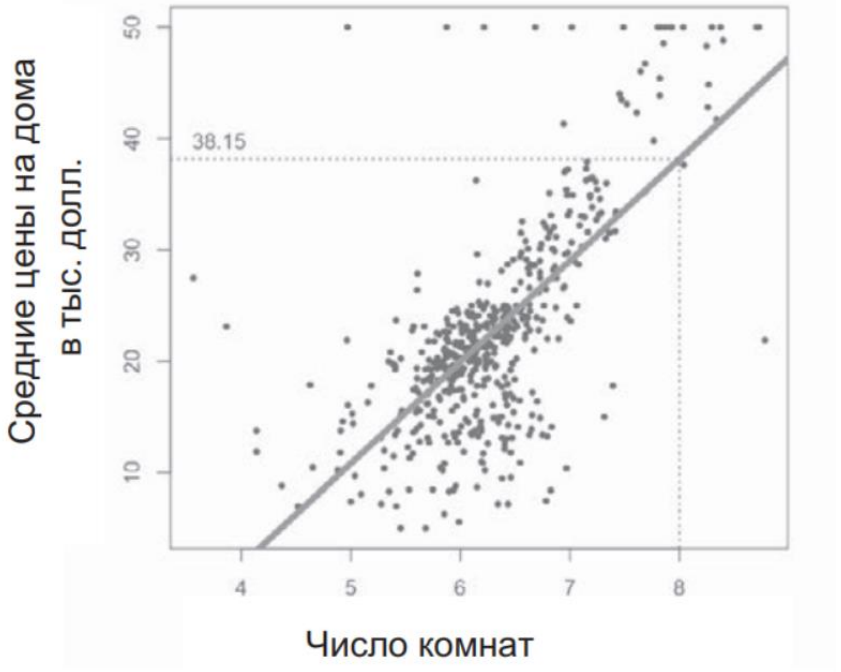


Рис. 1. Цены на дома в сравнении с числом комнат

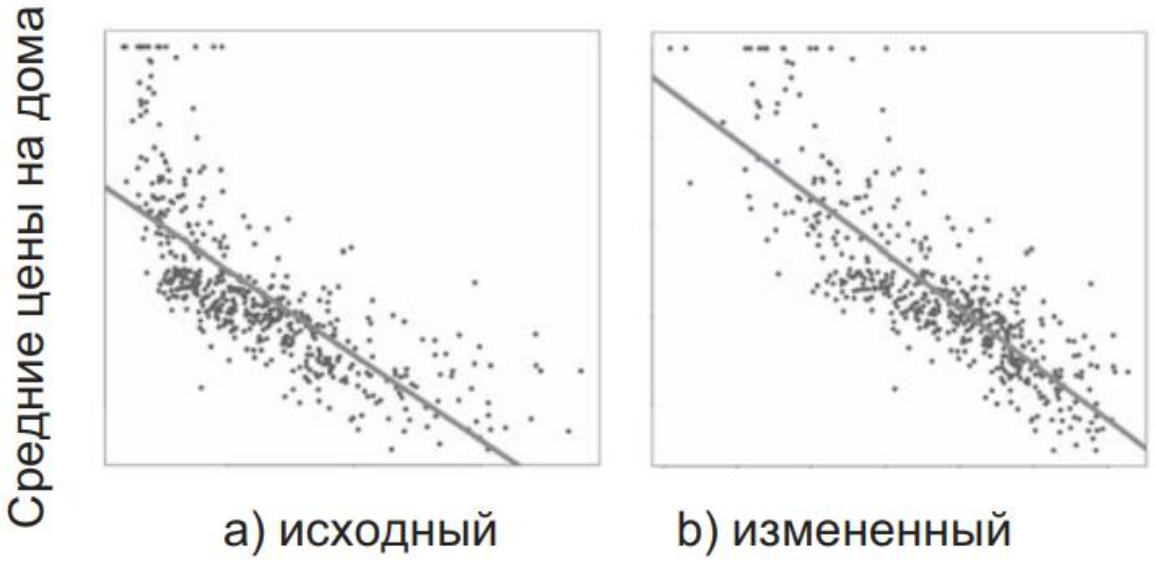


Рис. 2. Цены на дома в сравнении с долей соседей с низким доходом

Регрессионный анализ

Коэффициенты корреляции показывают абсолютную силу отдельных предикторов и, следовательно, являются более надежным способом их ранжирования, чем **коэффициенты регрессии**.

Несмотря на то что регрессионный анализ информативен и не требует долгих вычислений, он имеет **недостатки:**

- **Чувствительность к резко отклоняющимся значениям**. Регрессионный анализ одинаково учитывает все предоставленные элементы данных. Если среди них будет хотя бы несколько элементов с крайними значениями, это может значительно исказить линию тренда. Чтобы избежать этого, можно использовать диаграмму рассеяния для предварительного выявления таких резко отклоняющихся значений.
- **Искажение веса при корреляции предикторов**. Включение в регрессионную модель высококоррелирующих предикторов исказит интерпретацию их веса. Эта проблема называется **мультиколлинеарностью**. Для преодоления мультиколлинеарности нужно либо исключить из анализа коррелирующие предикторы, либо воспользоваться более продвинутым методом, таким как **лассо** или **риджрегрессия** (или гребневая регрессия).

Регрессионный анализ

Несмотря на то что регрессионный анализ информативен и не требует долгих вычислений, он имеет **недостатки:**

- **Криволинейные тренды.** В нашем примере тренды отображались прямой линией. Тем не менее некоторые тренды могут быть криволинейными, как на рис. 2, а. В этом случае потребуется преобразовать значения предикторов или использовать альтернативные алгоритмы, такие как метод опорных векторов.
- **Корреляция не говорит о причинности.** Предположим, была обнаружена положительная корреляция между стоимостью дома и наличием собаки. Понятно, что если просто завести собаку, цена дома от этого не изменится, однако можно предположить, что те, кто могут позволить себе содержать собак, располагают в среднем большим доходом и, вероятно, проживают в районах, где дома стоят дороже.

Несмотря на эти ограничения, регрессионный анализ остается одним из основных, простых в использовании и интуитивно-понятных методов для прогнозирования. Внимательное отношение к способу интерпретации результатов – залог уверенности в точности выводов.

Регрессионный анализ

- Регрессионный анализ находит линию наилучшего соответствия, тяготеющую к максимально возможному числу элементов данных.
- Линия тренда выводится на основании уравновешенной комбинации предикторов. Вес предиктора называется *коэффициентом регрессии*. Он показывает силу одного предиктора в присутствии других.
- Регрессионный анализ хорошо работает в условиях низкой корреляции между предикторами, отсутствия резко отклоняющихся значений и там, где линия тренда ожидается в виде прямой линии.