Методы анализа данных

Якушева И.А.
irkinok@mail.ru

Раздел 1. Введение в анализ данных

Тема 1. Данные

Раздел 1. Введение в анализ данных
Тема 2. Статистические методы анализа данных

2.1 MS Excel как инструмент статистического анализа

2.2 Корреляционный анализ

2.3 Регрессионный анализ

2.4 Дискриминантный анализ

2.5 Анализ табличных данных

Раздел 1. Введение в анализ данных

Тема 3. Хранение данных

3.1 Технология баз данных
3.2 OLTP-системы. Требования к OLTP-системам
3.3 Неэффективность использования OLTP-систем для анализа данных
3.4 СППР-системы
3.5 Хранилища данных. Витрины данных
3.6 OLAP-системы. Гиперкубы.

Раздел 2. Интеллектуальный анализ данных
Тема 1. Основные положения Data Mining

1.1 Сравнение Data Mining с другими методами анализа данных

1.2 Сфера применения Data Mining

1.3 Перспективы технологии Data Mining

1.4 Задачи Data Mining

Раздел 2. Интеллектуальный анализ данных

Тема 2. Методы классификации и прогнозирования

2.1 Задача классификации
2.2 Задача прогнозирования
2.3 Метод k-ближайших соседей
2.4 Метод опорных векторов
2.5 Метод деревьев решений
2.6 “Случайные леса”

Раздел 2. Интеллектуальный анализ данных

Тема 3. Методы кластеризации

3.1 Задача кластеризации

3.2 Кластеризация методом k-средних

3.3 Метод главных компонент

3.5 Анализ социальных сетей

Раздел 2. Интеллектуальный анализ данных
Тема 4. Поиск ассоциативных правил

Раздел 2. Интеллектуальный анализ данных

Тема 5. Нейронные сети

5.1 Задачи, которые ставятся перед нейронными сетями

5.2 Как работает нейронная сеть

5.3 Общая схема анализа данных с помощью нейронных сетей

Раздел 2. Интеллектуальный анализ данных
Тема 6. Технология Big Data (Анализ больших данных)

6.1 Характеристики Big Data

6.2 Источники больших данных

6.3 Методы и техники анализа, применимые к большим данным

6.4 Технологии. Аппаратные решения

Раздел 2. Интеллектуальный анализ данных

7. Лабораторный практикум

7.1 Лабораторная работа 1.Знакомство с аналитической платформой Deductor Studio (2 часа)

7.2 Лабораторная работа 2. Поиск ассоциативных правил в Deductor Studio (6 часов)

1.1. Информация. Данные. Знания

Каждое из этих понятий имеет свое собственное определение

Информация – это любые сведения о событии, явлении, факте, концепции и так далее.

Данные – информация, представленная в виде, пригодном для ее обработки некоторым устройством. В широком понимании данные представляют собой факты текст, графики, картинки, звуки, аналоговые или цифровые видео-сегменты.

Знания – информация, проверенная опытом.

1.2 Шкалы измерений

Данные получаются в результате **измерений**.

Существует пять типов **шкал измерений**: номинальная, порядковая, интервальная, относительная и дихотомическая.

Номинальная шкала (nominal scale) - шкала, содержащая только категории; данные в ней не могут упорядочиваться, с ними не могут быть произведены никакие арифметические действия.

Порядковая шкала (ordinal scale) - шкала, в которой числа присваивают объектам для обозначения относительной позиции объектов, но не величины различий между ними.

1.2 Шкалы измерений

Данные получаются в результате **измерений**.

Существует пять типов **шкал измерений**: номинальная, порядковая, интервальная, относительная и дихотомическая.

Интервальная шкала (interval scale) - шкала, разности между значениями которой могут быть вычислены, однако их отношения не имеют смысла.

Относительная шкала (ratio scale) - шкала, в которой есть определенная точка отсчета и возможны отношения между значениями шкалы.

Дихотомическая шкала (dichotomous scale) - шкала, содержащая только две категории.

1.3 Классификация данных

Данные могут являться числовыми либо символьными.

Числовые данные, в свою очередь, могут быть **дискретными и непрерывными**.

.

Дискретные данные являются значениями признака, общее число которых конечно либо бесконечно, но может быть подсчитано при помощи натуральных чисел от одного до бесконечности.

Непрерывные данные - данные, значения которых могут принимать какое угодно значение в некотором интервале. Измерение непрерывных данных предполагает большую точность.

1.3 Классификация данных

По критерию постоянства своих значений в ходе решения задачи данные могут быть:

- переменными;
- постоянными;
- условно-постоянными.

.

Переменные данные - это такие данные, которые изменяют свои значения в процессе решения задачи.

Постоянные данные - это такие данные, которые сохраняют свои значения в процессе решения задачи (математические константы, координаты неподвижных объектов) и не зависят от внешних факторов.

Условно-постоянные данные - это такие данные, которые могут иногда изменять свои значения, но эти изменения не зависят от процесса решения задачи, а определяются внешними факторами.

1.3 Классификация данных

Следует различать данные за период и точечные данные. Эти различия важны при проектировании системы сбора информации, а также в процессе измерений:

- данные за период;
- точечные данные.

.

Данные за период характеризуют некоторый период времени. Примером данных за период могут быть: прибыль предприятия за месяц, средняя температура за месяц.

Точечные данные представляют значение некоторой переменной в конкретный момент времени.

1.3 Классификация данных

Данные бывают первичными и вторичными.

Вторичные данные - это данные, которые являются результатом определенных вычислений, примененных к **первичным** данным.

Вторичные данные, как правило, приводят к ускоренному получению ответа на запрос пользователя за счет увеличения объема хранимой информации.

1.4 Подготовка данных

Если качество данных низкое, то результаты даже самого изощренного анализа окажутся не очень-то и хорошими.

Формат данных

Обычно для анализа данных используют табличное представление. Каждая строка таблицы представляет собой *элемент данных* с описанием отдельного наблюдения, а каждый столбец несет *переменную* для его описания. **Переменные** так же называются атрибутами или признаками, или размерностями.

В зависимости от цели можно изменить представленный в строках тип наблюдений.

1.4 Подготовка данных

Типы переменных

Есть 4 главных типа переменных:

- **Бинарная**. Это простейший тип переменных только с двумя вариантами значений. В таблице первой бинарная переменная показывает, брал ли покупатель рыбу.
- **Категориальная**. Если вариантов больше двух, информация может быть представлена категориальной переменной (вид покупателя).
- **Целочисленная.** Такой тип используется, когда информация может быть представлена целым числом (количество купленных каждым покупателем фруктов.
- **Непрерывная** (количественная). Это самая подробная переменная. Она содержит числа со знаками после запятой (количество потраченных покупателем денег.

1.4 Подготовка данных

Выбор переменных

В нашем первоначальном наборе данных (генеральная совокупность) может быть много разных переменных, применение в алгоритме слишком большого их числа ведет к замедлению вычислений или к ошибочным предсказаниям из-за информационного шума. Поэтому имеет смысл остановиться на коротком списке важнейших переменных.

Выбор переменных часто делается методом проб и ошибок. Переменные имеет смысл добавлять и убирать, учитывая промежуточные результаты.

1.4 Подготовка данных

Конструирование признаков

Иногда хорошие переменные требуется сконструировать. Например, если мы хотим предсказать, кто из покупателей первой таблицы не будет брать рыбу, то можем посмотреть на переменную их вида, заключив, что кролики, лошади и жирафы рыбу не покупают. А если мы группируем виды покупателей в более широкие категории – травоядных, хищников и всеядных, то получим более универсальный вывод – травоядные рыбу не берут.

Конструировать переменные можно и другими способами, мы их рассмотрим дальше относительно конкретных методов анализа данных.

1.4 Подготовка данных

Неполные данные

Мы не всегда располагаем полными данными. Иногда значения каких-то признаков бывает попросту неизвестно. Неполные данные мешают анализу и при любой возможности с ними надо разобраться одним из следующих способов:

Приближение. Если пропущено значение бинарного или категориального типа, то его можно заменить самым типичным значением (модой) переменной. Для целочисленных и непрерывных переменных используется медиана.

Вычисление. Пропущенные значения могут быть вычислены с применением более продвинутых алгоритмов обучения с учителем. Такие вычисления требуют времени, но обычно приводят к более точным оценкам неполных значений

Удаление. В качестве последнего средства строки с неполными значениями могут быть удалены. Этого обычно избегают, чтобы не уменьшать объем данных, доступных для анализа.

1.4 Подготовка данных

Подготовка данных является первым шагом в исследовании Data Science.

Следующие шаги:

2) – **выбор алгоритмов** для моделирования этих данных;

3) – **настройка алгоритмов** для оптимизации моделей;

4) – **оценка моделей**, основанная на их точности.

3.1 Технология баз данных

Для решения задач анализа данных и поиска решений необходимо накопление и хранение достаточно больших объемов данных. Этим целям служат базы данных.

- **База данных** – это совокупность специальным образом организованных и взаимосвязанных данных. Системы, предоставляющие средства для работы с базами данных, называются системы управления базами данных (**СУБД**).
- По сути база данных отражает некоторую модель предметной области (части реального мира). Структура БД должна максимально соответствовать модели заданной предметной области.

3.1 Технология баз данных

Транзакция – это последовательность операций над БД, которую СУБД рассматривает как единое целое. Транзакция переводит БД из одного целостного состояния в другое целостное состояние.

Развитый механизм управления транзакциями в современных СУБД сделал их инструментом для построения OLTP-систем (**On-Line Transactional Processing**), основной задачей которых является обеспечение выполнения операций над базой данных.

OLTP-системы характеризуются большим количеством изменений, одновременным обращением многих пользователей к одним и тем же данным для выполнения операций чтения, записи, удаления или модификации данных.

Требования к OLTP-системам и системам анализа данных разные.

3.1 Технология баз данных

Общая идея хранилищ данных заключается в разделении БД для OLTP-систем и для выполнения анализа.

Хранилище данных – это предметно-ориентированный, неизменчивый, поддерживающий хронологию набор данных, организованный для целей поддержки принятия решений.

Свойства ХД отражены в определении:

- предметная ориентация,
- интеграция,
- неизменяемость,
- поддержка хронологии.

Снижения затрат на создание ХД можно добиться, создавая его упрощенный вариант – **витрину данных** (Data Mart).

3.1 Технология баз данных

Для анализа информации наиболее удобной является **многомерная модель** или **гиперкуб**, ребрами которых являются измерения. **Измерение** – это последовательность значений одного из анализируемых параметров.

На пересечениях осей измерений располагаются данные, количественно характеризующие анализируемые факты - **меры**, это могут быть объемы продаж, выраженные в единицах продукции или в денежном выражении, остатки на складе, издержки и так далее.

Число измерений (осей) гиперкуба определяется факторами, важными для деятельности предприятия. При этом в качестве осей могут использоваться, например, категории услуг, тарифы, география и объемы подключений, классы абонентов, время и т.д.

3.1 Технология баз данных

С концепцией многомерного анализа тесно связан оперативный анализ, который выполняется средствами OLAP-систем.

OLAP (On-Line Analytical Processing) - технология оперативной аналитической обработки данных, использующая методы и средства для сбора, хранения и анализа многомерных данных в целях поддержки процессов принятия решений.

OLAP-система включает в себя два основных компонента:

- **OLAP-сер**вер – обеспечивает хранение данных, выполнение над ними необходимых операций и формирование многомерной модели. В настоящее время OLAP-серверы объединяют с ХД или ВД.
- **OLAP-клиент** – представляют пользователю интерфейс к многомерной модели данных.

1.1. Сравнение Data Mining с другими методами анализа данных

Data Mining - это процесс обнаружения в сырых данных **ранее неизвестных**, **нетривиальных**, **практически полезных** и **доступных интерпретации** знаний, необходимых для принятия решений в различных сферах человеческой деятельности.

Одно из основных положений Data Mining - поиск неочевидных закономерностей.

Data Mining оперирует реальными значениями

Data Mining опирается на ретроспективные данные для получения ответов на вопросы о будущем.

1.2 Сфера применения Data Mining

Data Mining представляет большую ценность для руководителей и аналитиков в их повседневной деятельности.

Некоторые бизнес-приложения Data Mining

Розничная торговля

Банковское дело

Страхование

. Data Mining для научных исследований

Биоинформатика

Медицина

Фармацевтика

Молекулярная генетика и генная инженерия

Химия

1.3 Перспективы технологии Data Mining

Потенциал Data Mining дает "зеленый свет" для расширения границ применения технологии. Относительно перспектив Data Mining возможны следующие **направления развития**:

• выделение типов предметных областей с соответствующими им эвристиками, формализация которых облегчит решение соответствующих задач Data Mining

• создание формальных языков и логических средств

• создание методов Data Mining, способных не только извлекать из данных закономерности, но и формировать некие теории,

• преодоление существенного отставания возможностей инструментальных средств Data Mining от теоретических достижений в этой области.

1.4 Задачи Data Mining

Методы Data Mining помогают решить многие задачи, с которыми сталкивается аналитик.

Задача классификации сводится к определению класса (сущности) объекта по его характеристикам. Множество классов, к которым может быть отнесен объект, заранее известно

Задача регрессии тоже позволяет по известным характеристикам объекта получить значение некоторого параметра объекта. В отличие от задачи классификации значением параметра объекта является не конечное множество классов, а множество действительных чисел.

1.4 Задачи Data Mining

Методы Data Mining помогают решить многие задачи, с которыми сталкивается аналитик.

При **поиске ассоциативных правил** целью является нахождение зависимостей между объектами или событиями. Найденные зависимости представляются в виде правил и могут использоваться для лучшего понимания природы анализируемых объектов или для предсказания появления каких-то событий.

Задача кластеризации заключается в поиске независимых групп (кластеров) и их характеристик среди множества всех анализируемых данных. Группировка позволяет сократить общее число данных, и таким образом облегчить анализ.

1.4 Задачи Data Mining

По способам решения задачи разделяют на supervised learning (**обучение с учителем**) и unsupervised learning (**обучение без учителя**).

Unsupervised learning - так называют алгоритмы, используемые тогда, когда мы не знаем, какие закономерности искать, и предоставляем их поиск самим алгоритмам.

Supervised learning – алгоритмы, предсказания которых основаны на уже существующих шаблонах.

Обучение с подкреплением

В отличие от обучения с учителем и без, где модели проходят обучение и после применяются без дальнейших изменений, модель обучения с подкреплением постоянно развивается, используя результаты обратной связи.

2.1 Задача классификации

В Data Mining задачу классификации рассматривают как задачу определения значения одного из параметров анализируемого объекта на основании значений других параметров.

Определяемый параметр называют зависимой переменной, а параметры, которые участвуют в его определении – независимыми переменными..

Выделяется так называемая обучающая выборка. В нее входят объекты, для которых известны значения как зависимых, так и независимых переменных.

На основе обучающей выборки строится модель определения значений зависимой переменной. Ее часто называют **функцией классификации.**

2.1 Задача классификации

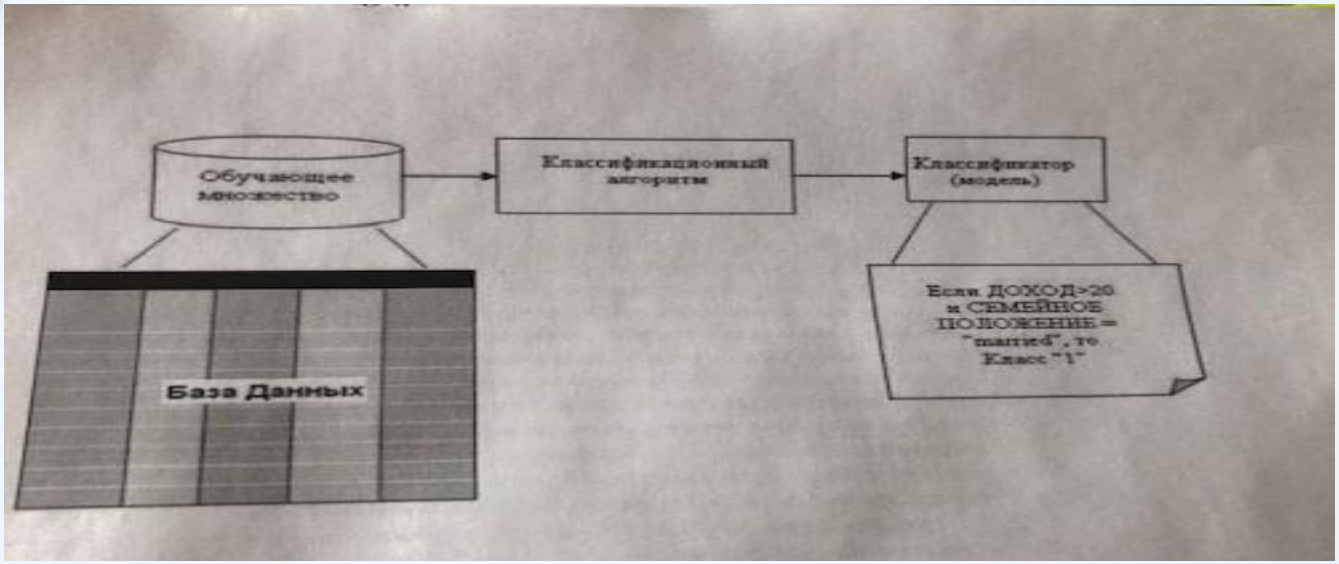
Для получения максимально точной функции к обучающей выборке предъявляются такие требования:

.

- Количество объектов, входящих в выборку, должно быть достаточно большим;
- В выборку должны входить объекты, представляющие все возможные классы в случае задачи классификации или всю область значений в случае задачи регрессии;
- Для каждого класса в задаче классификации или каждого интервала значений в задаче регрессии выборка должна содержать достаточное количество объектов.

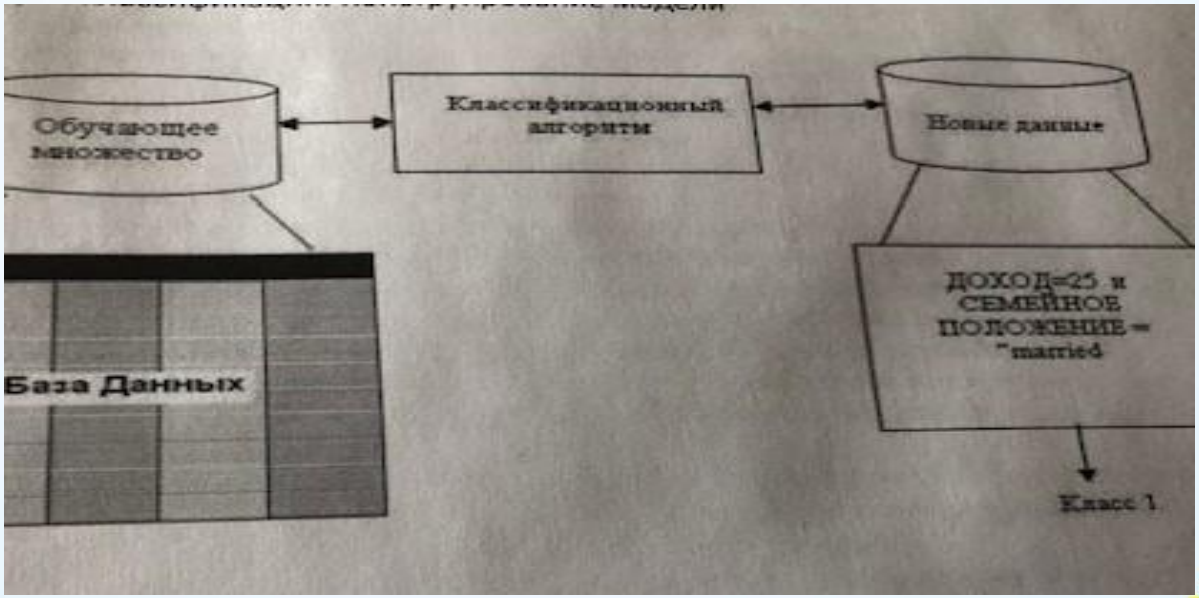
2.1 Задача классификации

На рисунке показано конструирование модели классификации



2.1 Задача классификации

На втором этапе построенную модель применяют к анализируемым объектам.
Использование модели на рисунке:



2.1 Задача классификации

Оценивание классификационных методов следует проводить, исходя из следующих характеристик:.

Скорость характеризует время, которое требуется на создание модели и ее использование.

Робастность означает возможность работы с зашумленными данными и пропущенными значениями в данных.

Интерпретируемость обеспечивает возможность понимания модели аналитиком.

Надежность методов классификации предусматривает возможность работы этих методов при наличии в наборе данных шумов и выбросов.

2.2 Задача прогнозирования

Задачи прогнозирования решаются в самых разнообразных областях человеческой деятельности, таких как наука, экономика, производство и множество других сфер.

Задача прогнозирования одна из наиболее сложных задач Data Mining, она требует тщательного исследования исходного набора данных и методов, подходящих для анализа. **Прогнозирование** - установление функциональной зависимости между зависимыми и независимыми переменными.

2.2 Задача прогнозирования

Перед началом прогнозирования необходимо ответить на следующие вопросы:

1. Что нужно прогнозировать? Определяем переменные, которые будут прогнозироваться.
2. В каких временных элементах (параметрах)? Определяем следующие параметры:
   - периода прогнозирования;
   - горизонта прогнозирования;
   - интервала прогнозирования.
3. С какой точностью прогноз?

2.2 Задача прогнозирования

Виды прогнозов. Прогноз может быть краткосрочным, среднесрочным и долгосрочным.

1. Краткосрочный прогноз представляет собой прогноз на несколько шагов вперед, то есть осуществляется построение прогноза не более чем на 3% от объема наблюдений
2. Среднесрочный прогноз - это прогноз на 3-5% от объема наблюдений, но не более 7-12 шагов вперед; так
3. Долгосрочный прогноз - это прогноз более чем на 5% от объема наблюдений.

2.2 Задача прогнозирования

Прогнозирование сходно с задачей классификации. Многие методы Data Mining используются для решения задач классификации и прогнозирования.

При решении обеих задач используется двухэтапный процесс построения модели на основе обучающего набора и ее использования для предсказания неизвестных значений зависимой переменной.

Различие задач классификации и прогнозирования состоит в том, что в первой задаче предсказывается класс зависимой переменной, а во второй - числовые значения зависимой переменной, относящиеся к будущему.

2.3 Метод k-ближайших соседей

Метод "ближайшего соседа" (“Nearest Neighbour") относится к классу методов, работа которых основывается на хранении данных в памяти для сравнения с новыми элементами.

При появлении новой записи для прогнозирования находятся отклонения между этой записью и подобными наборами данных, и наиболее подобная (или ближний сосед) идентифицируется.
Например, множество "ближайших соседей" потенциального клиента банка выбирается на основании ближайшего значения дохода, возраста и т.д.

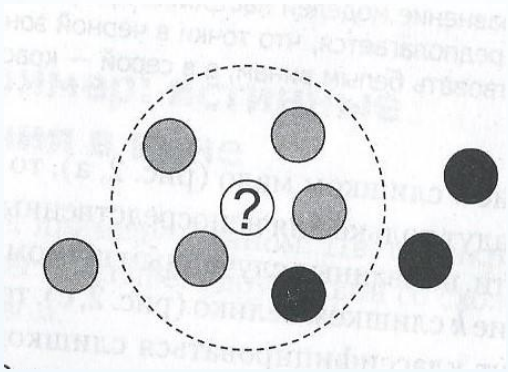
2.3 Метод k-ближайших соседей

При таком подходе используется термин **"k-ближайших соседей"** ("k-Nearest Neighbors").

Метод k-ближайших соседей – это алгоритм, который классифицирует элементы данных, исходя из класса соседних.
В названии метода параметр k означает количество ближайших соседей, которое нужно учитывать в расчетах.

2.3 Метод k-ближайших соседей

Пример. **Пять ближайших соседей (k = 5)**
Элемент данных в середине будет сочтен серым, поскольку именно этот цвет преобладает среди его ближайших соседей



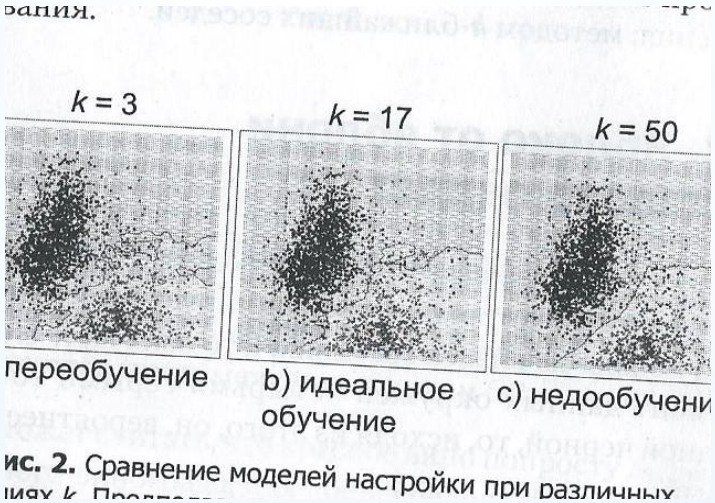
2.3 Метод k-ближайших соседей

Данный метод по своей сути относится к категории **"обучение без учителя",** то есть является "самообучающейся" технологией,

Выбор правильного значения k является примером настройки параметра и критически важен для точности классификации или прогнозирования.
Если значение k выбрано не точно (слишком мало или слишком велико), то результаты проведенного анализа будут неверными

2.3 Метод k-ближайших соседей

Пример
Сравнение моделей настройки при различных значениях k.



ис. 2. Сравнение моделей настройки при различных

2.3 Метод k-ближайших соседей

Поскольку не всегда удобно хранить все данные, иногда хранится только **множество "типичных" случаев.** В таком случае используемый метод называют рассуждением по аналогии (Case Based Reasoning, CBR), рассуждением по прецедентам.

Прецедент - это описание ситуации в сочетании с подробным указанием действий, предпринимаемых в данной ситуации.

2.3 Метод k-ближайших соседей

Подход, основанный на прецедентах, условно можно поделить на следующие этапы:

.

- **сбор** подробной информации о поставленной задаче;
- **сопоставление** этой информации с деталями прецедентов, хранящихся в базе, для выявления аналогичных случаев;
- **выбор прецедента**, наиболее близкого к текущей проблеме, из базы прецедентов;
- **адаптация** выбранного решения к текущей проблеме, если это необходимо;
- **проверка** корректности каждого вновь полученного решения;
- **занесение** детальной информации о новом прецеденте в базу прецедентов.

2.3 Метод k-ближайших соседей

Метод, основанный на прецедентах, представляет собой такой метод анализа данных, который делает заключения относительно данной ситуации **по результатам поиска аналогий**, хранящихся в базе прецедентов.

Разработка баз прецедентов по конкретной предметной области происходит на естественном для человека языке, следовательно, может быть выполнена наиболее опытными сотрудниками компании - экспертами или аналитиками, работающими в данной предметной области.

Однако это не означает, что CBR-системы самостоятельно могут принимать решения. Последнее всегда остается за человеком, данный метод лишь предлагает возможные варианты решения и указывает на самый "разумный" с ее точки зрения.

2.3 Метод k-ближайших соседей

Преимущества метода

.

• Простота использования полученных результатов.
• Решения не уникальны для конкретной ситуации, возможно их использование для других случаев.
• Целью поиска является не гарантированно верное решение, а лучшее из возможных.

2.3 Метод k-ближайших соседей

Недостатки метода "ближайшего соседа"

.

• Данный метод не создает каких-либо моделей или правил, обобщающих предыдущий опыт, - в выборе решения они основываются на всем массиве доступных исторических данных, поэтому невозможно сказать, на каком основании строятся ответы.

• Существует сложность выбора меры "близости" (метрики). От этой меры главным образом зависит объем множества записей, которые нужно хранить в памяти для достижения удовлетворительной классификации или прогноза. Также существует высокая зависимость результатов классификации от выбранной метрики.

2.3 Метод k-ближайших соседей

Недостатки метода "ближайшего соседа"

.

• При использовании метода возникает необходимость полного перебора обучающей выборки при распознавании, следствие этого - вычислительная трудоемкость.

• Типичные задачи данного метода - это задачи небольшой размерности по количеству классов и переменных.

2.4 Метод опорных векторов

Метод опорных векторов (Support Vector Machine - SVM) относится к группе граничных методов. Она определяет классы при помощи границ областей. При помощи данного метода решаются задачи бинарной классификации.

В основе метода лежит понятие **плоскостей решений**. Плоскость (plane) решения разделяет объекты с разной классовой принадлежностью.

Цель метода опорных векторов - найти плоскость, разделяющую два множества объектов.

Опорными векторами называются объекты множества, лежащие на границах областей.

Классификация считается хорошей, если область между границами пуста.

2.4 Метод опорных векторов

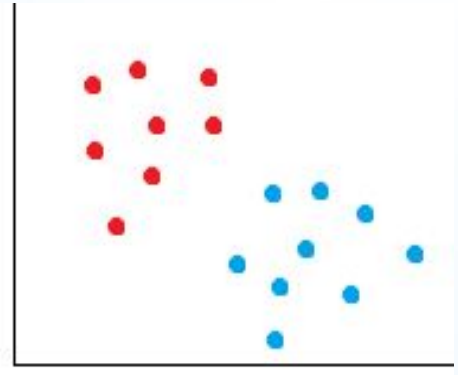
В чем же состоит идея метода опорных векторов? Давайте, сначала рассмотрим очень простой случай. .

Предположим, что у нас есть **множество точек на плоскости**, часть которых относится к **классу А**, а другая часть к **классу В** и есть точки, класс которых нужно определить. Задачу такого рода можно определить как задачу классификации, причем, для её решения, нужен алгоритм обучения "с учителем". Метод опорных векторов как раз подходит именно для таких задач.

2.4 Метод опорных векторов

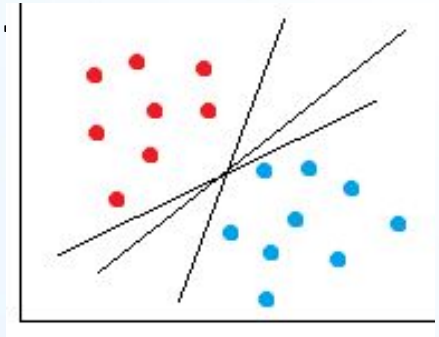
Рассмотрим упрощенный случай. Пусть точки, принадлежащие разным классам, можно разделить с помощью прямой линии:

.



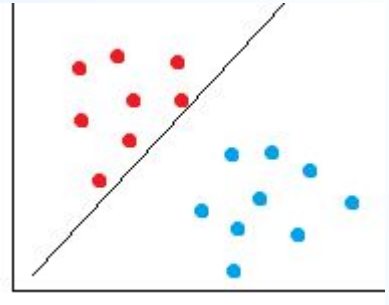
2.4 Метод опорных векторов

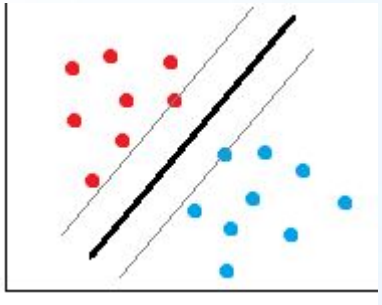
Очевидный способ решения задачи: провести прямую так, чтобы все точки одного класса лежали по одну сторону от этой прямой, а все точки другого класса были на противоположной стороне. Тогда чтобы классифицировать неизвестные точки нам нужно просто посмотреть с какой стороны прямой они окажутся



2.4 Метод опорных векторов

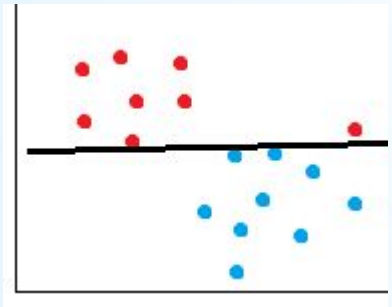
Какую из прямых выбрать? Интуитивно понятно, что нам бы хотелось прямую где-нибудь по центру. Прямая слева, очевидно, не является лучшим выбором. Лучше всего выбрать прямую, максимально удаленную от имеющихся точек.





2.4 Метод опорных векторов

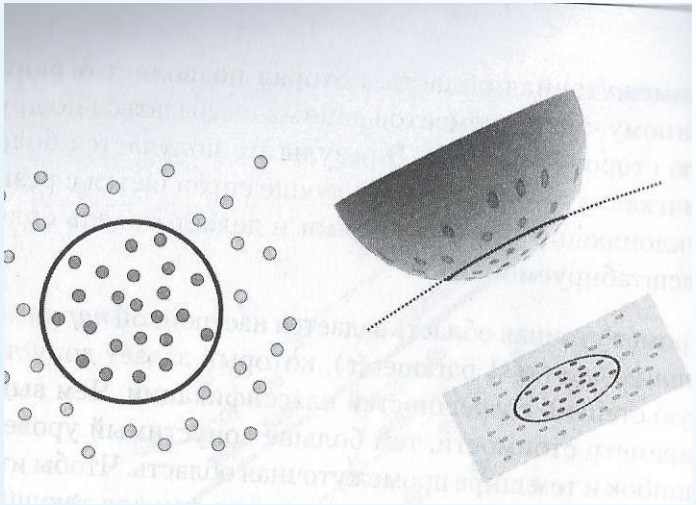
В реальной жизни, к сожалению, данные далеко не всегда можно разделить линейно. И, даже когда это можно, возникают ситуации, в которых мы не хотим этого делать. Например в таком случае:



Красная точка справа является **выбросом** и было бы лучше её игнорировать. В этом случае на помощь приходит **алгоритм с мягким зазором** (soft marging). Математически это выражается введением дополнительных параметров в систему уравнений, а по сути, мы просто назначаем некий **штраф**, за каждую точку, оказавшуюся на чужой стороне. Используя такой подход мы можем работать даже с линейно неразделимыми данными.

2.4 Метод опорных векторов

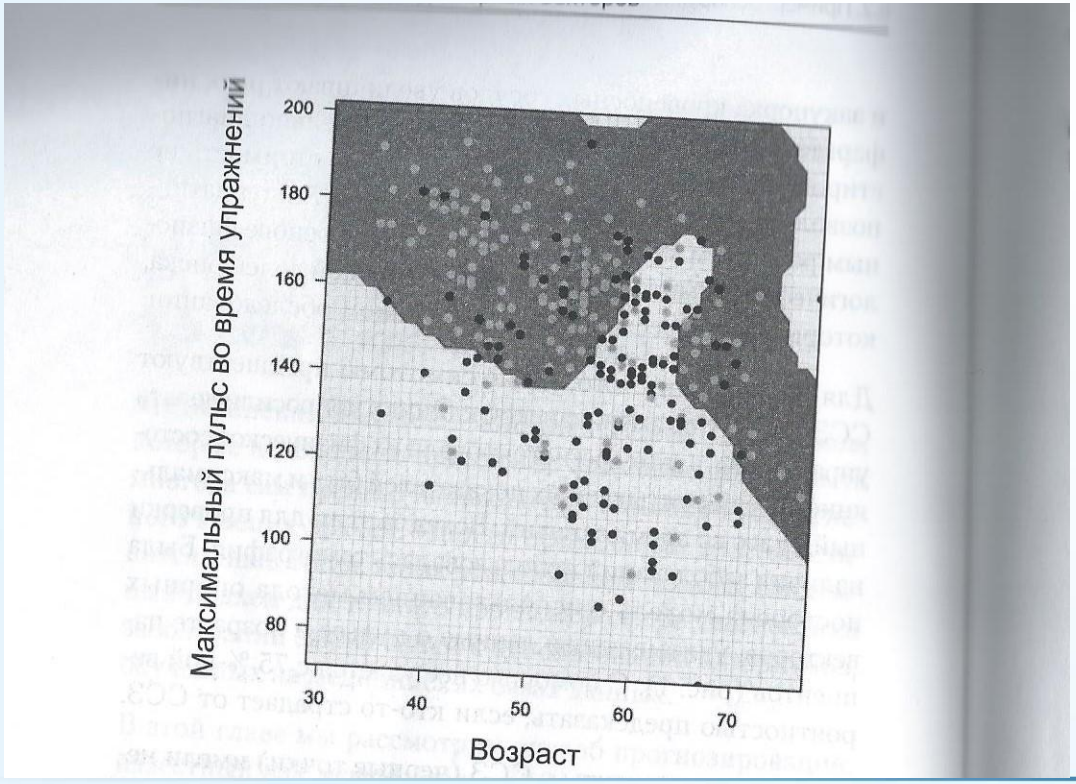
Существенное достоинство метода опорных векторов – способность обнаруживать в данных **криволинейные паттерны.**



Вместо того чтобы сразу прочертить границу на плоскости данных, метод опорных векторов сначала проецирует их на дополнительное измерение, которое может быть отделено прямой линией. Эти прямые линии легче как вычислять, так и преобразовывать в кривые при возврате к изначальной размерности.

2.4 Метод опорных векторов

Пример: обнаружение сердечно-сосудистых заболеваний.



Пациентов американской клиники попросили делать упражнения, а затем регистрировали их физическое состояние, среди показателей был и максимальный пульс во время занятий. Затем использовалась томография.
Была построена модель с использованием метода опорных векторов,

2.4 Метод опорных векторов

Метод опорных векторов является адаптивным и быстрым инструментом. Тем не менее, он может не подходить в следующих случаях:

Малые наборы данных. Поскольку для определения границ метод опирается на опорные вектора, то небольшой набор данных сокращает их число и отрицательно влияет на точность расчета.

Множество групп. Метод опорных векторов способен классифицировать данные только на две группы за раз. Если групп три и более, то надо применять метод, который называется *многоклассовая классификация*

2.4 Метод опорных векторов

Метод опорных векторов является адаптивным и быстрым инструментом. Тем не менее, он может не подходить в следующих случаях:

Большое перекрытие данных. Метод опорных векторов классифицирует элементы данных исходя из того, с какой стороны границы они оказались. Когда 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. Тем не менее, для оценки точности классификации отдельного элемента можно ориентироваться на расстояние от него до границы разделения

2.5 Метод деревьев решений

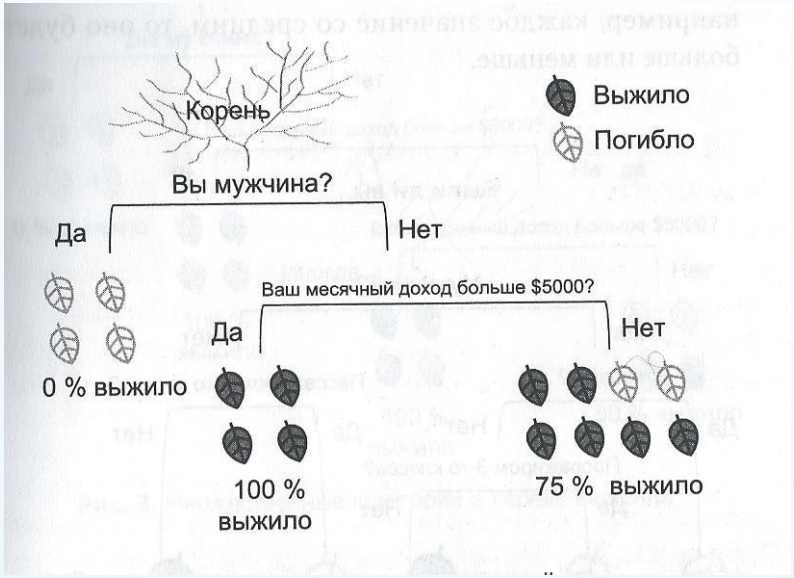
Метод деревьев решений (decision trees) является одним из наиболее популярных методов решения задач классификации и прогнозирования. Иногда этот метод Data Mining также называют деревьями решающих правил, деревьями классификации и регрессии.

Если зависимая, то есть целевая переменная принимает дискретные значения, при помощи метода дерева решений решается **задача классификации**. Если же зависимая переменная принимает непрерывные значения, то дерево решений устанавливает зависимость этой переменной от независимых переменных, т.е. решает **задачу численного прогнозирования**.

2.5 Метод деревьев решений

В наиболее простом виде дерево решений - это способ представления правил в иерархической, последовательной структуре. Основа такой структуры - ответы "Да" или "Нет" на ряд вопросов.

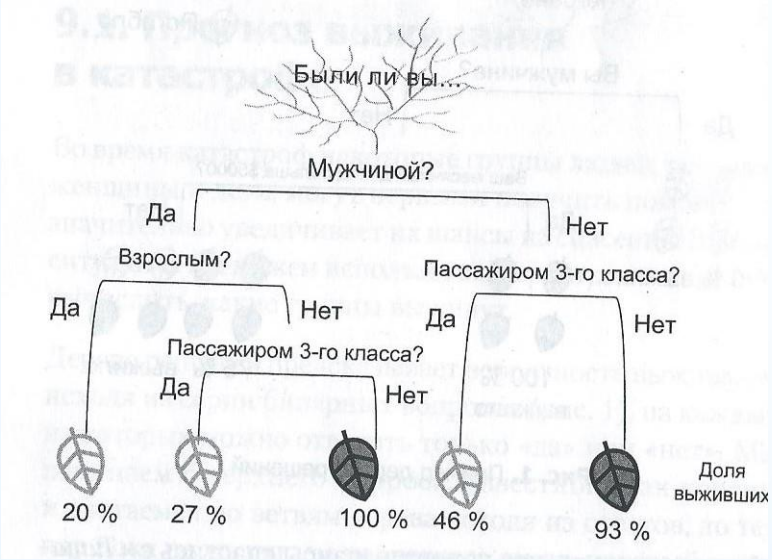
.



2.5 Метод деревьев решений

Пример – дерево решений для оценки шансов пассажиров на выживание. “Титаник”

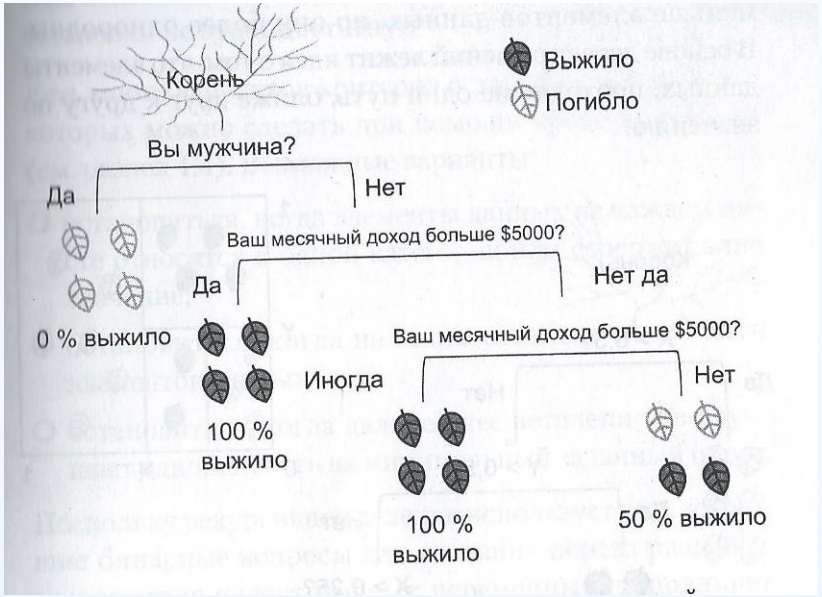
.



2.5 Метод деревьев решений

В обычных деревьях решений есть только два возможных ответа на каждом ветвлении. Если нужно учесть три и более варианта ответа (“да”, “нет”, “иногда”), то можно добавить больше ветвлений

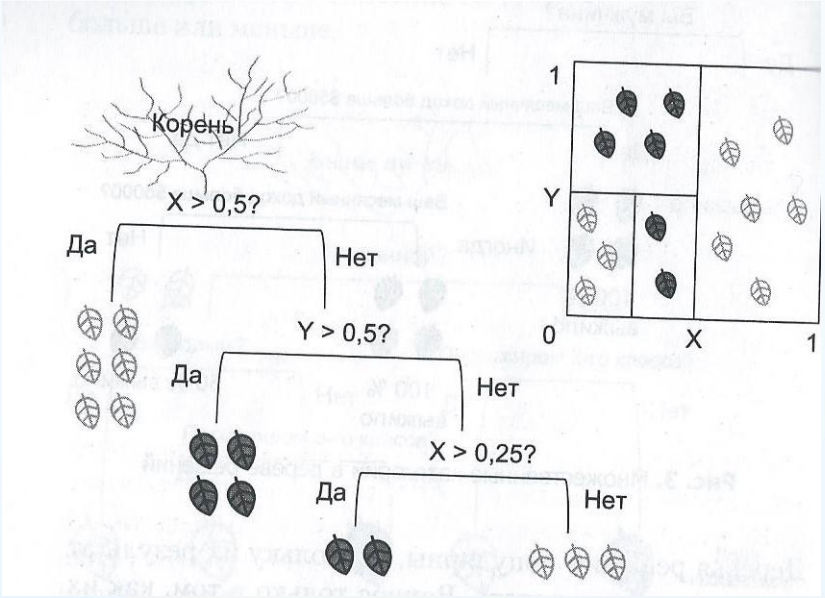
.



2.5 Метод деревьев решений

Дерево решений вырастает из разделения элементов данных на две группы так, чтобы похожие элементы оказались вместе. Далее этот процесс продолжается для каждой группы

.



2.5 Метод деревьев решений

Метод деревьев решений часто называют "наивным" подходом. Но благодаря целому ряду преимуществ, данный метод является одним из наиболее популярных для решения задач классификации. Рассмотрим эти **преимущества**.

Интуитивность деревьев решений. Классификационная модель, представленная в виде дерева решений, является интуитивной и упрощает понимание решаемой задачи. Результат работы алгоритмов конструирования деревьев решений, в отличие, например, от нейронных сетей, представляющих собой "черные ящики", легко интерпретируется пользователем.

Алгоритм конструирования дерева решений **не требует от пользователя выбора входных атрибутов** (независимых переменных). На вход алгоритма можно подавать все существующие атрибуты, алгоритм сам выберет наиболее значимые среди них, и только они будут использованы для построения дерева.

2.5 Метод деревьев решений

Метод деревьев решений часто называют "наивным" подходом. Но благодаря целому ряду преимуществ, данный метод является одним из наиболее популярных для решения задач классификации. Рассмотрим эти **преимущества**.

Точность моделей, созданных при помощи деревьев решений, сопоставима с другими методами построения классификационных моделей (статистические методы, нейронные сети).

Разработан ряд **масштабируемых алгоритмов**, которые могут быть использованы для построения деревьев решения на сверхбольших базах данных; масштабируемость здесь означает, что с ростом числа примеров или записей базы данных время, затрачиваемое на обучение, т.е. построение деревьев решений, растет линейно. Примеры таких алгоритмов: SLIQ, SPRINT.

2.5 Метод деревьев решений

Метод деревьев решений часто называют "наивным" подходом. Но благодаря целому ряду преимуществ, данный метод является одним из наиболее популярных для решения задач классификации. Рассмотрим эти **преимущества**.

Быстрый процесс обучения. На построение классификационных моделей при помощи алгоритмов конструирования деревьев решений требуется значительно меньше времени, чем, например, на обучение нейронных сетей.

Большинство алгоритмов конструирования деревьев решений имеют **возможность специальной обработки пропущенных значений.** Многие классические статистические методы, при помощи которых решаются задачи классификации, могут работать только с числовыми данными, в то время как **деревья решений работают и с числовыми, и с категориальными типами данных.**

2.5 Метод деревьев решений

Метод деревьев решений часто называют "наивным" подходом. Но благодаря целому ряду преимуществ, данный метод является одним из наиболее популярных для решения задач классификации. Рассмотрим эти **преимущества**.

Многие статистические методы являются параметрическими, и пользователь должен заранее владеть определенной информацией, например, знать вид модели, иметь гипотезу о виде зависимости между переменными, предполагать, какой вид распределения имеют данные. Деревья решений, в отличие от таких методов, **строят непараметрические модели**. Таким образом, деревья решений способны решать такие задачи Data Mining, в которых отсутствует априорная информация о виде зависимости между исследуемыми данными.

2.5 Метод деревьев решений

Несмотря на легкость интерпретации, деревья решений также имеют свои недостатки.

Нестабильность. Поскольку деревья решений строятся путем разделения элементов данных на однородные группы, небольшое изменение в этих данных способно повлиять на то, как будет выглядеть все дерево. Поскольку деревья решений стремятся к наилучшему способу разделения элементов данных, они восприимчивы к переобучению.

Неточность. Использование наилучшего бинарного вопроса ля разбивки данных не всегда ведет к точным предсказаниям. Иногда для лучшего прогнозирования нужны менее эффективные первоначальные разделения

2.6 Случайные леса

Случайный лес – это ансамбль деревьев решений

Ансамблирование – это способ комбинирования моделей для улучшения точности прогноза или классификации.

Хорошим примером ансамблей считается теорема Кондорсе «о жюри присяжных» (1784). Если каждый член жюри присяжных имеет независимое мнение, и если вероятность правильного решения члена жюри больше 0.5, то тогда вероятность правильного решения присяжных в целом возрастает с увеличением количества членов жюри и стремится к единице. Если же вероятность быть правым у каждого из членов жюри меньше 0.5, то вероятность принятия правильного решения присяжными в целом монотонно уменьшается и стремится к нулю с увеличением количества присяжных.

2.6 Случайные леса

Случайный лес – это ансамбль деревьев решений

Ансамблирование – это способ комбинирования моделей для улучшения точности прогноза или классификации.

Пример ансамблей — "Мудрость толпы".
Фрэнсис Гальтон в 1906 году посетил рынок, где проводилась некая лотерея для крестьян.
Их собралось около 800 человек, и они пытались угадать вес быка, который стоял перед ними. Бык весил 1198 фунтов. Ни один крестьянин не угадал точный вес быка, но если посчитать среднее от их предсказаний, то получим 1197 фунтов.
Эту идею уменьшения ошибки применили в теории анализа данных, а в дальнейшем в машинном обучении

2.6 Случайные леса

Пример: предсказание криминальной активности

Открытая сводка от полицейского управления Сан-Франциско представляет информацию о месте, времени и тяжести преступлений, совершенных в городе с 2014 по 2016 год.

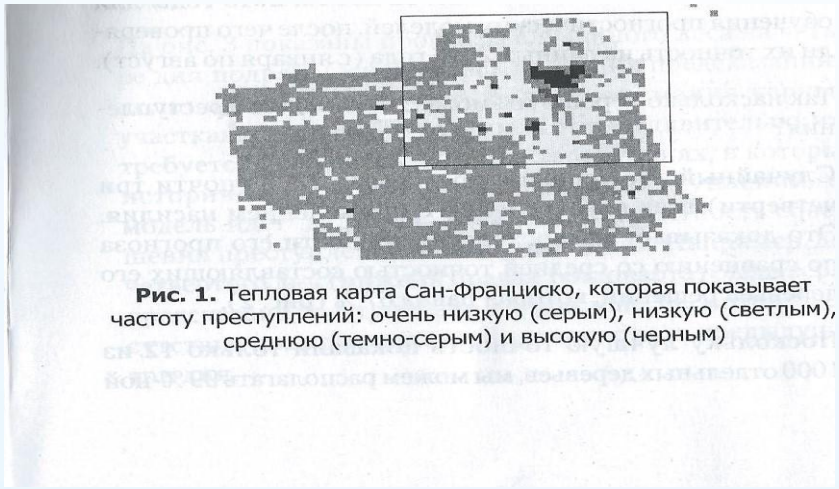
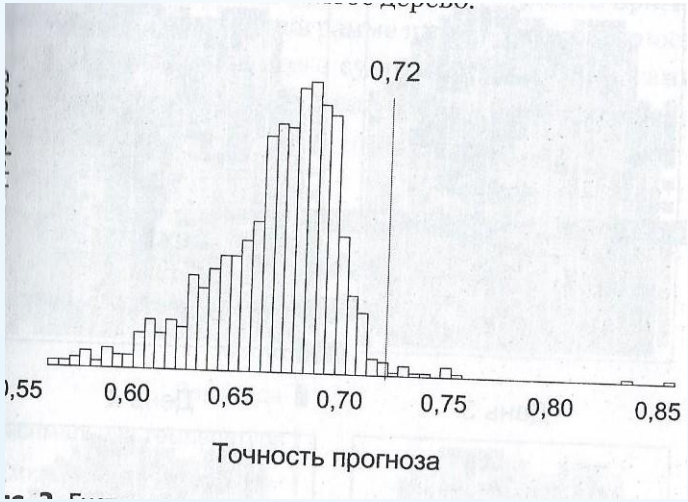


Рис. 1. Тепловая карта Сан-Франциско, которая показывает частоту преступлений: очень низкую (серым), низкую (светлым), среднюю (темно-серым) и высокую (черным)

2.6 Случайные леса

Пример: предсказание криминальной активности

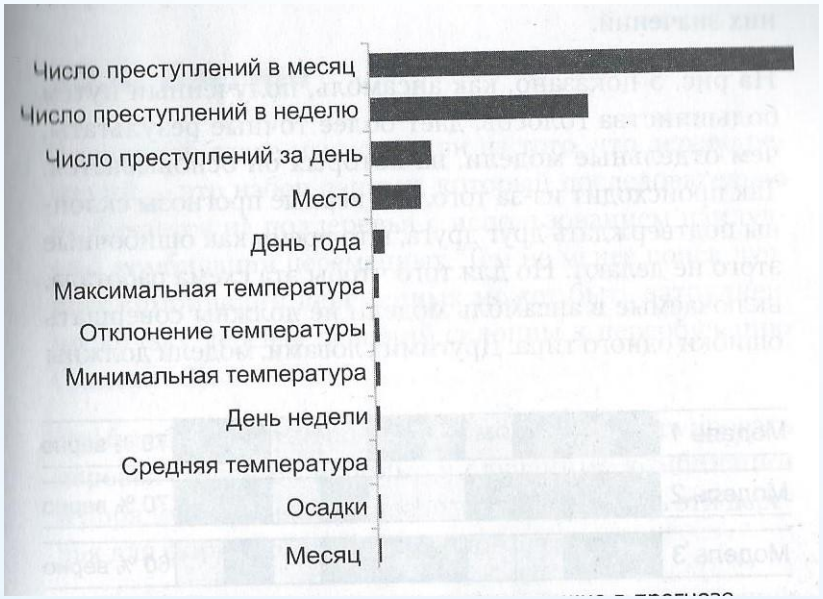
Были созданы 1000 возможных деревьев решений, которые учитывали данные по преступности и погоде. После этого построили на их основе случайный лес.



2.6 Случайные леса

Пример: предсказание криминальной активности

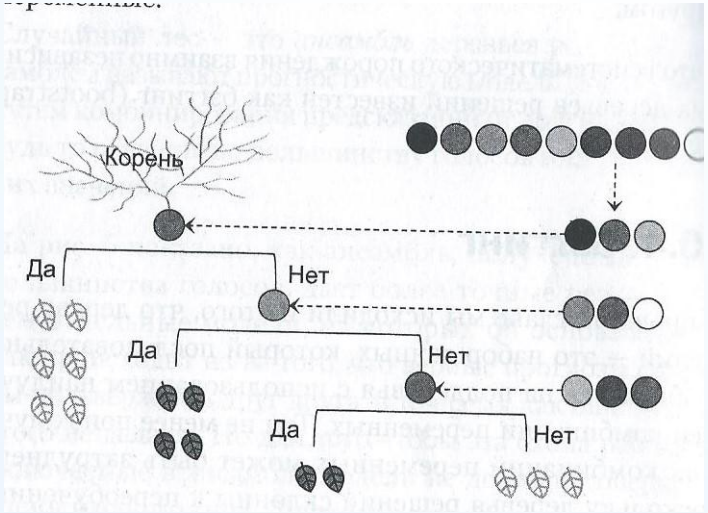
Модель случайного леса также позволяет увидеть, какие переменные сыграли наибольшую роль в прогнозировании.



2.6 Случайные леса

Методы формирования ансамблей моделей – **бустинг** и **бэггинг.**

В применении к деревьям решений бэггинг – метод систематического порождения взаимно-независимых деревьев решений.



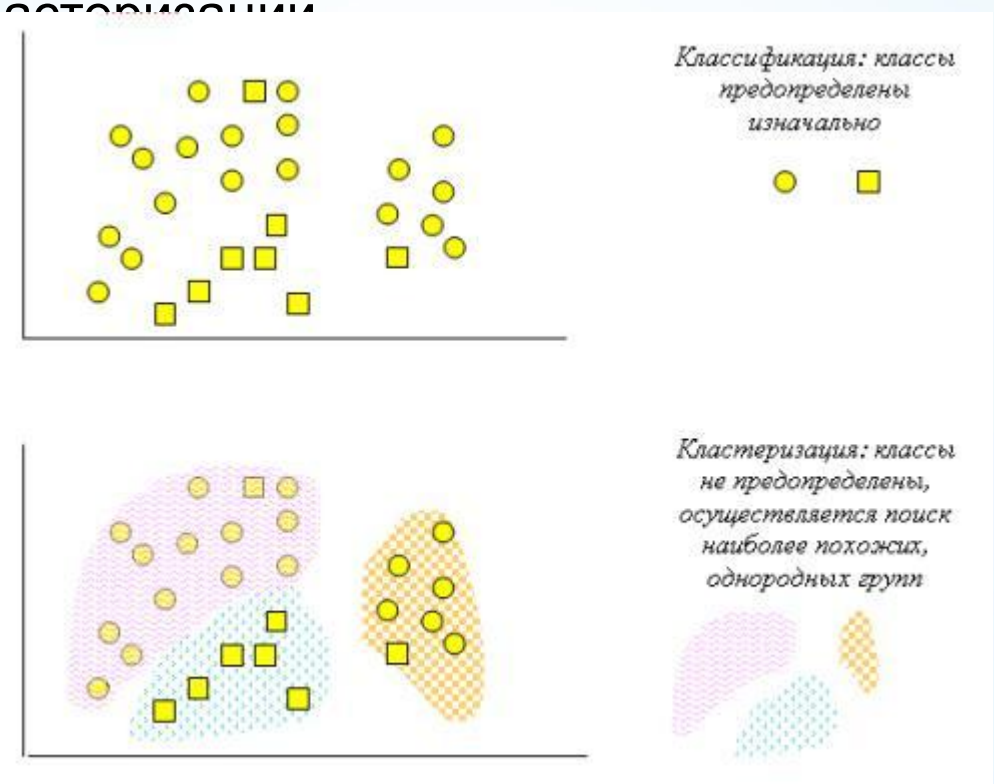
2.6 Случайные леса

Ограничения на модель

Случайные леса считаются **черными ящиками**, так как они состоят из случайно сгенерированных деревьев решений, которые не основаны на ясных прогностических принципах.
Единственное, что мы знаем, это то, что к такому заключению пришло большинство составляющих его деревьев решений.
Тем не менее случайные леса широко используются, поскольку их легко получить.
Они **очень эффективны** в ситуациях, когда точность результатов важнее их интерпретируемости.

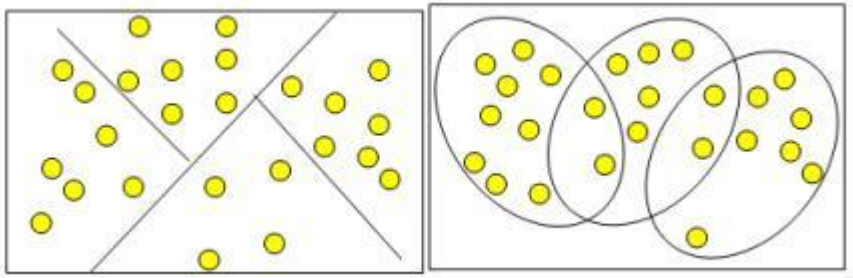
3.1 Задача кластеризации

Сравним задачи классификации и задачи кластеризации



3.1 Задача кластеризации

Кластеры могут быть непересекающимися, или эксклюзивными (non-overlapping, exclusive), и пересекающимися (overlapping)



В результате применения различных методов кластерного анализа могут быть получены кластеры различной формы. Например, возможны кластеры "цепочного" типа, когда кластеры представлены длинными "цепочками", кластеры удлиненной формы и т.д., а некоторые методы могут создавать кластеры произвольной формы.

3.1 Задача кластеризации

Термин «**кластерный анализ**» (впервые ввел Tryon в 1939 г.) в действительности включает в себя набор различных алгоритмов кластеризации

Кластерный анализ не требует априорных предположений о наборе данных, не накладывает ограничения на представление исследуемых объектов, позволяет анализировать показатели различных типов данных (интервальные данные, частоты, бинарные данные). При этом необходимо помнить, что переменные должны измеряться в сравнимых шкалах. Общий вопрос, задаваемый исследователями во многих областях, состоит в том, как разбить данные на группы с близкими значениями параметров.

3.1 Задача кластеризации

Пример. Сегментация рынка

Можно кластеризовать потребителей по двум параметрам — цены и качества.
Допустим, компания — производитель автомобилей - провела опрос потребителей, в котором задавала два вопроса: «За какую цену Вы готовы купить автомобиль?» и «Оцените качество автомобиля X по 50-балльной шкале»

3.1 Задача кластеризации

Пример. Сегментация рынка

№ участника опроса	Цена, тыс.$	Качество автомобиля X
1	27	19
2	11	46
3	25	15
4	36	27
5	35	25
6	10	43
7	11	44
8	36	24
9	26	14
10	26	14
11	9	45
12	33	23
13	27	16
14	10	47

3.1 Задача кластеризации

Если посмотреть на диаграмму (**диаграмма рассеяния**) «цена — качество», представленную на рисунке, то сразу будут видны **группы потребителей**:

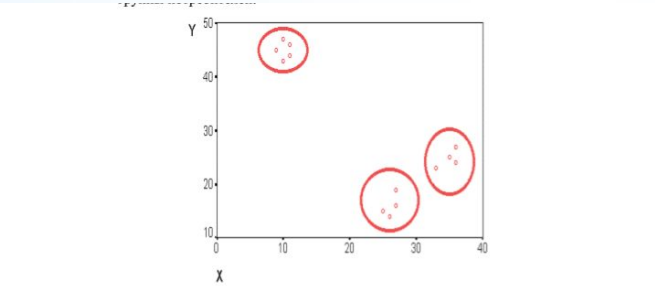


Рис. 1. Соотношение цены — качества

3.1 Задача кластеризации

В реальной жизни кластеры, различимые глазом, встречаются нечасто, гораздо чаще бывают ситуации, когда все результирующие параметры смешиваются в одну «кучу»

Рис. 2. Диаграмма рассеяния

3.1 Задача кластеризации

Для проведения кластерного анализа, кроме сбора данных, необходимо определить две вещи: на какое **количество кластеров** необходимо разделить данные и как определить **меру сходства** в данных.

Например, все предприятия России можно кластеризовать по географическому признаку на 10 кластеров. Тогда мера сходства будет определяться коммуникационной близостью предприятий друг к другу. В более сложных случаях можно применять другие меры сходства, которые подробно описаны в литературе по кластерному анализу. Существует много разных мер сходства, наиболее употребительны из них порядка десяти

3.1 Задача кластеризации

Методы кластерного анализа можно разделить на две группы:

- **иерархические;**
- **неиерархические**.

Каждая из групп включает множество подходов и алгоритмов. Используя различные методы кластерного анализа, аналитик может получить различные решения для одних и тех же данных.

3.1 Задача кластеризации

Иерархические методы кластерного анализа

Суть иерархической кластеризации состоит в последовательном объединении меньших кластеров в большие или разделении больших кластеров на меньшие.

Иерархические агломеративные методы (Agglomerative Nesting, AGNES)

Эта группа методов характеризуется последовательным объединением исходных элементов и соответствующим уменьшением числа кластеров.

В начале работы алгоритма все объекты являются отдельными кластерами. На первом шаге наиболее похожие объекты объединяются в кластер. На последующих шагах объединение продолжается до тех пор, пока все объекты не будут составлять один кластер.

3.1 Задача кластеризации

Иерархические методы кластерного анализа

Иерархические дивизимные (делимые) методы (DIvisive ANAlysis, DIANA)

Эти методы являются логической противоположностью агломеративным методам. В начале работы алгоритма все объекты принадлежат одному кластеру, который на последующих шагах делится на меньшие кластеры, в результате образуется последовательность расщепляющих групп.

Программная реализация алгоритмов кластерного анализа широко представлена в различных инструментах Data Mining, которые позволяют решать задачи достаточно большой размерности.

3.1 Задача кластеризации

Иерархические методы кластерного анализа

Иерархические методы кластеризации различаются правилами построения кластеров. В качестве правил выступают критерии, которые используются при решении вопроса о "схожести" объектов при их объединении в группу (агломеративные методы) либо разделения на группы (дивизимные методы).
Иерархические методы кластерного анализа используются при небольших объемах наборов данных.
Преимуществом иерархических методов кластеризации является их наглядность.

3.1 Задача кластеризации

Неиерархические методы кластерного анализа

При большом количестве наблюдений используют **неиерархические методы**, которые представляют собой итеративные методы дробления исходной совокупности. В процессе деления новые кластеры формируются до тех пор, пока не будет выполнено правило остановки.

Такая неиерархическая кластеризация состоит в разделении набора данных на определенное количество отдельных кластеров. Существует два подхода. Первый заключается в определении границ кластеров как наиболее плотных участков в многомерном пространстве исходных данных, т.е. определение кластера там, где имеется большое "сгущение точек". Второй подход заключается в минимизации меры различия объектов

3.2 Кластеризация методом k-средних

Наиболее распространен среди неиерархических методов алгоритм k-средних, также называемый быстрым кластерным анализом.

Алгоритм k-средних строит k кластеров, расположенных на возможно больших расстояниях друг от друга.
Выбор числа k может базироваться на результатах предшествующих исследований, теоретических соображениях или интуиции.
Общая идея алгоритма: заданное фиксированное число k кластеров наблюдения сопоставляются кластерам так, что средние в кластере (для всех переменных) максимально возможно отличаются друг от друга.

3.2 Кластеризация методом k-средних

Пример. Кинопредпочтения зрителей

Пользователей Facebook пригласили пройти опрос, чтобы распределить их, исходя из четырех свойств: экстраверсии (насколько их радуют социальные взаимодействия), добросовестности (насколько они трудолюбивы0, эмоциональности (как часто они испытывают стресс) и открытости (насколько восприимчивы к новому).
Первичный опрос показал наличие связи между этими личностными особенностями.
Для лучшей визуализации некоторые свойства были объединены.

3.2 Кластеризация методом k-средних

Рис. 1. Персональные профили кинозрител

Пример.
Кинопредпочтения зрителей

Суммарные очки черт характера были соотнесы с информацией о связанных с фильмами страницах, которые пользователь “лайкнул” .

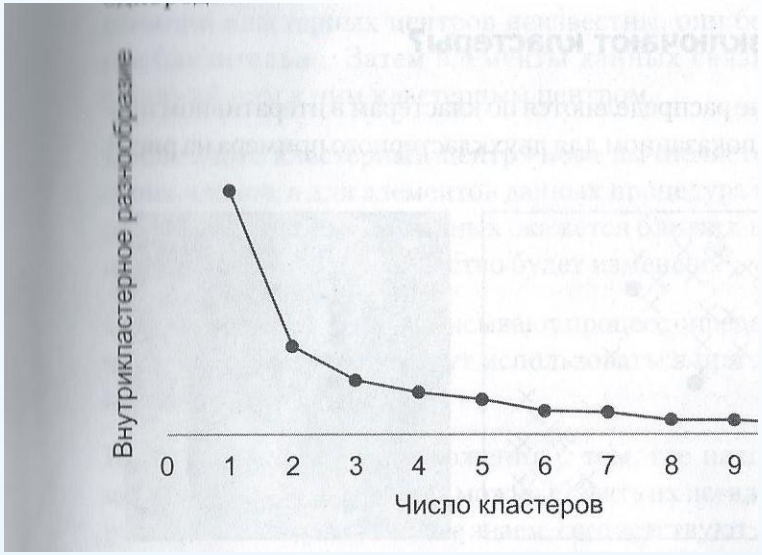
3.2 Кластеризация методом k-средних

При определении кластеров нужно ответить на два вопроса:

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

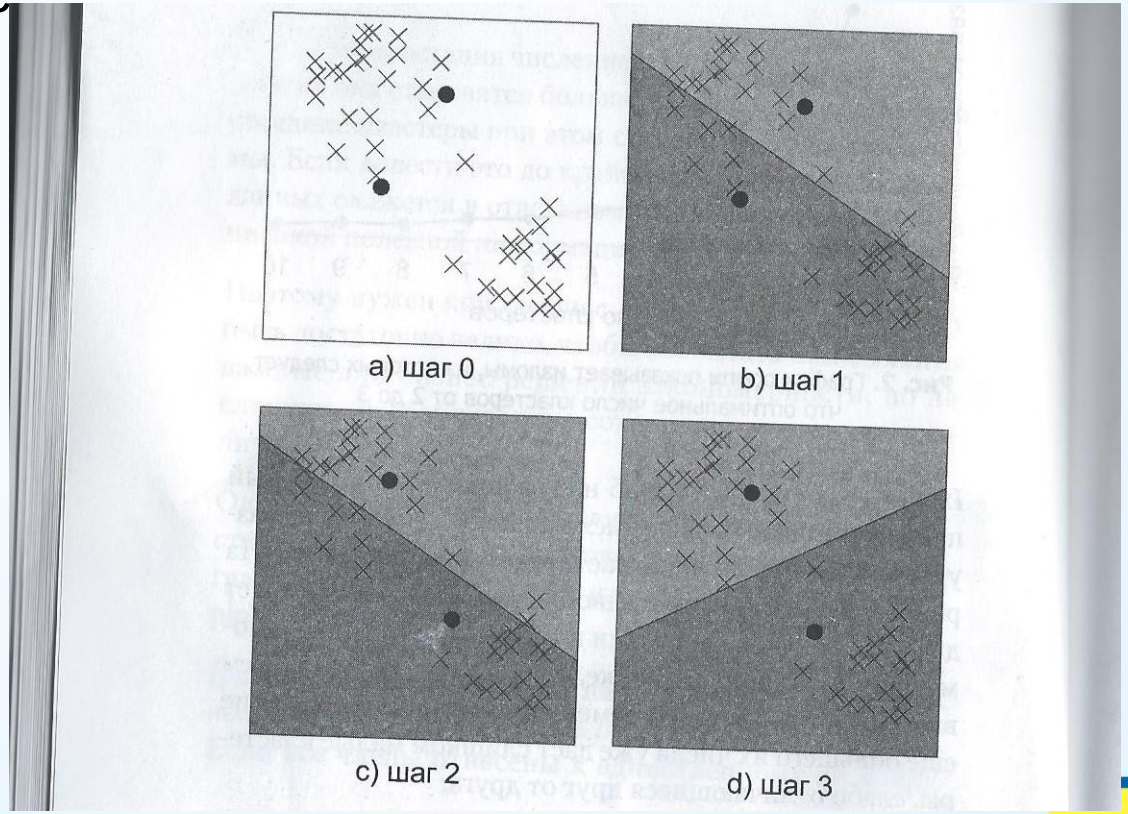
3.2 Кластеризация методом k-средних

Одним из способов определить оптимальное количество кластеров является использование так называемого **графика каменистой осыпи** или графика Кеттела



3.2 Кластеризация методом k-средних

Данные распределяются по кластерам в итеративном процессе



3.2 Кластеризация методом k-средних

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

1. Каждый элемент данных может быть связан **только с одним кластером.**
2. Предполагается, что **кластеры сферичны**. Итеративный процесс поиска ближайшего кластерного центра для элементов данных ограничен его радиусом, поэтому итоговый кластер похож на сферу. Это может стать проблемой.
3. Кластеры предполагаются **цельными**. Метод не допускает того, чтобы они пересекались или были вложены друг в друга

3.3 Метод главных компонент

Метод главных компонент – это способ нахождения основополагающих переменных (главные компоненты), которые дифференцируют ваши элементы данных оптимальным образом.

Знание о главных компонентах может иметь несколько применений:

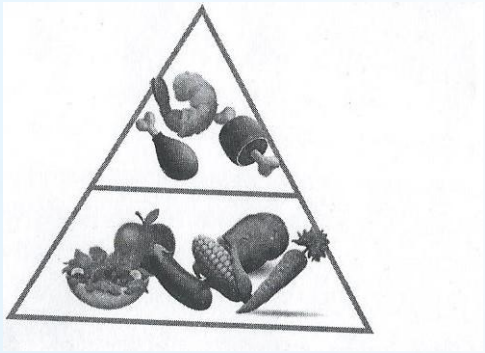
Визуализация. Отображение элементов на графике с подходящей шкалой может дать их лучшее понимание.

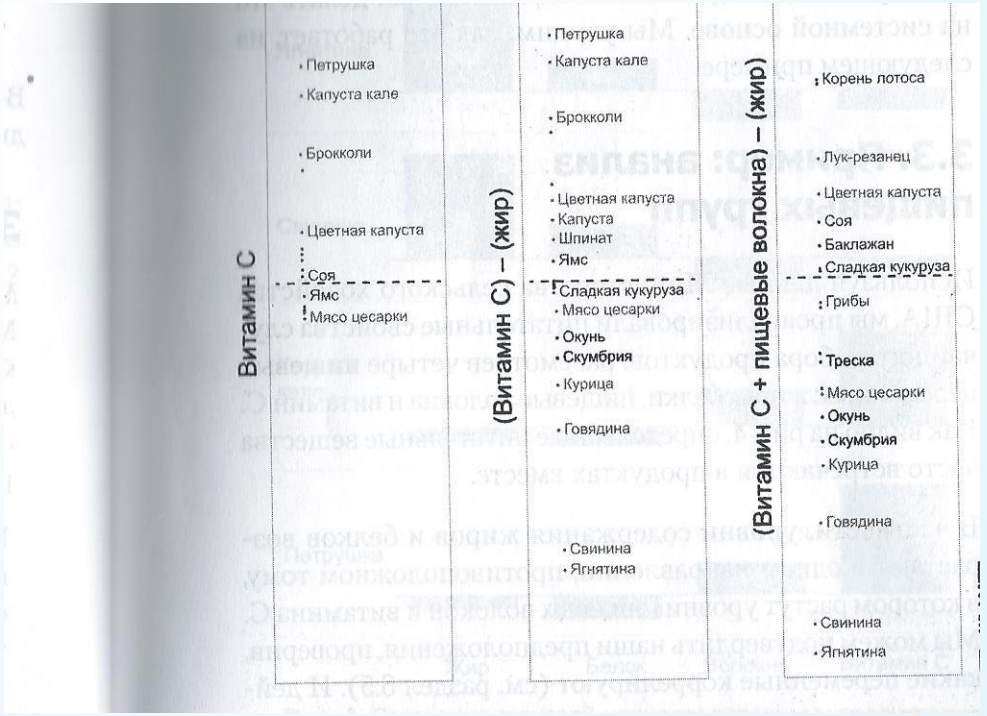
Обнаружение кластеров. При хорошей визуализации могут быть обнаружены скрытые категории или кластеры.

3.3 Метод главных компонент

Пример. Диетология.

Как лучше всего дифференцировать пищевые продукты?





3.3 Метод главных компонент

Пример. Диетология.

Как лучше всего дифференцировать пищевые продукты?

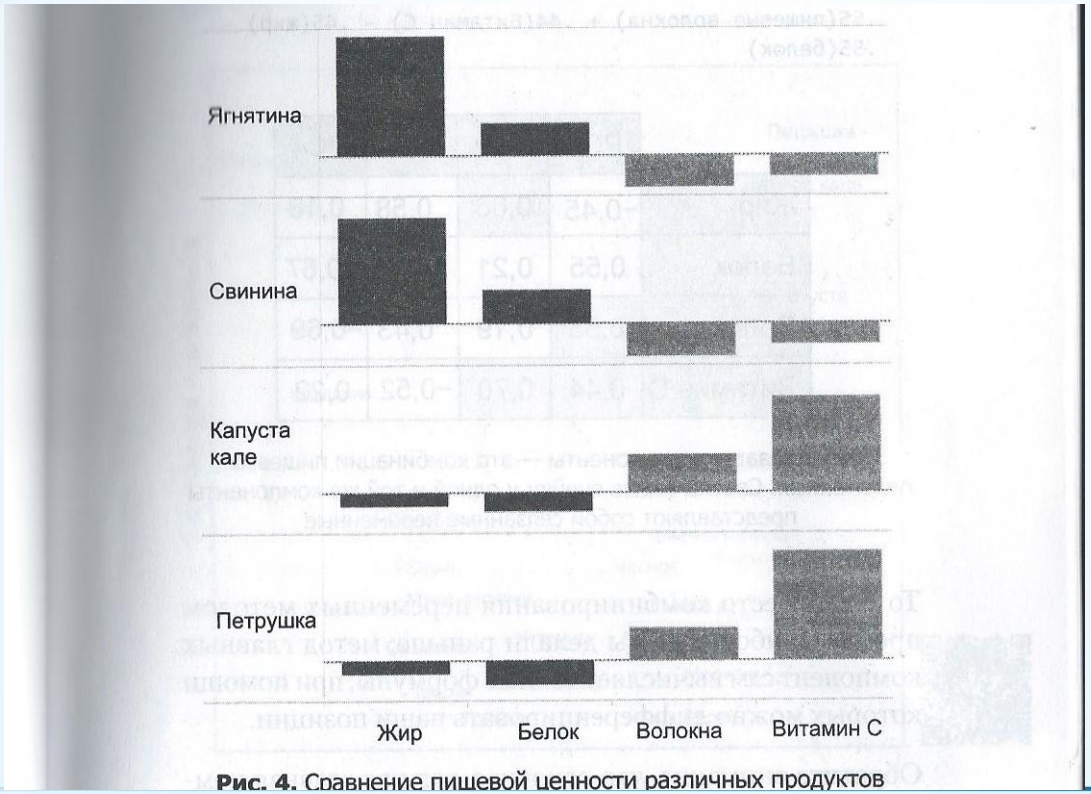


Рис. 4. Сравнение пищевой ценности различных продуктов

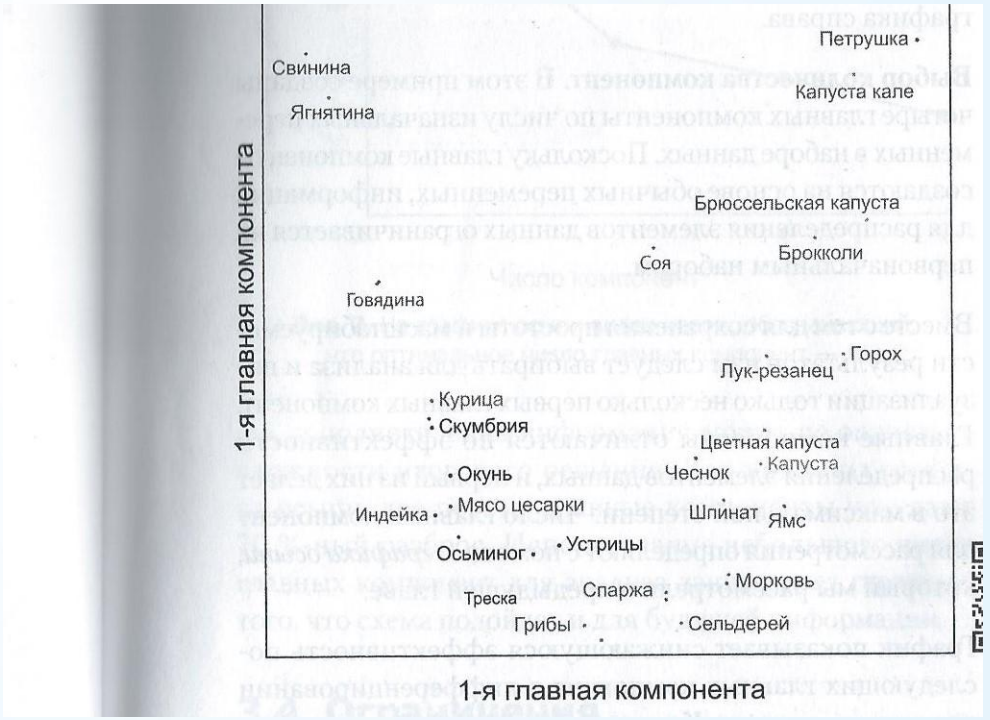
3.3 Метод главных компонент

Пример. Диетология.

Как лучше всего дифференцировать пищевые продукты?

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



3.3 Метод главных компонент

Пример. Диетология.

Как лучше всего дифференцировать пищевые продукты?

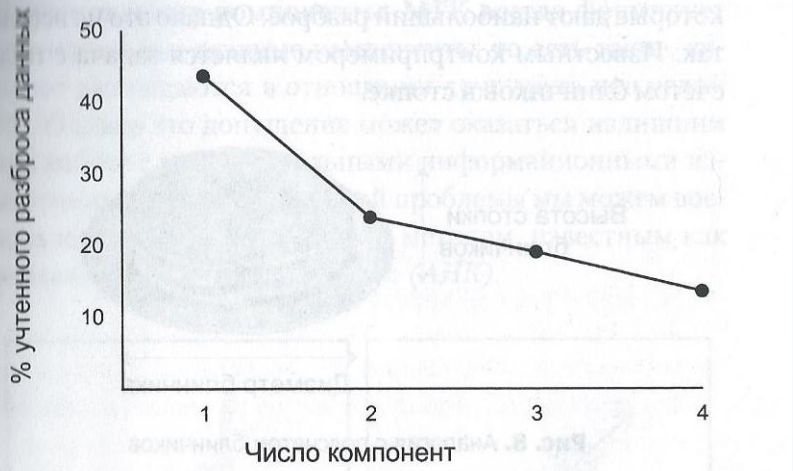


Рис. 7. На графике осыпи виден излом, обозначающий, что оптимальное число главных компонент — две

3.3 Метод главных компонент

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

Максимизация распределения. Метод исходит из такого допущения, что наиболее полезны те измерения, которые имеют наибольший разброс. Это не всегда так.

Интерпретация компонент. В методе необходима интерпретация сгенерированных компонент, иногда это сделать сложно.

Ортогональные компоненты. Это допущение может оказаться излишним при работе с неортогональными информационными измерениями

4.1 Ассоциативные правила

Целью поиска ассоциативных правил (association rule) является нахождение закономерностей между связанными событиями в базах данных.

Впервые задача поиска ассоциативных правил (association rule mining) была предложена для нахождения типичных шаблонов покупок, совершаемых в супермаркетах, поэтому иногда ее еще называют анализом рыночной корзины (market basket analysis). **Рыночная корзина** - это набор товаров, приобретенных покупателем в рамках одной отдельно взятой транзакции.

4.1 Ассоциативные правила

Понимание закономерностей рыночной корзины поможет увеличить продажи 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 другого;
- товары X и Y могут быть скомбинированы в некий новый продукт, такой как X со вкусом Y

4.1 Ассоциативные правила

Регистрируя все бизнес-операции в течение всего времени своей деятельности, торговые компании накапливают огромные собрания транзакций. Каждая такая транзакция представляет собой набор товаров, купленных покупателем за один визит.

Транзакционная или операционная база данных (Transaction database) представляет собой двумерную таблицу, которая состоит из номера транзакции (TID) и перечня покупок, приобретенных во время этой транзакции.
TID - уникальный идентификатор, определяющий каждую сделку или транзакцию.

4.1 Ассоциативные правила

Пример транзакционной базы данных, состоящей из покупательских транзакций

TID	Приобретенные покупки
100	хлеб, молоко, печенье
200	молоко, сметана
300	молоко, хлеб, сметана, печенье
400	колбаса, сметана
500	хлеб, молоко, печенье, сметана
600	Конфеты

4.1 Ассоциативные правила

Существуют три основные меры для определения ассоциаций: **поддержка**, **достоверность** и **лифт**

Имеется транзакционная база данных D. Присвоим значениям товаров переменные.
Хлеб = a Молоко = b Печенье = c Сметана = d Колбаса = e Конфеты = f
Тогда по базе данных имеем
100 a, b, c
200 b, d
300 b, a, d, c
400 e, d
500 a, b, c, d
600 f

4.1 Ассоциативные правила

Поддержкой называют количество или процент транзакций, содержащих определенный набор данных.

Поддержка показывает то, **как часто данный товарный набор появляется**, что измеряется долей покупок, в которых он присутствует. Рассмотрим набор товаров (Itemset), включающий, например, {хлеб, молоко, печенье}. Выразим этот набор с помощью переменных: abc={a,b,c}

Этот набор товаров встречается в нашей базе данных **три раза**, то есть поддержка этого набора товаров равна 3: SUP(abc)=3. При минимальном уровне поддержки, равной трем, набор товаров abc является часто встречающимся шаблоном.

Для данного набора товаров поддержка, выраженная в процентном отношении, равна 50%. SUP(abc)=(3/6)*100%=50%

4.1 Ассоциативные правила

Достоверность показывает, как часто товар Y появляется вместе с товаром X, что выражается как {X -> Y}

Рассмотрим правило "из покупки молока следует покупка печенья" для базы данных, которая была приведена выше в таблице.

Правило "Из A следует B" справедливо с достоверностью c, если c% транзакций из всего множества, содержащих набор элементов A, также содержат набор элементов B.

Число транзакций, содержащих молоко, равно четырем, число транзакций, содержащих печенье, равно трем, достоверность правила равна (3/4)*100%, т.е. 75%.

Достоверность правила "из покупки молока следует покупка печенья" равна 75%, т.е. 75% транзакций, содержащих товар A, также содержат товар B.

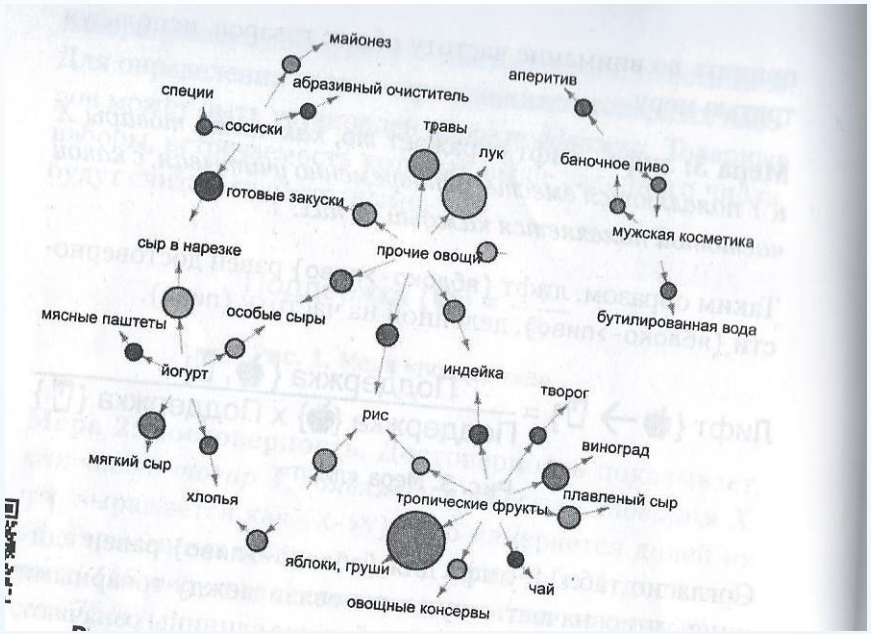
4.1 Ассоциативные правила

Лифт отражает то, как часто товары X и Y появляются вместе, одновременно учитывая, с какой частотой появляется каждый из них

Согласно определению, можно сказать, что лифт некоторого набора {X, Y} равен достоверности {X -> Y}, деленной на частоту {Y}.
Достоверность "из покупки молока следует покупка печенья" для базы данных, которая была приведена выше в таблице, может быть выражена как
поддержка {молоко, печенье} / поддержка {молоко}
Тогда
лифт {молоко -> печенье} = достоверность {молоко -> печенье} / поддержка {печенье}
В нашем случае лифт {молоко -> печенье} = ?

4.1 Ассоциативные правила

Пример: ведение продуктовых продаж. Проанализированы данные одного продуктового магазина за 30 дней.



4.2 Методы поиска ассоциативных правил

Перечислим некоторые методы и алгоритмы

Алгоритм AIS. Первый алгоритм поиска ассоциативных правил, называвшийся AIS [62], (предложенный Agrawal, Imielinski and Swami) был разработан сотрудниками исследовательского центра IBM Almaden в 1993 году.
В алгоритме AIS кандидаты множества наборов генерируются и подсчитываются "на лету", во время сканирования базы данных.
Алгоритм SETM. Создание этого алгоритма было мотивировано желанием использовать язык SQL для вычисления часто встречающихся наборов товаров. Как и алгоритм AIS, SETM также формирует кандидатов "на лету", основываясь на преобразованиях базы данных.

4.2 Методы поиска ассоциативных правил

Для улучшения работы представленных выше алгоритмов был предложен **алгоритм Apriori**

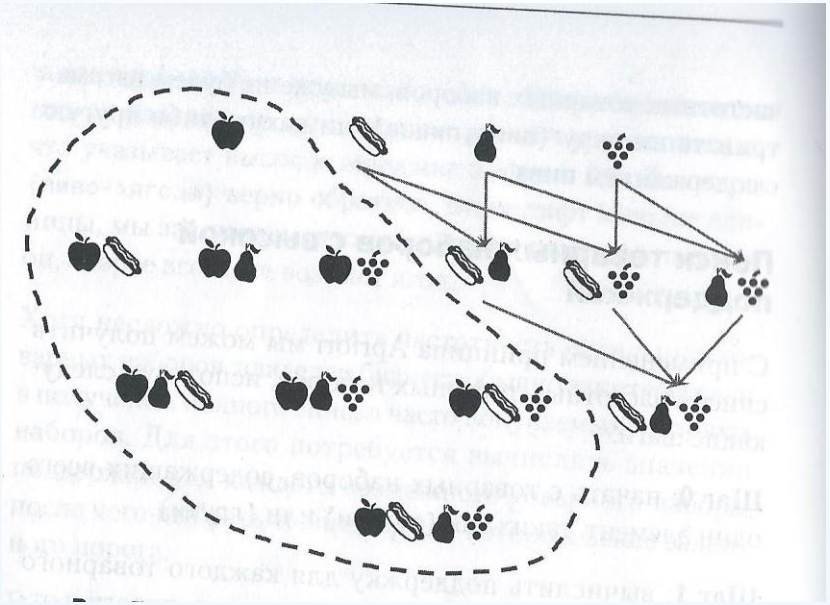
Принцип Apriori утверждает, что если какой-то товарный набор редкий, то и бОльшие наборы, которые его включают, тоже должны быть редки. Это означает, что если редким является, скажем, {пиво}, то редким должно быть и сочетание {пиво, пицца}. Таким образом, составляя список частотных товарных наборов, мы уже не будем рассматривать ни пару {пиво, пицца}, ни какую-либо другую с содержанием пива.

Работа данного алгоритма состоит из нескольких этапов, каждый из этапов состоит из следующих шагов:

- ***формирование кандидатов;***
- ***подсчет кандидатов***.

4.2 Методы поиска ассоциативных правил

Пример использования **алгоритма Apriori**



4.3 Приложения с применением ассоциативных правил

Часто встречающиеся приложения с применением ассоциативных правил:

- **розничная торговл**я: определение товаров, которые стоит продвигать совместно; выбор местоположения товара в магазине; анализ потребительской корзины; прогнозирование спроса;
- **перекрестные продажи**: если есть информация о том, что клиенты приобрели продукты А, Б и В, то какие из них вероятнее всего купят продукт Г?
- **маркетинг**: поиск рыночных сегментов, тенденций покупательского поведения;
- **сегментация клиентов**: выявление общих характеристик клиентов компании, выявление групп покупателей;
- **оформление каталогов**, **анализ сбытовых кампаний фирмы**, **определение последовательностей покупок** клиентов (какая покупка последует за покупкой товара А);
- **анализ Web-логов**.

4.4 Ограничение методов поиска ассоциативных правил

Ограничения существуют несмотря на большой диапазон использования

- **требует долгих вычислений.** Принцип Априори снижает число потенциальных товарных наборов для рассмотрения, но вычислений может быть много, если список товаров большой или указан низкий порог поддержки
- **ложные ассоциации.** В больших наборах данных ассоциации могут быть чистой случайностью. Чтобы убедиться, что обнаруженные ассоциации масштабируемы, их нужно уметь оценивать.

Лабораторная работа. Знакомство с платформой Deductor

Целью выполнения данной лабораторной работы является:

- получение первоначальных сведений о возможностях аналитической платформы;
- изучение основных модулей; работа с мастерами импорта, экспорта, обработки и визуализации данных.

Лабораторная работа. Знакомство с платформой Deductor

Теоретическая часть

АП «Deductor» применима для решения **задач распознавания и обработки данных,** таких как парциальная обработка данных (подготовка к анализу) прогнозирование, поиск закономерностей и так далее. Платформа применима в задачах, где требуется консолидация и отображение данных различными способами, построение моделей и последующее применение полученных моделей к новым данным.

Лабораторная работа. Знакомство с платформой Deductor

Задачи, решаемые АП:

- **Системы корпоративной отчетности**. Готовое хранилище данных и гибкие механизмы предобработки, очистки, загрузки, визуализации позволяют быстро создавать законченные системы отчетности в сжатые сроки.
- **Обработка нерегламентированных запросов**. Конечный пользователь может получить ответ на вопросы типа "Сколько было продаж товара по группам за прошлый год с разбивкой по месяцам?" и просмотреть результаты наиболее удобным для него способом.

Лабораторная работа. Знакомство с платформой Deductor

Задачи, решаемые АП:

- **Анализ тенденций и закономерностей, планирование, ранжирование.** Простота использования и интуитивно понятная модель данных позволяет вам проводить анализ по принципу «Что, если...?», соотносить ваши гипотезы со сведениями, хранящимися в базе данных, находить аномальные значения, оценивать последствия принятия бизнес-решений.
- **Прогнозирование**. Построив модель на исторических примерах, можно использовать ее для прогнозирования ситуации в будущем. По мере изменения ситуации нет необходимости перестраивать все, необходимо всего лишь дообучить модель**.**

Лабораторная работа. Знакомство с платформой Deductor

Задачи, решаемые АП:

- **Управление рисками**. Реализованные в системе алгоритмы дают возможность достаточно точно определиться с тем, какие характеристики объектов и как влияют на риски, благодаря чему можно прогнозировать наступление рискового события и заблаговременно принимать необходимые меры к снижению размера возможных неблагоприятных последствий.
- **Анализ данных маркетинговых и социологических исследований**. Анализируя сведения о потребителях, можно определить, кто является вашим клиентом и почему. Как изменяются их пристрастия в зависимости от возраста, образования, социального положения, материального состояния и множества других показателей.

Лабораторная работа. Знакомство с платформой Deductor

Задачи, решаемые АП:

- **Диагностика.** Механизмы анализа, имеющиеся в системе Deductor, с успехом применяются в медицинской диагностике и диагностике сложного оборудования. Например, можно построить модель на основе сведений об отказах. При ее помощи быстро локализовать проблемы и находить причины сбоев.
- **Обнаружение объектов на основе нечетких критериев**. Часто встречается ситуация, когда необходимо обнаружить объект, основываясь не на таких четких критериях, как стоимость, технические характеристики продукта, а на размытых формулировках, например, найти продукты, похожие на ваши с точки зрения потребителя.

Лабораторная работа. Знакомство с платформой 1С:Предприятие

160698	КЕТЧУПЫ,СОУСЫ,АДЖИКА
160698	МАКАРОНЫ
160698	ЧАЙ
160747	МАКАРОНЫ
160747	МЕД
160747	ЧАЙ
161217	КЕТЧУПЫ,СОУСЫ,АДЖИКА
161217	МАКАРОНЫ
161217	СЫРЫ
161243	КЕТЧУПЫ,СОУСЫ,АДЖИКА
161243	МАКАРОНЫ

5.1 Задачи, которые ставятся перед нейронными сетями

Начало нейронным сетям как инструменту анализа данных было положено в начале 40-х годов в работе МакКаллока и Питтса. В этой работе предлагалась модель искусственного нейрона.

Предполагалось, что, моделируя нейронную структуру мозга, возможно приблизиться к искусственному интеллекту. К тому времени уже было известно, что мозг человека состоит из особых биологических клеток – **нейронов**, и казалось, что построение сетей из нейронов позволит решать сложные задачи, которые ежедневно решает мозг человека.
С тех пор интерес к нейронным сетям периодически то возрастал, то спадал, что обуславливалось новыми разработками в этой области, и сейчас нейронные сети являются одним из достаточно популярных инструментов анализа данных

5.1 Задачи, которые ставятся перед нейронными сетями

По мнению Anil K. Jain из Мичиганского государственного университета и специалистов Исследовательского центра IBM Jianchang Mao и K. M. Mohiuddin, список задач для нейронных сетей можно классифицировать следующим образом.

Классификация образов. К известным приложениям относятся *распознавание букв*, *распознавание речи*, *классификация сигнала электрокардиограммы*, *классификация клеток крови*, *обеспечение деятельности биометрических сканеров* и т. п.
Кластеризация / категоризация. Кластеризация применяется для *извлечения знаний*, *сжатия данных* и *исследования свойств данных*.

5.1 Задачи, которые ставятся перед нейронными сетями

По мнению Anil K. Jain из Мичиганского государственного университета и специалистов Исследовательского центра IBM Jianchang Mao и K. M. Mohiuddin, список задач для нейронных сетей можно классифицировать следующим образом.

Аппроксимация функций. Типичным примером является *шумоподавление при приеме сигнала различной природы*, вне зависимости от передаваемой информации.

Предсказание / прогноз. В качестве примера можно привести *предсказание цен на фондовой бирже* и *прогноз погоды*.

Оптимизация. *Назначение штата работников по ряду умений и факторов* являются классическими примерами задач оптимизации.

.

5.1 Задачи, которые ставятся перед нейронными сетями

По мнению Anil K. Jain из Мичиганского государственного университета и специалистов Исследовательского центра IBM Jianchang Mao и K. M. Mohiuddin, список задач для нейронных сетей можно классифицировать следующим образом.

Память, адресуемая по содержанию (ассоциативная память). Ассоциативная память доступна по указанию заданного содержания. Содержимое памяти может быть вызвано даже по частичному входу или искаженному содержанию. Ассоциативная память может найти применение при создании *мультимедийных информационных баз данных*.
Управление. Примером является *оптимальное управление двигателем, рулевое управление на кораблях, самолетах*.

5.2 Как работает нейронная сеть

Предположим, что нам даются **наборы чисел** (входные векторы), и для каждого из них нам сообщают **значение функции**, которое она имеет на данном наборе.

Пример: значением является обменный курс некоторой валюты на следующий день, **вход** – уровень этого курса и некоторых других финансовых показателей за, скажем, последний месяц.
Другой пример: **входной вектор** – характеристики заемщика банка (возраст фирмы, капитал, количество занятых, подвергался ли судимости директор и т. п.), **результат** – выполнил ли клиент условия возврата кредита. В обоих случаях речь идет пока об исторических данных.

.

5.2 Как работает нейронная сеть

Затем нам предъявляют уже новые данные: значения финансовых показателей по сегодняшний день включительно или данные о новом клиенте, обратившемся с просьбой о предоставлении кредита.

Результат теперь неизвестен, и мы должны его (хотя бы приближенно) найти: *каким будет обменный курс завтра, перспективен ли для банка данный клиент*.

.

5.2 Как работает нейронная сеть

Как действует в этой ситуации нейронная сеть?

Элементарная операция, которую она производит с данными, состоит в следующем: берется «взвешенная» сумма входных величин (сумма, взятая с некоторыми коэффициентами, которые называются весами). Затем полученная величина преобразуется с помощью нелинейной монотонной функции (функции активации) так, чтобы получившееся в результате значение лежало в интервале от 0 до 1. Описанная конструкция называется **искусственным нейроном**. Сеть состоит из многих таких нейронов, причем часть из них обрабатывает непосредственно входные данные (*первый слой нейронов*), другие – сигналы, полученные на выходе с нейронов первого слоя и т. д. (*скрытые слои нейронов*), и, наконец, есть *единственный выходной нейрон*, который и выдает нам результат.

.

5.2 Как работает нейронная сеть

Пример **распознавания рукописных цифр**. Здесь нейронная сеть использует несколько слоев нейронов, чтобы строить прогноз на основании вводимых изображений

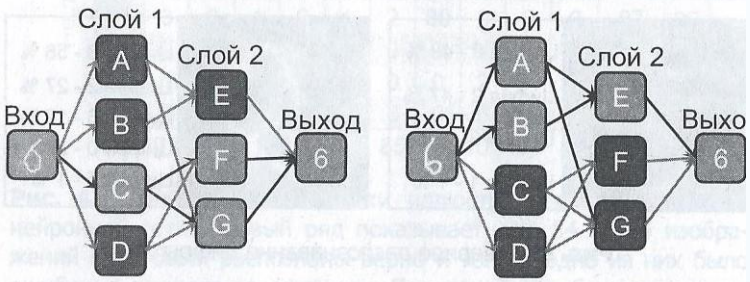
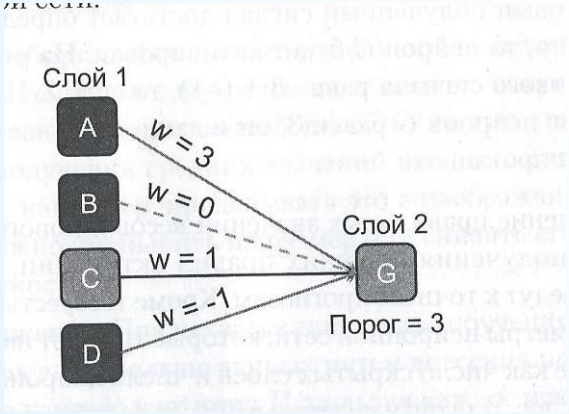


Рис. 7. Пример нейронной сети: разные входные данные с одними выходными. Активные нейроны выделены темным

5.2 Как работает нейронная сеть

Чтобы построить прогноз, нейроны, в свою очередь, должны быть **активированы** на протяжении нейронного пути. Активация каждого нейрона управляется правилом активации, которое определяет источник и силу входного сигнала, получаемого нейроном перед активацией



5.3 Общая схема анализа данных с помощью нейронных сетей

Общая схема анализа данных с помощью нейронных сетей состоит из 5 этапов.

Выбор типологии сети. Существует 9 типов сетей, на этом этапе подбирается наиболее подходящий под задачу тип сети.
Экспериментальный подбор характеристик сети. После выбора типа необходимо подобрать структуру сети (количество нейронов, их веса, взаимосвязи и так далее).
Экспериментальный подбор параметров обучения. Далее необходимо экспериментально определить параметры обучения: максимальное время обучения, количество данных, максимально допустимую ошибку и так далее.

.

5.3 Общая схема анализа данных с помощью нейронных сетей

Общая схема анализа данных с помощью нейронных сетей состоит из 5 этапов.

Обучение сети. По обучающей выборке проводится обучение сети. Предполагается, что обучающая выборка содержит в себе информацию, которая характеризует данные в целом.
Проверка адекватности обучения. Проводится анализ полученных результатов на данных, которые не входили в обучающую выборку. Осуществляется ручной контроль результатов работы нейронной сети.

.

6.1 Характеристики Big Data

Big Data — обозначение структурированных и неструктурированных данных огромных объёмов и значительного многообразия, эффективно обрабатываемых горизонтально масштабируемыми программными инструментами

В широком смысле о «больших данных» говорят как о социально-экономическом феномене, связанном с появлением технологических возможностей анализировать огромные массивы данных, в некоторых проблемных областях — весь мировой объём данных, и вытекающих из этого трансформационных последствий.

.

6.1 Характеристики Big Data

В качестве определяющих характеристик для больших данных традиционно выделяют «три V»:

Объём (англ. volume, в смысле величины физического объёма), **скорость** (velocity в смыслах как скорости прироста, так и необходимости высокоскоростной обработки и получения результатов), **многообразие** (variety, в смысле возможности одновременной обработки различных типов структурированных и полуструктурированных данных); в дальнейшем возникли различные вариации и интерпретации этого признака.

6.1 Характеристики Big Data

Набор признаков VVV (volume, velocity, variety) изначально выработан Meta Group в 2001 году вне контекста представлений о больших данных как об определённой серии информационно-технологических методов и инструментов

В дальнейшем появились интерпретации с «четырьмя V» (добавлялась veracity — достоверность, использовалась в рекламных материалах IBM), «пятью V» (в этом варианте прибавляли viability — жизнеспособность, и value — ценность), и даже «семью V» (кроме всего, добавляли также variability — переменчивость, и visualization). IDC интерпретирует «четвёртое V» как value с точки зрения важности экономической целесообразности обработки соответствующих объёмов в соответствующих условиях

6.2 Источники больших данных

Есть классические источники данных. Есть источники данных, которые образовались в последнее время

Классическими источниками больших данных признаются интернет вещей и социальные медиа, считается также, что большие данные могут происходить из внутренней информации предприятий и организаций (генерируемой в информационных средах, но ранее не сохранявшейся и не анализировавшейся), из сфер медицины и биоинформатики, из астрономических наблюдений.

6.2 Источники больших данных

Есть классические источники данных. Есть источники данных, которые образовались в последнее время

В качестве примеров источников возникновения больших данных приводятся непрерывно поступающие данные с измерительных устройств, события от радиочастотных идентификаторов, потоки сообщений из социальных сетей, метеорологические данные, данные дистанционного зондирования Земли, потоки данных о местонахождении абонентов сетей сотовой связи, устройств аудио- и видеорегистрации. Ожидается, что развитие и начало широкого использования этих источников инициирует проникновение технологий больших данных как в научно-исследовательскую деятельность, так и в коммерческий сектор и сферу государственного управления.

6.3 Методы и техники анализа, применимые к большим данным

Перечислим эти методы и техники:

► методы класса Data Mining: обучение ассоциативным правилам (англ. association rule learning), классификация (методы категоризации новых данных на основе принципов, ранее применённых к уже наличествующим данным), кластерный анализ, регрессионный анализ;
► краудсорсинг — категоризация и обогащение данных силами широкого, неопределённого круга лиц, привлечённых на основании публичной оферты, без вступления в трудовые отношения;
► смешение и интеграция данных (англ. data fusion and integration) — набор техник, позволяющих интегрировать разнородные данные из разнообразных источников для возможности глубинного анализа, в качестве примеров таких техник, составляющих этот класс методов приводятся цифровая обработка сигналов и обработка естественного языка (включая тональный анализ);

6.3 Методы и техники анализа, применимые к большим данным

Перечислим эти методы и техники:

► машинное обучение, включая обучение с учителем и без учителя, а также Ensemble learning (англ.) — использование моделей, построенных на базе статистического анализа или машинного обучения для получения комплексных прогнозов на основе базовых моделей (англ. constituent models, ср. со статистическим ансамблем в статистической механике);
► искусственные нейронные сети, сетевой анализ, оптимизация, в том числе генетические алгоритмы;
► распознавание образов;
► прогнозная аналитика;

6.3 Методы и техники анализа, применимые к большим данным

Перечислим эти методы и техники:

► имитационное моделирование;
► пространственный анализ (англ. Spatial analysis) — класс методов, использующих топологическую, геометрическую и географическую информацию в данных;
► статистический анализ, в качестве примеров методов приводятся A/B-тестирование и анализ временных рядов;
► визуализация аналитических данных — представление информации в виде рисунков, диаграмм, с использованием интерактивных возможностей и анимации как для получения результатов, так и для использования в качестве исходных данных для дальнейшего анализа.

6.4 Технологии. Аппаратные решения

Существует ряд аппаратно-программных комплексов, предоставляющих предконфигурированные решения для обработки больших данных:

- Aster MapReduce appliance (корпорации Teradata),
- Oracle Big Data appliance,
- Greenplum appliance (корпорации EMC, на основе решений поглощённой компании Greenplum).

Эти комплексы поставляются как готовые к установке в центры обработки данных телекоммуникационные шкафы, содержащие кластер серверов и управляющее программное обеспечение для массово-параллельной обработки.

6.4 Технологии. Аппаратные решения

К решениям из области больших данных иногда относят такие решения, несмотря на то, что такая обработка изначально не является массово-параллельной, а объёмы оперативной памяти одного узла ограничиваются несколькими терабайтами:

- Аппаратные решения для резидентных вычислений, прежде всего, для баз данных в оперативной памяти и аналитики в оперативной памяти, в частности, предлагаемой аппаратно-программными комплексами Hana (предконфигурированное аппаратно-программное решение компании SAP) и
- Exalytics (комплекс компании Oracle на основе реляционной системы Timesten (англ.) и многомерной Essbase),

6.4 Технологии. Аппаратные решения

Иногда к решениям для больших данных относят и аппаратно-программные комплексы на основе традиционных реляционных систем управления базами данных:

- Netezza,
- eradata,
- Exadata, как способные эффективно обрабатывать терабайты и эксабайты структурированной информации, решая задачи быстрой поисковой и аналитической обработки огромных объёмов структурированных данных.

Первыми массово-параллельными аппаратно-программными решениями для обработки сверхбольших объёмов данных были машины компаний Britton Lee (англ.), впервые выпущенные в 1983 году, и Teradata (начали выпускаться в 1984 году, притом в 1990 году Teradata поглотила Britton Lee).

6.4 Технологии. Аппаратные решения

Аппаратные решения DAS — систем хранения данных, напрямую присоединённых к узлам —

в условиях независимости узлов обработки в SN-архитектуре также иногда относят к технологиям больших данных. Именно с появлением концепции больших данных связывают всплеск интереса к DAS-решениям в начале 2010-х годов, после вытеснения их в 2000-е годы сетевыми решениями классов NAS и SAN.

Контрольные вопросы

1.
2.
3.
4.