Санкт-Петербургский
Государственный
Политехнический
Университет

Институт компьютерных
наук и технологий

курс: Решение прикладных задач методами машинного обучения

Лекция 6

Технологии решения прикладных задач МО.

19 октября 2022 г.

Что обсуждали на прошлой лекции

- В общем случае задача машинного обучения это **некорректная по Адамару** прикладная задача, так как решение ее

  не единственно и не непрерывно зависит от входных данных (фактически зависят от данных, которые в явном виде условиях задачи отсутствуют)

- Поэтому, все задачи МО нуждаются в т.н. регуляризации, чтобы избежать т.н. переобучения

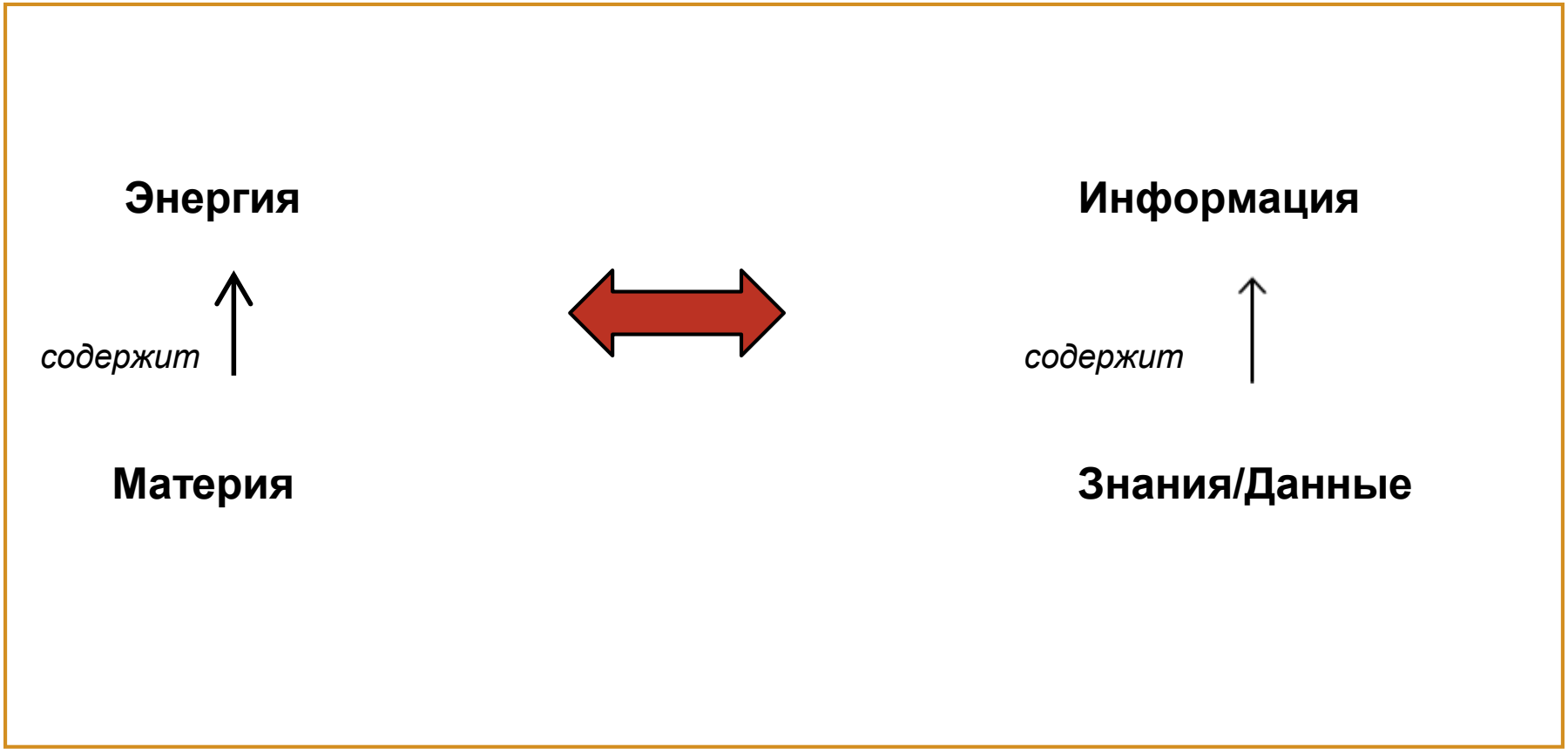
Суть **задачи обучения** …можно описать так

Построение отображения (оператора)**, продолжаемого до гомоморфизма, непрерывного отображения** счетного множества входного данных («облако данных») в конечное множество «предметных» понятий, характеризующих существенные свойства объектов реальности.

Чтобы формализовать задачу обучения надо определить, что есть

1. «объект» машинного обучения
2. «предмет» машинного обучения

«квадрат» фундаментальных понятий



Существует ли физика «процесса обучения» ?

Факты:

Учитель не может напрямую «вложить» знания непосредственно в сознание студента, Учитель может только «дать» информацию, указав признаки различимости объектов-понятий, которые обучаемый может превратить непосредственно в «свои знания»: отобразить входной поток данных на множество мыслимых понятий.

Существующие трактовки :

процесса обучения как построение «отображения» с использование :

- Статистических (корреляции между данными, которые образуют «неделимые» символы)
- Алгебраических (факторизация множеств входных данных на классы эквивалентности с сохранением предикатов)
- Топологических (использование дискрипторов персистетности при изменении параметров масштаба)

подходов

Обучение как субъективное «естествознание»

В результате обучения «картина мира» определяется не столько свойствами самого этого мира, сколько характеристиками субъекта познания (человека), а именно его «концептуальными взглядами».

Выводы:

Исключить субъективное начало из процесса обучения полностью невозможно.

Причинность, которая с точки зрения физики, есть объективное начало для всех научных знаний, (в физике, например, причина понимается «механистически» как **внешняя сила *или* вероятность –** в квантовой механике**)** в процессе обучения обретает новое статистическое, алгебраическое и топологическое содержание

Обучение как основа согласованного соучастия объективного и субъективного начал естествознания

Прикладная задача обучения: построение обладающей потенциальной полнотой совокупности реакций субъекта на водные сигналы в пределах существующих физических ограничений (сверх-задача обучения гораздо шире)

Тезис 1. Любая способная к «обученную» система является «само организованной активной системой», которая существует, адаптируется и развивается (эволюционирует) благодаря тому, что реализуют «согласованные действия», используя как физические (причинные), так и «информационные» (знания) ресурсы о той среде, где система в данны существует .

«Первый закон самоорганизации» (применим к микро и макро масштабам) :

Необходимым условием сохранения целостности «само организованной активной системой» в неравновесных и динамически изменяющихся условиях является согласованное поведение всех элементов такой системы - т.е. реализация процесса структурогенеза.

Структурогенез.

- Структурогенез - процесс формообразования, обусловленный предопределенной последовательностью изменения структурных характеристик строения объекта с учетом влияния среды.

Создание структурогенез рациональных действий

Сознание – это способность субъекта формировать рациональные решения (в том числе с учетом морального контекста)

Процесс структурогенеза описывается через понятие супервéнтности (англ. Supervenience) — отношение детерминированности (полной определенности) состояния одной системы состоянием другой системы.

Считается, что набор свойств одной системы супервентен относительно набора свойств другой системы в том случае, если существование различия между двумя фактами в свойствах первой системы невозможно без существования такого же различия между двумя фактами в свойствах второй системы.

Через свойство супервентности возможно описание объективной зависимости ментальных явлений от физических явлений,

Примеры супервентности, которые важны для практики

«supervenience» – это

- Детерминированность (отсутствие различий) в ментальных свойствах при отсутствии различий в физических свойствах;
- Детерминированность (отсутствие различий) в компьютерной программе при отсутствии различий в аппаратной конфигурации компьютера;
- Детерминированность (отсутствие различий) в экономике при отсутствии различий в поведении экономических агентов.

Физическое vs информационное

Информация материальна: связана с физическими объектами (носителями)
Материя информационна: информация атрибут структуры материи, которая является носителем отличий для физических объектов

Тезис 1. Если имеется два физически тождественных состояния системы, но в одном случае нет информации о состоянии системы, а во втором имеем какую-то информацию об этом, то эти два состояния системы различаются фундаментально.

Суть различия: Во втором случае мы можем «заставить» систему совершать «работу», а в первом – нет!

Тезис. 2 (антропный принцип). То что называется объективные характеристики физического мира связано с существованием наблюдателя, который обладает знаниями о признаках (воспринимает информацию), характеризующих различие между объектами материального мира или информацию

О проблеме «материализации » опыта и знаний

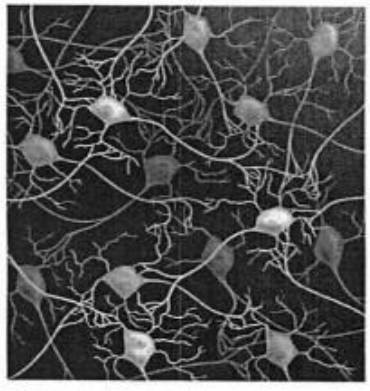
Согласно гипотезе др. Диспензе, вся информация о прошлых событиях «записана» в нейросетях мозга, которые формируют то, как мы воспринимаем и ощущаем мир в целом и его конкретные объекты в частности.

Большинство реакций «обученного» субъекта запрограммировано на уровне устойчивых нейронных связей в мозгу. Каждый объект (стимул) активирует ту или иную нейронную сеть, которая в свою очередь вызывает набор определенных химических реакций в организме.

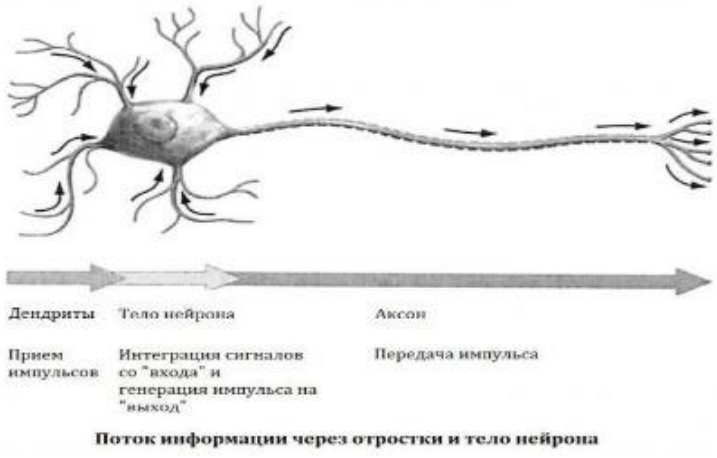
Эти химические реакции заставляют человека действовать или чувствовать себя определенным образом

Выводы:

- все эмоциональные реакции – не более чем результат химических процессов, обусловленных сложившимися нейросетями, и основываются они на прошлом опыте.
- в 99% случаев мы воспринимаем реальность не такой, какая она есть, а интерпретируем ее на основе готовых образов из прошлого.



Нейронные сети



Поток информации через отростки и тело нейрона

Типы прикладных задач МО

1)Задача регрессии – прогноз значений (на основе прецендентов) с использованием ранее полученной выборки данных . Результатом решения задачи является вещественное число (2, 35, 76.454 и др.), которое к примеру естьё: цена квартиры, стоимость ценной бумаги по прошествии полугода, ожидаемый доход магазина на следующий месяц, качество вина при слепом тестировании пр.

2)Задача классификации – получение категориального ответа на основе набора признаков. Имеет конечное количество ответов (как правило, в формате «да» или «нет»): есть ли на фотографии кот, является ли изображение человеческим лицом, болен ли пациент раком.

3)Задача кластеризации – распределение данных на группы: разделение всех клиентов мобильного оператора по уровню платёжеспособности, отнесение космических объектов к той или иной категории (планета, звёзда, чёрная дыра и т. п.).

4)Задача уменьшения размерности – сведение большого числа признаков к меньшему числу (обычно 2–3) для удобства их последующей визуализации (например, для оценки данных).

5)Задача выявления аномалий – отделение аномалий от стандартных случаев. На первый взгляд она совпадает с задачей классификации, но есть одно существенное отличие: аномалии – явление редкое, и обучающих примеров, на которых можно натаскать машинно-обучающуюся модель на выявление таких объектов, либо исчезающе мало, либо просто нет, поэтому методы классификации здесь не работают. На практике такой задачей является, например, выявление мошеннических действий с банковскими картами.

Особенности данных, которые используются в прикладных задачах МО

- разнородные (признаки измерены в разных шкалах)
- неполные (измерены не все, имеются пропуски)
- неточные (измерены с погрешностями)
- противоречивые (объекты одинаковые, ответы разные)
- избыточные (сверхбольшие, не помещаются в память)
- недостаточные (объектов меньше, чем признаков)
- неструктурированные (нет признаковых описаний)

Риски, связанные с постановкой задачи:

- «грязные» данные
  (заказчик не обеспечивает качество данных)
- неясные критерии качества модели
  (заказчик не определился с целями или индикаторами KPI)

Особенности данных и постановок прикладных задач

- разнородные (признаки измерены в разных шкалах)
- неполные (измерены не все, имеются пропуски)
- неточные (измерены с погрешностями)
- противоречивые (объекты одинаковые, ответы разные)
- избыточные (сверхбольшие, не помещаются в память)
- недостаточные (объектов меньше, чем признаков)
- неструктурированные (нет признаковых описаний)

Риски, связанные с постановкой задачи:

- «грязные» данные
  (заказчик не обеспечивает качество данных)
- неясные критерии качества модели
  (заказчик не определился с целями или индикаторами KPI)

Примеры: адачи медицинской диагностики

Объект - пациент в определённый момент времени.

Классы: диагноз или способ лечения или исход заболевания.

Примеры признаков:

- бинарные: пол, головная боль, слабость, тошнота, и т. д.
- порядковые: тяжесть состояния, желтушность, и т. д.
- количественные: возраст, пульс, артериальное давление, содержание гемоглобина в крови, доза препарата, и т. д.

Особенности задачи:

- обычно много «пропусков» в данных;
- нужен интерпретируемый алгоритм классификации;
- нужно выделять синдромы - сочетания симптомов;
- нужна оценка вероятности отрицательного исхода.

Задачи распознавания месторождений

Объект - геологический район (рудное поле).

Классы - есть или нет полезное ископаемое.

Примеры признаков:

- бинарные: присутствие крупных зон смятия и рассланцевания, и т. д.
- порядковые: минеральное разнообразие; мнения экспертов о наличии полезного ископаемого, и т. д.
- количественные: содержания сурьмы, присутствие в рудах антимонита, и т. д.

Особенности задачи:

- проблема «малых данных» - для редких типов месторождений объектов много меньше, чем признаков.

Задача кредитного скоринга

Объект - заявка на выдачу банком кредита.

Классы - Ьad или good.

Примеры признаков:

- бинарные: пол, наличие телефона, и т. д.
- номинальные: место проживания, профессия, работодатель, и т. д.
- порядковые: образование, должность, и т. д.
- количественные: возраст, зарплата, стаж работы, доход семьи, сумма кредита, и т. д.

Особенности задачи:

- нужно оценивать вероятность дефолта $P(y(x) = \text{Ьad})$.

Задача предсказания оттока клиентов

Объект - абонент в определённый момент времени.

Классы - уйдёт или не уйдёт в следующем месяце.

Примеры признаков:

- бинарные: корпоративный клиент, включение услуг, и т. д.
- номинальные: тарифный план, регион проживания, и т. д.
- количественные: длительность разговоров (входящих, исходящих, СМС, и т. д.), частота оплаты, и т. д.

Особенности задачи:

- нужно оценивать вероятность ухода;
- сверхбольшие выборки;
- признаки приходится вычислять по «сырым» данным.

Задача категоризации текстовых документов

Объект - текстовый документ.

Классы - рубрики иерархического тематического каталога.

Примеры признаков:

- номинальные: автор, издание, год, и т. д.
- количественные: для каждого термина - частота в тексте, в заголовках, в аннотации, и т. д.

Особенности задачи:

- лишь небольшая часть документов имеют метки y_i;
- документ может относиться к нескольким рубрикам;
- в каждом ребре дерева свой классификатор на 2 класса.

Задачи биометрической идентификации личности

Идентификация личности по отпечаткам пальцев

Идентификация личности по радужной оболочке глаза

Особенности задач:

- нетривиальная предобработка для извлечения признаков; высочайшие требования к точности.

Задача прогнозирования стоимости недвижимости

Объект - квартира в Москве.

Примеры признаков:

- бинарные: наличие балкона, лифта, мусоропровода, охраны, и т. д.
- номинальные: район города, тип дома (кирпичный/панельный/блочный/монолит), и т. д.
- количественные: число комнат, жилая площадь, расстояние до центра, до метро, возраст дома, и т. д.

Особенности задачи:

- выборка неоднородна, стоимость меняется со временем;
- разнотипные признаки;
- для линейной модели нужны преобразования признаков;

Задача прогнозирования объёмов продаж

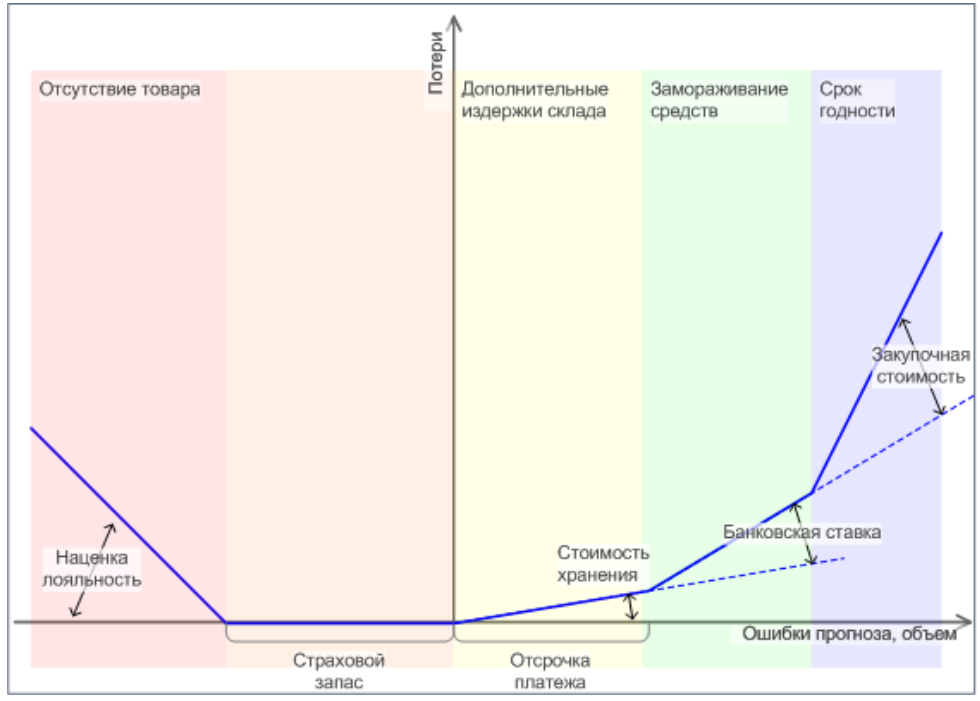
Объект - тройка (товар, магазин, день).

Примеры признаков:

- бинарные: выходной день, праздник, промоакция, и т. д.
- количественные: объёмы продаж в предшествующие дни.

Особенности задачи:

- функция потерь не квадратична и даже не симметрична;
- разреженные данные.



Оценка Revenue Prediction

Объект - место для открытия нового ресторана.

Предсказать - прибыль от ресторана через год.

Примеры признаков:

- демографические данные: возраст, достаток и т.д.,
- цены на недвижимость поблизости,
- маркетинговые данные: наличие школ, офисов и т.д.

Особенности задачи:

- мало объектов, много признаков;
- разнотипные признаки;
- есть выбросы;
- разнородные объекты (возможно, имеет смысл строить разные модели для мелких и крупных городов).

Задача ранжирования поисковой выдачи

Объект - пара (короткий текстовый запрос, документ).

Классы - релевантен или не релевантен,

разметка делается людьми - асессорами.

Примеры количественных признаков:

- частота слов запроса в документе,
- число ссылок на документ,
- число кликов на документ: всего, по данному запросу.

Особенности задачи:

- сверхбольшие выборки документов;
- оптимизируется не число ошибок, а качество ранжирования;
- проблема конструирования признаков по сырым данным.

Конкурс kaggle.com: Avito Context Ad Clicks Prediction

Объект - тройка (пользователь, объявление, баннер).

Предсказать - кликнет ли пользователь по контекстной рекламе, которую показали в ответ на его запрос на avito.ru.

Сырые данные:

- все действия пользователя на сайте,
- профиль пользователя (бразуер, устройство и т. д.),
- история показов и кликов других пользователей по баннеру,
- . . . всего 10 таблиц данных.

Особенности задачи:

- признаки надо придумывать;
- данных много - сотни миллионов показов;
- основной критерий качества - доход рекламной площадки;

Машинное обучение на данных сложной структуры

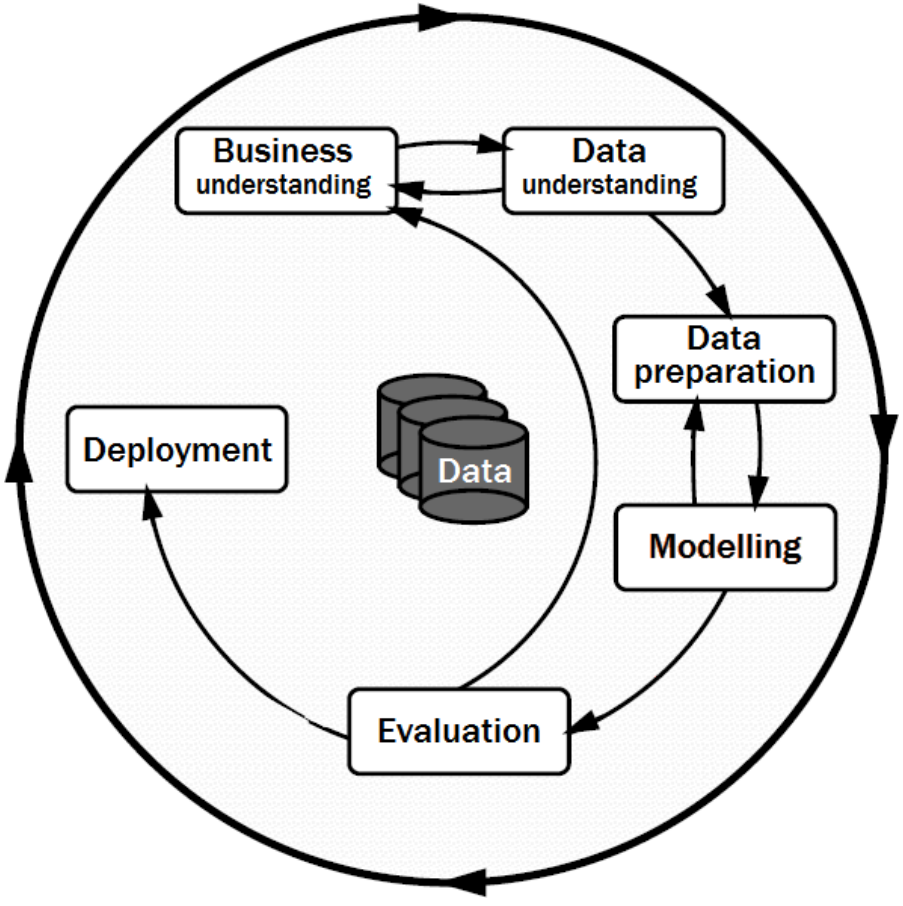
- Статистический машинный перевод:
  объект - предложение на естественном языке
  ответ - его перевод на другой язык
- Перевод речи в текст:
  объект - аудиозапись речи человека
  ответ - текстовая запись речи
- Компьютерное зрение:
  объект - изображение или видеопоследовательность
  ответ - решение (объехать, остановиться, игнорировать)

Предпосылки успешного решения задач со сложными данными:

- Большие и чистые данные (Big Data)
- Глубокие нейросетевые архитектуры (Deep Learning)
- Методы оптимизации для задач большой размерности
  Рост вычислительных мощностей (закон Мура, GPU)

Межотраслевой стандарт интеллектуального анализа данных

CRISP-DM: CRoss Industry Standard Process for Data Mining (1999)



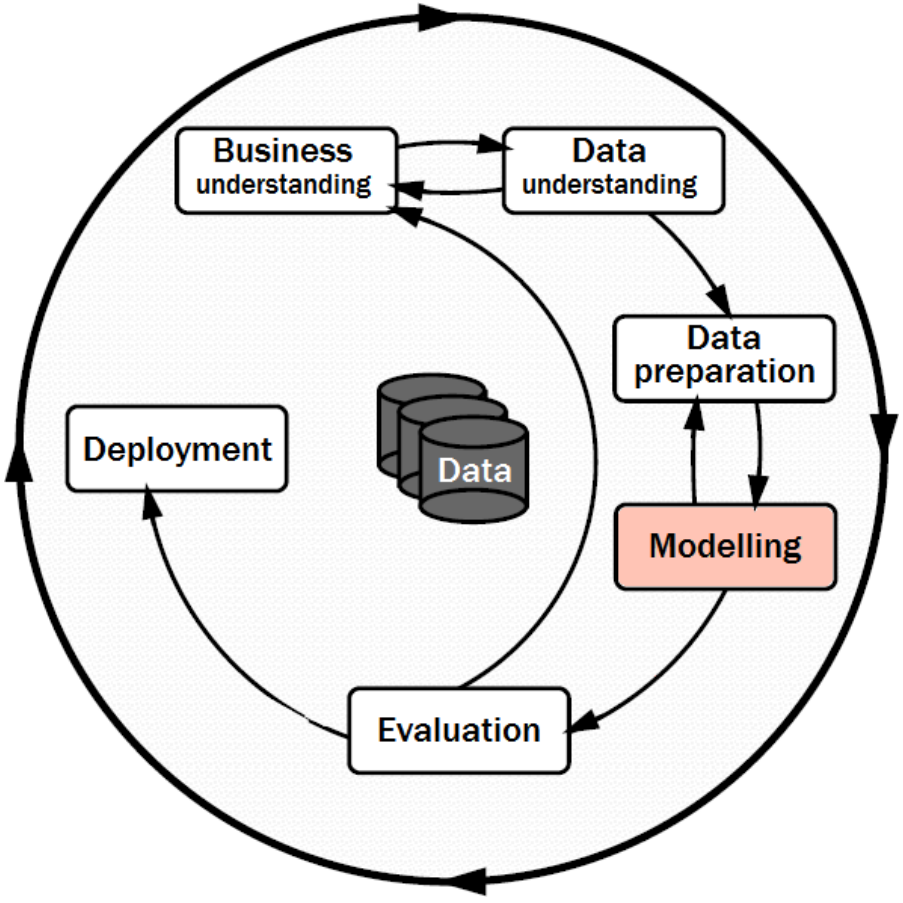
Шаги процесса:

- понимание бизнеса понимание данных
- предобработка данных и инженерия признаков разработка моделей и
- настройка параметров
- оценивание качества внедрение

Понимание эволюции ИИ как автоматизации шагов

CRISP-DM

CRISP-DM: CRoss Industry Standard Process for Data Mining (1999)



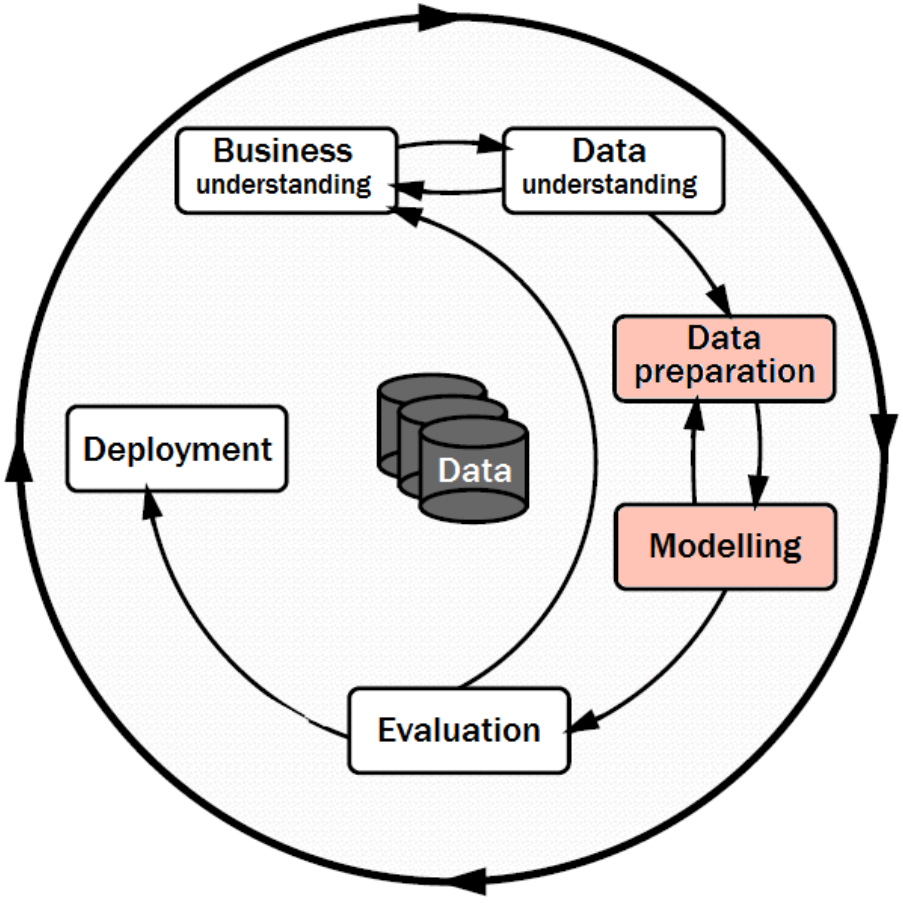
Эволюция ИИ:

- Expert Systems: жесткие модели, основанные на правилах
- Machine Learning: параметрические модели, обучаемые по данным

Понимание эволюции ИИ как автоматизации шагов

CRISP-DM

CRISP-DM: CRoss Industry Standard Process for Data Mining (1999)



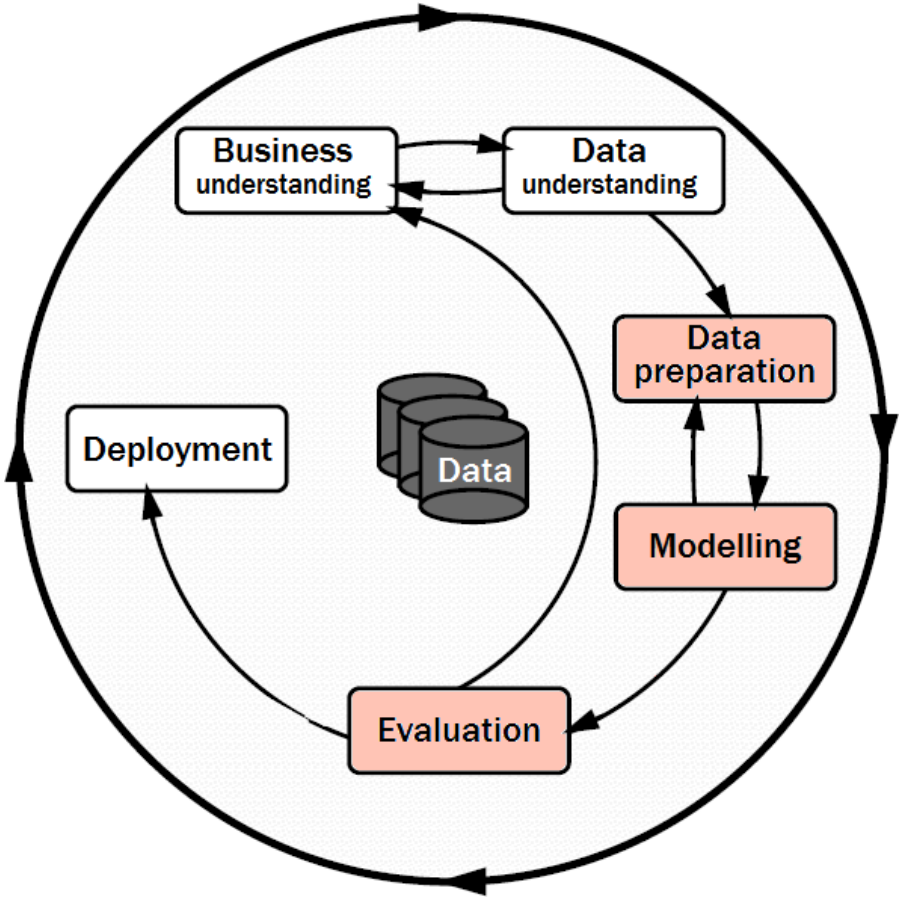
Эволюция ИИ:

- Expert Systems: жесткие модели, основанные на правилах
- Machine Learning: параметрические модели, обучаемые по данным
- Deep Learning: модели с обучаемой векторизацией данных

Понимание эволюции ИИ как автоматизации шагов

CRISP-DM

CRISP-DM: CRoss Industry Standard Process for Data Mining (1999)



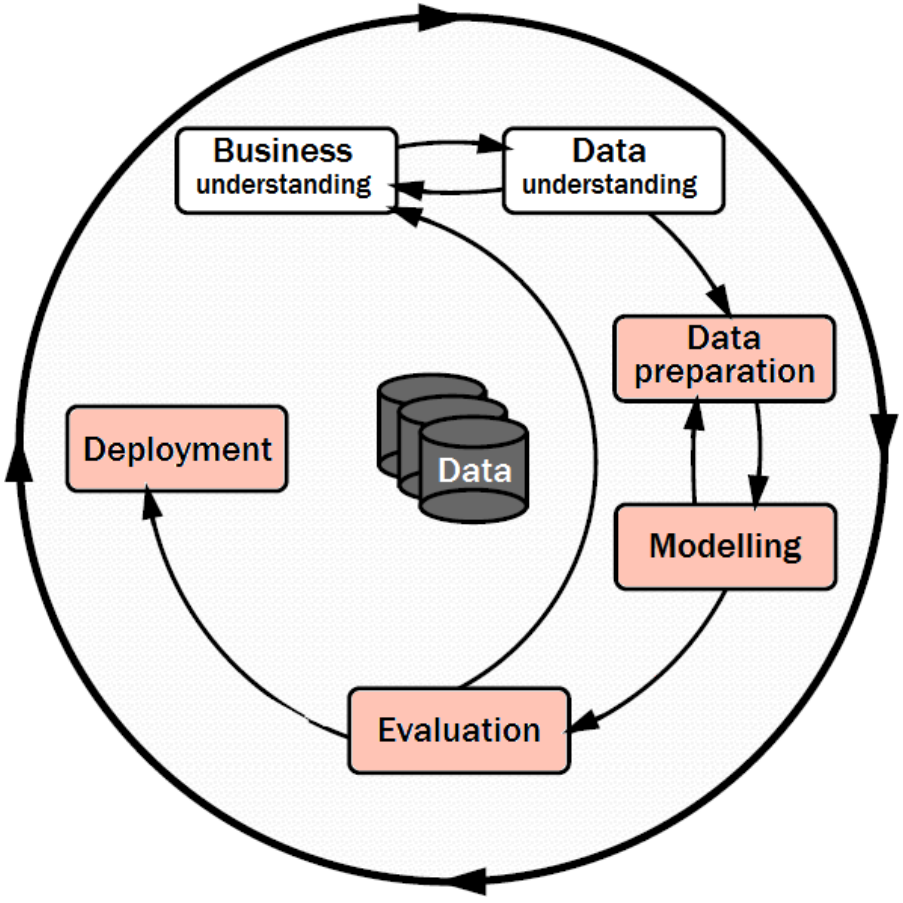
Эволюция ИИ:

- Expert Systems: жесткие модели, основанные на правилах
- Machine Learning: параметрические модели, обучаемые по данным
- Deep Learning: модели с обучаемой векторизацией данных
- AutoML: автоматический выбор моделей и архитектур

Понимание эволюции ИИ как автоматизации шагов

CRISP-DM

CRISP-DM: CRoss Industry Standard Process for Data Mining (1999)



Эволюция ИИ:

- Expert Systems: жесткие модели, основанные на правилах
- Machine Learning: параметрические модели, обучаемые по данным
- Deep Learning: модели с обучаемой векторизацией данных
- AutoML: автоматический выбор моделей и архитектур
- Lifelong Learning: бесшовная интеграция обучения и выбора моделей в бизнес-процесс

Эксперименты на реальных данных

Эксперименты на конкретной прикладной задаче:

- цель - решить задачу как можно лучше
- важно понимание задачи и данных
- важно придумывать информативные признаки
- конкурсы по анализу данных: http://www.kaggle.com
- отечественная платформа: http://DataRing.ru

Эксперименты на наборах прикладных задач:

- цель - протестировать метод в разнообразных условиях
- нет необходимости (и времени) разбираться в сути задач :(
- признаки, как правило, уже кем-то придуманы

  репозиторий UC Irvine Machine Learning Repository
  http://archive.ics.uci.edu/ml (588 задач, 2021-09-03)

Эксперименты на синтетических данных

Используются для тестирования новых методов обучения.
Преимущество - мы знаем истинную $y(x)$ (ground truth)

Эксперименты на синтетических данных:

- цель - отладить метод, выявить границы применимости
- объекты x_i из придуманного распределения (часто 2D)
- ответы $y_i = y(x_i)$ для придуманной функции $y(x)$
- двумерные данные + визуализация выборки

Эксперименты на полу-синтетических данных:

- цель - протестировать помехоустойчивость модели
- объекты x_i из реальной задачи (признаки + шум)
- ответы $y_i = y(x_i)$ для придуманной функции $y(x)$ (+ шум)

Выводы

- Основные понятия машинного обучения: объект, ответ, признак, алгоритм, модель алгоритмов, метод обучения, эмпирический риск, переобучение
- Постановка задачи: Дано, Найти, Критерий
- Этапы решения задач машинного обучения:
  - понимание задачи и данных
  - предобработка данных и изобретение признаков
  - построение модели
  - сведение обучения к оптимизации
  - решение проблем оптимизации и переобучения
  - оценивание качества
  - внедрение и эксплуатация
- Прикладные задачи машинного обучения: очень много, очень разных, во всех областях бизнеса, науки, производства