Машинное обучение

Лекция 5 (1 час)

Емельянова М.Г.

Понятие машинного обучения

«Четвёртая технологическая революция строится на вездесущем мобильном Интернете, искусственном интеллекте и машинном обучении» (2016) – Клаус Мартин Шваб, президент Всемирного экономического форума.

Машинное обучение – это ...

- одна из ключевых информационных технологий будущего;
- наиболее успешное направление искусственного интеллекта, вытесняющее экспертные системы и инженерию знаний;
- проведение функции через заданные точки в сложно устроенных пространствах;
- математическое моделирование, когда данных много, знаний мало;
- тысячи алгоритмов;
- около 100 000 научных публикаций в год.

Понятие машинного обучения

Цель машинного обучения – предсказать результат по входным данным. Чем разнообразнее входные данные, тем проще машине найти закономерности и тем точнее результат.

Машинное обучение – это класс алгоритмов, обучающихся предсказывать неизвестные данные на основе известных.

Если нет данных, то нет и машинного обучения!

Машинное обучение – это раздел искусственного интеллекта.

Нейронные сети – один из видов машинного обучения.

Глубокое обучение – архитектура нейронных сетей, один из подходов к их построению и обучению.

Понятие машинного обучения

Примеры задач машинного обучения

Медицинская диагностика:

объект – данные о пациенте на текущий момент

ответ – диагноз / лечение / риск исхода

Поиск месторождений полезных ископаемых:

объект – данные о геологии района

ответ – есть/нет месторождение

Управление технологическими процессами:

объект – данные о сырье и управляющих параметрах

ответ – количество/качество полезного продукта

Примеры задач машинного обучения

Кредитный скоринг:
объект – данные о заёмщике
ответ – вероятность дефолта, решение по кредиту

Предсказание оттока клиентов:
объект – данные о клиенте на момент времени t
ответ – уйдёт ли клиент к моменту времени $t + \Delta$

Прогнозирование объёмов продаж:
объект – данные о продажах на момент времени t
ответ – объём спроса в интервале от t до $t + \Delta$

Продажа рекламы в Интернете:
объект – данные о тройке «пользователь, страница, баннер»
ответ – оценка вероятности клика

Примеры задач машинного обучения

Информационный поиск в Интернете:

объект – данные о паре «запрос и документ»

ответ – оценка релевантности документа запросу

Статистический машинный перевод:

объект – предложение на естественном языке

ответ – его перевод на другой язык

Компьютерное зрение:

объект – изображение предмета в видеопоследовательности

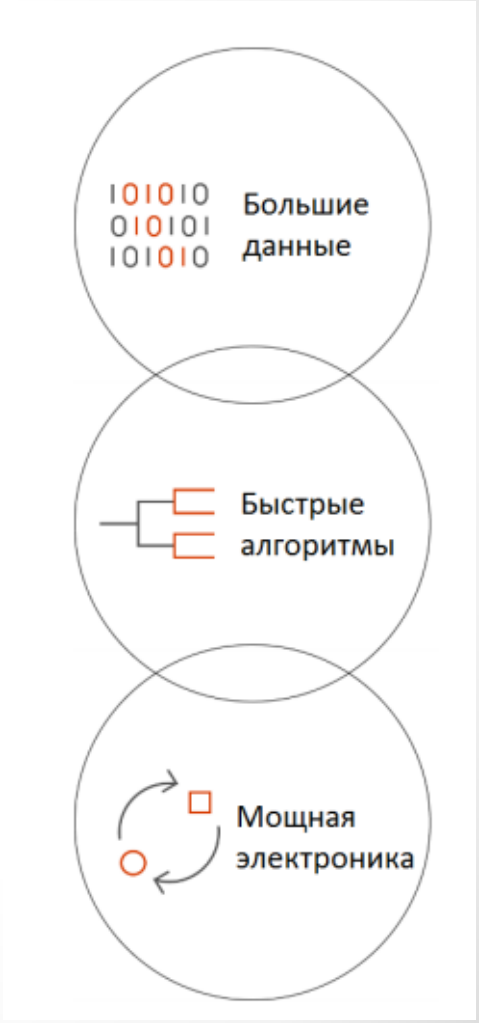
ответ – решение (объехать, остановиться, игнорировать)

Предпосылки бума искусственного интеллекта

Повсеместное применение компьютерных технологий, накопление больших выборок данных.

Развитие математических методов и алгоритмов.

Достижения микроэлектроники.



Составляющие машинного обучения

1. Данные (Data).

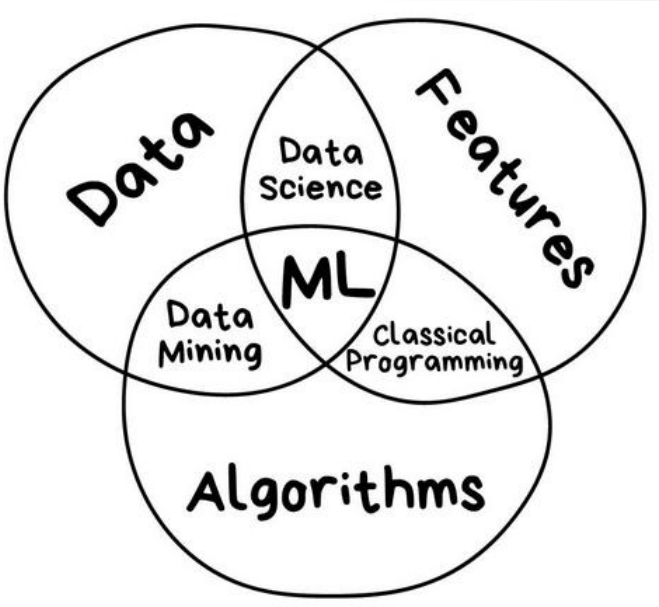
Для определения спама нужны примеры спам-писем, для предсказания курса акций нужна история цен, для определения интересов пользователя нужны его лайки или посты. Данных нужно как можно больше!

2. Признаки (Features).

Каждый объект X характеризуется набором признаков $x_1, x_2, \ldots, x_n$

Объект: сообщение в электронной почте

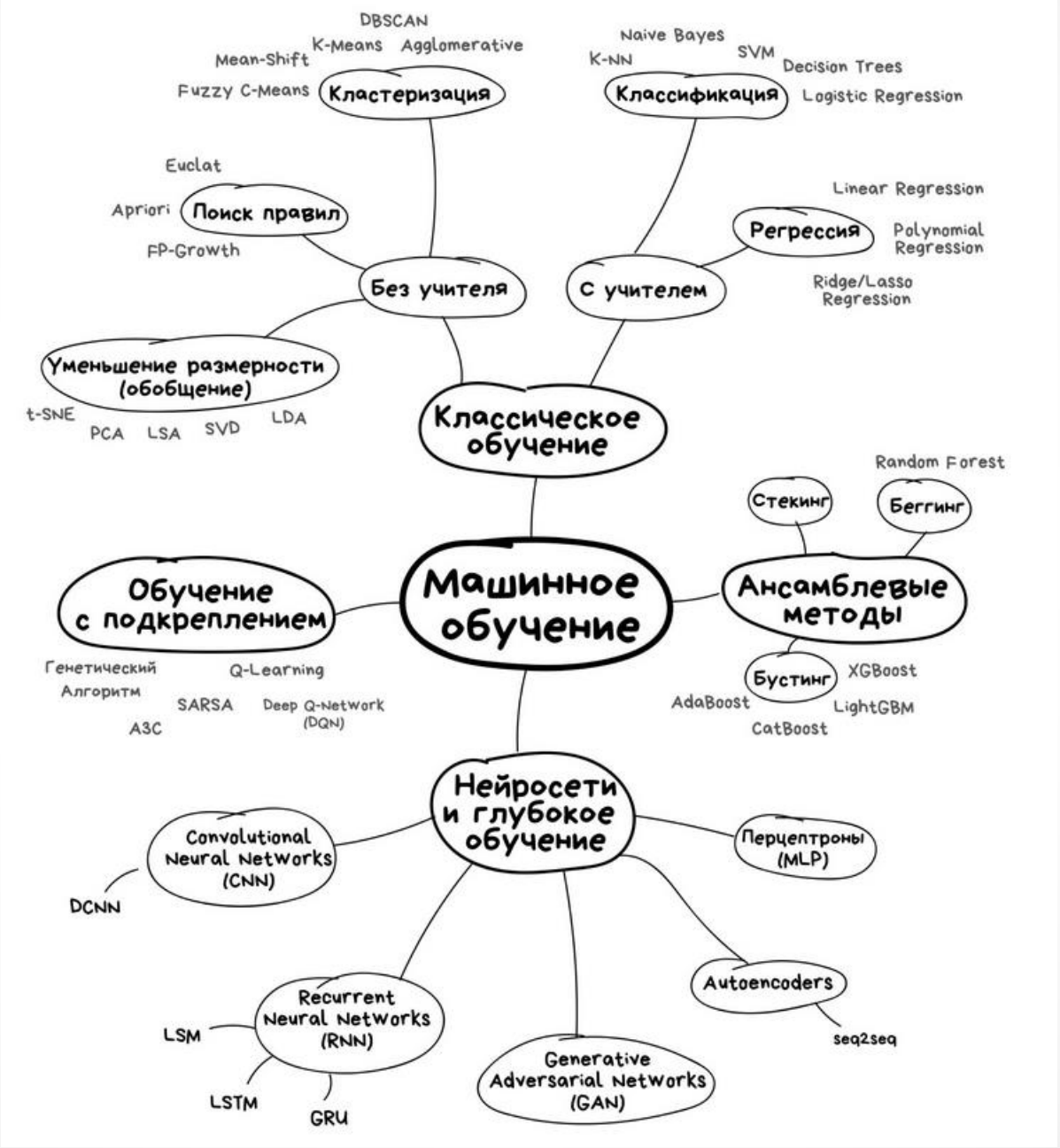
Признаки: набор слов, длина, дата,
отправитель, получатель, язык, частота
сообщения от данного адресата и т.д.



3. Алгоритмы (Algorithms).

От выбора метода зависит точность,
скорость работы и размер готовой модели.

Классификация алгоритмов машинного обучения



Системы машинного обучения и анализа данных

Orange (freely available)

Weka (freely available)

Knime (community edition for free)

RapidMiner (community edition for free)

Deductor (Loginom) (бесплатная версия для обучения)

QuDA (freely available)

https://soware.ru/categories/machine-learning-systems

Библиотеки машинного обучения и анализа данных

scikit-learn (freely available Machine Learning in Python)

MALLET – Machine Learning for LanguagE Toolkit (freely available)

Accord.NET Framework (.NET machine learning framework combined with audio and image processing libraries completely written in C#)

Infer.NET (framework for running Bayesian inference in graphical models)

R (free software environment for statistical computing and graphics+many packages for ML&DM)

Классическое машинное обучение:

- обучение с учителем (supervised learning);
- обучение без учителя (unsupervised learning).

Обучение с учителем.

Данные, подготовленные для анализа, изначально содержат правильный ответ, поэтому цель алгоритма – не ответить, а понять, «Почему именно так?» путём выявления взаимосвязей. Результатом становится способность выстраивать корректные прогнозы и модели.

Классификация, регрессия.

Обучение без учителя.

Для данного типа обучения ключевым понятием является паттерн – обрабатывая значительные массивы данных, алгоритм должен сперва самостоятельно выявлять закономерности. На следующем этапе на основе выявленных закономерностей машина интерпретирует и систематизирует данные.

Кластеризация, ассоциация.

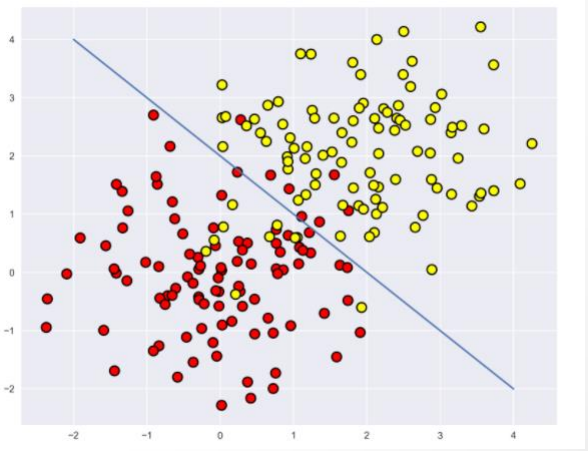
Классификация

В задачах классификации алгоритм предсказывает дискретные значения (y), соответствующие классам, к которым принадлежат объекты.

Есть обучающая выборка, в которой представлены объекты в виде их признакового описания (вектор признаков) и метки класса. Надо найти алгоритм, который для каждого нового объекта (его признакового описания) определит метку класса этого объекта. Это эквивалентно построению разделяющей поверхности в многомерном признаковом пространстве.

Используют:

- спам-фильтры;
- определение языка;
- поиск похожих документов;
- распознавание рукописных букв и цифр;
- определение подозрительных транзакций.



Популярные алгоритмы: наивный Байес, деревья решений, логистическая регрессия, K-ближайших соседей, машина опорных векторов, нейронные сети.

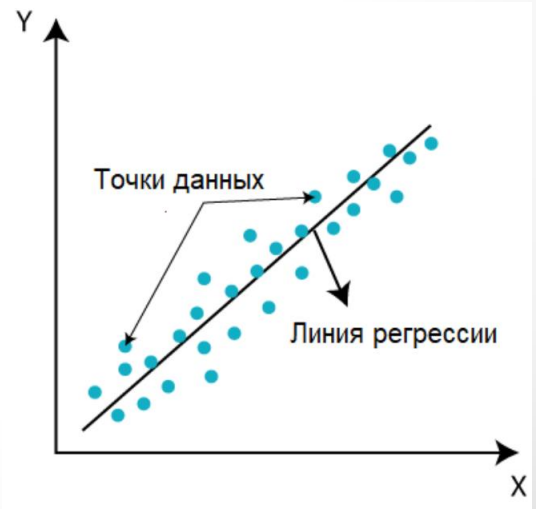
Регрессия

Значения y непрерывны (принимают любое значение из диапазона).

Есть обучающая выборка, в которой представлены объекты в виде их признакового описания (вектор признаков) и значения целевой переменной (непрерывной в отличие от классификации).

Надо найти алгоритм, который для каждого нового объекта (его признакового описания) спрогнозирует значение целевой переменной.

Геометрически определяет прямую (в случае линейной регрессии), наиболее близко проходящую ко всем точкам.



Используют:

- прогноз стоимости ценных бумаг;
- анализ спроса, объёма продаж;
- любые зависимости числа от времени.

Кластеризация

Разделяет объекты по неизвестному признаку.

Машина сама решает как лучше.

Поиск похожих объектов и объединение их в кластеры.

Найти разбиение исходного множества объектов на группы (кластеры).

Объекты внутри одного кластера обладают высоким сходством.

Объекты из разных кластеров сильно различаются.

Используют:

- сегментация рынка

(например, типов покупателей);

- сжатие изображений;
- детекторы аномального поведения.

Ассоциация (поиск правил)

Поиск закономерностей в данных.

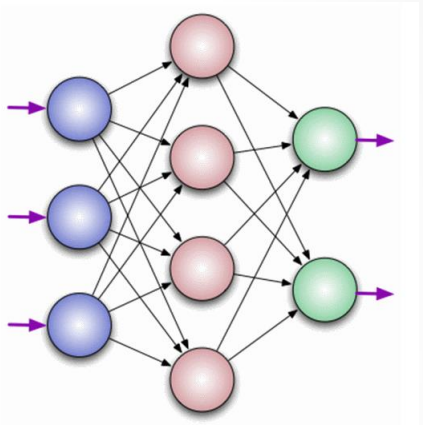
Используют:
анализ товаров, покупаемых вместе;
расстановка товаров на полках;
анализ паттернов поведения на веб-сайтах.

Популярные алгоритмы: Apriori, Euclat, FP-growth

Нейронные сети и глубокое обучение

Любая нейронная сеть – это набор нейронов и связей между ними.

Нейрон – это вычислительная единица, которая получает информацию, производит над ней простые вычисления и передаёт её дальше.



Глубокое обучение – это разновидность машинного обучения на основе искусственных нейронных сетей. Процесс обучения называется глубоким, так как структура искусственных нейронных сетей состоит из нескольких входных, выходных и скрытых слоев. Каждый слой содержит единицы, преобразующие входные данные в сведения, которые следующий слой может использовать для определенной задачи прогнозирования. Благодаря этой структуре компьютер может обучаться с помощью собственной обработки данных.

Нейронные сети и глубокое обучение

Используют:

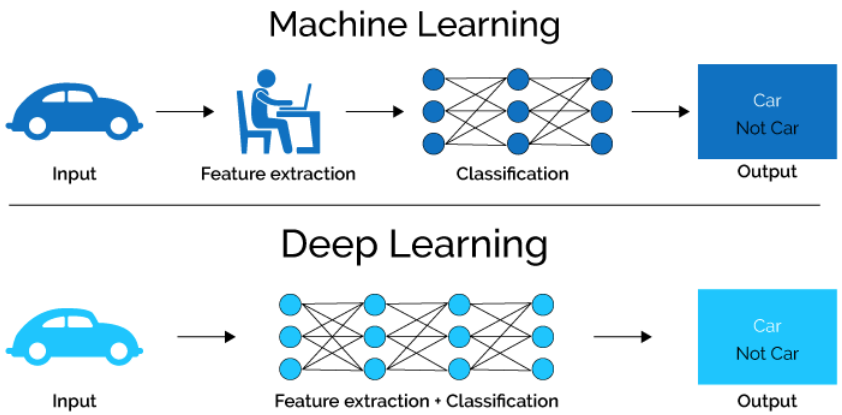
- вместо алгоритмов классического обучения;
- определение объектов на фото и видео;
- распознавание и синтез речи;
- обработка изображений;
- машинный перевод.

Популярные архитектуры: перцептрон, свёрточные сети (CNN), рекуррентные сети (RNN).

Машинное обучение и глубокое обучение

Основное различие между глубоким обучением и машинным обучением обусловлено тем, как данные представляются в систему.

Алгоритмы машинного обучения почти всегда требуют структурированных данных, в то время как сети глубокого обучения полагаются на слои нейронных сетей.



Вопросы для проверки

1. Что такое машинное обучение?
2. Что является составляющими машинного обучения?
3. Какие задачи решаются на основе обучения с учителем?
4. Какие задачи решаются на основе обучения без учителя?
5. Что такое нейронная сеть?
6. Что такое глубокое обучение?
7. В чём отличие классического машинного обучения от глубокого обучения?