**Лекция №1. Введение в машинное обучение**

Сегодня объемы информации и данных быстро растут. Они содержат потенциал для извлечения ценной информации и принятия осмысленных решений. Однако, для использования этих данных необходимы инструменты, способные извлекать скрытые закономерности и прогнозировать будущие события. Здесь на сцену выходит машинное обучение (ML). Суть его заключается в создании алгоритмов и моделей, которые способны автоматически извлекать знания из данных и решать задачи или предсказывать результаты на их основе.

***Машинное обучение*** – это область искусственного интеллекта (AI), занимающаяся разработкой алгоритмов и моделей, которые способны обучаться, используя данные, составлять прогнозы, а также принимать решения без программирования.

**Принципы машинного обучения**

Машинное обучение основано на ряде принципов, которые обеспечивают его функционирование:

* **Данные**. В основе ML лежит использование данных. Обучающие данные предоставляют модели информацию о входных признаках и соответствующих правильных ответах. Чем более разнообразными, качественными и представительными являются данные, тем лучше модель сможет обучиться, распознавать образцы и осуществлять правильные прогнозы на новых данных.
* **Модель**. Представляет собой алгоритм или математическую функцию, которая преобразует входные данные в выходные. Модель выбирается в зависимости от задачи и типа данных. Она может быть линейной, деревом решений, нейронной сетью и т. д. Одна из ключевых целей машинного обучения – создание моделей, которые способны выдавать точные предсказания для новых данных, которые ранее не применялись в процессе обучения.
* **Обучение**. Процесс обучения состоит в подгонке модели к обучающим данным. Модель анализирует данные, выявляет закономерности и корректирует свои внутренние параметры так, чтобы минимизировать ошибку между предсказаниями модели и правильными ответами. Обучение может происходить с учителем (с правильными ответами), без учителя (без правильных ответов) или с подкреплением (с вознаграждениями или наказаниями). Вместо явного программирования модели получают знания из данных и корректируют свои параметры для достижения производительности.
* **Автоматизация**. ML стремится к автоматизации процессов и принятия решений, основанных на данных, без необходимости явного вмешательства человека. Алгоритмы ML способны выполнять сложные задачи с большой скоростью и точностью.
* **Оценка и тестирование**. После обучения модели необходимо оценить ее производительность на новых данных. Для этого используется тестовый набор данных, который модель не видела во время обучения. Оценка производится с помощью метрик, которые измеряют точность, полноту, F1-меру и другие характеристики модели. Это позволяет оценить, как модель справляется с задачей и определить необходимость дальнейшей доработки.
* **Обобщение**. Модель в ML должна быть способна делать точные прогнозы или принимать решения на новых, ранее неизвестных данных. Это свойство называется обобщением. Хорошая модель способна обобщать знания, выявлять общие закономерности и применять их к новым ситуациям.
* **Регуляризация и управление сложностью**. Когда модель становится сложной, существует риск переобучения, когда модель хорошо адаптируется к обучающим данным, но плохо обобщает на новые данные. Для контроля сложности моделей используются методы регуляризации, такие как L1 и L2 регуляризация.

**Различия между ML, AI и DL**

Машинное обучение, искусственный интеллект и глубокое обучение тесно связаны, но имеют разные характеристики:

***Искусственный интеллект*** (**Artificial Intelligence**) охватывает более широкий спектр технологий и методов, направленных на создание интеллектуальных систем, способных выполнять задачи, требующие интеллектуальных способностей, которые обычно связываются с человеком. Машинное обучение является одной из техник, используемых в искусственном интеллекте.

***Машинное обучение*** (**Machine Learning**) – это область, включающая алгоритмы и методы, позволяющие компьютерным системам учиться из данных и делать прогнозы или принимать решения. Машинное обучение является подмножеством искусственного интеллекта.

***Глубокое обучение*** (**Deep Learning**) является подразделом ML, который использует искусственные нейронные сети с большим числом слоев для извлечения высокоуровневых признаков из данных. Оно обычно применяется к задачам, таким как распознавание образов, обработка естественного языка и автоматическое принятие решений.

**Примеры использования машинного обучения**

Машинное обучение имеет применяется как в повседневной жизни, так и в различных отраслях. Разберем несколько примеров.

Например, в повседневной жизни всем известны голосовые помощники Siri, Google Assistant и Алиса, которые используют машинное обучение для распознавания и понимания голосовых команд. Также во многих смартфонах сегодня есть функция автоматического распознавания лиц на фотографиях, которая позволяет организовать и классифицировать снимки по людям и создавать веселые видео на основе этих данных. Также одним из примеров машинного обучения являются рекомендательные системы в онлайн-платформах, такие как YouTube, КиноПоиск, Netflix, Яндекс Музыка, Spotify) которые предлагают персонализированные рекомендации по фильмам, музыке, книгам и пр.

Все чаще ML встречается в сфере здравоохранения, например в диагностике заболеваний на основе медицинских изображений или даже данных о голосе, включая COVID, поражения мозга, рак и другие патологии. Также используется для мониторинга и прогнозирования состояния пациентов в реальном времени с использованием носимых устройств и сенсоров. Машинное обучение вносит большой вклад в разработку новых лекарственных препаратов и поиск потенциальных лекарственных соединений.

Сфера, где мы сталкивается с машинным обучением почти каждый день – это финансы. Анализ финансовых данных для прогнозирования рыночных трендов и принятия решений о вложениях, определение мошеннических транзакций на основе аномалий в поведении клиентов и исторических данных, а также кредитный скоринг и оценка платежеспособности клиентов на основе их финансовой истории и других факторов.

Машинное обучение является ключевым компонентом в разработке автономных транспортных средств, позволяя им анализировать окружающую среду и принимать решения на основе данных с датчиков.

Машинное обучение также применяется в промышленности для оптимизации производственных процессов, предсказания отказов оборудования и улучшения качества продукции.

**Типы машинного обучения**

Каким бывает машинное обучение? Большинство задач можно разделить на два вида машинного обучения: ***обучение с учителем (supervised learning)*** и ***обучение без учителя (unsupervised learning)***. В этих методах “учителем” может выступать программист, который задает правила и контролирует работу алгоритма, но это не является обязательным. В контексте машинного обучения “учителем” может быть любое вмешательство человека в процесс обработки информации. В обоих случаях алгоритму предоставляются исходные данные, которые он должен проанализировать и найти в них закономерности. Главное отличие между обучением с учителем и обучением без учителя заключается в наличии или отсутствии предоставленных гипотез, которые требуется проверить или подтвердить. Также есть третий тип – обучение с подкреплением, где модель обучается принимать решения и выполнять действия в определенной среде, с целью максимизировать некоторую награду или накопленную полезность. Давайте рассмотрим подробнее эти типы.



**Обучение с учителем**

Обучение с учителем – это тип ML, при котором модель обучается на основе маркированных данных, где каждый пример данных имеет соответствующую целевую переменную или метку. Модель стремится найти зависимости и общие закономерности между входными данными и соответствующими выходными метками. Примеры алгоритмов обучения с учителем включают линейную регрессию, метод опорных векторов, случайные леса и нейронные сети.

**Примеры задач:**

* Регрессия – прогнозирование непрерывной целевой переменной. Например, предсказание цены недвижимости на основе ее характеристик.
* Классификация – определение принадлежности объекта к конкретному классу. Например, классификация электронных писем на спам.

**Обучение без учителя**

Обучение без учителя – это тип ML, при котором модель обучается на немаркированных данных без явной целевой переменной. Вместо этого, модель ищет скрытые структуры, паттерны и группы в данных. Алгоритмы обучения без учителя используются для кластеризации данных, снижения размерности, ассоциативного анализа и генерации новых признаков.

**Примеры задач:**

* Кластеризация – группировка схожих объектов внутри данных. Например, сегментация покупателей на основе их покупательского поведения.
* Снижение размерности – уменьшение размерности данных, сохраняя важные признаки и устраняя шум. Например, сжатие изображений без значительной потери информации.

**Обучение с подкреплением**

Обучение с подкреплением – это тип ML, в котором модель обучается, взаимодействуя со средой и получая обратную связь в виде награды или штрафа. Модель самостоятельно принимает решения и настраивает свое поведение на основе полученной награды. Его активно используют в робототехнике, играх и управлении автономными системами.

**Примеры задач:**

* Управление роботом – обучение робота совершать определенные действия в окружающей среде для достижения поставленных целей.
* Игры – обучение агента играть в игры, например, шахматы или видеоигры, чтобы достичь наивысшего возможного счета.

**Необходимые навыки и образование для машинного обучения**

Машинное обучение требует глубокого понимания алгоритмов и методов, а также умение применять их на практике.

**1. Программирование**

Один из самых важных навыков для работы с машинным обучением — это программирование. Программирование позволяет реализовывать и применять алгоритмы машинного обучения. Два наиболее популярных языка программирования в области машинного обучения – это Python и R. Python обладает широким набором библиотек и инструментов, таких как NumPy, Pandas и TensorFlow, которые упрощают разработку и экспериментирование с моделями машинного обучения. R является мощным инструментом для статистического анализа и визуализации данных. Знание и опыт работы с Python и R являются важными навыками для работы с машинным обучением.

**2. Статистика**

Статистика помогает в анализе данных, оценке моделей и принятии статистически обоснованных выводов. Знание основных понятий статистики, таких как распределения вероятностей, статистические тесты и регрессионный анализ, позволяет более глубоко понимать данные и их связи.

**3. Алгебра и математический анализ**

Линейная алгебра, матрицы, векторы и операции с ними используются во многих алгоритмах и моделях машинного обучения. Знание алгебры позволяет понимать и работать с основными понятиями, такими как скалярное произведение, собственные значения и векторы. Математический анализ включает в себя дифференциальное и интегральное исчисление, которые могут быть полезны при оптимизации моделей и обучении на основе градиентного спуска.

**4. Теория вероятностей**

Теория вероятностей – это фундаментальный инструмент в машинном обучении. Знание основных понятий и методов теории вероятностей помогает строить вероятностные модели, оценивать вероятности событий и работать с вероятностными распределениями. Оно необходимо для понимания стохастических процессов и байесовского подхода в машинном обучении.

**5. Обработка естественного языка и компьютерное зрение**

Обработка естественного языка (Natural Language Processing, NLP) – это область машинного обучения, которая связана с анализом и обработкой текстовой информации, таких как новости, отзывы, социальные медиа и т.д. Для работы с NLP необходимо знание алгоритмов и методов для токенизации, лемматизации, классификации, анализа тональности и машинного перевода. Эти навыки позволяют анализировать и понимать естественный язык.

**6. Компьютерное зрение**

Компьютерное зрение – это область, которая связана с анализом и интерпретацией визуальных данных, таких как изображения и видео. Для работы с компьютерным зрением необходимы навыки обработки изображений, распознавания образов, сегментации и классификации изображений. Знание алгоритмов и методов компьютерного зрения, таких как сверточные нейронные сети и методы обработки изображений, позволяет решать разнообразные таски, связанные с визуальными данными.