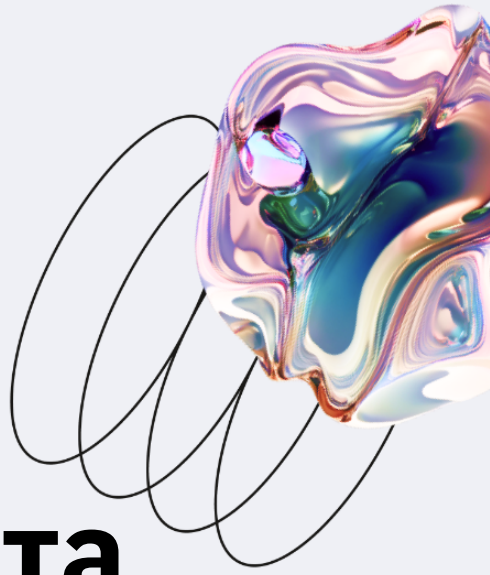
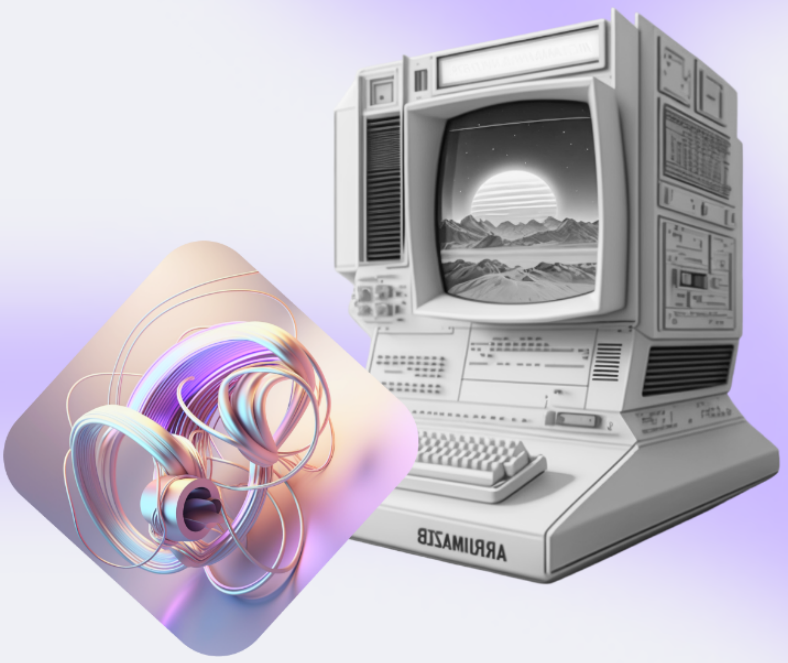
GeekBrains



Создание проекта машинного обучения

Курс «Искусственный интеллект»



Оглавление

Введение 2

Термины, используемые в лекции 3

Определение проблемы и целей 3

- Как определить проблему? 3
- Что значит SMART? 5

Сбор данных 7

Разведочный анализ данных 8

Выбор алгоритмов и моделей 11

Обучение и оценка модели 13

Развёртывание (deploying) модели машинного обучения 15

Сравнение производительности модели с производительностью человека 17

Команда разработчиков проекта машинного обучения 18

Домашнее задание 20

Что можно почитать ещё? 21

Введение

Всем привет и добро пожаловать на очередную лекцию курса «Искусственный интеллект», которая посвящена процессу создания проекта машинного обучения.

В прошлой лекции мы выяснили, что машинное обучение является составляющей искусственного интеллекта и в последние годы внедрено во многие отрасли и используется для решения широкого спектра проблем, от выявления мошенничества до улучшения качества обслуживания клиентов.

В этой лекции мы рассмотрим основные этапы создания проекта машинного обучения, от понимания проблемы и сбора данных до обучения и развёртывания конечной модели.

Так что начнём!

Термины, используемые в лекции

Алгоритм — конечная совокупность точно заданных правил решения некоторого класса задач или набор инструкций, описывающих порядок действий исполнителя для решения определённой задачи.

Определение проблемы и целей

Прежде чем приступать к созданию проекта машинного обучения, важно чётко понимать проблему, которую вы пытаетесь решить. Мы начнём с рассмотрения первого шага создания проекта машинного обучения: определение проблемы и постановка чёткоих целей проекта.

Определение проблемы, которую вы пытаетесь решить, является одним из ключевых шагов в создании проекта машинного обучения. Это может показаться простой задачей, но на самом деле это одна из самых важных частей процесса. Чёткое понимание проблемы поможет выбрать алгоритмы, модели и данные и определить направление проекта.

Постановка проблемы определяет цель проекта и направляет процесс принятия решений. Плохо сформулированная проблема может также привести к некачественному выполнению проекта, некачественным результатам и большой трате времени и ресурсов.

Вот почему так важно иметь чёткое представление о проблеме до начала построения модели машинного обучения. Это обеспечит целенаправленность и последовательность проекта, а также позволит получить в итоге модель, дающую значимые результаты.

Как определить проблему?

Сначала сбор информации о проблеме. Для этого можно провести исследование, поговорить с заинтересованными сторонами или изучить имеющиеся данные.

Существует несколько способов сбора информации о проблеме и целевой аудитории:

- **Проведение исследований:** может включать в себя изучение существующей литературы, данных или отчётов, связанных с проблемой. Важно учитывать как первичные, так и вторичные источники, чтобы получить полное представление о проблеме.

- **Интервью с заинтересованными сторонами:** к ним относятся ключевые лица, такие как эксперты в предметной области или клиенты. Задача — узнать их точку зрения на проблему, а также их потребности и требования.
- **Анализ существующих данных:** включает в себя анализ существующих наборов данных или источников данных, связанных с проблемой, таких как данные о клиентах или данные о продажах. Целью этого является выявление тенденций и закономерностей и получение представления о проблеме.

Посмотрим на конкретном примере. Предположим, вы работаете над проектом по разработке модели машинного обучения для прогнозирования того, какие товары с наибольшей вероятностью будут распроданы в ближайшем будущем. Чтобы собрать информацию о проблеме, вы можете предпринять следующие шаги:

- **Проведение исследования:** изучение существующих отчётов по управлению запасами и прогнозированию продаж товаров в розничной торговле, а также данные других компаний в том же секторе.
- **Беседа с заинтересованными сторонами:** проведение интервью с менеджерами по продажам, командами цепочки поставок и другими ключевыми лицами в компании, чтобы узнать их точку зрения на проблему, а также их потребности и требования к модели прогнозирования продаж продукции.
- **Анализ существующих данных:** проведение анализа данных о продажах компании, включая информацию о популярности продукции, истории покупок и сезонных тенденциях, чтобы выявить закономерности и получить представление о проблеме.

Выполняя эти действия, вы получите более глубокое понимание проблемы прогнозирования продаж и целевой аудитории, что поможет вам в принятии решений и определит направление проекта.

Собрав информацию о проблеме, вы будете лучше подготовлены к формулировке целей проекта. Эта информация также поможет выбрать алгоритмы и модели и определить направление проекта.

Цель проекта — это то, чего вы надеетесь достичь. Определение целей проекта — важная часть создания проекта машинного обучения. Чёткоие, хорошо сформулированные цели помогут вам определить направление проекта и убедиться, что вы работаете над достижением конкретного результата.

Определение целей проекта включает в себя:

- Определение проблемы: как мы уже говорили ранее, важно чётко понимать проблему, которую вы пытаетесь решить.
- Определение целевой аудитории: Кто будет использовать создаваемое вами решение? Понимание нужд и потребностей целевой аудитории очень важно для формирования целей проекта.

Продолжая пример с проектом для прогнозирования того, какие товары с наибольшей вероятностью будут распроданы в ближайшем будущем, допустим, мы определили, что ключевые люди в компании, которые будут использовать модель — это продакт-менеджеры. Мы проводим опрос с ними и определяем их потребности, а также их требования к нашей модели машинного обучения. Например, вы можете узнать, что менеджеры магазинов нуждаются в обновлении информации о наличии товара в режиме реального времени, а командам цепочек поставок для принятия обоснованных решений требуется более подробная информация о тенденциях спроса на товар.

После того как у вас есть чёткое понимание проблемы и целевой аудитории, пришло время установить конкретные, измеримые, достижимые, актуальные и привязанные ко времени (SMART) цели. Эти цели должны быть непосредственно связаны с проблемой и соответствовать потребностям целевой аудитории.

Что значит SMART?

S — специфическая: разработать модель машинного обучения, которая предсказывает, какие продукты с наибольшей вероятностью будут распроданы в ближайшем будущем.

M — измеримая: результат можно измерить по точности модели при составлении прогнозов, которую можно отследить с помощью различных метрик.

A — достижимая: цель достижима, поскольку компания располагает множеством данных о продажах, которые могут быть использованы для обучения модели, и уже существуют алгоритмы и методики построения подобных моделей.

R — актуальность: цель соответствует бизнес-целям компании, поскольку она поможет улучшить доступность продукции и сократить количество отходов, что повысит доход и удовлетворённость клиентов.

T — ограниченная по времени: компания намерена завершить проект в течение 6 месяцев и внедрить модель в магазинах в течение следующих 12 месяцев.

Постановка чётких целей проекта очень важна, поскольку она помогает вам оставаться сосредоточенными и идти по намеченному пути на протяжении всего

проекта. Это также помогает убедиться, что вы работаете над достижением конкретного результата, который отвечает потребностям целевой аудитории и решает поставленную проблему.

Что может пойти не так на этом этапе? Допустим, перед группой специалистов по обработке данных была поставлена задача создать модель машинного обучения для повышения продаж компании, занимающейся розничной торговлей продуктами. Генеральный директор компании с нетерпением ждёт результатов и поставил перед командой задачу разработать модель как можно быстрее. Команда с головой окунулась в проект, не потратив времени на чёткое определение проблемы, которую они пытались решить. Они просто предположили, что проблема заключается в увеличении продаж, не рассматривая первопричину того, почему продажи не соответствуют ожиданиям. В результате они собрали данные и построили модель, которая должна была предсказать, какие продукты будут наиболее популярными и обеспечат наибольший объём продаж.

Однако, когда модель была развёрнута, результаты оказались неутешительными. Модель не улучшила продажи, а наоборот, ухудшила их. Здесь команда понимает, что постановка задачи была слишком широкой и что они не учли другие факторы, влияющие на продажи, такие как конкуренция, изменение потребительских предпочтений и экономические условия.

Приходится возвращаться к планированию и тратить время на то, чтобы чётко определить проблему, которую они пытались решить. Например, собрать данные о потребительских предпочтениях, тенденциях рынка и экономическом климате. Побеседовать с клиентами и торговым персоналом, чтобы лучше понять, что является движущей силой продаж.

Отсюда следует вывод, что время, потраченное на чёткое определение проблемы, имеет решающее значение для успеха проекта машинного обучения.

Этот пример демонстрирует, как важно уделять время чёткому определению проблемы в проекте машинного обучения. Определив чёткоую постановку проблемы, вы сможете убедиться, что ваш проект сфокусирован и что ваши усилия направлены на решение правильной проблемы. Игнорируя постановку проблемы, вы рискуете построить модель, которая не будет целевой и даже может иметь негативные последствия.

Сбор данных

Сбор данных — важнейший этап создания проекта машинного обучения. Качество ваших данных напрямую влияет на точность и эффективность вашей модели.

Сбор данных начинается с определения того, а какие собственно данные нужны. Для построения успешной модели важно определить правильные данные для сбора. Это включает в себя рассмотрение типа данных, которые будут наиболее релевантны для постановки вашей проблемы, и типа данных, которые будут необходимы для обучения вашей модели.

Компания может рассмотреть следующие вопросы:

- Какой тип данных будет наиболее релевантным для проблемы прогнозирования покупательского поведения клиентов? Это могут быть данные о покупателе, такие как демографическая информация, история покупок и поведение при просмотре сайтов, а также данные о самих товарах, такие как категории товаров, цены и рейтинги.
- Какие источники данных доступны компании? Это могут быть внутренние источники данных, такие как базы данных клиентов и файлы журналов, а также внешние источники данных, такие как демографические данные из государственных баз данных или данные о продукции с сайтов конкурентов.
- Какое количество данных необходимо для решения задачи? Если у вас небольшая компания и в вашем доступе имеется информация об ограниченном количестве клиентов, то небольшого датасета может оказаться недостаточно для построения модели.

Источники данных. Существует множество различных источников данных, которые могут быть использованы для проектов машинного обучения. Эти источники могут включать внутренние данные, такие как базы данных клиентов, или внешние источники, такие как общедоступные наборы данных. При сборе данных важно учитывать качество и актуальность данных, а также методы, используемые для их получения.

Подготовка данных. После сбора данных их обычно необходимо очистить, предварительно обработать и преобразовать, чтобы сделать пригодными для использования в проекте машинного обучения. Это может включать такие задачи, как удаление неактуальных данных, обработка отсутствующих значений и некоторые математические преобразования для расчёта статистических показателей.

Выборка данных. В зависимости от размера набора данных, для построения модели может потребоваться выборка данных. Выборка — это процесс отбора подмножества данных для использования при обучении, проверке и тестировании. Цель состоит в том, чтобы выбрать репрезентативную выборку, которая точно отражает общее распределение данных.

Это важно, поскольку использование слишком большого количества данных может сделать модель медленной и трудной для обучения, в то время как использование слишком малого количества данных может привести к недостаточно обученной модели, которая будет работать плохо.

Продолжая наш пример, компания может начать с определения объёма данных, которые можно использовать для проекта. Например, в их базе данных могут быть миллионы покупок клиентов. Затем компания должна решить, какой объём данных она хочет использовать для проекта. Например, они могут решить использовать 10% от общего объёма данных, что составит сотни тысяч покупок клиентов.

Чтобы убедиться, что выборка является репрезентативной для всего набора данных, компания может использовать метод случайной выборки для отбора данных, которые будут использоваться для проекта. Например, они могут использовать генератор случайных чисел для отбора подмножества покупок клиентов.

Конфиденциальность и безопасность данных. При сборе и хранении данных важно учитывать вопросы конфиденциальности и безопасности. Это включает в себя обеспечение безопасного хранения персональных данных и их использование только в тех целях, для которых они были собраны.

Разведочный анализ данных

Разведочный анализ данных (Exploratory Data Analysis, EDA) — это важный шаг в процессе анализа данных, который включает в себя визуальное и статистическое изучение наборов данных для понимания их основных характеристик, выявления закономерностей и обнаружения потенциальных проблем. Почему EDA имеет важное значение:

Понимание структуры данных. EDA помогает понять структуру и организацию набора данных, включая количество наблюдений, переменных и их типы (например, числовые, категориальные). Это важно для выбора подходящих методов анализа и обеспечения точности результатов.

Выявление закономерностей и взаимосвязей. EDA позволяет выявить взаимосвязи между переменными, тенденции и закономерности в данных, что позволяет аналитикам выдвигать обоснованные гипотезы о глубинных процессах и структурах, лежащих в основе данных, которые могут быть исследованы с помощью более продвинутых методов анализа.

Обнаружение аномалий. Например: выбросы, недостающие значения и ошибки в данных. Раннее обнаружение таких проблем имеет решающее значение для обеспечения точности анализа и предотвращения неверных выводов. Выявление аномалий также помогает улучшить качество данных путём исправления ошибок или восполнения недостающих значений.

Принятие более эффективных решений. Понимая характеристики и взаимосвязи данных, аналитики могут сделать более обоснованный выбор, на каких переменных сосредоточиться, какие модели использовать и как интерпретировать результаты.

Обоснование дальнейшего анализа. Здесь имеется в виду выбор наиболее подходящих статистических методов и моделей машинного обучения для дальнейшего анализа. Понимая распределение и взаимосвязи данных, аналитики могут выбрать правильные инструменты для изучения своих исследовательских вопросов или эффективного построения прогностических моделей.

Представление результатов. Визуальные представления данных, такие как графики и диаграммы, являются важной частью EDA. Они помогают легче донести сложную информацию до нетехнической аудитории, облегчая заинтересованным лицам понимание результатов и принятие обоснованных решений на основе анализа.

Пример **разведочного анализа данных** для проекта машинного обучения для компании, занимающейся розничной продажей товаров:

Одномерный анализ. Первым шагом является проведение одномерного анализа для каждой переменной в наборе данных. Например, компания розничной торговли может проанализировать распределение цен на товары, ассортимент категорий товаров и количество проданных товаров в день. Примером может служить такой статистический показатель, как среднее значение.

Многомерный анализ. Далее компания может провести многомерный анализ, чтобы понять взаимосвязь между ценой товара и количеством проданных товаров. Для визуализации взаимосвязи между этими двумя переменными можно построить диаграмму рассеяния. Это может помочь компании выявить любые корреляции между ценой продукта и продажами, что может послужить основой для разработки модели машинного обучения.

Визуализация данных. Компания может использовать методы визуализации данных, такие как графики и гистограммы, чтобы лучше понять распределение цен на продукцию и продаж. Например, график может быть использован для отображения количества проданных товаров за определённое время, а гистограмма — для отображения общего количества проданных товаров по категориям.

Преобразование данных. Компания также может выполнить преобразование данных, чтобы лучше понять взаимосвязь между ценой продукта и продажами. Например, они могут преобразовать переменную цены продукта в категории, такие как низкая, средняя и высокая, и проанализировать связь между ценовой категорией продукта и продажами.

Выполняя разведочный анализ данных, компания розничной торговли может получить представление о данных и выявить взаимосвязи между переменными, которые могут послужить основой для разработки модели машинного обучения. Это может помочь компании принимать более обоснованные решения о ценообразовании, размещении товаров и рекламных акциях для увеличения продаж.

Другой пример, **анализ данных для проекта** глубокого обучения для прогнозирования рака лёгких по изображениям компьютерной томографии (КТ):

Первым шагом является сбор большого набора данных КТ-изображений лёгких, как с раком, так и без него. Эти данные могут быть получены из баз данных медицинских изображений или непосредственно из больниц. Перед проведением разведывательного анализа данных изображения необходимо предварительно обработать. Это может быть изменение размера изображений до однотипного, нормализация значений пикселей и преобразование изображений в оттенки серого.

Одномерный анализ. Далее для каждого изображения может быть проведён одномерный анализ. Например, можно рассчитать среднее значение, стандартное отклонение и диапазон значений пикселей для каждого изображения, чтобы понять распределение значений пикселей внутри каждого изображения.

Многомерный анализ. Для понимания взаимосвязи между несколькими изображениями можно провести многомерный анализ. Например, можно построить диаграмму рассеяния для визуализации взаимосвязи между средними значениями пикселей двух изображений.

Визуализация данных. Визуализация данных является важным шагом в анализе данных для проектов глубокого обучения. Изображения можно визуализировать с

помощью таких инструментов, как гистограммы и тепловые карты, чтобы лучше понять распределение значений пикселей внутри каждого изображения и выявить любые корреляции между изображениями.

Сегментация изображений. В некоторых случаях может потребоваться сегментация изображений, которая предполагает разделение изображений на различные области, представляющие интерес. Это может помочь определить области на изображениях, которые могут быть важны для прогнозирования рака лёгких.

Что, если пропустить этот важный этап? Почему бы не взять и просто ввести данные в алгоритм глубокого обучения и получить точные результаты. На самом деле так не получится, потому что модель практически со 100% вероятностью будет выдавать неточные прогнозы. Что может пойти не так?

Например, набор данных может быть сильно несбалансирован, например, с непропорционально много изображений без рака. В таком случае модель будет склонна к предсказанию отсутствия рака, это приведёт к большому количеству ложноотрицательных результатов.

Если не провести предварительную обработку изображений, что приведёт к большим различиям в значениях пикселей на изображениях. Из-за этого модели глубокого обучения будет сложно точно определить закономерности в данных.

Кроме того, модель будет обучаться лучше, если ей указать важные регионы на изображениях, которые важны для прогнозирования рака лёгких.

Выбор алгоритмов и моделей

При создании проекта машинного обучения очень важно выбрать соответствующие алгоритмы и модели, которые лучше всего подходят для ваших данных и проблемы.

Что такое алгоритм и модель? Допустим, вы хотите испечь печенье. Рецепт выпечки печенья похож на алгоритм. Он даёт вам пошаговые инструкции о том, какие ингредиенты использовать, как их смешивать и как долго выпекать.

Модель машинного обучения — это как умная компьютерная программа, которая может выполнять определённые задачи. Например, она может помочь вам определить объекты на картинках, например, узнавать животных или отличить автомобиль от пешехода, или помочь вам перевести слова с одного языка на другой.

Здесь важно понимать, что один и тот же алгоритм машинного обучения используется для создания совершенно разных моделей машинного обучения. Это происходит, когда один и тот же алгоритм машинного обучения обучается на разных данных.

Например, представьте, что у вас есть алгоритм для распознавания различных типов цветов. Если вы обучите алгоритм на наборе данных красных роз, он создаст модель, которая будет хорошо распознавать красные розы. Но если вы обучите тот же алгоритм на наборе данных жёлтых маргариток, то получите другую модель, которая будет хорошо распознавать жёлтые маргаритки, но плохо красные розы.

Таким образом, несмотря на то, что алгоритм в обоих случаях один и тот же, модели, которые он создаёт, отличаются, потому что обучающие данные разные. Алгоритм получает данные и использует их для создания модели, которая лучше всего подходит для данной конкретной задачи.

Именно поэтому модели машинного обучения могут быть точно настроены и адаптированы для конкретных случаев использования путём обучения на правильных данных.

Выбор правильного алгоритма и тестирование модели — это итерационный процесс, и вам может потребоваться попробовать несколько вариантов, прежде чем вы найдёте наилучший вариант для ваших данных и проблемы. Кроме того, некоторые алгоритмы и модели могут работать лучше на определённых типах данных или проблем, поэтому всегда полезно экспериментировать и оценивать различные варианты.

Приведу пример некоторых популярных моделей машинного обучения:

- **Линейная регрессия** — это простая модель, которая предсказывает числовое значение на основе взаимосвязи между входными характеристиками (независимые переменные) и выходными данными (зависимая переменная). Она предполагает линейную зависимость между этими переменными. Например, прогнозирование цен на дома на основе размера, местоположения и возраста дома.

- **Деревья решений или решающие деревья** — это графические модели, которые принимают решения на основе серии вопросов о входных данных. Они разбивают данные на подмножества, основанные на определённых условиях, создавая древовидную структуру. Например, решение о том, играть в теннис или нет, принимается на основе погодных условий, таких как температура, влажность и скорость ветра.

- **Метод опорных векторов (SVM)** — это модель, которая находит наилучшую границу (или линию) для разделения различных классов в данных. Она работает путём максимизации расстояния между классами. Например, классификация электронной почты как спама или не спама на основе слов и фраз, присутствующих в письме.

Это лишь несколько примеров из множества доступных моделей машинного обучения. Каждая модель имеет свои сильные и слабые стороны и подходит для разных типов задач. Выбор модели зависит от конкретной проблемы, типа данных и желаемого результата.

Обучение и оценка модели

Обучение и оценка модели — важнейший этап в создании проекта машинного обучения. Именно здесь алгоритм используется для обучения на основе данных и создания модели, способной выполнить задачу, для которой она была разработана.

Чтобы обучить модель, данные подаются в алгоритм, и алгоритм использует их для обновления **параметров модели**.

Что такое параметры модели? Представьте, что у вас есть игрушечный робот, который может сортировать игрушечные кубики по цвету. У робота есть камера, с помощью которой он фотографирует кубики и определяет их цвет. Чтобы правильно сортировать кубики, робот должен уметь точно определять их цвет.

В этом случае параметрами модели машинного обучения робота могут быть настройки регуляторов (как регулятор громкости телевизора) для определения красного, зелёного, синего и других цветов. Положение регуляторов определяют, как робот различает разные цвета. Например, если положение регулятора для красного цвета установлено слишком низко, робот может классифицировать розовый блок как красный, а если положение установлен слишком высоко, робот может классифицировать красный блок как коричневый.

Таким образом, изменяя положение регуляторов, мы можем изменить поведение робота и сделать его лучше в определении цветов. В процессе обучения роботу даётся набор блоков разных цветов, а пороговые значения настраиваются до тех пор, пока робот не научится правильно определять цвета блоков. Окончательный набор положений регуляторов будет представлять собой модель или знания робота о том, как выглядят красный, зелёный, синий и другие цвета.

Это простой пример, но та же идея применима и к более сложным моделям машинного обучения. Процесс поиска оптимальных параметров повторяется несколько раз, причём алгоритм каждый раз вносит коррективы, пока производительность модели не достигнет приемлемого уровня. Цель состоит в том, чтобы найти оптимальный набор параметров, дающий наиболее точные результаты.

Кроме параметров модели, существует такое понятие как **гиперпараметры** — это внешние настройки или регуляторы, которые помогают управлять процессом обучения самого алгоритма. Они не обучаются на данных, а задаются пользователем до начала обучения. Они определяют общую структуру и поведение модели, например, количество слоёв в нейронной сети, глубину дерева решений.

Разница между параметрами и гиперпараметрами:

- **Обучение**: параметры настраиваются на основе данных в процессе обучения, а гиперпараметры задаются пользователем до начала обучения.
- **Цель**: параметры напрямую влияют на прогнозы модели, тогда как гиперпараметры контролируют процесс обучения и общее поведение модели.
- **Настройка**: параметры итеративно обновляются алгоритмом обучения, в то время как гиперпараметры часто настраиваются пользователем с использованием различных методов, таких как поиск по сетке или случайный поиск.

В нашем примере с роботом параметрами являются настройки управления для определения определённых цветов, а гиперпараметрами могут быть такие параметры, как скорость обучения (как быстро робот настраивает свои органы управления) или максимальное количество итераций, допустимое для обучения робота. Путём точной настройки этих гиперпараметров можно оптимизировать процесс обучения робота для достижения лучших результатов в определении цветов.

Как уже было сказано, в машинном обучении собственно обучение происходит на основе набора данных, который называется **датасет.** Датасет обычно делится на две (иногда более) части: тренировочный датасет и тестовый датасет. Эти датасеты служат разным целям в процессе построения и оценки моделей машинного обучения.

Тренировочный датасет — это основной набор данных, используемый для обучения модели машинного обучения. Модель изучает основные закономерности и взаимосвязи в данных, настраивая свои параметры так, чтобы минимизировать ошибку между своими прогнозами и фактическими целевыми значениями.

Тестовый датасет — это отдельный набор данных, который используется для оценки производительности обученной модели машинного обучения. Очень важно оценить работу модели на данных, которые она никогда раньше не видела, чтобы убедиться, что она хорошо обобщает новые данные, которые она ранее не «видела». Эти данные не используются в процессе обучения. Работа модели на тестовом датасете даёт оценку того, насколько хорошо модель будет работать на реальных данных.

Основная цель разделения данных на обучающий и тестовый наборы состоит в том, чтобы предотвратить **переобучение**. Переобучение происходит, когда модель настолько хорошо усваивает учебные данные, что улавливает шум или случайные колебания в данных, что приводит к низкой производительности на новых, невидимых данных. Оценивая модель на тестовом наборе данных, мы можем получить объективную оценку способности модели к обобщению на новые данные, что является критически важным аспектом успешной модели машинного обучения.

Для оценки модели могут использоваться различные **метрики**. Эти метрики позволяют количественно оценить работу модели и сравнить её с другими моделями. По результатам оценки может потребоваться дальнейшая доработка модели, например, корректировка её параметров или использование другого алгоритма.

Важно помнить, что оценка — это лишь моментальный снимок работы модели, и она может измениться по мере поступления новых данных. Чтобы модель оставалась точной, её следует периодически переоценивать и обновлять по мере необходимости.

Развёртывание (deploying) модели машинного обучения

Развёртывание модели машинного обучения в реальном сценарии — это заключительный этап создания проекта машинного обучения. На этом этапе модель используется для выполнения задачи, для которой она была разработана.

При развёртывании модели машинного обучения в реальном мире необходимо учитывать несколько факторов. Одним из наиболее важных является то, как модель будет получать входные данные и выдавать выходные результаты. Это может включать в себя интеграцию модели в более крупную систему, такую как веб-приложение, мобильное приложение или база данных.

Другим важным фактором являются вычислительные ресурсы, необходимые для работы модели. Модели машинного обучения могут требовать значительной вычислительной мощности и памяти, поэтому важно продумать, как будет работать модель и откуда будут браться вычислительные ресурсы.

Кроме того, важно учитывать последствия развёртывания модели машинного обучения для безопасности и конфиденциальности. Например, модели могут обрабатывать персональные данные, поэтому важно обеспечить защиту данных и безопасную работу модели.

Наконец, важно следить за производительностью модели после развёртывания. Это может включать в себя создание системы для отслеживания работы модели, например, протоколирование входных данных и выходных результатов, или отслеживание точности. Мониторинг производительности модели может помочь выявить проблемы, которые необходимо решить, например, изменения во входных данных или необходимость в дополнительных данных для обучения.

Примером развёртывания модели машинного обучения в реальном мире может быть развёртывание чат-бота для обслуживания клиентов крупной компанией, занимающейся электронной коммерцией.

Допустим, некая компания имеет большую клиентскую базу, которая обращается в центр обслуживания клиентов с самыми разными вопросами и проблемами. Чтобы улучшить качество обслуживания клиентов и сократить время ожидания, компания решила развернуть чат-бот на базе машинного обучения, который мог бы обрабатывать основные запросы клиентов круглосуточно и без выходных.

Компания создала модель машинного обучения, которая была обучена на большом наборе данных о взаимодействии с клиентами, включая как текстовые, так и голосовые запросы. Модель была разработана таким образом, чтобы отвечать на запросы клиентов на обычном (естественном) языке, предоставляя быстрые и точные ответы на вопросы клиентов.

Затем модель была интегрирована в центр обслуживания клиентов компании, что позволило клиентам взаимодействовать с чат-ботом через веб-сайт компании, мобильное приложение и колл-центр. Чат-бот способен обрабатывать широкий спектр запросов клиентов, включая вопросы о заказах, доставке, возврате и информации о продукции.

Однако вскоре после развёртывания компания обнаружила, что вычислительных ресурсов, необходимых для работы модели машинного обучения, недостаточно. Это привело к медленному времени отклика и ухудшению общего впечатления клиентов. Компания быстро поняла, что модель машинного обучения была слишком

сложной для первоначально выделенных ресурсов. Это означало, что модель требовала слишком много времени для обработки запросов клиентов, что приводило к замедлению времени ответа и недовольству клиентов.

Чтобы решить эту проблему, компании пришлось выделить дополнительные вычислительные ресурсы для модели машинного обучения. Это могло потребовать модернизацию серверной инфраструктуры или переноса модели на более мощную вычислительную платформу. Компания также оптимизировала модель, чтобы сделать её более эффективной, сократив количество вычислительных ресурсов, необходимых для работы модели.

Кроме того, чтобы убедиться, что модель продолжает работать хорошо, компания отслеживает её работу в течение долгого времени и при необходимости вносит коррективы. Например, если модель с трудом даёт точные ответы на определённый тип запросов или даёт неточные ответы, компания может добавить дополнительные обучающие данные, чтобы повысить эффективность модели.

В этом примере развёртывание модели машинного обучения в реальном сценарии позволяет улучшить качество обслуживания клиентов компании, занимающейся электронной коммерцией.

Следует отметить, что развёртывание модели машинного обучения в реальном мире требует тщательного рассмотрения нескольких факторов, включая вычислительные ресурсы, производительность модели и ограничения реального мира. Оценив эти факторы и предприняв необходимые шаги для решения любых проблем, организации смогут успешно развернуть модели машинного обучения, которые обеспечат реальные решения реальных проблем. Учитывая растущее значение машинного обучения в современном бизнесе, понимание того, как эффективно развёртывать модели машинного обучения, является необходимым навыком для профессионалов в этой области.

Сравнение производительности модели с производительностью человека

При создании проекта машинного обучения важно сравнить производительность модели с производительностью человека. Это сравнение помогает понять сильные и слабые стороны модели, а также даёт представление об областях, требующих улучшения.

Производительность на уровне человека можно рассматривать как уровень точности и эффективности, которого мог бы достичь человек-эксперт при решении конкретной задачи. Например, в случае классификации изображений, производительность на уровне человека — это точность, с которой человек-эксперт может правильно отнести изображение к определённой категории.

По сравнению с этим, модели машинного обучения могут обеспечить гораздо более высокую точность и эффективность, но у них также есть свои ограничения. Например, модели машинного обучения могут работать со сложными или неструктурированными данными, и они не всегда могут дать полное представление о том, почему было принято то или иное решение.

Сравнивая производительность модели машинного обучения с производительностью на уровне человека, организации могут определить области, в которых модель работает хорошо, и области, в которых она нуждается в улучшении. Эта информация может быть использована для принятия обоснованных решений о том, как улучшить модель и оптимизировать её работу.

Одним из примеров сравнения производительности с человеческим уровнем являются многочисленные исследования в контексте диагностики рака молочной железы. Исследования показывают, что различные модели машинного обучения, обученные на маммографических изображениях (рентген-снимки молочной железы), достигают точности порядка 90% в выявлении рака молочной железы, что сопоставимо с точностью людей-рентгенологов.

Сравнение производительности моделей машинного обучения с производительностью человека подчёркивают потенциал моделей машинного обучения в области медицины.

Команда разработчиков проекта машинного обучения

Персонал команды разработчиков проекта машинного обучения имеет решающее значение для успеха проекта. Есть несколько ключевых ролей, которые должны быть заполнены для обеспечения гладкого и эффективного проекта.

Менеджер проекта (**Project Manager**). Менеджер проекта отвечает за контроль над общим развитием проекта. Он отвечает за создание плана проекта, распределение ресурсов и обеспечение того, чтобы проект не отставал от графика.

Data Scientist отвечает за анализ данных, выбор подходящих алгоритмов и построение моделей машинного обучения. Он должен хорошо разбираться в статистике, математике и программировании.

Инженер-программист (**Software Engineer**). Инженер-программист отвечает за внедрение моделей машинного обучения в программное обеспечение. Он должен обладать сильными навыками программирования и должен быть знаком с библиотеками машинного обучения.

Инженер по данным (**Data Engineer**). Инженер по данным отвечает за управление инфраструктурой данных и обеспечение надлежащего сбора, хранения и обработки данных. Он должен хорошо разбираться в базах данных и хранилищах данных.

Команда обеспечения качества (тестировщики) отвечает за тестирование программного обеспечения и обеспечение его соответствия требованиям и спецификациям. Они должны иметь опыт тестирования программного обеспечения и быть знакомы с инструментами автоматизированного тестирования.

Пример команды, разрабатывающей модель машинного обучения для прогнозирования продаж в интернет-магазине, может выглядеть следующим образом:

Инженер по данным собирает и обрабатывает большие объёмы данных о продажах в интернет-магазине, включая информацию о поведении покупателей, информацию о товарах и исторические данные о продажах. Эти данные используются для обучения модели машинного обучения. Data Scientist анализирует данные, чтобы выявить тенденции и закономерности в продажах. Он также проведёт очистку и предварительную обработку данных, чтобы убедиться, что они пригодны для обучения.

Затем специалист по анализу данных выбирает наиболее подходящие алгоритмы машинного обучения и модели, такие как деревья решений или нейронные сети. Он также разделят данные на обучающий и тестовый наборы, чтобы оценить эффективность модели.

Затем инженер-программист реализует выбранные модели в коде и обучает модель на обработанных данных. Они также проводят точную настройку модели путём корректировки параметров и тестирования различных подходов. После обучения и тестирования модели инженер-программист внедряет её в производственную среду. Он также интегрируют модель в существующую программную систему интернет-магазина, например, в систему рекомендаций или панель прогнозной аналитики.

Data Scientist и инженер-программист будут постоянно следить за эффективностью модели и при необходимости вносить обновления, чтобы она оставалась точной и эффективной.

Параллели с разработкой обычной программы можно увидеть в этапах планирования, проектирования, реализации и сопровождения. В обоих случаях команда людей с различными наборами навыков работает вместе для достижения общей цели. В случае обычной программы команда может быть сосредоточена на создании программного приложения, в то время как в проекте машинного обучения команда сосредоточена на создании модели, которая может обучаться и делать прогнозы. Однако общий процесс сбора требований, разработки решения, его внедрения и поддержания в рабочем состоянии в обоих случаях схож.

Разумеется, состав команды, задачи и роли будут меняться в зависимости от сферы, однако общая структура остаётся примерно как в приведённом примере.

В заключение следует отметить, что создание проекта машинного обучения требует чёткого понимания проблемы, которую вы пытаетесь решить, данных, которые вы будете использовать, и соответствующих алгоритмов. Также требуется тщательная предварительная обработка и очистка данных, а также оценка и сравнение эффективности различных моделей.

В целом создание приложений искусственного интеллекта имеет много параллелей с программированием, однако имеет особенности, которые мы обсудили. При тщательном планировании и исполнении машинное обучение может стать мощным инструментом для получения информации и составления прогнозов. Ключ к успеху любого проекта машинного обучения — это научный и экспериментальный подход, постоянная итерация и тонкая настройка моделей до достижения желаемого уровня производительности.

Спасибо за внимание и увидимся на следующей лекции!

Домашнее задание

1. Исследуйте и найдите реальную проблему, которую, по вашему мнению, можно решить с помощью машинного обучения.
2. Напишите краткий отчёт о проблеме, включая её историю, почему она важна и какие проблемы она представляет.

3. Поищите информацию о том, как машинное обучение в настоящее время используется для решения подобных проблем, и обобщите свои выводы в отчёте.

4. Определите одно или два потенциальных применения машинного обучения для решения этой проблемы и опишите, как оно может помочь решить проблему значимым образом.

Пример решения:

1. Проблема пробок в городских районах.

2. Пробки на дорогах являются распространённой проблемой в городских районах, вызывая задержки и увеличивая загрязнение воздуха. Это очень важно, поскольку влияет на повседневную жизнь людей, приезжающих на работу, способствует снижению производительности труда и усугубляет экологические проблемы. Основные задачи в решении этой проблемы заключаются в поиске способов оптимизации транспортного потока и сокращения узких мест.

3. В настоящее время машинное обучение используется для прогнозирования трафика, оптимизации маршрутов для грузовиков и улучшения транспортного потока в режиме реального времени.

4. Одно из потенциальных применений машинного обучения для решения этой проблемы — использование данных с GPS и датчиков движения для прогнозирования трафика и оптимизации маршрутизации транспортных средств на дороге. Другое применение — использование алгоритмов машинного обучения для анализа данных о транспортном потоке и выявления узких мест, что может помочь градостроителям принимать обоснованные решения о том, как улучшить транспортный поток и уменьшить заторы.

Что можно почитать ещё?

1. «Верховный алгоритм» Педро Домингос.

2. «Машинное обучение» Хенрик Бринк, Джозеф Ричардс, Марк Феверолф.

3. «Машинное обучение. Наука и искусство построения алгоритмов, которые извлекают знания из данных» Петер Флах.