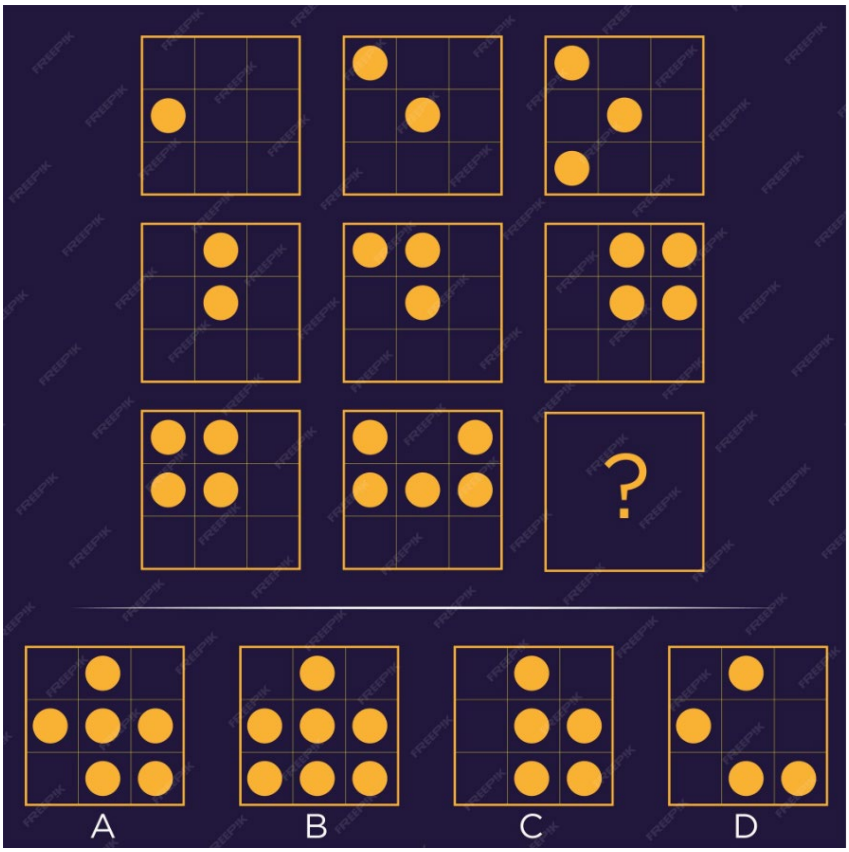

DROFS
2024

КОНЦЕПЦИЯ МАШИННОГО ОБУЧЕНИЯ

Многие из вас наверняка проходили тесты на IQ и вам знакомы задания, в которых необходимо найти закономерность среди нескольких картинок. Вот типовой пример таких задач:

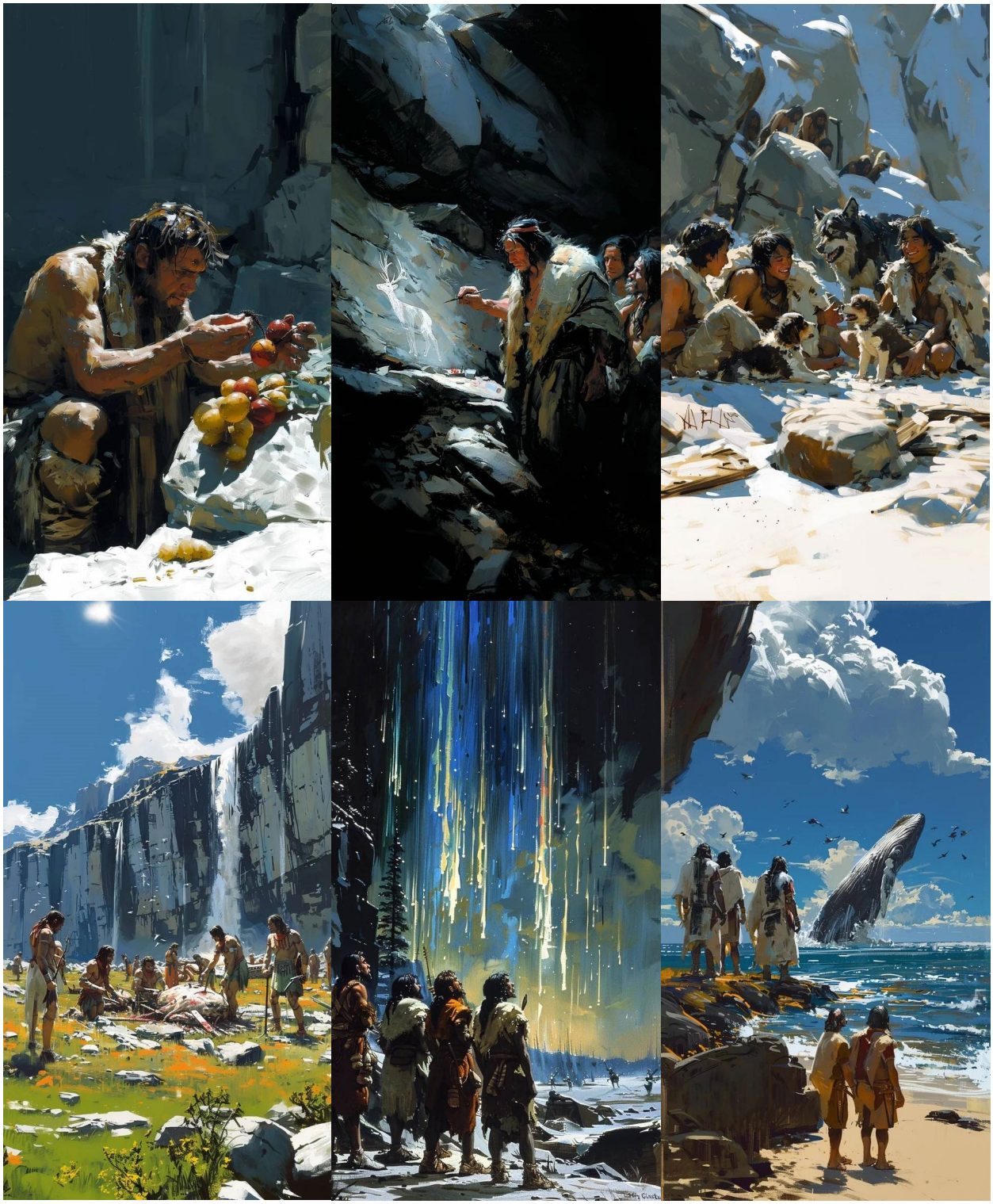


Обычно в таких тестах оценивается скорость, с которой вы находите правильный ответ: чем быстрее вы справитесь, тем выше ваш результат. Однако, как и в нашем примере, поиск правильной закономерности не всегда очевиден и порой требуется изрядно поднапрячься чтобы отыскать необходимую комбинацию.

Подобные задачи развивают способность распознавать не тривиальные паттерны, тренируют наш мозг находить порядок в хаосе, а это уже немаловажный навык и давайте разбираться почему. Почему поиск закономерностей так важен?

Представьте себе древнего охотника, который внимательно следит за поведением животных и природными явлениями. Он замечает, что стада мигрируют в определенные сезоны, растения в основном плодоносят ближе к

осени, а днем, когда солнце находится в зените, температура выше, чем когда оно садится или встает. Этот процесс наблюдения и анализа лежит в основе его благополучия - он помогает ему проще находить пропитание для своего племени и избегать смертельных опасностей.



Размышляя аналогичным образом, отметим, что любопытство и стремление к пониманию вещей кажущимися на первый взгляд случайными, лежат в основе нашего выживания. А то, что мы сегодня называем “наукой” - выступает в роли борца с отведенным для нас временем.

Наука — это ничто иное как система накопленных знаний, которая облегчает поиск и описание природных закономерностей.

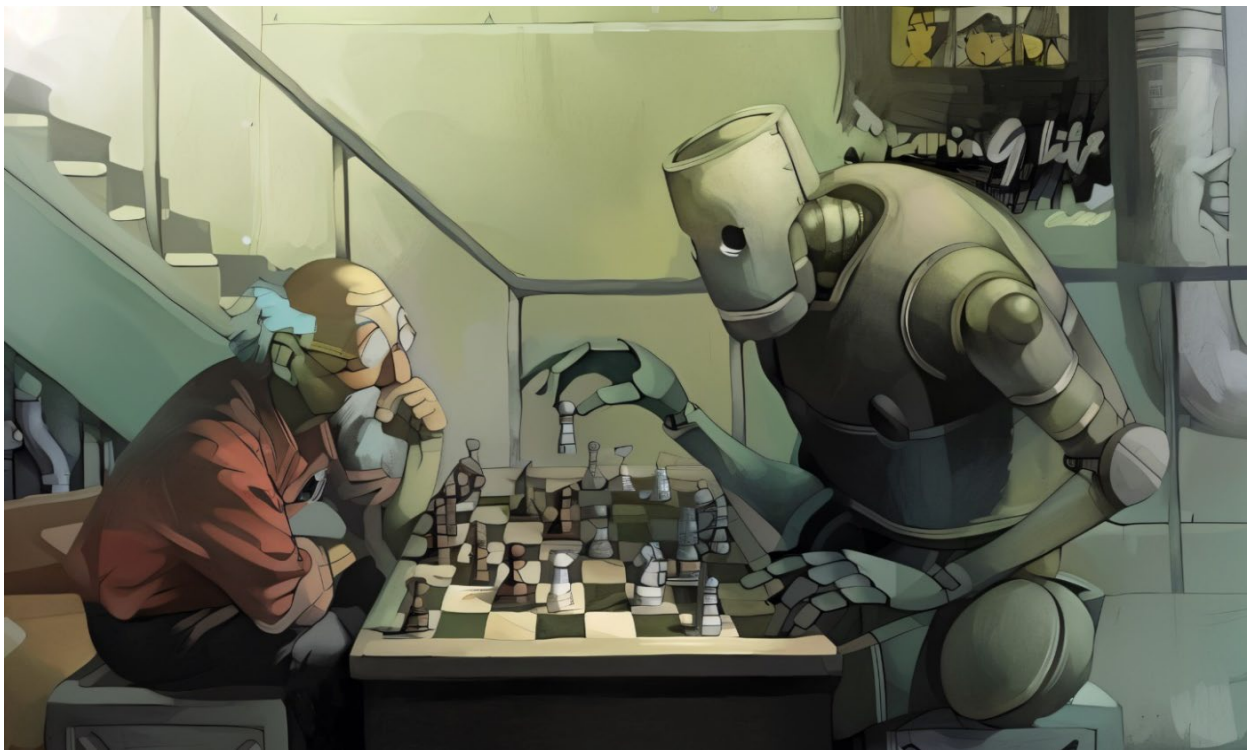
Взглянув на ретроспективу событий не трудно заметить, что эта система неумолимо растет. Наблюдая за историей развития человечества, прослеживается тот факт, что мы куда-то движемся, и тенденция нашего роста инерционна - с каждым годом этот процесс все ускоряется и ускоряется, накопленных знаний становится все больше и больше.

И так как наука постоянно растет можно предположить, что со временем появятся такие задачи, которые будут непосильны человеку. Что если для обнаружения правильной закономерности потребуется проанализировать и запомнить непомерное число элементов? Представьте, что вам нужно найти закономерность не среди десяти, а среди десяти тысяч картинок, внимательно посмотрев и запомнив каждую из них. Удержать подобный объем информации, быстро проанализировать его и дать правильный ответ – физически невыполнимая задача.

И что, если человек, неготовый мериться со своими слабостями и преисполненный горящим нетерпением познания этого мира, неосознанно придумывает механизмы способные помочь ему в этом. Машины, которые без устали готовы перемалывать тонны информации, выявлять скрытые закономерности и находить контринтуитивные решения.

Если это так, то получается единственное, что теперь будет требоваться от человека – это *обучить машину,* дать ей алгоритм, а еще лучше сделать так чтобы она сама искала алгоритм по поиску закономерностей.

Эта идея, так скажем в первом приближении, она возможно не совсем точна, и со временем мы наделим ее ещё большим смыслом, но пока она дает нам понять суть машинного обучения как такового.



Давайте попробуем выяснить из каких основных этапов должно строиться машинное обучение, то есть без чего оно не может существовать. И здесь, как ни странно, достаточно задаться вопросом: "А чего мы, собственно говоря, хотим получить от машин?" Мы хотим: по входным дынным – на основе какого-то алгоритма (обучения) – получить результат. Схематично эту концепцию можно представить следующим образом (см. **рисунок 3**):

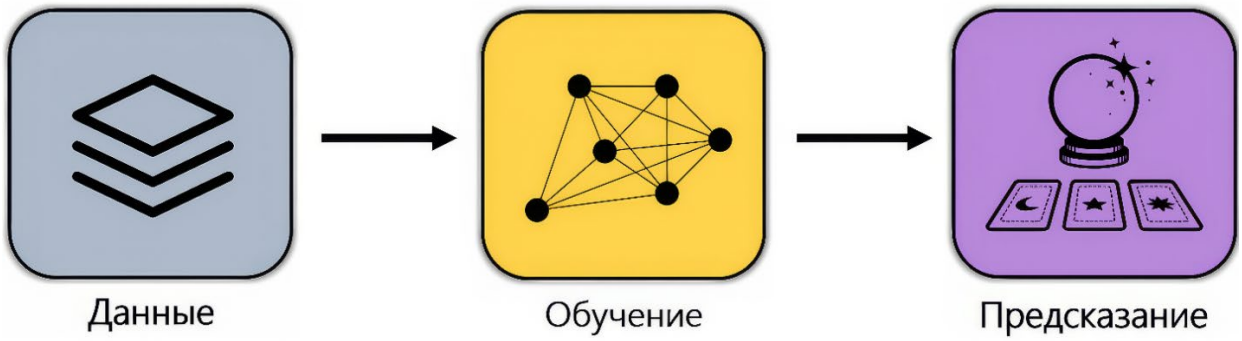


Рисунок – 3 Схематичное изображение концепции машинного обучения

Как и всякая обобщенная схема, представленная концепция имеет свои плюсы и минусы, но на данном этапе нас это вообще не волнует. Наша первостепенная задача — это уловить основную идею и только после ее осознания, вдаваться в детали.

И поскольку мы с вами занимаемся наукой, то, как это принято при изучении любой новой дисциплины, нам необходимо обзавестись базовыми понятиями и ввести общепринятую терминологию. Для этого давайте снимем первый слой и вкратце поговорим про каждый отдельно взятый аспект.

Данные (Data, Dataset). Представьте, что вы хотите создать модель, которая будет играть в шахматы лучше любого гроссмейстера. В шахматах огромное количестве всевозможных ходов и показав машине лишь пару партий она вряд ли научиться побеждать. Ей, как и человеку необходимо разобрать огромное количество комбинаций, чтобы понимать, какие ходы ведут к выигрышу, а какие – к поражению.

Еще пример - представьте, что вы хотите стать экспертом по винам. Если вы попробуете только парочку бутылок, вам будет сложно отличить тонкие нотки и особенности различных сортов. Но если вы будете пробовать множество вин, со временем вы научитесь распознавать разницу между благородным вином урожая 1818 года и обычным столовым, купленным из ближайшей забегаловки.

И еще один пример - представьте, что вам нужно научить машину распознавать котов на фотографиях. Вы можете подумать: "Ну, у меня есть десяток снимков моего кота, этого должно в принципе хватить, верно?" К сожалению, это не так. Машины, в отличие от людей, не могут учиться на небольшом количестве примеров. Им нужно много данных, чтобы понять, что именно делает кота котом.

Данные можно собираться *вручную* – то есть вы можете бегать по улицам и фоткать всех дворовых котов. Сведения, собранные таким способом, получатся более качественными, ведь вы прекрасно понимаете кто является кошкой, а кто нет, но сам процесс сбора будет более трудоемким и затратным.

Но вы также можете попытаться *автоматизировать* этот процесс – подключиться к системе видеонаблюдения в вашем районе, и в местах, где часто можно увидеть кошек, например около мусорных контейнеров, делать снимки каждый раз при регистрации любого движения. Такой метод

обеспечит вам высокую скорость сбора данных и позволит сформировать большую базу фотографий, но качество их будет низким из-за большого количества ложных срабатываний – на снимках могут оказаться люди, выкидывающие мусор, птички, мышки и другие объекты.

Двигаемся дальше. Процесс присвоения меток данным – называется **разметка данных** (Data Labeling).

Разметка данных = разбиение на классы. Разметка, по сути, тоже самое что и классификация – это процесс присвоение имен множеству объектов, на основе отличительных характеристик. Метки (labels) могут обозначать, по сути, все что угодно, например настроение людей – счастливый, грустный, злой, наличие или отсутствие перелома на рентгеновском снимке, тип музыки – джас, рок, классика и так далее.

Например, если на изображении кошка, мы присваиваем метку «кошка». Если на изображении нет кошки, присваиваем метку «не кошка».

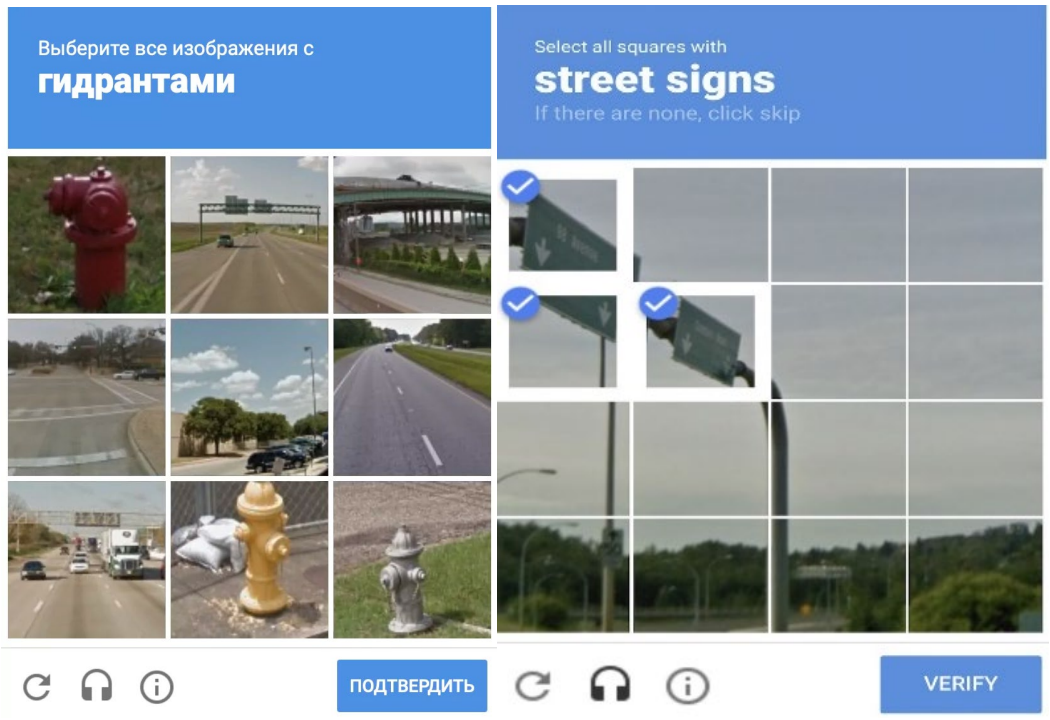
“КОШКА”

“НЕ КОШКА”

Однако процесс разбиения на самом деле не всегда так очевиден, как кажется. Например, к какому классу, вы бы отнесли вот эту картинку: музыкант, саксофон, джаз или силуэт человека.

Очевидно, что тут трудно дать однозначный ответ и нужно больше контекстных данных для правильной классификации.

Разметка данных является важным этапом в машинном обучении. Некоторые компании мучаются, собирая и обрабатывая данные вручную, некоторые просто сливает машине все что нашлось и верят в лучшее, а самые хитрые типа Google используют своих же пользователей для бесплатной разметки. Вспомните проверку ReCaptcha, которая иногда требует найти от вас на фотографии все дорожные знаки или гидранты — это она и есть.



Итак, мы разобрались, что для успешного машинного обучения нам необходимы большие объемы классифицированных данных.

Рискуя повториться, отметим, что для людей упорядочивание объектов по определенным категориям на основе их отличительных признаков является довольно простой задачей по сравнению с машинами. Именно поэтому в машинном обучении вводится отдельный термин, характеризующий уникальные свойства объекта - **фича** (features). Фичи – это признаки, которые помогают машине понять, на что обращать внимание при анализе данных.

Чтобы лучше запомнить, что такое фичи – представьте себе следующую ситуацию:

Раздался глухой удар грома, и гулкое эхо прокатилось по коридорам старого полицейского участка. Дождь безжалостно барабанил по окнам, словно пытаясь смыть накопившийся за месяцы кошмар. Детектив Сэм Белл сидел за своим столом, окруженный горами бумаг и фотографий с десятков мест преступлений. Он не спал уже третий день, отчаянно пытаясь найти хоть какую-то зацепку в череде жестоких убийств. Все его усилия были тщетны, найденные улики не складывались в единую картину. Казалось, преступник был призраком, оставлявшим за собой лишь следы ужаса.

Внезапно, вместе со вспышкой молнии, двери его кабинета распахнулись. На пороге стоял офицер, который поддерживал молодую девушку, дрожащую то ли от холода, то ли от леденящего страха. Промокшая до нитки, она судорожно поправляла волосы, прилипшие к ее лицу. Взгляд ее был парализован, казалось, что он застыл где-то на границе между отчаянием и надеждой.

Сэм медленно подошел к девушке и усадил ее на стул. В воздухе повисло напряжение, ощущаемое почти физически. После небольшой паузы она начала говорить. Её голос был тихим и едва удерживал остатки рассудка.

"Он... он пришёл за мной," - прошептала она, взгляд её метался по комнате, словно ища убежище от невидимого врага. "Я видела его глаза... такие пустые и холодные. Я думала, что смогу спрятаться, но он нашёл меня.

У него были ледяные голубые глаза и взгляд, пронизывающий насквозь. Нос был кривым, как после драки. На левой щеке виднелся глубокий шрам, как будто его кто-то ударил ножом. На его руках были татуировки, но они выглядели изуродованными, словно он пытался их удалить, оставляя страшные ожоги и рубцы".

Все эти детали образа преступника – и есть фичи. Собрав точный фоторобот мерзавца, Сэм наконец, после месяцев мучений и безрезультатных поисков, сумел распознать и найти серийного убийцу. Им как мы уже все знаем был дворецкий.

Выбор фич можно сравнить со сбором улик. Мы, как детективы, должны понимать, что не каждая из найденных зацепок приведет к раскрытию дела. Некоторые улики могут быть ложными следами, отвлекающими нас от истинного пути. И нам важно уметь выделять те фичи, которые действительно имеют значение, и игнорировать те, которые могут замедлить процесс.

Представьте, что у нас есть данные о студентах, и мы хотим предсказать их итоговую оценку по курсу. У нас есть следующие фичи:

1. Возраст студента
2. Пол студента
3. Количество пропущенных занятий
4. Количество выполненных домашних заданий
5. Время на самоподготовку

Все эти данные могут быть полезны, но важно понять, какие из них действительно влияют на успеваемость студента.

Представим, что мы передали в модель все фичи, включая незначимые признаки, такие как возраст и пол. Это приведет к тому, что модель будет обрабатывать больше информации, что замедлит её работу и снизит точность предсказаний.

Но если мы сконцентрируемся только на значимых признаках, таких как — количество пропущенных занятий, количество выполненных домашек и уделенному времени на самоподготовку — модель будет работать быстрее и

точнее, потому что она будет анализировать только те данные, которые действительно влияют на итоговую оценку.

Мы рассмотрели с вами случай, когда у нас в принципе не так много данных и достаточно понятные свойства. А что делать, если у нас гигабайты данных, и мы не знаем, как лучше описать их особенности? Мы упираемся в то, что нам как людям, видно и интуитивно понятно какие признаки отличают одни объекты от других, но формально описать их очень сложно. Например, у котов и собак есть такие общие черты, как ушки, хвост и лапки, но их внешний вид значительно различается. Поэтому зачастую отбор правильных фич занимает больше времени чем все остальное обучение.

Обучение (Learning) — это процесс анализа и обработки данных, включающий в себя: *выбор оптимальной модели*, ее *тренировку* и последующее *тестирование*.

Чтобы быть уверенными в том, что мы правильно понимаем друг друга, определим значения следующих терминов:

Не для кого не секрет что одну и туже задачу можно решить различными способами и от выбора метода будет зависеть как точность, так и скорость выполнения работы. Однако найти лучшее решении не так уж и просто, ведь поле возможностей – слишком огромно, а количество комбинаций –

зашкаливает. В этом контексте термин "лучший" становится для нас синонимом слова "оптимальный". *Оптимальный* – наилучший возможный вариант в рамках заданных условий.

Модель — это упрощенное, абстрактное представление реальных объектов или процессов, созданное для их анализа, понимания и оптимизации. Модели используются для предсказания и улучшения реальных ситуаций, превращая абстрактные образы в конкретные действия и результаты.

Для нас модель машинного обучения пока что представляет собой нечто вроде "черного ящика", в который мы закидываем данные, они там каким-то образом обрабатываются, и мы получаем предсказание. Магия, не иначе. И нашей последующей задачей как раз таки будет научиться этому волшебству. Мы начнем с простых заклинаний и со временем станем настоящими волшебниками в мире заколдованных алгоритмов и цифровых данных.

Предсказание (Prediction). Под предсказанием следует понимать не процесс, а конечный результат машинного обучения.

Представим себе следующую ситуацию. Вы работаете в компании, которая занимается доставкой товаров. Ваши клиенты постоянно

интересуются, когда именно будет доставлен их заказ. Понятное дело что указать точное время доставки невозможно, но зато возможно выполнить прогноз. Используя данные о прошлых заказах - время отправки, расстояние до клиента, текущие погодные условия, и загруженность дорог можно построить модель, которая будет предсказывать время доставки с высокой степенью точности.

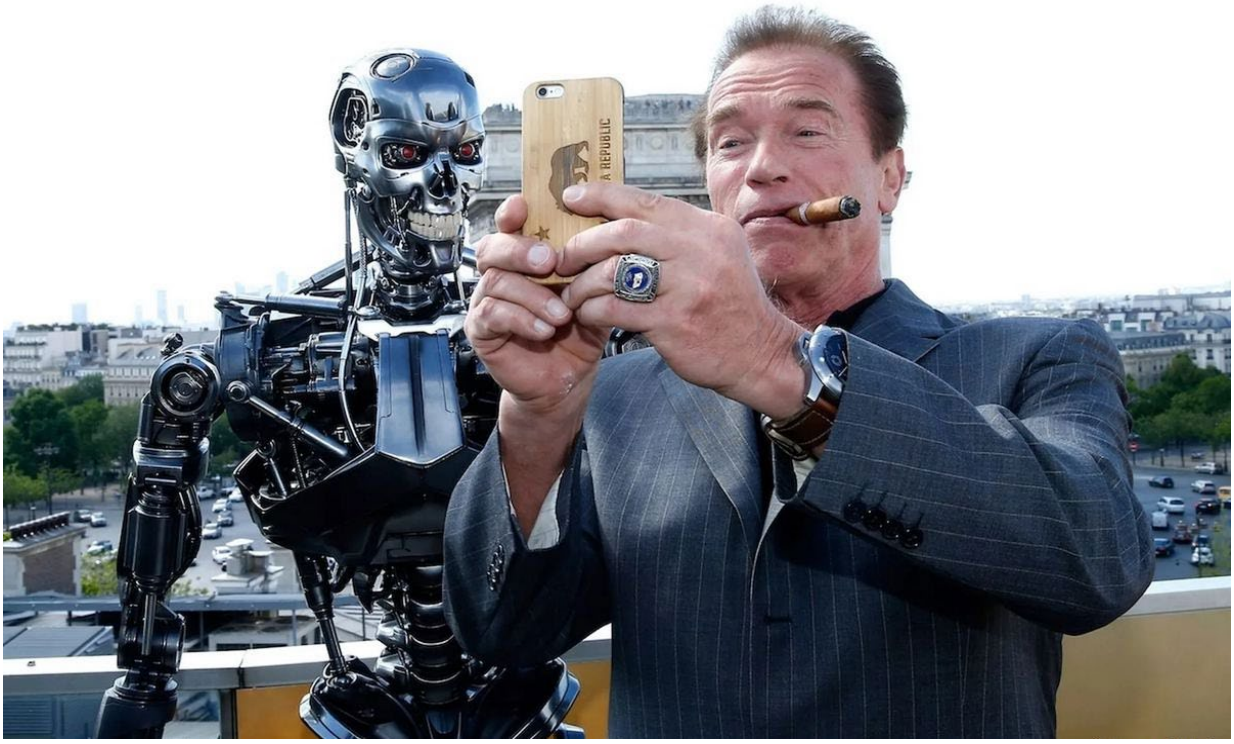
Еще пример. Вы не поверите, но машинное обучение очень широко используется в спорте. Если вы смотрели фильм "Человек, который изменил всё", то хорошо понимаете, о чем идет речь. Фильм основан на реальных событиях и рассказывает историю генерального менеджера команды "Окленд Атлетикс", который в рамках ограниченного бюджета, используя статистические данные о всех бейсболистах лиги, собрал успешную команду из недооцененных игроков и провел выдающийся сезон.

И еще один пример. Часто такое случается, что мы не знаем какой бы нам фильм посмотреть сегодня вечерком и тратим несколько часов на поиск, а не на просмотр киноленты. Мы ищем любимых актеров, читаем отзывы, выбираем жанры и, если повезет находим увлекательную картину, а если нет, то пытаемся спасти вечер купленными чипсиками или попкорном.

Сегодня многие киноплатформы смогли решить эту проблему используя машинное обучение. Сайты запоминают фильмы, которые вы смотрели, ваши

любимые жанры, оценки, которые вы ставили, и анализируя эту накопленную информацию о ваших предпочтениях, выдают персонализированные рекомендации.

Эти рекомендации не гарантируют стопроцентного попадания в ваш вкус, и не факт, что все предложенные фильмы подойдут вам, но они скорее всего помогут вам сузить круг поиска и найти что-то по-настоящему стоящее.



Некоторые результаты машинного обучения стали нам настолько привычны что мы их даже не замечаем – например, когда мы фотографируемся то камера телефона автоматически выделяет наши лица квадратиками или когда мы допускаем ошибки в тексте они начинают подсвечиваться красным цветом или автоматически исправляются.

Машины все больше и больше проникают в нашу жизнь и от этого процесса уже никуда не деться. Мы можем сопротивляться этим изменениям, считая их опасными, можем просто довольствоваться готовыми результатами, а можем принимать непосредственное участие и быть архитекторами будущего. Выбор за вами.

Фантастика и наука – две крайности одной и той же сущности. Вспомним, что нашей основной целью является создание искусственного интеллекта, давайте более конкретно постараемся себе представить, что это такое и чего мы, собственно говоря, хотим достичь. Не вдаваясь глубоко в понятие интеллект скажем: в пределе мы стремимся создать такую систему, которая была бы на уровень выше человеческого интеллекта. Достижение машинами уровня человеческих возможностей, является лишь промежуточным этапом, над которым, кстати говоря, мы активно трудимся прямо сейчас.

Обратите внимание на то, что мы постепенно воссоздаем свои собственные способности. Все то, чем мы занимаемся в рамках данной науки — это стараемся объяснить машинам то, как мы видим, слышим, говорим, чувствуем и так далее. Мы описываем процессы, которые характеризуют нас как людей. И по сути, сейчас мы занимаемся построением наших органов, но в цифровом виде. Мы на языке математики описываем их функциональность, их внутренние процессы, а затем стараемся объяснить это машинам. Искусственный интеллект – это анатомия XXI века.

Рассуждая аналогичным образом, становится ясно, что нейросети — это попытка воссоздать аналог человеческого мозга, разработка двигателей – это имитация человеческого сердца, а компьютерное зрение — работа глаз. И как это ни странно, все то, чем мы занимаемся опять-таки сводится к описанию природы, и, в частности, в контексте искусственного интеллекта к описанию природы человека.

Посмею предположить, что изучение инженерами биологии, зоологии и анатомии многократно ускорит процесс копирования природы ну или сказав более мягче – ускорит процесс создания чего-то нового.

Но когда, а главное, как мы поймем, что создали настоящий искусственный интеллект? Это достаточно спорный вопрос, на который пока нет однозначного ответа. Одни утверждают, что это просто невозможно, другие ссылаются на тест Тьюринга, а мне думается так:

В детстве я очень любил играть в мобильную игру под названием "Алхимия". В начале игры вам давали всего четыре элемента: огонь, воду, землю и воздух. Объединяя их друг с другом, вы получали новые объекты, которые далее могли комбинировать с другими, открывая всё больше и больше новых соединений. Например, смешивая воду и землю, получалась глина, обжигая эту глину огнем – получался кирпич. Так вот, я думаю, что наша жизнь во многом построена именно по такому принципу. Природа нам дает эти элементы и как бы невзначай показывает, что можно получить если совместить один элемент с другим. Нашей задачей остается понять, как это сделать (очень яркий пример такого мышления — это обнаружение новых химических соединений в условиях сверх аномальных давлений).

Мне видится, что отличительной характеристикой искусственного интеллекта будет скорость нахождения таких связок причем без необходимости подтверждения их на практике.

Давайте поясню, мы с вами играем в игру алхимия только в более масштабную версию, которую называется - "Жизнь". Мы играем в нее уже несколько тысяч лет и за это время насоздавали огромное количество различных объектов, комбинируя исходные элементы. Но для вселенских масштабов – это только начало игры. Представьте какое количество комбинаций еще не испробовано, и сколько всего еще можно наобъединять.

Мы создали объекты, опираясь на нашу логику, опыт и насмотренность. Однако порой инновационные решения оказываются неочевидными, и ответы приходят из совершенно неожиданных областей. Искусственный интеллект, видя всю поляну целиком, без лишних усилий сможет находить наилучшие из возможных решений.

Таким образом искусственный интеллект – это не просто инструмент в руках человека, а его новая способность познания мира. Не уверен насчёт эволюции, но то, что искусственный интеллект представляет собой революцию человеческого сознания - бесспорно.

ВОПРОСЫ ДЛЯ САМОКОНТРОЛЯ

1. Объясните своими словами, что значит обучить машину?
2. Какова основная концепция машинного обучения?
3. Как мы помогаем Google, когда отмечаем на картинках гидранты, дорожные знаки и мосты?
4. Какими способами можно собирать данные? Плюсы и минусы.
5. Что такое фича тремя словами? Приведите пример.
6. Приведите примеры результатов машинного обучения.
7. Почему наличие большого объема данных является важным аспектом машинного обучения?
8. Поиграйте в игру https://quickdraw.withgoogle.com/
9. Что такое модель?

ЗАДАЧИ НА ЛОГИКУ

Вы разрабатываете систему фильтрации назойливых спам-сообщений для электронной почты. Каждый день вам поступают тысячи сообщений, и ваша задача — обучить модель распознавать спам.

Какую разметку данных вы бы использовали для этой задачи?

Представьте, что у вас есть возможность создать любую модель машинного обучения, которая может решить глобальную проблему.

Опишите, какую глобальную проблему вы бы выбрали для решения и как ваша модель машинного обучения могла бы помочь в её решении.

ДОПОЛНИТЕЛЬНЫЙ МАТЕРИАЛ

Артем Оганов. "Запрещённая" химия, или как школьные двоечники оказались правы – https://www.youtube.com/watch?v=R0zwwbcWcNY