



**3D МОДЕЛИРОВАНИЕ И
ПРОГРАММИРОВАНИЕ**



приоритет2030⁺

лидерами становятся

ТЕМА:

Машинное обучение. Введение

КОНСПЕКТ ЛЕКЦИИ

Преподаватель:

**Шарипов Ильдар Курбангалиевич
к.т.н., доцент, заведующий кафедрой
электроснабжения и эксплуатации
электрооборудования.**





Тема 1. Машинное обучение. Введение.

Вопрос 1. Понятие машинного обучения, задачи и алгоритмы.

Вопрос 2. Сферы применения.

Вопрос 3. Проблемы и недостатки машинного обучения.



Вопрос 1. Понятие машинного обучения, задачи и алгоритмы.

Термин «машинное обучение» был впервые введён пионером в области компьютерных игр и искусственного интеллекта Артуром Самюэлем в 1952 году. Артур Сэмюэль занимался созданием программы для игры в шашки, и в ходе этой работы была создана программа «Checkers-playing», которая смогла «научиться» играть в шашки лучше, чем её создатель. Таким образом, программа, продемонстрировавшая способность к самообучению на основе своего предыдущего опыта, опровергла суждение о том, что компьютеры способны выполнять только строго заданные для них алгоритмы. Артур Сэмюэль определил машинное обучение, как «методы, позволяющие компьютерам учиться без непосредственного их программирования».

Более формальное определение машинного обучения дал американский учёный в области науки о данных Том Митчелл: «Говорят, что компьютерная программа обучается на основе опыта E по отношению к некоторому классу задач T и меры качества P , если качество решения задач из T , измеренное на основе P , улучшается с приобретением опыта E ».

Таким образом, машинное обучение представляет собой подраздел искусственного интеллекта, стоящий на стыке таких дисциплин, как математика, статистика, теория вероятностей, теория графов и изучающий алгоритмы, способные самостоятельно обучаться на основе опыта.

Машинное обучения позволяет решать ряд задач, самыми распространёнными из которых являются:

1. Классификация
2. Регрессия
3. Кластеризация
4. Фильтрация выбросов



Задача классификации является задачей обучения с учителем. Все вводимые данные разделяются на два и более классов, а задача обучающегося алгоритма состоит в том, чтобы определять новые вводимые данные в один из этих классов.

Главным отличием задачи регрессии от задачи классификации является то, что при регрессии выходные данные являются непрерывными, а не дискретными.

При прогнозировании оценок студента перед алгоритмом стоят именно эти задачи, так как на входе имеются данные о «ситуации», то есть наборе характеристик студента, и «решении», то есть оценке студента, при этом количество оценок может быть ограничено: например, разделение на два класса «зачёт» и «незачёт», 4 класса «отлично», «хорошо», «удовлетворительно» и «неудовлетворительно» или большее количество классов, а может быть и непрерывно для более глубокой детализации оценки.

Кластеризация является задачей обучения без учителя, целью которой является группировка объектов в кластеры на основании характеристик каждого объекта.

Задача фильтрации выбросов заключается в поиске аномальных измерений, выделяющихся на фоне всей выборки.

Существуют различные алгоритмы машинного обучения, на основе которых строится модель. Во многом выбор подходящего алгоритма зависит от характеристик набора данных, таких как объем, структура и качество. Также на выбор алгоритма влияет желаемый результат (двухклассовая или многоклассовая классификация, регрессия или фильтрация выбросов), требуемая точность предсказания и время, необходимое для обучения модели.

Вопрос 2. Сферы применения.



На данный момент технологии машинного обучения уже широко применяются в таких сферах, как:

- Оптимизация поиска в интернете
- Обнаружение спама
- Распознавание речи, жестов и образов
- Рекомендательные системы
- Персонализированный маркетинг
- Кредитный скоринг
- Выявление мошенничества
- Страхование
- Медицинская диагностика

Сферой применения машинного обучения, с которой практически каждый современный человек сталкивается на ежедневной основе, является оптимизация поиска в интернете. Пользователь заходит в поисковик и задаёт свой вопрос, задача же поисковой системы заключается в том, чтобы произвести ранжирование миллионов подходящих вариантов и выдать вначале те, которые наиболее актуальны для каждого конкретного человека. Алгоритм отслеживает действия пользователя: перешёл ли он по первой ссылке, как долго он задержался на этом сайте, перешёл ли к следующей странице поиска или ввёл новый запрос. Алгоритм может отслеживать не только действия одного конкретного человека, но и учитывать поведение пользователей со схожими запросами. Таким образом, поисковый алгоритм может улучшать качество своей работы.

Классическим примером решения задачи классификации является обнаружение спама в электронных письмах. Для каждого входящего письма алгоритм должен определить, является ли оно спамом или нет, то есть отнести объект к одному из двух классов. Для обучения алгоритм может



использовать различные характеристики письма: кем и откуда было отправлено письмо, длину письма, наличие в нём каких-либо специфических слов. При этом алгоритм может переобучаться с помощью человека, который отмечает спамом письма, пропущенные алгоритмом, или наоборот достаёт письма по ошибке попавшие в спам.

Распознавание речи уже давно широко используется в повседневной жизни: люди дают своим смартфонам различные голосовые команды, такие как создать напоминание, поставить будильник, написать сообщение или позвонить определённому человеку.

Одним из развивающихся трендов современного мира являются беспилотные автомобили - транспортные средства, оборудованные системой автоматического управления, которые могут передвигаться без помощи человека. Для того, чтобы автомобиль «видел», куда он едет, используются алгоритмы распознавания образов.

Ещё один пример применения методов машинного обучения, с которым многие люди сталкиваются в повседневной жизни, это персонализированный маркетинг. Чем больше компания знает о своих клиентах, тем эффективнее она может нацеливать рекламу, обеспечивая лучший сервис клиентам и большие прибыли себе. Компании могут регулировать, какие рекламные письма, скидки и купоны должны быть показаны потенциальному клиенту, учитывая его предыдущее поведение в интернете.

На этих же принципах основываются и рекомендательные системы. Эти системы могут определить, в какой ценовой категории необходимы товары, какие характеристики товара важны, а какие нет, покупает ли человек данную вещь для себя или в подарок. Некоторые системы будут обучаться только на информации, собранной с одного пользователя, другие же могут учитывать предпочтения людей со схожим поведением.



Банки начинают использовать методы машинного обучения для кредитного скоринга – оценки платежеспособности клиента банка. Опираясь на накопленный опыт, алгоритм может позволить банку оценить уровень кредитного риска заёмщика, выявить вероятность потенциального мошенничества, предсказать финансовое поведение (наличие или отсутствие отсрочек по платежам, заинтересованность в других услугах и продуктах банка), а также определить, какие действия более эффективно применять к каждому конкретному должнику.

Задача выявления финансового мошенничества решается с помощью метода фильтрации выбросов. Практически невозможно предсказать, как в данных будет выглядеть финансовое мошенничество, потому что вариантов очень много, а выборка, на которой должен обучаться алгоритм мала. Для решения этой задачи следует изучать нормальную активность, на фоне которой будут выделяться транзакции со значительными отличиями.

Технологии анализа больших данных уже успешно используется в некоторых областях страхования, например, на рынке страхования автомобилей. Страховые компании просят водителя установить в своей машине электронное устройство, которое собирает все данные и все параметры автомобиля. Это позволяет определить структуру вождения водителя, например, скорость его вождения, осуществляет ли он опасные маневры, как резко он трогается с места, время суток, в которое он чаще всего водит. Используя эти данные, компании могут назначать цену на страховое покрытие адекватную риску этого водителя. Страховые компании уже по-разному используют подобные данные, некоторые вводят систему скидок, если данные «говорят», что в течение какого-то отрезка времени водитель «вел себя хорошо». Другие компании определяют, будут ли они осуществлять страховую выплату в зависимости от того, нарушал ли водитель в момент аварии правила дорожного движения, например, превышал ли скорость.



Очевидно, что с помощью такого устройства, факт превышения скорости устанавливается очень легко. Находя различные поведенческие паттерны, страховые компании смогут определять, если люди имеющие определенные привычки поведения в интернете, выходящие в социальные сети в определенное время дня или выкладывающие определенные типы фотографий, имеют тенденцию к тому, чтобы жить меньше, чаще терять работу или иметь проблемы с законом.

Очень быстро развивающейся сферой применения машинного обучения является сфера медицинской диагностики.

Применение искусственного интеллекта может повысить скорость диагностирования заболевания, уменьшить количество ошибок при диагностировании, помочь назначить нужные анализы, тем самым оптимизировав расходы больницы.

Применение машинного обучения также позволяет увеличить производительность диагностирования, так как компьютер может хранить и анализировать гораздо больше данных, чем любой врач в своей голове. Машинное обучение уже используется для диагностирования некоторых форм рака и различных редких заболеваний и имеет высокий потенциал для использования в медицине.

Вопрос 3. Проблемы и недостатки машинного обучения.

Несмотря на все преимущества и возможности использования таких методов анализа, существуют некоторые проблемы и опасения, касающиеся моральной, этической и правовой использования результатов. Например, при диагностике редких заболеваний выборка, на которой происходит обучение алгоритма, является несбалансированной.



Это означает, что данных о людях, имеющих заболевание существенно меньше, чем о здоровых людях, что увеличивает вероятность неверного отнесения человека в класс «больных» или «здоровых».

Также алгоритмы могут выдавать неожиданные корреляции, которые повлияют на конечный результат. Например, машинное обучение используется для классификации людей по группам риска при страховании жизни. У различных национальностей и этнических групп есть свои отличительные особенности поведения в интернете.

Пример 1. При диагностике заболеваний по рентгеновским снимкам модель обучается и тестируется на данных, собранных с разных больниц. Рентгеновские аппараты в разных больницах выдают снимки с немного разной тональностью, при этом различия могут быть даже незаметны человеческим глазом. Модель может обучиться определять больницу на основе тональности снимка, а значит и вероятный диагноз (в разных больницах лежат люди с разными диагнозами), при этом вообще не анализируя изображение на снимке. Такая модель покажет отличную точность при тестировании, так как тестовые снимки взяты из той же выборки данных.

Пример 2. В задаче диагностики реальных и фейковых вакансий почти все фейковые вакансии относились к Европе. Обученная на этих данных модель может не считать подозрительными любые вакансии из других регионов. При применении на практике такая модель станет бесполезна, как только фейковые вакансии станут размещаться в других регионах. Конечно можно удалить регион из признаков, но тогда модель будет пытаться предсказать регион по другим признакам (скажем, по отдельным словам в описании и требованиях), и проблема сохранится.