

Тема 1. Введение в машинное обучение

План

1.1 Введение

1.2 Типы задач машинного обучения

1.3 Основные виды машинного обучения

1.4 Основные алгоритмы моделей машинного обучения

1.5 Примеры применения в реальной жизни

1.1 Введение

Благодаря машинному обучению программист не обязан писать инструкции, учитывающие все возможные проблемы и содержащие все решения. Вместо этого в отдельную программу закладывают алгоритм самостоятельного нахождения решений путём комплексного использования статистических данных, из которых выводятся закономерности и на основе которых делаются прогнозы.

Технология машинного обучения на основе анализа данных берёт начало в 1950 году, когда начали разрабатывать первые программы для игры в шашки. За прошедшие десятилетия общий принцип не изменился. Зато благодаря взрывному росту вычислительных мощностей компьютеров многократно усложнились закономерности и прогнозы, создаваемые ими, и расширился круг проблем и задач, решаемых с использованием машинного обучения.

Чтобы запустить процесс машинного обучения, для начала необходимо загрузить в компьютер Датасет (некоторое количество исходных данных), на которых алгоритм будет учиться обрабатывать запросы. Например, могут быть фотографии собак и кошек, на которых уже есть метки, обозначающие к кому они относятся. После процесса обучения, программа уже сама сможет распознавать собак и кошек на новых изображениях без содержания меток. Процесс обучения продолжается и после выданных прогнозов, чем больше данных мы проанализировали программой, тем более точно она распознает нужные изображения.

Благодаря машинному обучению компьютеры учатся распознавать на фотографиях и рисунках не только лица, но и пейзажи, предметы, текст и цифры. Что касается текста, то и здесь не обойтись без машинного обучения: функция проверки грамматики сейчас присутствует в любом текстовом редакторе и даже в телефонах. Причем учитывается не только написание слов, но и контекст, оттенки смысла и другие тонкие лингвистические аспекты. Более того, уже существует программное обеспечение, способное без участия человека писать новостные статьи (на тему экономики и, к примеру, спорта).

Машинное обучение часто рассматривается как часть сферы искусственного интеллекта. Исследования в области машинного обучения возникли на основе научных исследований в этой области, но в контексте приложения методов машинного обучения к науке о данных полезнее рассматривать машинное обучение как средство *создания моделей данных*.

Машинное обучение занимается построением математических моделей для

исследования данных. Задачи «обучения» начинаются с появлением у этих моделей *настраиваемых параметров*, которые можно приспособить для отражения наблюдаемых данных, таким образом, программа как бы обучается на данных. Как только эти модели обучатся на имеющихся данных наблюдений, их можно будет использовать для предсказания и понимания различных аспектов данных новых наблюдений.

1.2 Типы задач машинного обучения

Все задачи, решаемые с помощью ML, относятся к одной из следующих категорий.

1)Задача регрессии – прогноз на основе выборки объектов с различными признаками. На выходе должно получиться вещественное число (2, 35, 76.454 и др.), к примеру цена квартиры, стоимость ценной бумаги по прошествии полугода, ожидаемый доход магазина на следующий месяц, качество вина при слепом тестировании.

2)Задача классификации – получение категориального ответа на основе набора признаков. Имеет конечное количество ответов (как правило, в формате «да» или «нет»): есть ли на фотографии кот, является ли изображение человеческим лицом, болен ли пациент раком.

3)Задача кластеризации – распределение данных на группы: разделение всех клиентов мобильного оператора по уровню платёжеспособности, отнесение космических объектов к той или иной категории (планета, звезда, чёрная дыра и т. п.).

4)Задача уменьшения размерности – сведение большого числа признаков к меньшему (обычно 2–3) для удобства их последующей визуализации (например, сжатие данных).

5)Задача выявления аномалий – отделение аномалий от стандартных случаев. На первый взгляд она совпадает с задачей классификации, но есть одно существенное отличие: аномалии – явление редкое, и обучающих примеров, на которых можно натаскать машинно обучающуюся модель на выявление таких объектов, либо исчезающе мало, либо просто нет, поэтому методы классификации здесь не работают. На практике такой задачей является, например, выявление мошеннических действий с банковскими картами.

1.3 Основные виды машинного обучения

Основная масса задач, решаемых при помощи методов машинного обучения, относится к двум разным видам: обучение с учителем (supervised learning) либо без него (unsupervised learning). Однако этим учителем вовсе не обязательно является сам программист, который стоит над компьютером и контролирует каждое действие в программе. «Учитель» в терминах машинного обучения – это само вмешательство человека в процесс обработки информации.

В обоих видах обучения машине предоставляются исходные данные, которые ей предстоит проанализировать и найти закономерности. Различие лишь в том, что при обучении с учителем есть ряд гипотез, которые необходимо опровергнуть или подтвердить. Эту разницу легко понять на примерах.

Машинное обучение с учителем

Машинное обучение с учителем или *контролируемое обучение* (supervised learning) — включает моделирование признаков данных и соответствующих данным меток. После выбора модели ее можно использовать для присвоения меток новым, неизвестным ранее данным. Оно разделяется далее на *задачи классификации* и *задачи регрессии*. При классификации метки представляют собой дискретные категории, а при регрессии они являются непрерывными величинами.

Предположим, в нашем распоряжении оказались сведения о десяти тысячах московских квартир: площадь, этаж, район, наличие или отсутствие парковки у дома, расстояние от метро, цена квартиры и т. п. Нам необходимо создать модель, предсказывающую рыночную стоимость квартиры по её параметрам. Это идеальный пример машинного обучения с учителем: у нас есть исходные данные (количество квартир и их свойства, которые называются признаками) и готовый ответ по каждой из квартир – её стоимость. Программе предстоит решить задачу регрессии.

Ещё пример из практики: подтвердить или опровергнуть наличие рака у пациента, зная все его медицинские показатели. Выяснить, является ли входящее письмо спамом, проанализировав его текст. Это всё задачи на классификацию.

Машинное обучение без учителя

Машинное обучение без учителя (unsupervised learning) — включает моделирование признаков набора данных без каких-либо меток и описывается фразой «Пусть набор данных говорит сам за себя». Эти модели включают такие задачи, как *кластеризация* (clustering) и *понижение размерности* (dimensionality reduction). Алгоритмы кластеризации служат для выделения отдельных групп данных, в то время как алгоритмы понижения размерности предназначены для поиска более сжатых представлений данных.

В случае обучения без учителя, когда готовых «правильных ответов» системе не предоставлено, всё обстоит ещё интереснее. Например, у нас есть информация о весе и росте какого-то количества людей, и эти данные нужно

распределить по трём группам, для каждой из которых предстоит пошить рубашки подходящих размеров. Это задача кластеризации. В этом случае предстоит разделить все данные на 3 кластера (но, как правило, такого строгого и единственно возможного деления нет).

Если взять другую ситуацию, когда каждый из объектов в выборке обладает сотней различных признаков, то основной трудностью будет графическое отображение такой выборки. Поэтому количество признаков уменьшают до двух или трёх, и становится возможным визуализировать их на плоскости или в 3D. Это – задача уменьшения размерности.

Кроме того, существуют так называемые методы частичного обучения (semi-supervised learning), располагающиеся примерно посередине между машинным обучением с учителем и машинным обучением без учителя

1.4 Основные алгоритмы моделей машинного обучения

1. Дерево принятия решений

Это метод поддержки принятия решений, основанный на использовании древовидного графа: модели принятия решений, которая учитывает их потенциальные последствия (с расчётом вероятности наступления того или иного события), эффективность, ресурсозатратность.

Для бизнес-процессов это дерево складывается из минимального числа вопросов, предполагающих однозначный ответ — «да» или «нет». Последовательно дав ответы на все эти вопросы, мы приходим к правильному выбору. Методологические преимущества дерева принятия решений – в том, что оно структурирует и систематизирует проблему, а итоговое решение принимается на основе логических выводов.

2. Наивная байесовская классификация

Наивные байесовские классификаторы относятся к семейству простых вероятностных классификаторов и берут начало из теоремы Байеса, которая применительно к данному случаю рассматривает функции как независимые (это называется строгим, или наивным, предположением). На практике используется в следующих областях машинного обучения:

- определение спама, приходящего на электронную почту;
- автоматическая привязка новостных статей к тематическим рубрикам;
- выявление эмоциональной окраски текста;
- распознавание лиц и других паттернов на изображениях.

3. Метод наименьших квадратов

Всем, кто хоть немного изучал статистику, знакомо понятие линейной регрессии. К вариантам её реализации относятся и наименьшие квадраты. Обычно с помощью линейной регрессии решают задачи по подгонке прямой, которая проходит через множество точек. Вот как это делается с помощью метода наименьших квадратов: провести прямую, измерить расстояние от неё до каждой

из точек (точки и линию соединяют вертикальными отрезками), получившуюся сумму перенести наверх. В результате та кривая, в которой сумма расстояний будет наименьшей, и есть искомая (эта линия пройдет через точки с нормально распределённым отклонением от истинного значения).

Линейная функция обычно используется при подборе данных для машинного обучения, а метод наименьших квадратов – для сведения к минимуму погрешностей путем создания метрики ошибок.

4. Логистическая регрессия

Логистическая регрессия – это способ определения зависимости между переменными, одна из которых категориально зависима, а другие независимы. Для этого применяется логистическая функция (аккумулятивное логистическое распределение). Практическое значение логистической регрессии заключается в том, что она является мощным статистическим методом предсказания событий, который включает в себя одну или несколько независимых переменных. Это востребовано в следующих ситуациях:

- кредитный скоринг;
- замеры успешности проводимых рекламных кампаний;
- прогноз прибыли с определённого товара;
- оценка вероятности землетрясения в конкретную дату.

5. Метод опорных векторов (SVM)

Это целый набор алгоритмов, необходимых для решения задач на классификацию и регрессионный анализ. Исходя из того что объект, находящийся в N -мерном пространстве, относится к одному из двух классов, метод опорных векторов строит гиперплоскость с мерностью $(N - 1)$, чтобы все объекты оказались в одной из двух групп. На бумаге это можно изобразить так: есть точки двух разных видов, и их можно линейно разделить. Кроме сепарации точек, данный метод генерирует гиперплоскость таким образом, чтобы она была максимально удалена от самой близкой точки каждой группы.

SVM и его модификации помогают решать такие сложные задачи машинного обучения, как сплайсинг ДНК, определение пола человека по фотографии, вывод рекламных баннеров на сайты.

6. Метод ансамблей

Он базируется на алгоритмах машинного обучения, генерирующих множество классификаторов и разделяющих все объекты из вновь поступающих данных на основе их усреднения или итогов голосования. Изначально метод ансамблей был частным случаем байесовского усреднения, но затем усложнился и оброс дополнительными алгоритмами:

- бустинг (boosting) – преобразует слабые модели в сильные посредством формирования ансамбля классификаторов (с математической точки зрения это является улучшающим пересечением);
- бэггинг (bagging) – собирает усложнённые классификаторы, при этом параллельно обучая базовые (улучшающее объединение);
- корректирование ошибок выходного кодирования.

Метод ансамблей – более мощный инструмент по сравнению с отдельно стоящими моделями прогнозирования, поскольку:

- он сводит к минимуму влияние случайностей, усредняя ошибки каждого базового классификатора;
- уменьшает дисперсию, поскольку несколько разных моделей, исходящих из разных гипотез, имеют больше шансов прийти к правильному результату, чем одна отдельно взятая;
- исключает выход за рамки множества: если агрегированная гипотеза оказывается вне множества базовых гипотез, то на этапе формирования комбинированной гипотезы оно расширяется при помощи того или иного способа, и гипотеза уже входит в него.

7. Алгоритмы кластеризации

Кластеризация заключается в распределении множества объектов по категориям так, чтобы в каждой категории – кластере – оказались наиболее схожие между собой элементы.

Кластеризовать объекты можно по разным алгоритмам. Чаще всего используют следующие:

- на основе центра тяжести треугольника;
- на базе подключения;
- сокращения размерности;
- плотности (основанные на пространственной кластеризации);
- вероятностные;
- машинное обучение, в том числе нейронные сети.

Алгоритмы кластеризации используются в биологии (исследование взаимодействия генов в геноме, насчитывающем до нескольких тысяч элементов), социологии (обработка результатов социологических исследований методом Уорда, на выходе дающим кластеры с минимальной дисперсией и примерно одинакового размера) и информационных технологиях.

8. Метод главных компонент (PCA)

Метод главных компонент, или PCA, представляет собой статистическую операцию по ортогональному преобразованию, которая имеет своей целью перевод наблюдений за переменными, которые могут быть как-то взаимосвязаны между собой, в набор главных компонент – значений, которые линейно не коррелированы.

Практические задачи, в которых применяется PCA, – визуализация и большинство процедур сжатия, упрощения, минимизации данных для того, чтобы облегчить процесс обучения. Однако метод главных компонент не годится для ситуаций, когда исходные данные слабо упорядочены (то есть все компоненты метода характеризуются высокой дисперсией). Так что его применимость определяется тем, насколько хорошо изучена и описана предметная область.

9. Сингулярное разложение

В линейной алгебре сингулярное разложение, или SVD, определяется как разложение прямоугольной матрицы, состоящей из комплексных или вещественных чисел. Так, матрицу M размерностью $[m \times n]$ можно разложить таким образом, что $M = U\Sigma V$, где U и V будут унитарными матрицами, а Σ – диагональной.

Одним из частных случаев сингулярного разложения является метод главных компонент. Самые первые технологии компьютерного зрения разрабатывались на основе SVD и PCA и работали следующим образом: вначале лица (или другие паттерны, которые предстояло найти) представляли в виде суммы базисных компонент, затем уменьшали их размерность, после чего производили их сопоставление с изображениями из выборки. Современные алгоритмы сингулярного разложения в машинном обучении, конечно, значительно сложнее и изощрённее, чем их предшественники, но суть их в целом не изменилась.

10. Анализ независимых компонент (ICA)

Это один из статистических методов, который выявляет скрытые факторы, оказывающие влияние на случайные величины, сигналы и пр. ICA формирует порождающую модель для баз многофакторных данных. Переменные в модели содержат некоторые скрытые переменные, причем нет никакой информации о правилах их смешивания. Эти скрытые переменные являются независимыми компонентами выборки и считаются негауссовскими сигналами.

В отличие от анализа главных компонент, который связан с данным методом, анализ независимых компонент более эффективен, особенно в тех случаях, когда классические подходы оказываются бессильны. Он обнаруживает скрытые причины явлений и благодаря этому нашёл широкое применение в самых различных областях – от астрономии и медицины до распознавания речи, автоматического тестирования и анализа динамики финансовых показателей.

1.5 Примеры применения в реальной жизни

Пример 1. Диагностика заболеваний

Пациенты в данном случае являются объектами, а признаками – все наблюдающиеся у них симптомы, анамнез, результаты анализов, уже предпринятые лечебные меры (фактически вся история болезни, формализованная и разбитая на отдельные критерии). Некоторые признаки – пол,

наличие или отсутствие головной боли, кашля, сыпи и иные – рассматриваются как бинарные. Оценка тяжести состояния (крайне тяжёлое, средней тяжести и др.) является порядковым признаком, а многие другие – количественными: объём лекарственного препарата, уровень гемоглобина в крови, показатели артериального давления и пульса, возраст, вес. Собрав информацию о состоянии пациента, содержащую много таких признаков, можно загрузить её в компьютер и с помощью программы, способной к машинному обучению, решить следующие задачи:

- провести дифференциальную диагностику (определение вида заболевания);
- выбрать наиболее оптимальную стратегию лечения;
- спрогнозировать развитие болезни, её длительность и исход;
- просчитать риск возможных осложнений;
- выявить синдромы – наборы симптомов, сопутствующие данному заболеванию или нарушению.

Ни один врач не способен обработать весь массив информации по каждому пациенту мгновенно, обобщить большое количество других подобных историй болезни и сразу же выдать чёткий результат. Поэтому машинное обучение становится для врачей незаменимым помощником.

Пример 2. Поиск мест залегания полезных ископаемых

В роли признаков здесь выступают сведения, добытые при помощи геологической разведки: наличие на территории местности каких-либо пород (и это будет признаком бинарного типа), их физические и химические свойства (которые раскладываются на ряд количественных и качественных признаков).

Для обучающей выборки берутся 2 вида прецедентов: районы, где точно присутствуют месторождения полезных ископаемых, и районы с похожими характеристиками, где эти ископаемые не были обнаружены. Но добыча редких полезных ископаемых имеет свою специфику: во многих случаях количество признаков значительно превышает число объектов, и методы традиционной статистики плохо подходят для таких ситуаций. Поэтому при машинном обучении акцент делается на обнаружение закономерностей в уже собранном массиве данных. Для этого определяются небольшие и наиболее информативные совокупности признаков, которые максимально показательны для ответа на вопрос исследования – есть в указанной местности то или иное ископаемое или нет. Можно провести аналогию с медициной: у месторождений тоже можно выявить свои синдромы. Ценность применения машинного обучения в этой области заключается в том, что полученные результаты не только носят практический характер, но и представляют серьёзный научный интерес для геологов и геофизиков.

Пример 3. Оценка надёжности и платёжеспособности кандидатов на получение кредитов

С этой задачей ежедневно сталкиваются все банки, занимающиеся выдачей кредитов. Необходимость в автоматизации этого процесса назрела давно, ещё в 1960–1970-е годы, когда в США и других странах начался бум кредитных карт.

Лица, запрашивающие у банка заём, – это объекты, а вот признаки будут отличаться в зависимости от того, физическое это лицо или юридическое. Признаковое описание частного лица, претендующего на кредит, формируется на основе данных анкеты, которую оно заполняет. Затем анкета дополняется некоторыми другими сведениями о потенциальном клиенте, которые банк получает по своим каналам. Часть из них относятся к бинарным признакам (пол, наличие телефонного номера), другие — к порядковым (образование, должность), большинство же являются количественными (величина займа, общая сумма задолженностей по другим банкам, возраст, количество членов семьи, доход, трудовой стаж) или номинальными (имя, название фирмы-работодателя, профессия, адрес).

Для машинного обучения составляется выборка, в которую входят кредитополучатели, чья кредитная история известна. Все заёмщики делятся на классы, в простейшем случае их 2 – «хорошие» заёмщики и «плохие», и положительное решение о выдаче кредита принимается только в пользу «хороших».

Более сложный алгоритм машинного обучения, называемый кредитным скорингом, предусматривает начисление каждому заёмщику условных баллов за каждый признак, и решение о предоставлении кредита будет зависеть от суммы набранных баллов. Во время машинного обучения системы кредитного скоринга вначале назначают некоторое количество баллов каждому признаку, а затем определяют условия выдачи займа (срок, процентную ставку и остальные параметры, которые отражаются в кредитном договоре). Но существует также и другой алгоритм обучения системы – на основе прецедентов.

Резюме

Мы рассмотрели несколько простых примеров основных видов методов машинного обучения.

Мы рассмотрели:

- *обучение с учителем* — модели для предсказания меток на основе маркированных обучающих данных;
- *классификацию* — модели для предсказания меток из двух или более отдельных категорий;
- *регрессию* — модели для предсказания непрерывных меток;
- *обучение без учителя* — модели для распознавания структуры немаркированных данных;
- *кластеризацию* — модели для выявления и распознавания в данных отдельных групп;
- *понижение размерности* — модели для выявления и распознавания низкоразмерной структуры в высокоразмерных данных.

