Всем привет, поздравляю вас с началом третьего модуля. А конкретнее говоря, машинное обучение. Сегодняшняя тема занятия, основы машинного обучения

мы познакомимся с базовыми понятиями машинного обучения вы узнаете Как выглядят данные в машинном обучении и научитесь отличать регрессию от классификации

Что такое машинное обучение?

Машинное обучение находится на стыке математической статистики, методов оптимизации и классических математических дисциплин, но имеет также и собственную специфику, связанную с проблемами вычислительной эффективности и переобучения. Машинное обучение - не только математическая, но и практическая, инженерная дисциплина.

Общий термин «**Machine Learning**» или «**машинное обучение**» обозначает множество математических, статистических и вычислительных методов для разработки алгоритмов, способных решить задачу не прямым способом, а на основе поиска закономерностей в разнообразных входных данных.  Решение вычисляется не по четкой формуле, а по установленной зависимости результатов от конкретного набора признаков и их значений. Например, если каждый день в течении недели земля покрыта снегом и температура воздуха существенно ниже нуля, то вероятнее всего, наступила зима. Поэтому машинное обучение применяется для диагностики, прогнозирования, распознавания и принятия решений в различных прикладных сферах: от медицины до банковской деятельности.

ЦЕЛЬ МЛ

Цель Машинного обучения - предсказать результат до входным данным. Чем разнообразнее входные данные, тем проще машине найти закономерности и тем точнее результат. Если мы хотим обучить вычислительную машину, то нам нужны три вещи: данные, признаки, алгоритм.

Разберемся что подразумевается под данными

Данные

Если мы хотим определять письма как нежелательные или «спам», то нужно для обучения использовать обучающую выборку с примерами таких «спам-писем», если мы хотим предсказывать курс акций, то нужна история транзакций на фондовой бирже. Для качественного обучения необходимо как можно больше данных.

Хорошие наборы данных (датасеты), в которых скрываются закономерности (скрытые знания), стоят очень дорого, поэтому крупным компаниям проще раскрыть свои разработанные алгоритмы, а не передавать исходные данные.

Признаки

С английского языка features можно также перевести, как свойства, характеристики. Ими могут быть пробег автомобиля, пол пользователя, цена акций, счетчик частоты появления слова в тексте. Машина должна знать, какой признак конкретно использовать. Когда признаков много, модель работает медленно и неэффективно. Зачастую отбор правильных признаков занимает больше времени, чем все остальное обучение.

Алгоритм

Одну задачу можно решить разными методами. От выбора метода зависит точность, скорость работы и размер готовой модели. Но есть один нюанс: если данные плохие, даже самый лучший алгоритм не поможет.

Сегодня в Машинном обучении есть всего четыре основных направления разберемся вообще на какие виды у нас подразделяется машинное обучение: классическое обучение, обучение с подкреплением, Ансамбли, нейросети и глубокое обучение.

Обучение с учителем

Как следует из названия, для работы этого типа алгоритмов нам потребуется учитель, который и будет учить наш алгоритм. Так каким же образом происходит обучение алгоритма? Давайте представим процесс принятия решение в виде некоторого «черного ящика», который непонятно как работает внутри, для нас важен только результат его работы. Рассмотрим это на примере задачи распознавания образов, в которой нам нужно определить, относится некоторый объект к автомобилям или нет.

В этом случае мы выступаем в качестве эксперта, который формирует так называемое «обучающее множество», то есть набор примеров и правильных ответов с экспертной точки зрения. Далее мы передаем или, как говорят, «скармливаем» полученное множество с набором примеров, которые обозначим как Х, и решений, которые обозначим как y, в некий алгоритм, задачей которого является найти некоторую функцию f(X), преобразующую множество Х в множество y:

X -> f(X) -> y.

Такие пары множеств примеров и решений еще называют парами «объект-ответ», или «прецедентами». Процесс обучения с учителем также часто называют процессом обучения по прецедентам. Далее, используя найденную функцию, наш алгоритм пытается найти ответ для примера, которого не было в обучающем множестве.

Вот в двух словах то, как работает обучение по прецедентам.

Есть две основные задачи Машинного обучения с учителем: классификация (classification) и регрессия (regression).

Цель классификации состоит в том, чтобы спрогнозировать метку класса (целевую переменную) (class label), которая представляет собой выбор из заранее определенного списка возможных вариантов. Классификация иногда разделяется на бинарную классификацию (binary classification), которая является частным случаем разделения на два класса, и мультиклассовую классификацию (multiclass classification), когда в классификации участвует более двух классов. Бинарную классификацию можно представить как попытку ответить на поставленный вопрос в формате «да/нет».

Классификация электронных писем на «спам» и «не-спам» является примером бинарной классификации. В данной задаче бинарной классификации ответ «да/нет» дается на вопрос «является ли это электронное письмо спамом?»

Цель регрессии состоит в том, чтобы спрогнозировать непрерывное число или число с плавающей точкой (floating-point number), если использовать термины программирования, или вещественное число (real number), если говорить языком математических терминов. Прогнозирование годового дохода человека в зависимости от его образования, возраста и места жительства является примером регрессионной задачи. Прогнозируемое значение дохода представляет собой сумму (amount) и может быть любым числом в заданном диапазоне. Другой пример регрессионной задачи - прогнозирование объема урожая зерна на ферме в зависимости от таких атрибутов, как объем предыдущего урожая, погода, количество сотрудников, работающих на ферме. И снова объем урожая может быть любым числом.

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 ее 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 - ее стоимость. Программе предстоит решить задачу регрессии.

Довольно очевидно, что первые две линии не соответствуют нашим данным. Третья, похоже, лучше, чем две другие. Но как мы можем это проверить? Формально нам нужно выразить, насколько хорошо подходит линия, и мы можем это сделать, определив функцию потерь.

Функция потерь — метод наименьших квадратов

Функция потерь — это мера количества ошибок, которые наша линейная регрессия делает на наборе данных. Хотя есть разные функции потерь, все они вычисляют расстояние между предсказанным значением y(х) и его фактическим значением. Например, взяв строку из среднего примера выше, f(x)=−0.11⋅x+2.5, мы выделяем дистанцию ошибки между фактическими и прогнозируемыми значениями красными пунктирными линиями.

Одна очень распространенная функция потерь называется средней квадратичной ошибкой (MSE). Чтобы вычислить MSE, мы просто берем все значения ошибок, считаем их квадраты длин и усредняем.

Вычислим MSE для каждой из трех функций выше: первая функция дает MSE 0,17, вторая — 0,08, а третья — 0,02. Неудивительно, что третья функция имеет самую низкую MSE, подтверждая нашу догадку, что это линия наилучшего соответствия.

Еще пример из практики: подтвердить или опровергнуть наличие рака у пациента, зная все его медицинские показатели. Выяснить, является ли входящее письмо «спамом», проанализировав его текст.

В настоящее время разработано большое число алгоритмов обучения с учителем, каждый из которых имеет свои сильные и слабые стороны. Не существует единого алгоритма, который лучше всего подходит для всех задач анализа.

К числу алгоритмов обучения с учителем для решения задач классификации относятся:

﻿﻿деревья решений;

﻿﻿машины опорных векторов;

﻿﻿байесовский классификатор;

﻿﻿линейный дискриминантный анализ;

﻿﻿метод k-ближайших соседей.

Алгоритмами обучения с учителем для решения задачи регрессии являются:

﻿﻿линейная регрессия;

﻿﻿логистическая регрессия;

﻿﻿нейронные сети.

Это деление нестрогое, поскольку, например, нейронные сети могут быть адаптированы для классификации, а некоторые виды деревьев решений (к примеру, CART) позволяют производить численное предсказание.

Обучение без учителя

В этом случае, как следует из названия, алгоритмам приходится обучаться самостоятельно. Приведем пример. Если у вас есть своя страничка в социальной сети, то наверняка есть и какие-то «друзья», как-то с вами связанные. У этих друзей есть свои друзья и так далее.

Имея схему взаимосвязей и некоторую информацию по всем этим людям, возможно выявить какие-то общности, которые всех их связывают. Это могут быть, например, общие учебные заведения, место проживания, общие онлайн игры в которые эти люди играют, общие интересы и прочее. То есть мы можем выделить некоторые общности или группы/кластеры, о которых мы даже и не догадывались, соответственно, и не могли обучить их нахождению свой алгоритм.

Такая задача называется задачей кластеризации. В таком типе задач требуется обнаружить некоторые внутренние связи, зависимости, закономерности, которые существуют между исследуемыми объектами.

Подобные вопросы часто возникают в маркетинге, где требуется выделить какие-то сегменты рынков, целевые группы потребителей, в исследованиях астрономических данных, при организации компьютерных кластеров.

Основное применение обучения без учителя - построение моделей для кластеризации. Поскольку кластерная структура данных заранее неизвестна, а определяется в процессе обучения модели, использовать какие-либо целевые значения невозможно.

Кластерный анализ

В общих чертах дель кластеризации - найти разные группы в элементах данных. Для этого алгоритмы кластеризации находят структуру в данных так, что элементы одного кластера (или группы) больше похожи друг на друга, чем элементы из разных кластеров.

Визуально: представьте, что у нас есть набор данных фильмов и мы хотим их классифицировать. И мы можем их как-то поделить либо длительности, либо по жанру, если можно как-то описать наши фильмы

Эти неконтролируемые алгоритмы обучения имеют невероятно широкий спектр применений и весьма полезны для решения реальных задач, таких как группировка музыки, документов или фильмов и поиск клиентов с общими интересами на основе их покупок.

Вот некоторые из наиболее распространенных алгоритмов кластеризации:

﻿﻿к-средних;

﻿﻿иерархическая кластеризация;

- кластеризация сканирования на основе плотности (DBSCAN);

- гауссова модель кластеризации.

<https://habr.com/ru/company/ods/blog/325654/>

Обнаружение аномалий

Обнаружение аномалий – это обычно задача обучения без учителя, цель которой состоит в выявлении подозрительных наблюдений в данных. Ограничение состоит в стоимости неправильного обозначения нормальных точек как аномалий и невозможности найти фактические аномалии.

Поиск аномалий и выявление подозрительных операций широко применяется в клиентской аналитике, банковском аудите и других видах бизнес аналитики. Суть данной методики заключается в анализе больших объемов данных и выявлении поставщиков, клиентов, транзакций или иных активностей с крайне нетипичным поведением. Часто, такие аномалии являются индикатором мошенничества или поводом для более детального анализа подобных бизнес активностей.

Обучение с подкреплением

В обучении с подкреплением (reinforcement learning, RL) мы хотим построить алгоритм, моделирующий обучение методом проб и ошибок. Вместо получения обучающей выборки на вход такой алгоритм будет взаимодействовать с некоторой средой (environment), окружающим миром, а в роли «разметки» будет выступать награда (reward) — скалярная величина, которая выдаётся после каждого шага взаимодействия со средой и показывает, насколько хорошо алгоритм справляется с поставленной ему задачей. Например, если вы печёте тортики, то за каждый испечённый тортик вы получаете +1, а если вы пытаетесь кататься на велосипеде, то за каждое падение с велосипеда вам прилетает -1.

Награда не подсказывает, как именно нужно решать задачу и что вообще нужно делать;

Награда может быть отложенной во времени (вы нашли в пустыне сокровища, но чтобы получить заслуженные тортики, вам ещё понадобится куча времени, чтобы выбраться из пустыни; а награда приходит только за тортики) или сильно разреженной (большую часть времени давать агенту +0). Всё это сильно отличает задачу от обучения с учителем;

Награда предоставляет какой-то «сигнал» для обучения (хорошо/плохо), которого нет, например, в обучении без учителя.