

ML - машинное обучение

Большие данные и машинное обучение взаимосвязаны и часто работают в синергии, особенно в современных бизнес-приложениях. Допустим, есть задача: разослать маркетинговые предложения потенциальным клиентам для новой игрушки. Исходные данные для этой задачи могут быть огромными, включая информацию о миллионах пользователей.

В такой ситуации машинное обучение выступает как критически важный инструмент для анализа этих данных. Оно может помочь определить, какие клиенты наиболее вероятно заинтересуются предложением. Машинное обучение использует сложные алгоритмы для анализа больших данных и выявления закономерностей или трендов, которые человеку сложно или невозможно заметить вручную.

Однако успех машинного обучения зависит от качества исходных данных. Именно тут становится критически важной инженерия данных, которая включает в себя сбор, очистку и подготовку данных для дальнейшего анализа.

Большие данные предоставляют необходимое "сырье", а машинное обучение анализирует это сырье для получения ценных выводов и прогнозов. Эти две технологии вместе образуют мощную комбинацию, которая позволяет компаниям преобразовывать большие объемы данных в конкретные и ценные действия.

Основные этапы для работы с ML

Разработка модели машинного обучения — это итеративный процесс, который включает в себя несколько ключевых этапов, таких как предварительная обработка данных, обучение модели, её оценка и, наконец, применение для предсказаний. Каждый из этих этапов тесно связан с остальными, и изменения на одном этапе могут существенно повлиять на результаты других этапов.

Предварительная обработка данных

На этом начальном этапе аналитик данных проводит тщательную очистку и подготовку данных для дальнейшего анализа. Это включает в себя устранение пропущенных значений, обработку выбросов и, возможно, преобразование данных. Затем, данные разделяются на обучающий, валидационный и тестовый наборы. Пропорции разбиения могут варьироваться, но часто используется соотношение 70/10/20.

Выбор характеристик, или признаков, — это ещё одна важная задача на этом этапе. Характеристики — это атрибуты ваших данных, которые модель будет использовать для обучения. Они должны быть выбраны таким образом, чтобы максимально точно предсказывать целевую переменную (метку). Можно использовать методы сокращения размерности для выявления наиболее значимых характеристик.

Обучение модели

После подготовки данных следующим этапом является выбор алгоритма машинного обучения, который наилучшим образом подходит для решения вашей бизнес-задачи. На этом этапе, вы будете обучать модель на подготовленном обучающем наборе данных, экспериментируя с различными гиперпараметрами и техниками, чтобы добиться наилучшей производительности модели.

Оценка модели

После обучения модель нужно тщательно оценить. Обычно для этого используют валидационный набор данных, который был отложен на этапе предварительной обработки. Если модель не соответствует заранее определённым критериям точности, это значит, что нужно вернуться к предыдущим этапам и скорректировать как параметры модели, так и возможно, саму подготовку данных.

Прогнозирование

Как только модель прошла этап оценки и её производительность сочтена удовлетворительной, она может быть развернута в продакшн для работы с реальными данными. На этом этапе модель уже готова делать предсказания на новых данных. Это может быть как потоковая обработка данных в реальном времени, так и пакетная обработка.

Все эти этапы имеют тенденцию к взаимному влиянию. Например, изменение в методах предварительной обработки может потребовать изменения в алгоритме обучения, и наоборот. Поэтому, нахождение оптимального баланса между всеми этими факторами часто является результатом множества итераций и, в определенной степени, методом проб и ошибок.