Непрерывное машинное обучение (Continuous Machine Learning)

> Синонимы: Повторное обучение, CML, Retraining, Refitting
> Разделы: Бизнес-задачи

После того, как обучение ML-модели завершается, и устанавливается, что она приобрела достаточный уровень точности и обобщающей способности, принимается решение о ее передаче в промышленную эксплуатацию и производится развертывание у заказчика.

Однако тот факт, что ML-модель показывает высокую предсказательную эффективность на начальном этапе, вовсе не означает, что она сохранит ее на протяжении неопределенно долгого периода. Как правило, по мере увеличения времени, прошедшего после начала практического использования модели на реальных данных, ее предсказательная эффективность (точность и адекватность предсказаний, частота ошибок и т.д.) начинает постепенно снижаться.

Это явление известно как деградация моделей и вызвано изменениями в данных, происходящих под влиянием эволюции условий бизнес-окружения, в которых работает компания. Основными источниками таких изменений являются утечка и дрейф данных.

Если из-за указанных явлений предсказательная эффективность модели опускается ниже критического уровня, то ее дальнейшая эксплуатация становится нецелесообразной, поскольку неточные и неадекватные предсказания приводят к плохим управленческим решениям и потерям бизнеса. В этой ситуации возможны два решения:

1. **Полный вывод модели из эксплуатации и замена ее на новую**. Поскольку построение новой модели может оказаться достаточно длительным и затратным процессом, данное решение нежелательно. Однако если изменения в данных слишком велики и процесс деградации модели заходит слишком далеко, то этот выход оказывается безальтернативным.
2. **Адаптация модели к изменившимся распределениям и зависимостям в данных**. Для этого необходимо инициировать процесс повторного обучения модели с использованием, в качестве обучающих накопленные наборы данных, поступившие последними. Именно это и называется непрерывным машинным обучением, когда происходит перенастройка параметров модели для восстановления ее соответствия данным.

Иными словами, CML имеет место в том случае, когда процесс обучения модели не заканчивается на работе с обучающими данными, а продолжается в ходе ее практического использования.

Следует отметить, что в литературе процесс повторного обучения (retraining, refitting) ML-моделей иногда называют «переобучением». При этом не следует путать его с явлением переобучения, связанного с потерей обобщающей способности ML-моделей из-за слишком точной подгонки ее параметров к обучающим данным (overtraining, overfitting).

В зависимости от ситуации процесс непрерывного обучения может быть организован различными способами. По методу инициализации повторного обучения он может быть:

- **по запросу** — запускается по команде пользователя модели, когда тот посчитает, что в этом возникла необходимость;
- **автоматически** — запускается без участия пользователя в соответствии с некоторым регламентом, например, через заданные промежутки времени или по сигналу соответствующей системы мониторинга, встраиваемой в конвейер анализа данных, которая контролирует значимость изменения распределения данных или качества предсказаний модели.

Оба варианта имеют свои преимущества и недостатки. Вариант «по запросу», с одной стороны, позволяет избежать слишком частого повторного обучения, которое, может быть достаточно трудоемким и существенно нагружать информационную систему, где развернута модель.

С другой стороны, у пользователя модели возможно субъективное мнение о необходимости повторного обучения, или может оказаться недостаточно времени для отслеживания качества работы модели. Это потенциально приводит к «пропуску» ситуации, когда требуется повторное обучение.

Автоматический запуск в большинстве случаев является предпочтительным, поскольку позволяет избежать необходимости привлечения человека и связанного с этим влияния «человеческого фактора». С другой стороны, при автоматическом запуске возможны «ложные срабатывания», когда процесс переобучения может запускаться без необходимости.

При периодическом вызове, возможно, данные за промежуток времени не успеют достаточно измениться для того, чтобы переобучение имело смысл. А при запуске с помощью системы мониторинга возможны ложные сигналы тревоги. Кроме этого, регламент автоматического запуска может устареть (например интервалы запуска перестают соответствовать времени значимого изменения в данных), что так же требует его контроля.

Еще одной проблемой организации процесса CML является выбор формирования обучающих множеств для повторного обучения. Поскольку в общем случае характер (скорость, направление, значения) изменений в данных, к которым должна адаптироваться модель при повторном обучении, заранее неизвестны, возникает проблема неоднозначности выбора оптимального размера обучающего множества и параметров алгоритма повторного обучения.

Парадокс заключается в том, что при неудачном их выборе в результате повторного обучения работа модели может не только не улучшиться, но даже ухудшиться, т.е. произойдет своего рода «разобучение» модели. Поэтому приходится контролировать

вручную, что очень трудоемко, либо разрабатывать и интегрировать в процесс непрерывного обучения соответствующие алгоритмы, что также приводит к удорожанию модели и снижению скорости работы.

Остается открытым вопрос о формировании обучающего набора данных для CML. Это можно сделать следующими способами:

- **по размеру обучающего набора** — задается число наблюдений, используемых для переобучения, равное числу примеров, на которых модель обучалась впервые. Поскольку в процессе CML изменение структуры модели не предусматривается, то предположение о том, что если определенное число примеров дало хорошие результаты обучения один раз, то даст и в следующий, не лишено смысла. Недостатком подхода является то, что из-за неравномерности потока данных требуемое количество примеров может накапливаться слишком долго;
- **по размеру временного окна** — задается интервал времени, в течение которого происходит накопление наблюдений для повторного обучения перед очередным его запуском. Преимущество подхода в том, что глубину окна можно задать в соответствии с периодом актуальности данных (например, день, неделя, месяц). Это позволит избежать повторного обучения на устаревших данных. Недостаток в том, что в случае неравномерной плотности потока данных за период накопления может поступить число наблюдений недостаточное для качественного переобучения модели.

Таким образом проектирование и организация процесса непрерывного обучения могут оказаться достаточно сложными и затратными мероприятиями с неоднозначными результатами. Поэтому, принимая решение о внедрении CML в рабочий процесс анализа данных, необходимо сначала оценить связанные с этим все возможные выгоды и дополнительные затраты.