

Эталонные данные (Ground truth)

Синонимы: Основная истина, Фундаментальная истина

Разделы: [Источники данных](#)

Данные, которые извлекаются из различных источников и используются для машинного обучения, как правило отражают объективную реальность лишь с некоторой точностью и достоверностью. Именно поэтому модели машинного обучения (ML-модели) и оказываются столь полезными — они позволяют извлекать из данных зависимости и закономерности и обобщать их на новые наблюдения даже в тех случаях, когда более строгие численные и статистические методы оказываются бесполезными.

Поэтому при построении ML-моделей интереснее использовать данные, которые всегда содержат «правильный ответ». Именно они и называются эталонными. Обучающие примеры из наборов эталонных данных всегда содержат целевое значение (метку класса в классификации или число в регрессии), которое выдала бы модель, если бы работала идеально.

Получить эталонные данные из реальных источников, содержимое которых генерируется непосредственно бизнес-процессами, практически невозможно. Это связано с тем, что на данные всегда влияют случайные факторы, искажающие зависимости и закономерности, поиск которых и является целью анализа данных. Поэтому формирование эталонных данных как правило производится с помощью экспертов. При этом возможны два варианта:

- **размечивание** (маркирование) — когда имеется набор данных, наблюдения которого не содержат целевого значения. Его проставляет эксперт на основании своих знаний, опыта или расчета;
- **аннотирование** — когда целевые значения известны, эксперту остается только обозначить «метками достоверности» наблюдения, для которых они корректны.

Основным признаком размеченного или аннотированного набора данных является то, что гипотетическая идеальная модель показала бы на них нулевую выходную ошибку. Хотя в реальности это, конечно, вряд ли возможно, поскольку оценки экспертов тоже носят элемент субъективности и могут содержать ошибки. Поэтому эталонные данные скорее декларируются как достоверные, чем являются таковыми на самом деле.

Очевидным недостатком ручной подготовки эталонных данных является ее трудоемкость и длительность.

Рассмотрим пример. Пусть требуется построить модель, которая предсказывает, какие клиенты приобретут новый продукт компании в течение 7 дней. Данные, которые можно будет считать эталонными, появятся только через 7 дней, когда станет известно, кто из клиентов фактически купил продукт, а кто нет. Но если строить модель через 7 дней, то смысла в ней уже не будет, т.к. события, которые она должна предсказывать, уже произойдут.

Поэтому эксперт (или группа экспертов) возьмет клиентскую базу компании, произведет ее разметку вручную, после чего можно будет обучить соответствующую модель. По прошествии 7 дней, когда станет известно, кто из клиентов фактически купил, можно будет провести аннотирование — пометить как эталонные примеры, в которых предсказание модели совпало с фактическими событиями.

Применение эталонных данных зависит от особенности их получения. Так, по скорости их получения можно выделить три случая:

- **в реальном времени** — эталонные наблюдения становятся доступными сразу после своего появления. Это характерно для аналитических систем, работающих совместно с платформами электронной коммерции: когда клиент заходит на платформу и производит/не производит покупку, это становится известно мгновенно;
- **отсроченное** — когда эталонные данные становятся доступными с задержкой. Это наиболее распространенный случай;
- **эталонные данные недоступны** — эталонные данные не могут быть получены.

Можно выделить следующие проблемы формирования наборов эталонных данных:

- **сбор достаточного количества эталонных данных** — заранее неизвестно, сколько эталонных наблюдений потребуется для качественного решения и сколько удастся собрать;
- **обеспечение репрезентативности** — эталонные наблюдения могут оказаться сосредоточены только в отдельной области пространства признаков, поэтому их набор не будет достаточно репрезентативным;
- **высокие затраты** — необходимость маркирования или аннотирования большого числа наблюдений вручную представляет собой длительную и дорогостоящую процедуру;
- **ограниченность во времени** — сбор эталонных данных может занимать месяцы, за это время они могут утратить свою актуальность;
- **универсальность** — собранные эталонные данные могут подходить к любым видам ML-моделей и алгоритмам обучения или только к ограниченному их набору.

Применение наборов эталонных данных для повышения качества ML-моделей может быть разнообразным. Наиболее очевидное применение для тестирования и валидации моделей: результаты модели на обучающих данных сравниваются с целевыми значениями в эталонных, при этом чем больше совпадений, тем модель считается лучше.

Другим вариантом применения эталонных данных может быть их совместное использование с обычными данными при обучении моделей. При этом алгоритм обучения может быть настроен так, что вес обучающих примеров, помеченных как эталонные, будет выше, чем остальных при подстройке параметров модели.

Также интерес представляет подход, когда сначала модель обучается на эталонных данных, а затем постепенно в процесс вводятся реальные примеры в качестве своего рода шума. Это связано с тем, что зависимости между входными и выходными переменными, которые представляют эталонные данные, могут быть функциональными, что приводит к переобучению модели. В то же время ввод реальных данных позволяет повысить обобщающую способность.

Необходимо отметить, что хотя применение эталонных данных и способно принести значимую выгоду, подходить к их использованию следует осмотрительно, поскольку их **достоверность декларируется, а не удостоверяется**, и они тоже могут содержать ошибки.

В зарубежной литературе для обозначения понятия «эталонные данные» чаще всего используется термин «ground truth», что можно перевести как «наземная правда» или «базовая истина». Термин заимствован из метеорологии, картографии и дистанционного зондирования Земли, где наземными называют данные объективного контроля земной поверхности, которые затем используются для моделирования в соответствующей предметной области.