Выборка (Sample)

> Синонимы: Выборочная совокупность

В статистике и анализе данных выборка — это подмножество наблюдений генеральной совокупности, отобранных с целью изучения и анализа с помощью специальной процедуры (которая также называется выборкой), чтобы впоследствии обобщить полученные знания на всю совокупность. Выборки должны обладать свойством репрезентативности.

Причины, по которым анализируют выборки, а не всю совокупность, могут быть следующими:

- объем генеральной совокупности может быть очень велик, а ее анализ сложен в вычислительном плане (особенно, если нельзя использовать масштабируемые алгоритмы);
- получить доступ ко всем элементам совокупности очень сложно, или вообще невозможно (например, опросить население всего города — кто-то уехал, кто-то просто отвечать не хочет, поэтому проводят выборочный опрос);
- при использовании методов машинного обучения требуется использовать несколько множеств: обучающее, тестовое и валидационное, которые тоже являются выборками из исходного набора данных.

Выборки бывают:

- Смещенные и не смещенные. Смещенными называются выборки, структурные характеристики которых (среднее, математическое ожидание, дисперсия, среднеквадратическое отклонение) значимо отличаются (смещены) от соответствующих структурных характеристик совокупности. Значимость отличия проверяется специальными статистическими критериями (например, F-критерий Фишера). Использование смещенных выборок для исследования совокупности не имеет смысла. Чтобы получить несмещенную выборку нужно правильно выбрать алгоритм ее формирования.
- Случайными и детерминированными. В первом случае генерируется множество случайных значений и из совокупности извлекаются записи с соответствующими номерами. В детерминированной выборке извлекают сплошную последовательность наблюдений между заданными номерами, или удовлетворяющих некоторому условию (например, все клиенты с доходом больше 50 000 рублей). На практике чаще используют случайную выборку, поскольку она более соответствует вероятностному характеру большинства аналитических

моделей. Кроме этого используя детерминированный подход выше вероятность получить смещенную выборку.

- Сплошные и стратифицированные. В сплошной выборке наблюдения могут извлекаться из любой области генеральной совокупности. В стратифицированной выборке сначала делят совокупность на слои (называемые стратами) по какому-либо признаку, а затем производят выборку из каждого слоя независимо.
- С возвратом и без возврата. При выборке с возвратом, извлеченные наблюдения остаются в генеральной совокупности доступными для повторного выбора (в этом случае в выборке могут оказаться одинаковые наблюдения — дубликаты). В противном случае любое наблюдение может быть извлечено из совокупности только один раз.
- Зависимые и независимые. Если каждому наблюдению из одной выборки соответствует одно и только одно наблюдение из другой, то такие выборки называются зависимыми. Если это условие не выполняется, то выборки независимы. Очевидно, что зависимые выборки всегда должны иметь одинаковый объем, а для независимых это не обязательно.

К выборкам, используемым в машинном обучении, могут предъявляться дополнительные требования. Например, для обучения нейронных сетей требуется, чтобы число обучающих примеров было как минимум в два-три раза больше, чем число весов сети. При обучении классификаторов число примеров выборки должно быть много больше числа классов.

Если объем исходной совокупности недостаточен для формирования обучающей выборки требуемого объема, то применяются специальные методы отбора (например, перекрестная проверка).

В Loginom существует специализированный обработчик, который осуществляет отбор записей в выборку из исходного набора данных различными способами — Сэмплинг. Процесс отбора единиц наблюдения из генеральной совокупности с целью формирования выборки описан в статье «Методы и алгоритмы сэмплинга в анализе данных».