Сэмплинг (Sampling)

> Синонимы: Отбор
> Разделы: Бизнес-задачи, Алгоритмы

Сэмплинг — это процесс отбора из исходного набора данных выборки наблюдений, представляющей интерес для анализа. При реализации сэмплинга используются специальные методы отбора, которые должны обеспечить репрезентативность выборки с точки зрения решаемой аналитической задачи.

На практике в бизнес-аналитике применяются выборочные методы. Это обусловлено различными обстоятельствами, в том числе:

- Снижение трудоемкости алгоритмов анализа. При анализе сравнительно небольшого подмножества данных временные и вычислительных затраты значительно сокращаются.
- Коррекция распределений значений в выборке. В некоторых случаях исходное распределение значений факторов в наборе данных может негативно сказываться на процессе обучения модели. Типичный пример — несбалансированность классов в задаче кредитного скоринга. Коррекция распределений может заключаться, например, в увеличении числа объектов с требуемыми характеристиками (oversampling) или их сокращении (undersampling).

Различают следующие виды сэмплинга:

1. Случайный — выборка производится случайным образом из всей совокупности.
2. Равномерный — все наблюдения исходной совокупности разделяются на группы, в каждой из которых содержится их одинаковое число. Затем из каждой группы случайным образом выбирается одно наблюдение и помещается в результирующую выборку. Выборка, полученная в результате сэмплинга, будет состоять из наблюдений, случайным образом отобранных из каждой группы.
3. Стратификационный — применяется если исходная совокупность существенно неоднородна и случайный сэмплинг работает плохо. Тогда лучших результатов удается добиться, если разбить выборку на группы, и производить отбор наблюдений независимо от других групп. Выполняется в два этапа:
    - стратификация — группировка элементов исходной совокупности в относительно однородные подгруппы, которые называются стратами или слоями.
    - случайный отбор — случайная выборка из каждого слоя по отдельности.
4. Последовательный — наблюдения извлекаются по порядку от начала исходной совокупности к ее концу, и помещаются в выборку в том же порядке. Данный метод

имеет смысл использовать, если наблюдения в генеральной совокупности определенным образом упорядочены и их последовательность имеет значение с точки зрения анализа (например, временной ряд).

5. Со смещением — используется в ситуации, когда важные с точки зрения решаемой задачи данные представлены очень небольшим числом наблюдений, что не позволяет выполнить их достоверный анализ. В таких случаях применяется отбор со смещением — в выборку вносятся некоторые смещения значений признаков, делающих ее более репрезентативной. Например, изменяется баланс классов или значениям признаков устанавливаются некоторые веса.

В Loginom сэмплинг реализован как отдельный обработчик, который осуществляет отбор записей в выборку из исходного набора данных различными способами.

Больше алгоритмов и методов сэмплинга, правильный выбор и использование которых позволит сформировать выборки при решении конкретных задач, описано в статье «Методы и алгоритмы сэмплинга в анализе данных». А подробнее о методах балансировки классов можно узнать в статье «Сэмплинг в условиях несбалансированности классов».