Дополнительные данные

Цель

Разобрать метод генерирования синтетических точек данных SMOTE.

Генерирование новых данных

Ранее вы узнали о методе upsampling для выравнивания пропорции классов таргета.

Но его проблема в том, что данные просто копируются, и модель не получает никакой новой информации.

Генерирование новых данных

Вместо простого копирования точек данных можно попробовать сгенерированные данные, которые будут похожи на имеющиеся данные, но не будут совсем идентичными, чтобы модель машинного обучения получила какую-то новую информацию.

На этой идее и основан метод SMOTE. Давайте рассмотрим его пошаговый алгоритм.

Генерирование новых данных

Представим, что вы решаете задачу классификации следующего датасета, и хотим увеличить количество семплов минорного класса:

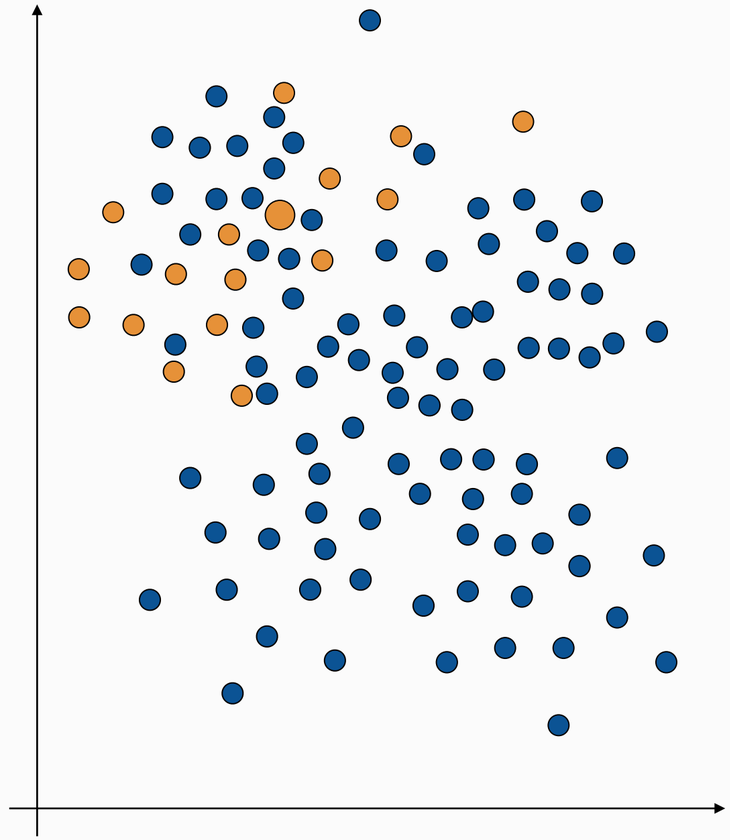

Схема: Андрей Мещеряков

Генерирование новых данных

Сначала выбирается случайная точка того класса, экземпляры которого нужно сгенерировать.

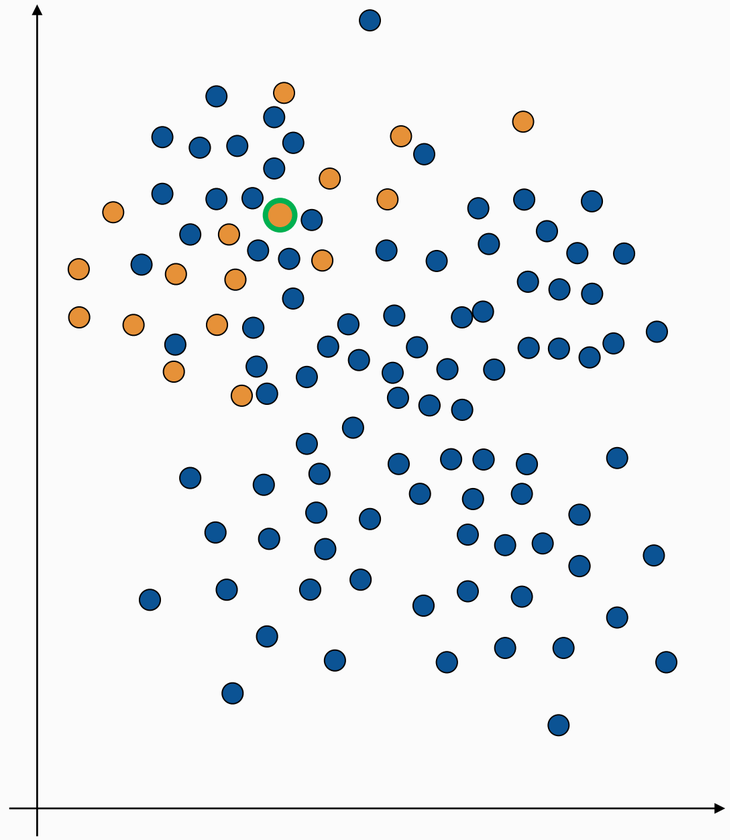

Схема: Андрей Мещеряков

Генерирование новых данных

Затем алгоритм ищет несколько ближайших соседей этой точки, принадлежащих тому же классу.

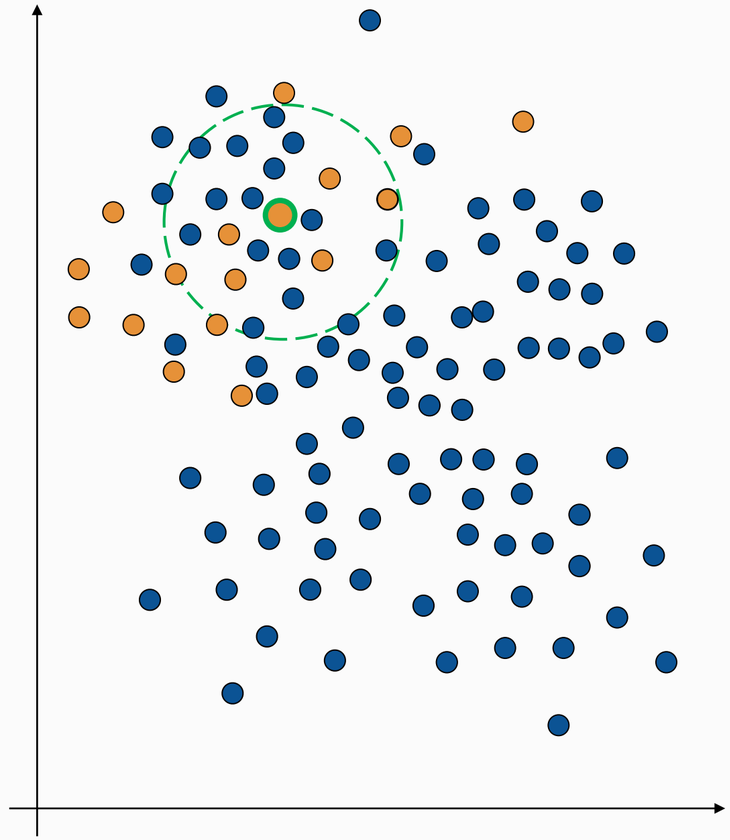

Схема: Андрей Мещеряков

Генерирование новых данных

Затем среди этих ближайших соседей
выбирается одна случайная точка,
и между двумя выбранными точками
проводится отрезок.

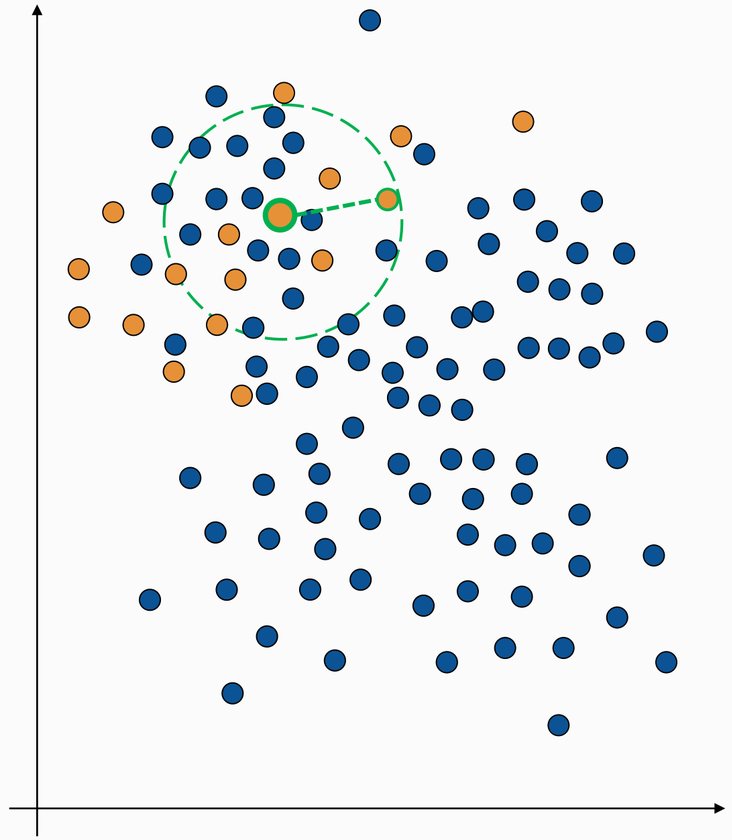

Схема: Андрей Мещеряков

Генерирование новых данных

В случайном месте этого отрезка размещается новая точка целевого класса.

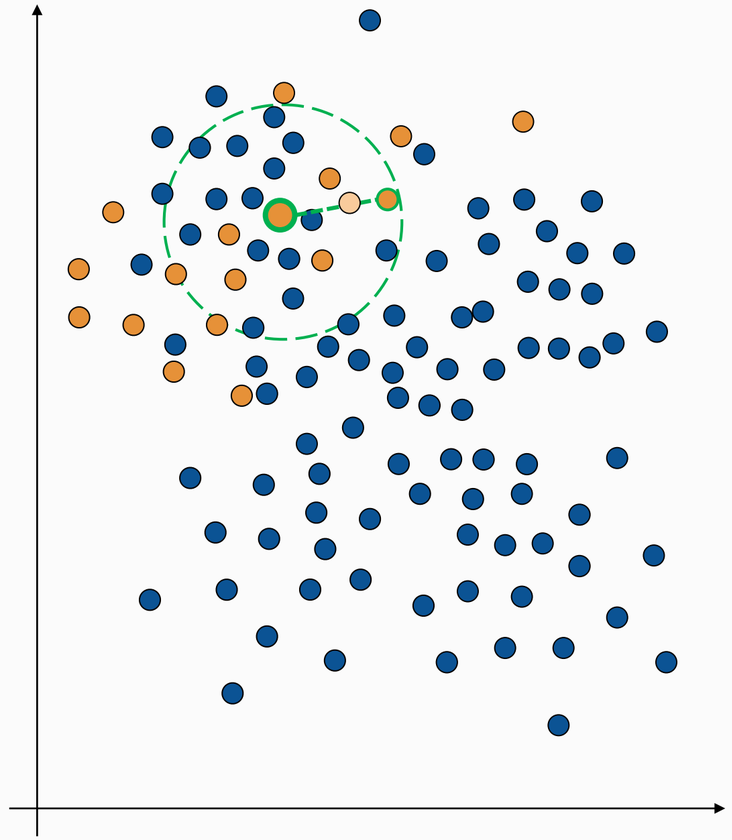

Схема: Андрей Мещеряков

Вывод

Вы разобрали метод SMOTE, который позволяет генерировать новые точки данных на основе информации об уже имеющихся данных.