Class Weight

Цель

Разобрать конфигурирование моделей
машинного обучения для работы
с несбалансированными выборками.

Сложности балансирования данных

Однако мы не всегда можем применять методы балансирования данных:

- downsampling, или генерирование новых данных, меняет распределение обучающих данных, поэтому ваши оценки качества таких моделей будут неправильными
- иногда из-за природы данных вы не сможете синтетически сгенерировать новые данные или выполнить downsampling

Сложности балансирования данных

В некоторых задачах необходима настройка процесса обучения моделей.

Перед запуском обучения разным значениям класса таргета присваиваются **веса классов**, которые говорят о важности экземпляров того или иного класса.

Эти веса классов затем используются в функциях потерь моделей машинного обучения. Всё станет понятнее на примере.

Веса классов

Предположим, что вы решаете задачу бинарной классификации со степенью дисбаланса 1:1000.

Веса классов

Предположим, что вы решаете задачу бинарной классификации со степенью дисбаланса 1:1000.

В качестве функции потерь, скорее всего, используется обычная бинарная кросс-энтропия:

$$L = -y\log(p) - (1-y)\log(1-p)$$

Веса классов

Предположим, что вы решаете задачу бинарной классификации со степенью дисбаланса 1:1000.

В качестве функции потерь, скорее всего, используется обычная бинарная кросс-энтропия:

$$L = -y\log(p) - (1-y)\log(1-p)$$

Пусть w1 — это вес положительного класса, а w0 — вес отрицательного класса, тогда функция потерь немного изменится:

$$L = -w1 \times y\log(p) - w0 \times (1-y)\log(1-p)$$

Веса классов

Пусть w1 — это вес положительного класса, а w0 — вес отрицательного класса, тогда функция потерь немного изменится:

$$L = -w1 \times y\log(p) - w0 \times (1-y)\log(1-p)$$

За неправильную классификацию точки данных положительного класса модель будет штрафоваться намного сильнее, чем за неправильную классификацию точки отрицательного класса.

А поскольку цель процесса обучения — минимизировать штраф модели, то в первую очередь модель будет стремиться правильно классифицировать объекты класса с наибольшим весом.

Вывод

Узнали о конфигурировании процесса обучения моделей на несбалансированных данных с помощью весов классов.