Несбалансированные выборки

Skillbox

Елена Кантонистова

Несбалансированные выборки

- Выборки в задачах классификации, в которых объектов одного класса сильно больше, чем объектов другого класса, называются несбалансированными

Примеры:

- задачи медицинской диагностики
- задачи поиска мошенников

Несбалансированные выборки

При дисбалансе классов некоторые модели могут давать плохое качество, потому что:

1. В обучающей выборке мало объектов редкого класса, и модели не хватает данных, чтобы хорошо обучиться находить этот класс

Несбалансированные выборки

При дисбалансе классов некоторые модели могут давать плохое качество, потому что:

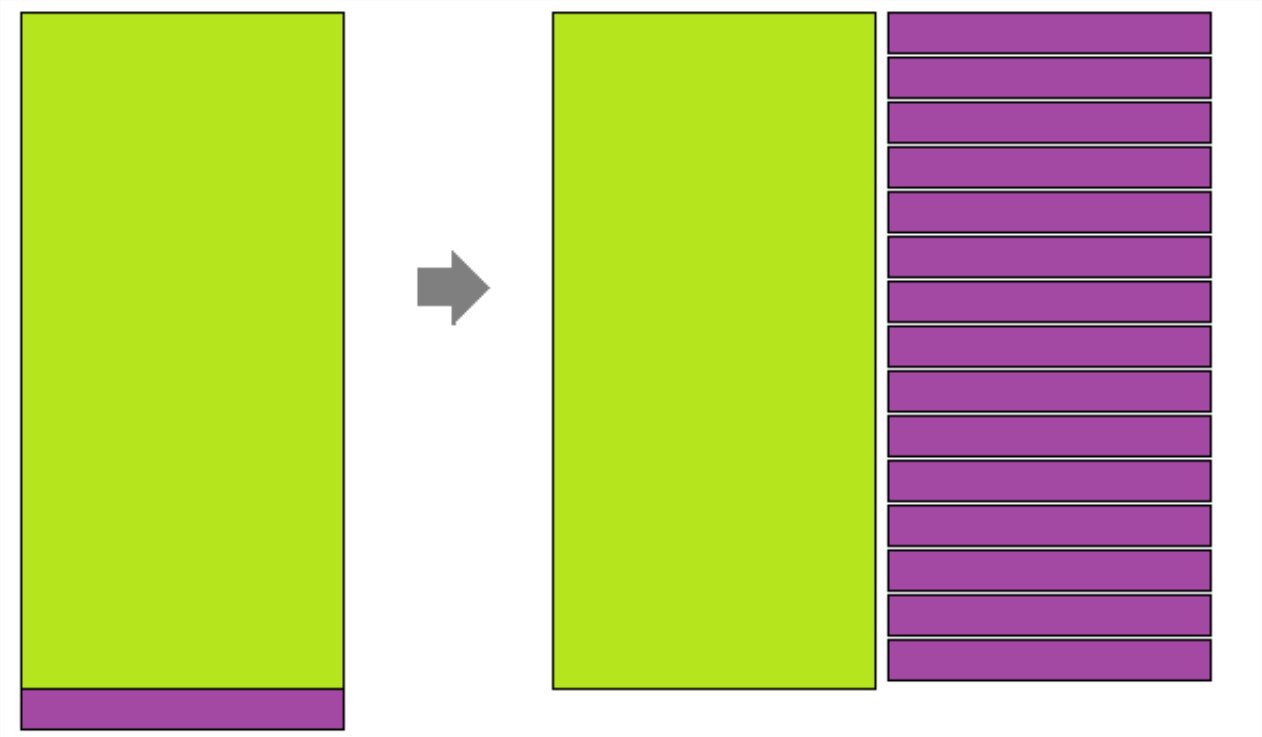
2. Запишем функцию потерь

$$Q(w) = \frac{1}{l}\sum_{i=1}^{l} L(\hat{y}_i, y_i) \to \min_w$$

где $L(\hat{y}_i, y_i)$ — ошибка на i-м объекте.
Слагаемых редкого класса мало, поэтому модель концентрируется на хороших предсказаниях на объектах большего класса при минимизации.

Решение: перебалансировка

- Перебалансировка выборки — добавление в выборку некоторого количества дубликатов объектов меньшего класса для достижения баланса классов



Решение: перебалансировка

- Альтернативным вариантом решения проблемы является добавление весов в функцию потерь таким образом (или перевзвешивание), чтобы ошибки на объектах меньшего класса весили больше, чем ошибки на объектах большего класса

Решение: перебалансировка

Пример:

- пусть объектов меньшего класса в 9 раз меньше, чем объектов большего класса
- тогда веса в функции потерь у объектов меньшего класса будут в 9 раз больше весов объектов старшего класса:

$$Q(w) = 1 \cdot \sum_{big\ class} L(\hat{y}_i, y_i) + 9 \cdot \sum_{small\ class} L(\hat{y}_i, y_i)$$

Тогда при оптимизации модель будет обращать одинаковое внимание на объекты обоих классов и, вероятно, лучше научится находить меньший класс.

Дисбаланс классов: итоги

Вы узнали:

- Что такое задача с несбалансированными классами
- Что дисбаланс классов затрудняет обучение моделей
- Что проблему можно решать с помощью перебалансировки объектов, а также с помощью перевзвешивания (class_weights)