Лекция № 3
Задача классификации

Задача классификации

Задача классификации

- Области применения алгоритмов классификации
- Формальное математическое определение

Несбалансированная классификация

Критерии качества классификации:

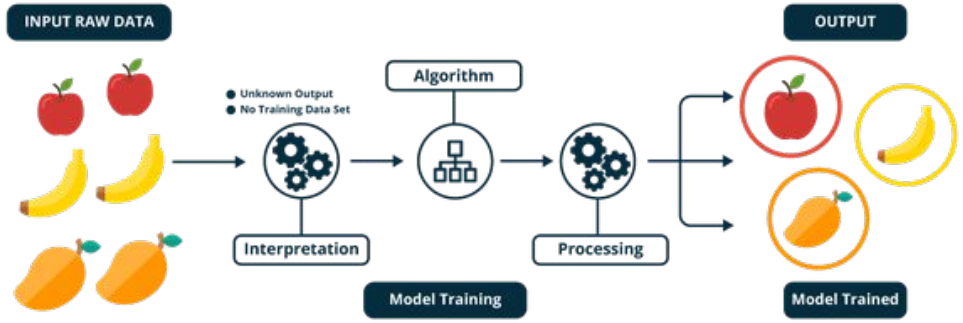
Precision, Recall, F1 score, ROC AUC

Области применения алгоритмов классификации

Обучение с учителем - область машинного обучения, при котором модель строится на основе имеющегося набора данных, называемого обучающей выборкой (training dataset).

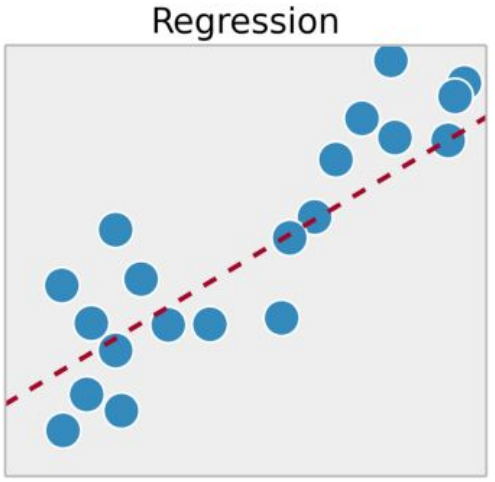
Обучающая выборка представлена парами «объект — ответ» (прецеденты), и предполагается, что существует функциональная зависимость между характеристиками объекта X и ответом y:

$$F: X \to y$$

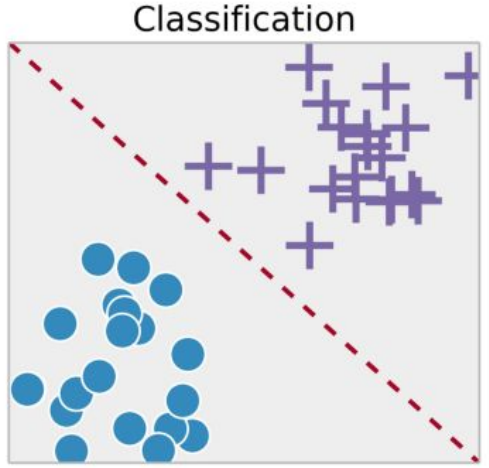


Области применения алгоритмов классификации

Регрессия - множество ответов бесконечно, так как они являются действительными числами или векторами действительных чисел.



Классификация - множество ответов дискретно и конечно. Решается задача классификации объектов в один или несколько классов.



Области применения алгоритмов классификации

- Оценка кредитоспособности заемщиков.
- Задачи медицинской диагностики
- Оптическое распознавание символов.
- Распознавание речи.
- Обнаружение спама.
- Классификация документов и т.д.

Формальное математическое определение

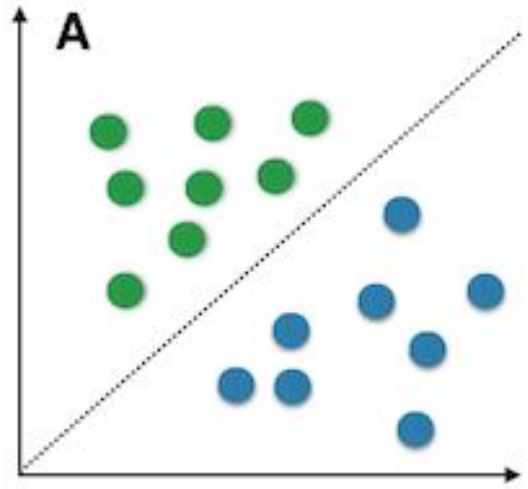
Обучающая выборка представляет собой набор отдельных объектов $X = \{x_i\}_{i=1}^{n}$, характеризующихся вектором вещественнозначных признаков $x_i = (x_{i,1}, \ldots, x_{i,d})$.

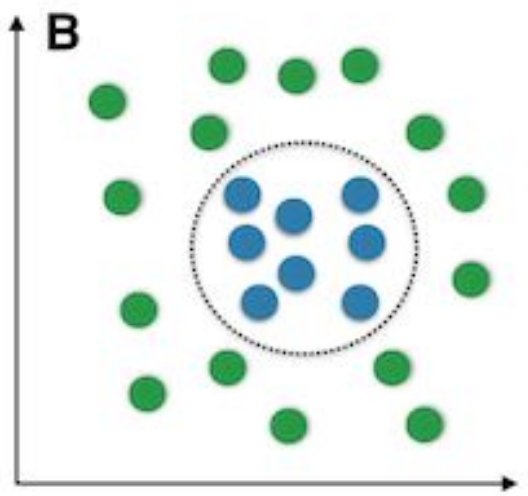
В качестве исхода объекта x фигурирует переменная y, принимающая конечное число значений, обычно из множества $Y = \{1, \ldots, k\}$.

Требуется построить алгоритм (классификатор), который по вектору признаков $\boldsymbol{x}$ вернул бы метку класса $\hat{y}$ или вектор оценок принадлежности (апостериорных вероятностей) к каждому из классов $\{p(s|\boldsymbol{x})\}_{s=1}^{k}$.

Формальное математическое определение

- Разделяющая гиперплоскость – это гиперплоскость, которая отделяет группы объектов, принадлежащим различным классам.
- Если такая гиперплоскость существует, то говорят о линейной разделимости выборки (A).
- Качество линейных методов классификации невысоко на линейно неразделимой выборке (B).



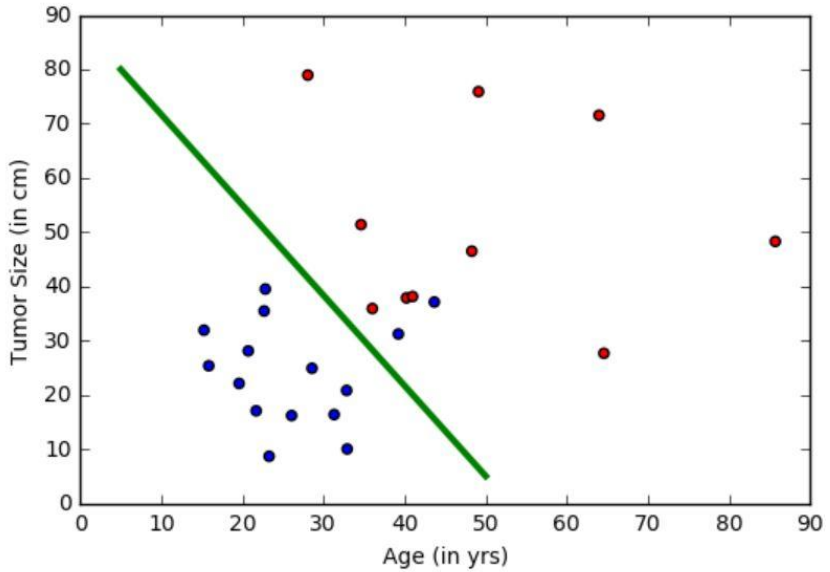


Формальное математическое определение

- Разделяющая гиперплоскость задается полиномом.
- Задача обучения линейных классификаторов состоит в нахождении коэффициентов данного полинома.

Полином $(w_1, \ldots, w_n, b)$ задаёт n-мерную гиперплоскость, уравнение которой:

$$w_1 x_1 + \cdots + w_n x_n + b = 0$$
$$w^T x + b = 0$$



Формальное математическое определение

- Знак полинома (линейной комбинации) позволяет отнести точку к верхнему или нижнему подпространству, на которые гиперплоскость делит гиперпространство:

  $w^T \mathrm{x} + b > 0$ — точка лежит «выше» гиперплоскости;

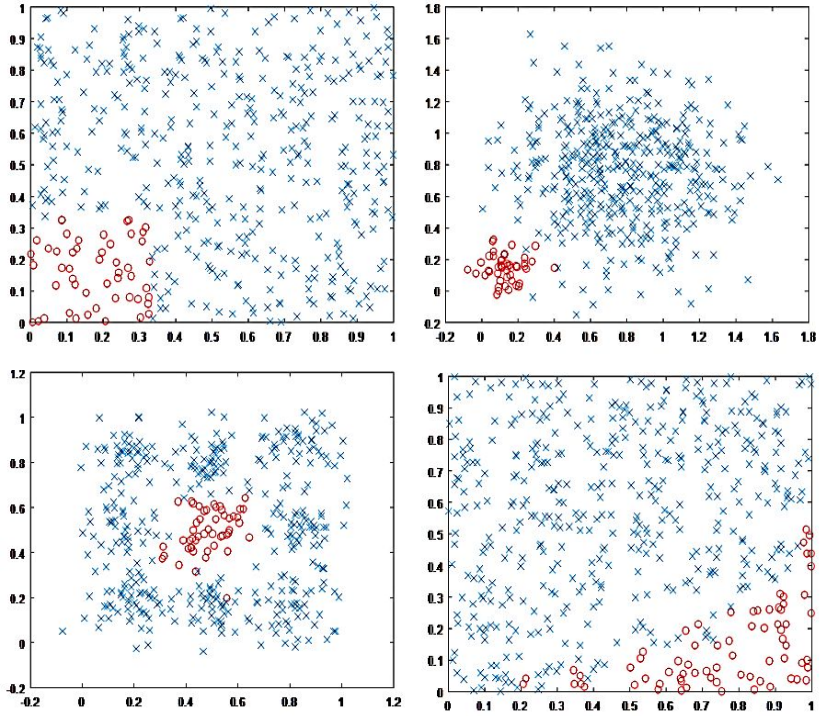
  $w^T \mathrm{x} + b < 0$ — точка лежит «ниже» гиперплоскости.

- Чем дальше точка от гиперплоскости, являющейся границей решений (decision boundary), тем выше вероятность, что образец (sample), определяемый этой точкой, попадает в тот или иной класс.

Несбалансированная классификация

Imbalanced Data : один из классов представлен значительно бóльшим количеством объектов, чем другой – мажоритарный и миноритарный классы.

- Классификация на подобных выборках может оказаться неэффективной, т.к. модель будет предвзятой и неточной. Причина: классификатор может полностью отнести объекты миноритарных классов к шуму.
- Алгоритмы машинного обучения обычно предназначены для повышения точности за счет уменьшения ошибки. Другими словами, классификатор настраивается на мажоритарный класс, получая высокую точность, не выделяя объекты миноритарного класса.



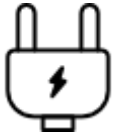
Несбалансированная классификация

Задачи, в которых несбалансированность данных не просто общая проблема, а ожидаема в силу специфики области применения:

- В медицинской диагностике объектам миноритарного класса соответствует наличие редкого заболевания.

- Прогнозирование природных катастроф.

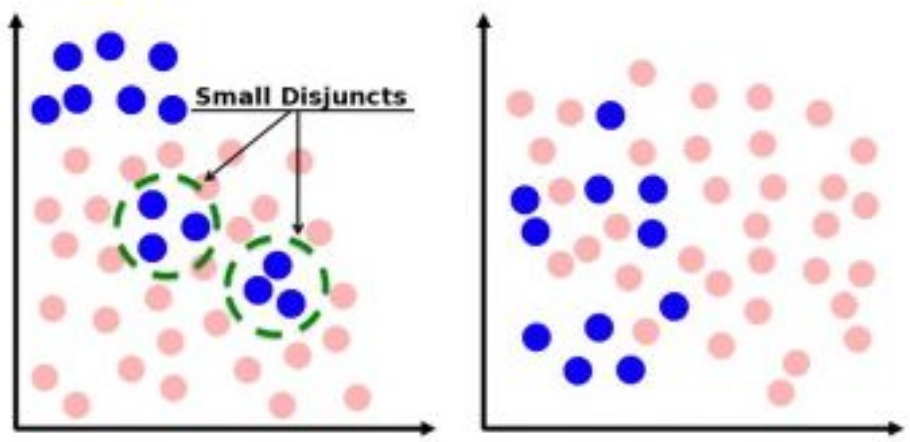
- Обнаружение аномалий в сценариях обнаружения кражи электроэнергии.



- Мошеннические транзакции – 1-2% транзакций, отличающихся от большинства.

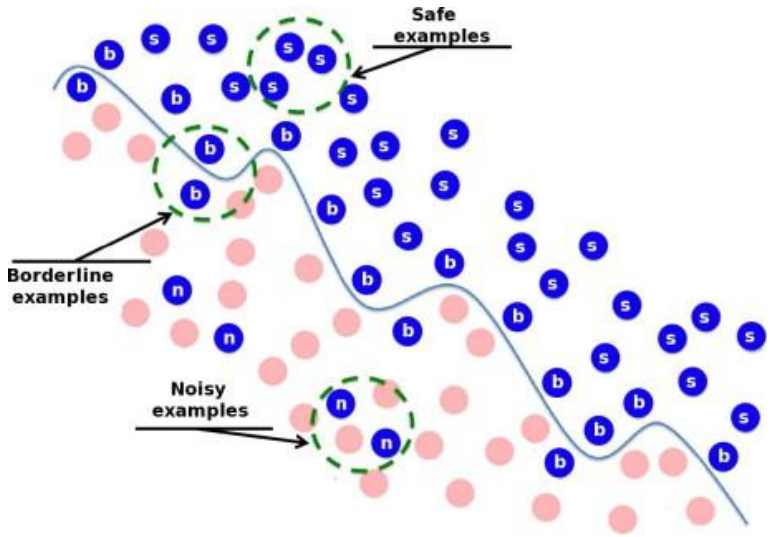
Несбалансированная классификация

Обучение на несбалансированных данных осложняется расположением отдельных примеров выборок:



Вкрапления

Наложения



s – чистые примеры класса (safe examples);

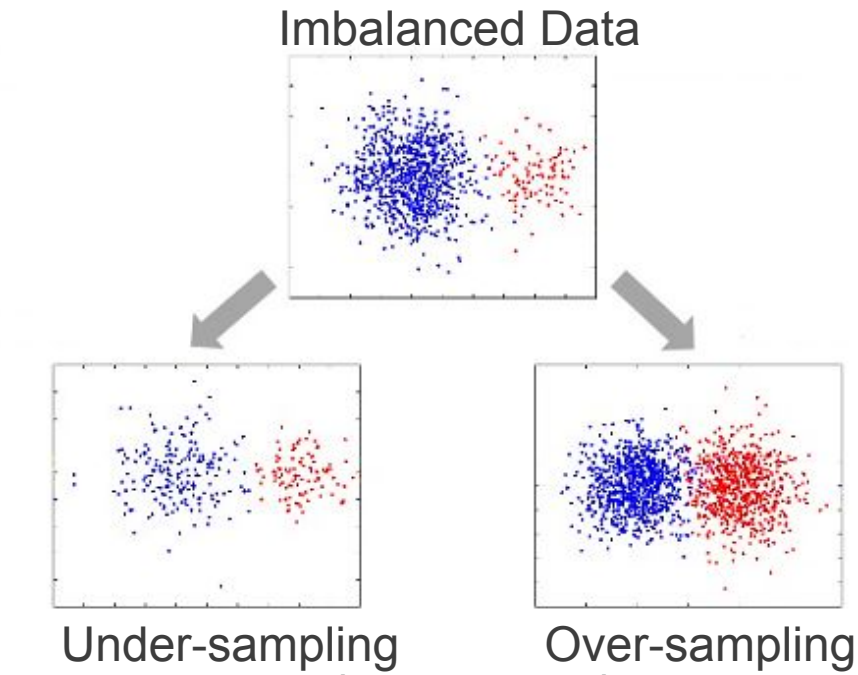
b – пограничные (borderline);

n – зашумляющие (noisy).

Несбалансированная классификация

Основные подходы к решению проблемы несбалансированных данных в классификации:

1. Сэмплинг (sampling)
    - Уменьшение большего класса
    - Увеличение меньшего класса
2. Изменение порога решения



Сэмплинг представляет собой выбор прецедентов таким образом, чтобы их количество для обоих классов уравнялось. Этот подход позволяет учесть распределение/соотношение классов.

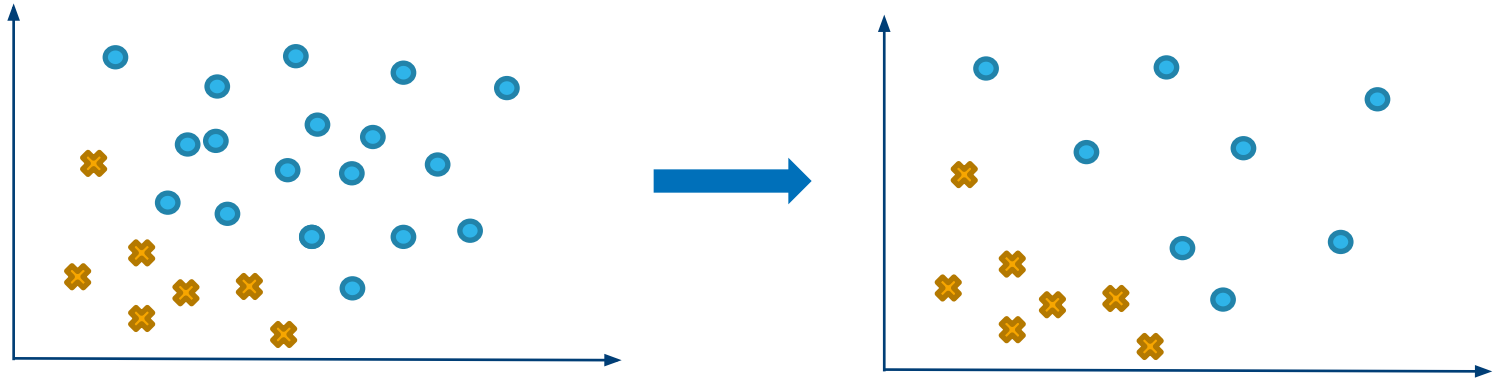
Несбалансированная классификация

Уменьшение большего класса (Undersampling)

Случайный или синтетический выбор прецедентов мажоритарного класса в обучающую выборку.

- Приводит к уменьшению тренировочной базы
- Возможно исключение важной информации и увеличение ошибки

Самый простой вариант — произвольный выбор прецедентов (Random Undersampling) — не учитывает положение прецедентов относительно друг друга и поверхности, разделяющей классы. Однако, на практике он оказывается наиболее эффективным.

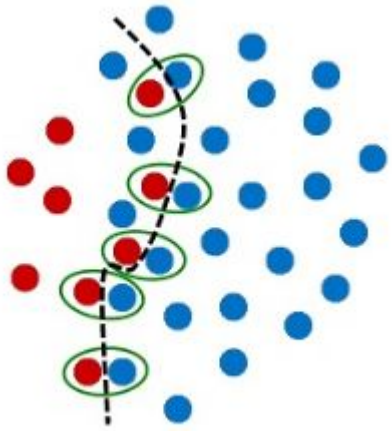


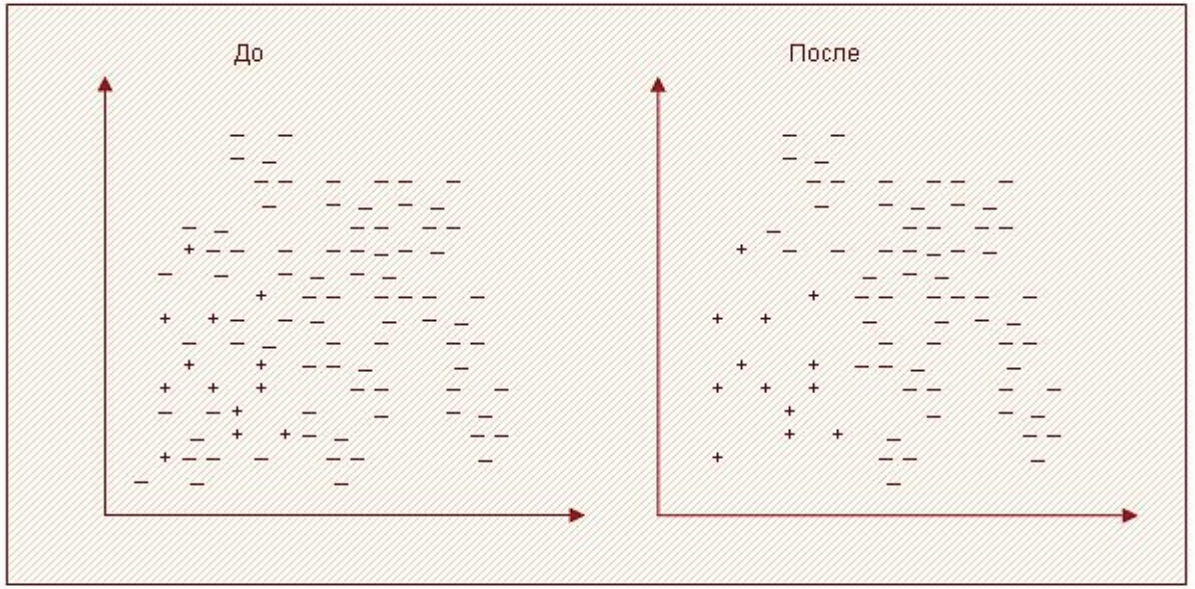
Несбалансированная классификация

Уменьшение большего класса (Undersampling)

Поиск связей Томека (Tomek Links)

Этот способ хорошо удаляет записи, которые можно рассматривать в качестве «зашумляющих».





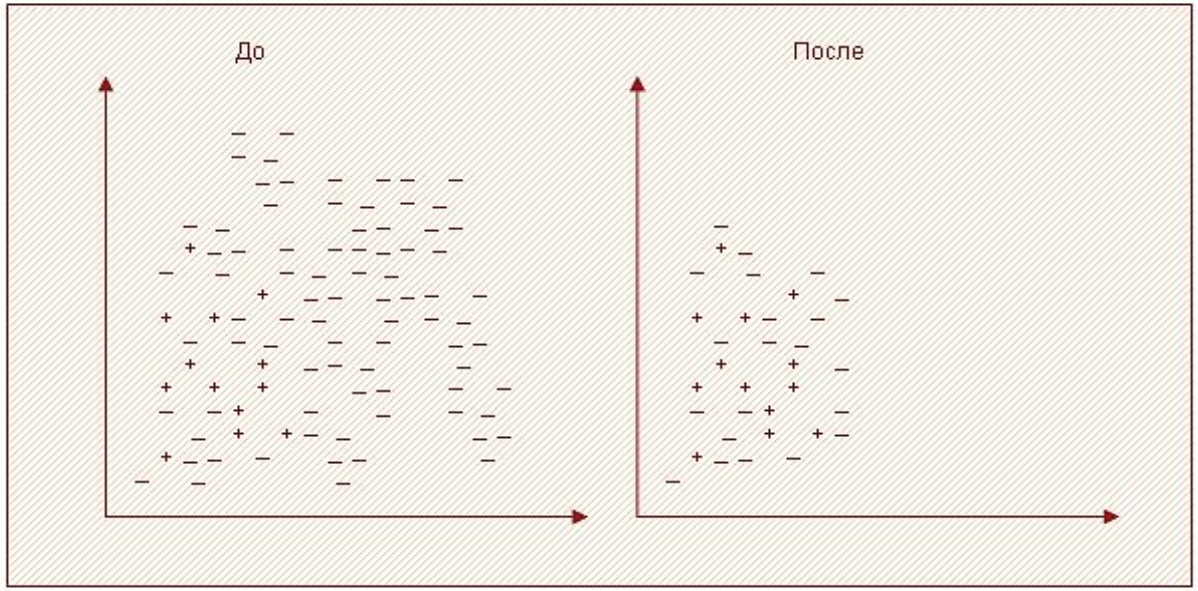
Несбалансированная классификация

Уменьшение большего класса (Undersampling)

Правило сосредоточенного ближайшего соседа

(Condensed Nearest Neighbor Rule)

Этот метод учит классификатор находить отличие между похожими примерами, но принадлежащими к разным классам.



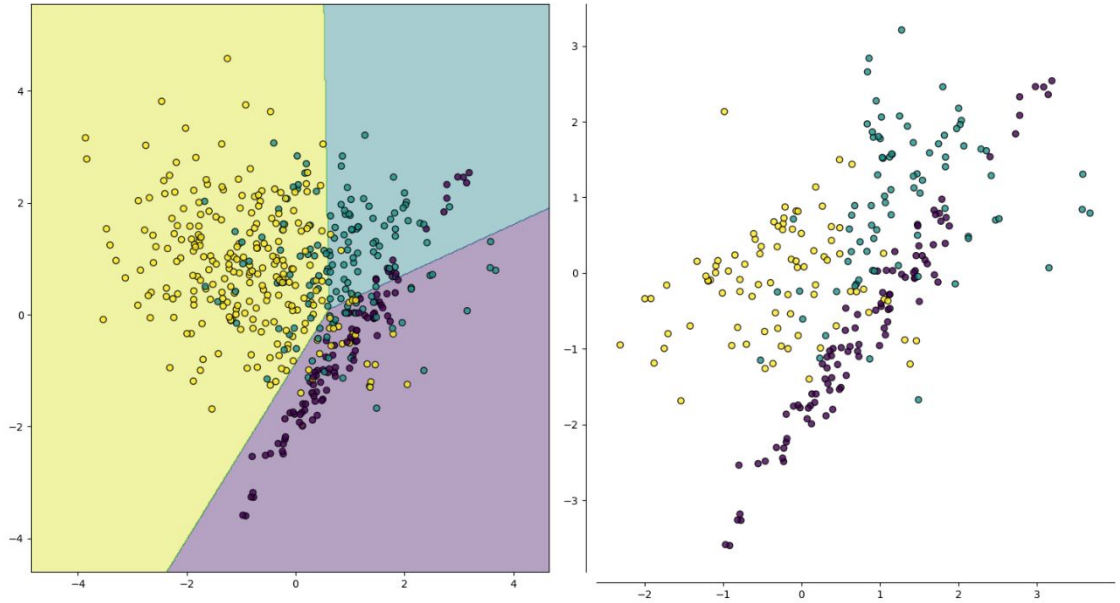
Несбалансированная классификация

Уменьшение большего класса (Undersampling)

Односторонний сэмплинг (One-side Sampling, One-sided Selection)

1. Применяется правило сосредоточенного ближайшего соседа.
2. Удаляются все мажоритарные примеры, участвующие в связях Томека.

Таким образом, удаляются большие «сгустки» мажоритарных примеров, а затем область пространства со скоплением миноритарных очищается от потенциальных шумовых эффектов.

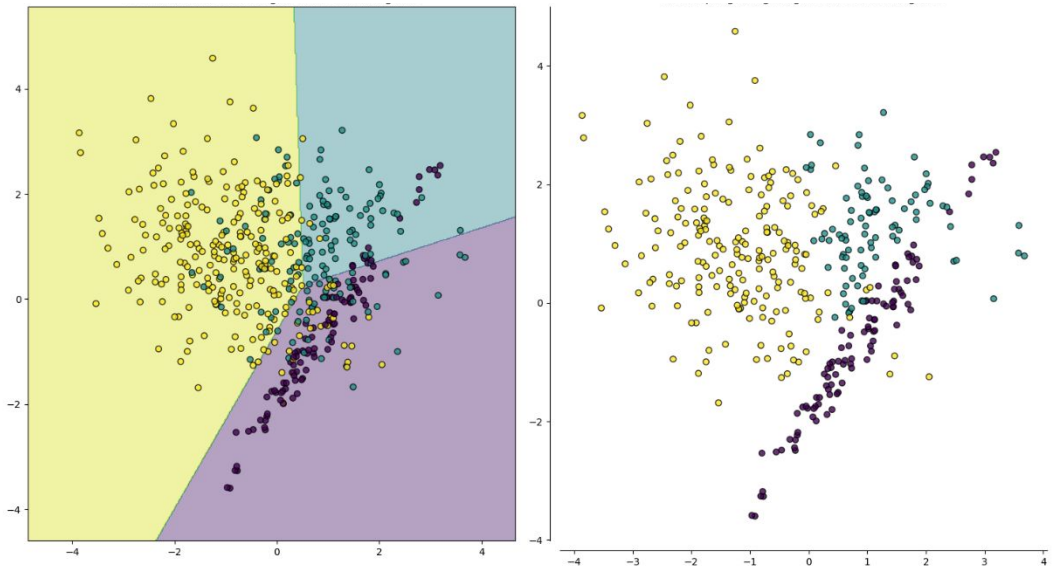


Несбалансированная классификация

Уменьшение большего класса (Undersampling)

Правило «очищающего» соседа (Neighborhood Cleaning Rule)

1. Все примеры классифицируются по правилу трех ближайших соседей.
2. Удаляются следующие мажоритарные примеры:
   - получившие верную метку класса;
   - являющиеся соседями миноритарных примеров, которые были неверно классифицированы.



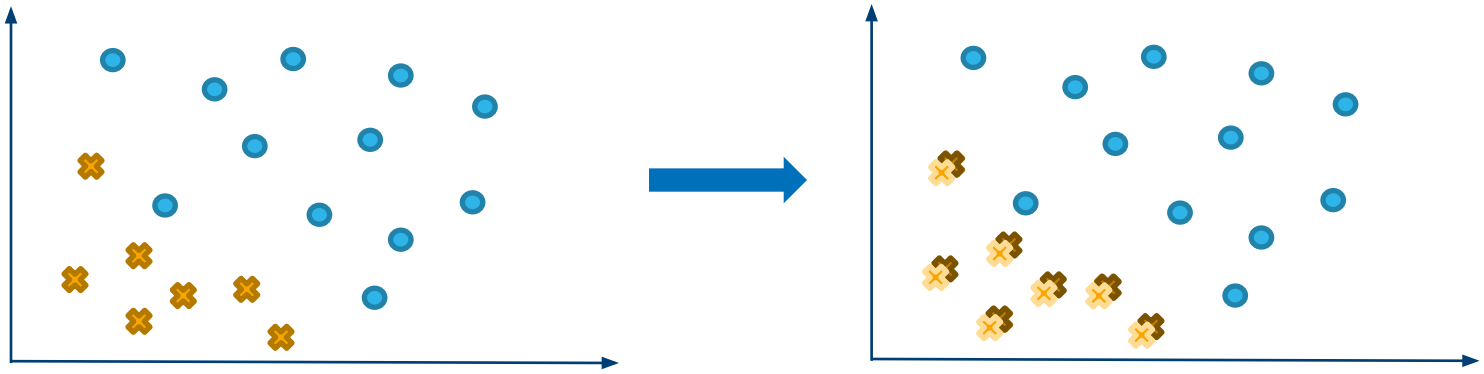
Эта стратегия также направлена на то, чтобы удалить те примеры, которые негативно влияют на исход классификации миноритарных.

Несбалансированная классификация

Увеличение меньшего класса (Oversampling)

Добавление прецедентов миноритарного класса позволяет сохранить всю имеющуюся информацию. Недостаток – увеличение размера тренировочной базы и, как следствие, большее время ее обработки.

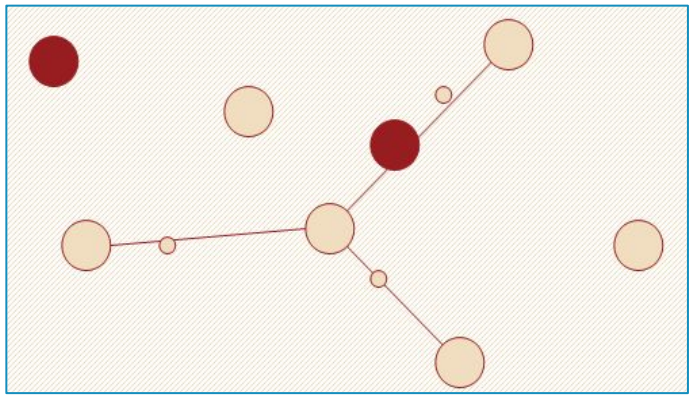
Самый простой вариант — дублирование случайных прецедентов меньшего класса, которое не добавляет лишней информации и не изменяет положение разделяющей поверхности.



Несбалансированная классификация

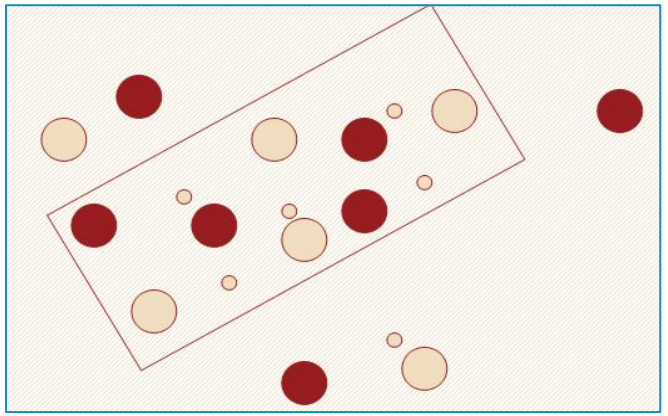
Увеличение меньшего класса (Oversampling)

Алгоритм SMOTE (Synthetic Minority Oversampling Technique) - генерация некоторого количества искусственных примеров, которые «похожи» на имеющиеся в миноритарном классе, но при этом не дублируют их.



Миноритарный пример
Мажоритарный пример
Искусственный пример

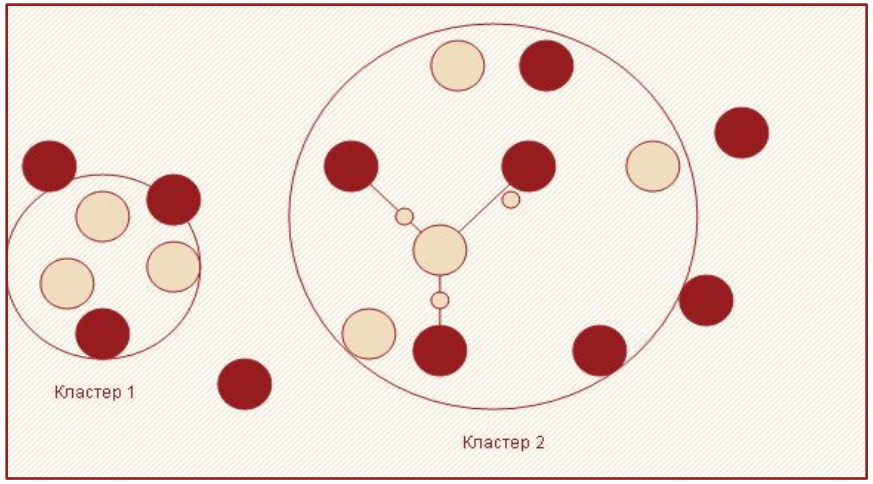
Алгоритм не подходит в случае, если миноритарные примеры равномерно распределены среди мажоритарных и имеют низкую плотность. Тогда SMOTE только сильнее перемешает классы.



Несбалансированная классификация

Увеличение меньшего класса (Oversampling)

Алгоритм ADASYN (Adaptive Synthetic Minority Oversampling) - использование функции плотности распределения как критерия для автоматического определения числа экземпляров, которые необходимо сгенерировать для каждого из объектов миноритарного класса, адаптивно меняя веса разных экземпляров миноритарного класса



Миноритарный пример
Мажоритарный пример
Искусственный пример

Несбалансированная классификация

Изменение порога решения (Changing Performance Metric)

Многие алгоритмы классификации определяют степень достоверности предсказания. При данном подходе, изменяя порог в решающем правиле, можно получать различные разделяющие поверхности.

- Такие методы довольно просты в реализации; однако, изменение порога не гарантирует точность формы границы, что может привести к значительному повышению общей ошибки.
- Существуют методы, предлагающие показатели эффективности, которые могут дать большее представление о точности модели, чем традиционные метрики.

Критерии качества классификации

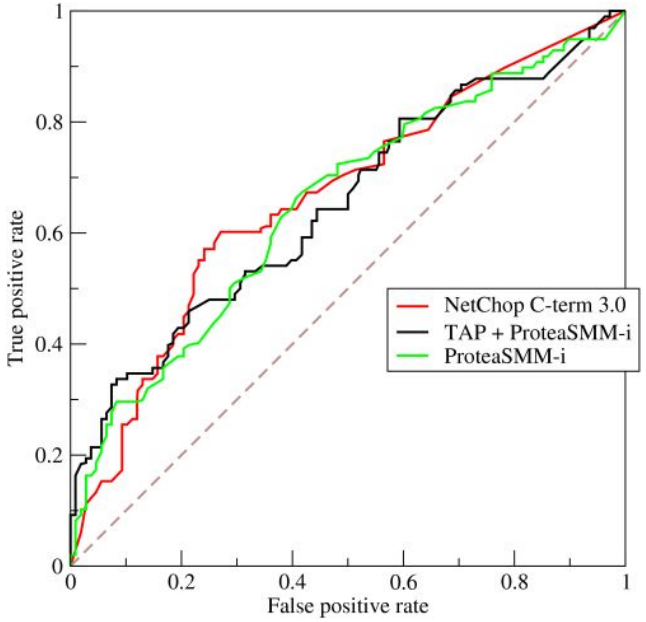
Определим ROC-кривую (receiver operating characteristic, рабочая характеристика приёмника)

Ось абсцисс: доля неправильных положительных предсказаний как функция w0

Ось ординат: доля правильных положительных предсказаний

AUC - площадь под кривой, используется для оценки точности классификации (AUC ≥ 0,5)

Пунктирная линия - наихудшая точность (случайное предсказание)



Критерии качества классификации

Точность и полнота (Precision and Recall) для случая бинарной классификации

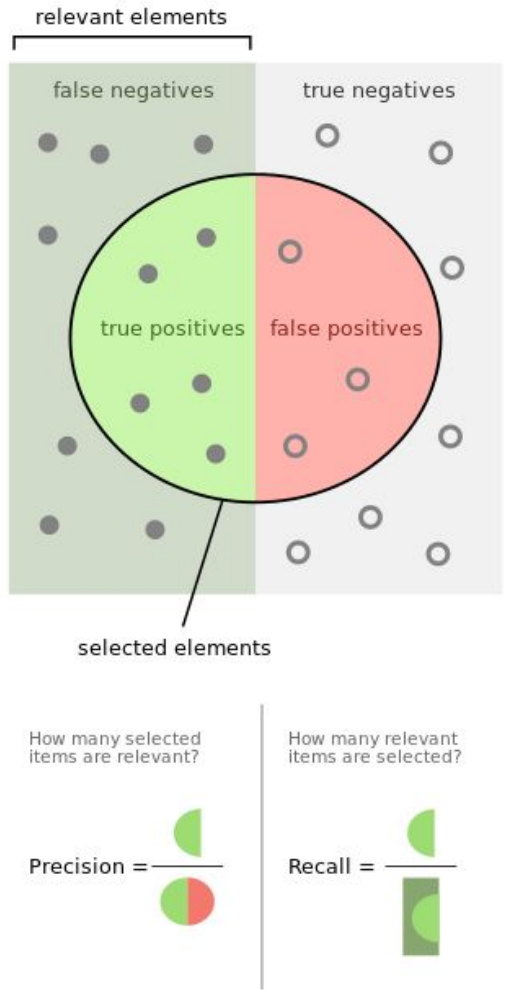
TP - правильные положительные предсказания

FP - неправильные положительные предсказания

FN - неправильные отрицательные предсказания

$$\text{Precision} : P = \frac{TP}{TP + FP}$$

$$\text{Recall} : R = \frac{TP}{TP + FN}$$



Критерии качества классификации

Точность и полнота (Precision and Recall) для случая многоклассовой классификации

Для каждого класса $y \in Y$

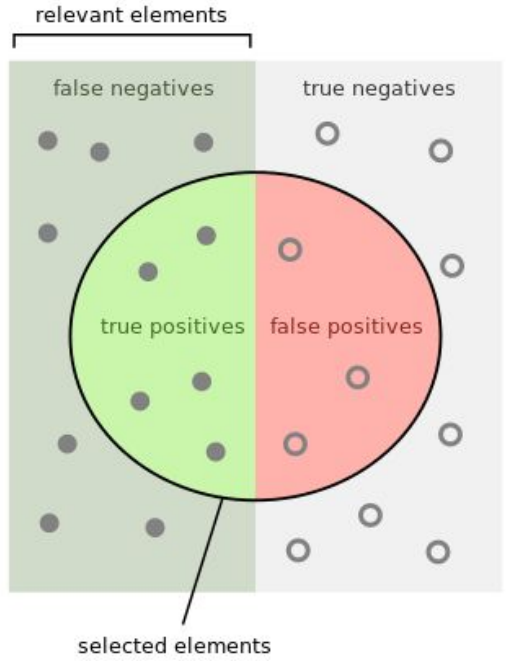
TP_y доля правильных положительных предсказаний

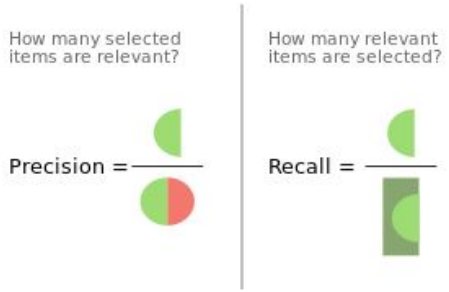
FP_y доля неправильных положительных предсказаний

FN_y доля неправильных отрицательных предсказаний

$$\text{Precision} : P = \frac{\sum_y TP_y}{\sum_y (TP_y + FP_y)}$$

$$\text{Recall} : R = \frac{\sum_y TP_y}{\sum_y (TP_y + FN_y)}$$





Ключевые концепции

- Задачи классификации оперируют множеством ответов с дискретными значениями, так как цель таких задач — определить, к какому классу относится пример.
- Чем дальше точка от гиперплоскости, являющейся границей решений (decision boundary), тем выше вероятность, что образец (sample), определяемый этой точкой, попадает в тот или иной класс.
- Imbalanced Data : один из классов представлен значительно бóльшим количеством объектов, чем другой – мажоритарный и миноритарный классы.
- Основные подходы к решению проблемы несбалансированных данных в классификации:

1. Сэмплинг
2. Изменение порога решения

- Многие алгоритмы классификации определяют степень достоверности предсказания. При данном подходе, изменяя порог в решающем правиле, можно получать различные разделяющие поверхности.