Деревья vs линейные методы

Елена Кантонистова

Skillbox

Елена Кантонистова

- Кандидат физико-математических наук
- Выпускница Школы Анализа Данных (ШАД) Яндекса
- Доцент факультета компьютерных наук ВШЭ
- Академический руководитель магистратуры «Машинное обучение и высоконагруженные системы» ФКН ВШЭ

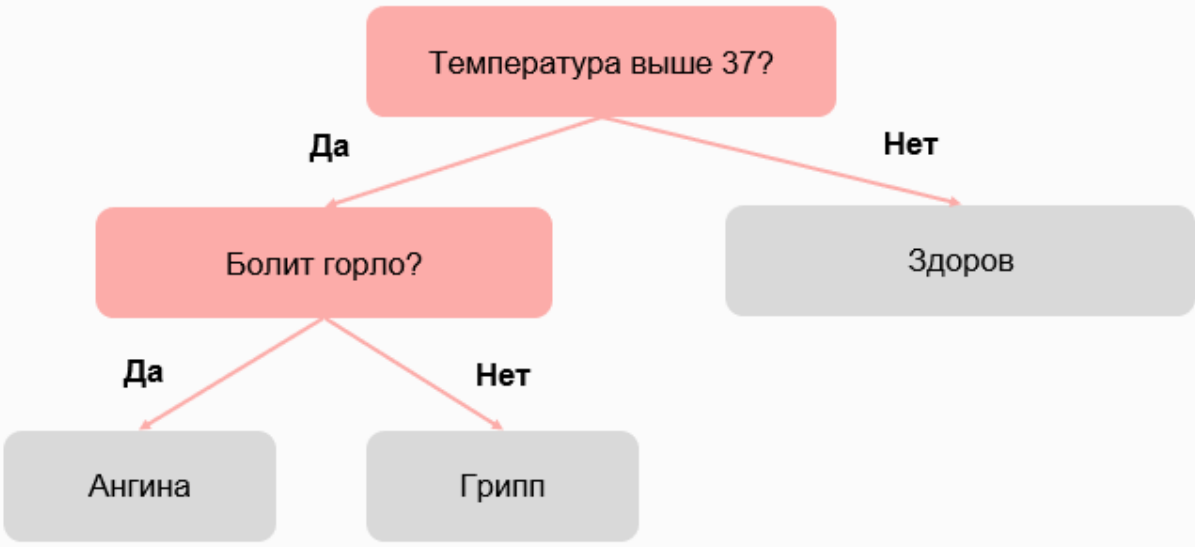
Предыдущий индустриальный опыт:

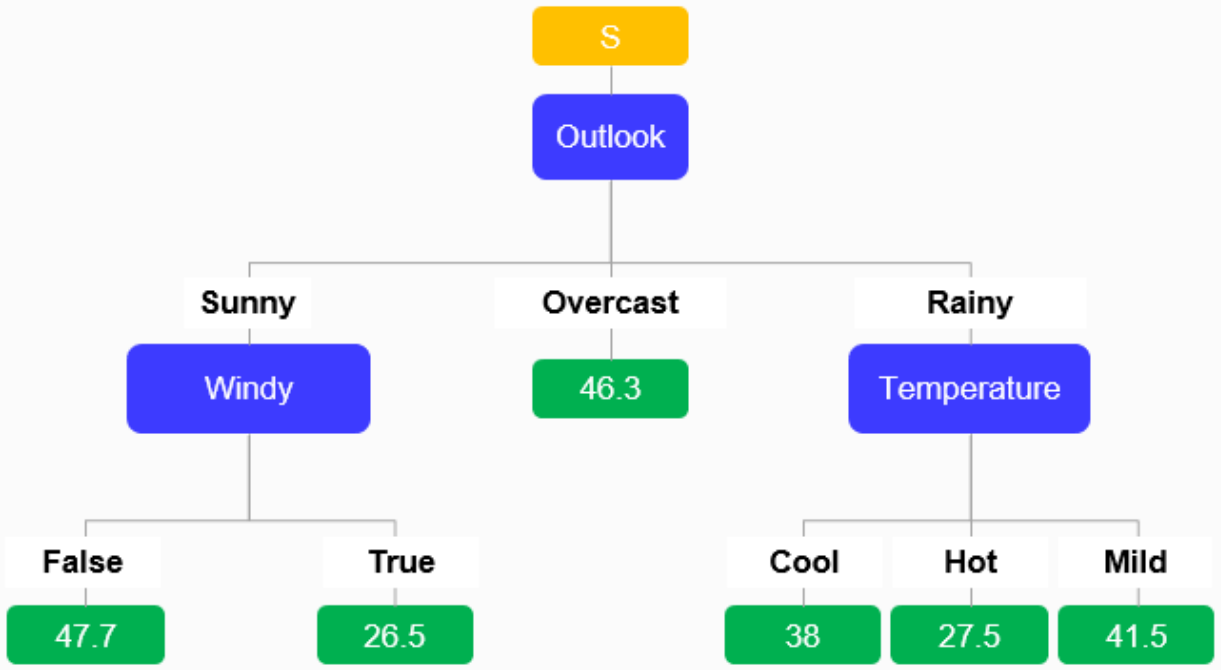
- Data scientist в Raxel Telematics
- Data scientist в United Consulting Group

В этом модуле вы…

- Узнаете, в каких задачах лучше использовать линейные модели, а в каких лучше себя показывают решающие деревья
- Изучите, как справляться с переобучением решающих деревьев
- Узнаете, что такое кросс-валидация и чем она может быть лучше, чем подход с отложенной выборкой
- Изучите модели бэггинга и случайного леса для решения задач классификации и регрессии
- Узнаете, как можно в полуавтоматическом режиме настраивать гиперпараметры моделей

Деревья





Автор: Кантонистова Елена, Дерево для регрессии

Деревья переобучаются

Если никак не ограничивать построение дерева, то оно идеально подгонится под данные и в большинстве случаев даст нулевую ошибку на обучающих данных и, значит, сильно переобучится.

Решение №1

Можно задавать ограничения на структуру дерева до его обучения.

Некоторые регулируемые гиперпараметры:

- max_depth
- min_samples_split
- min_samples_leaf
- max_features

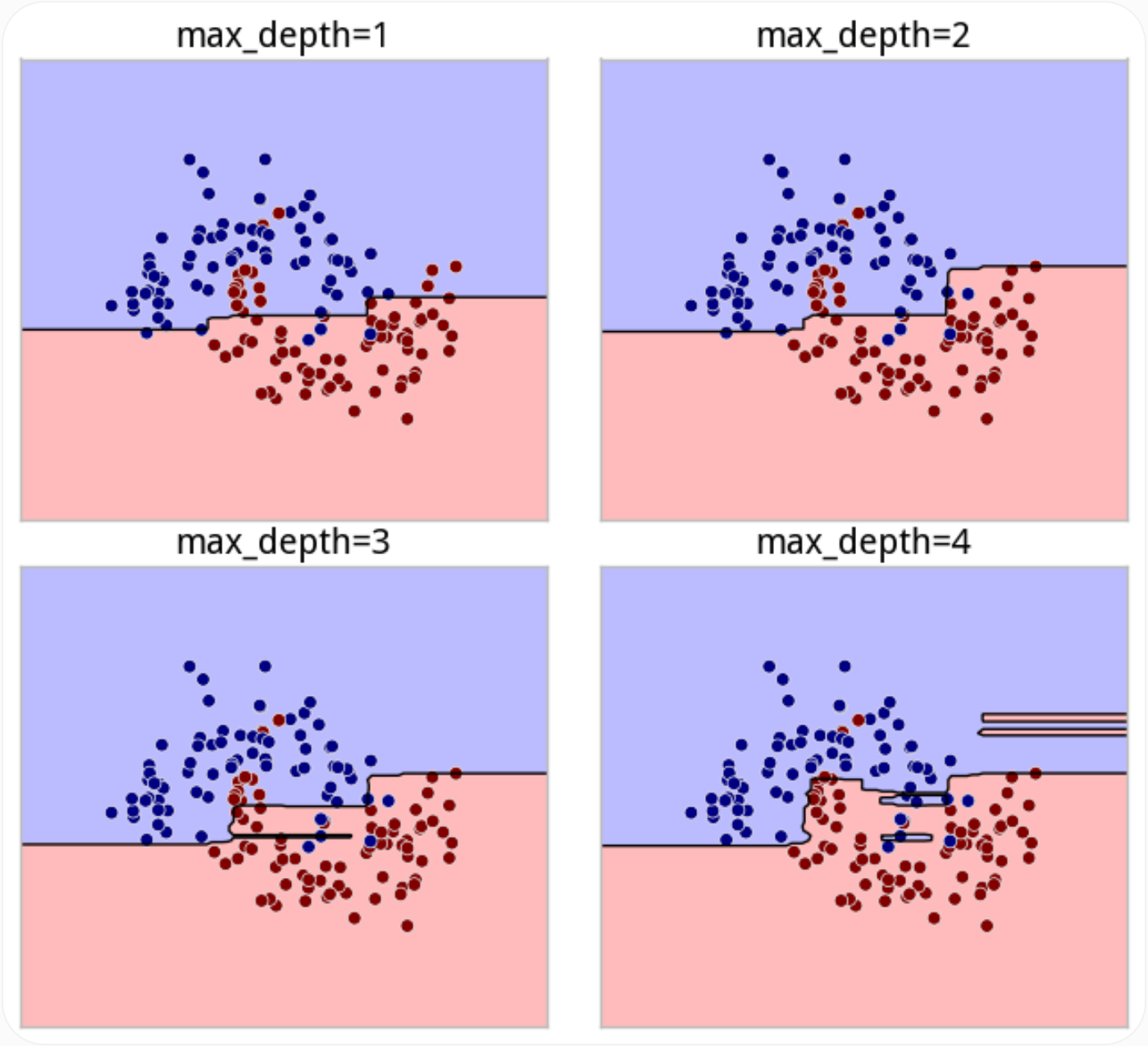
Решение №1

Можно задавать ограничения на структуру дерева до его обучения.

Некоторые регулируемые гиперпараметры:

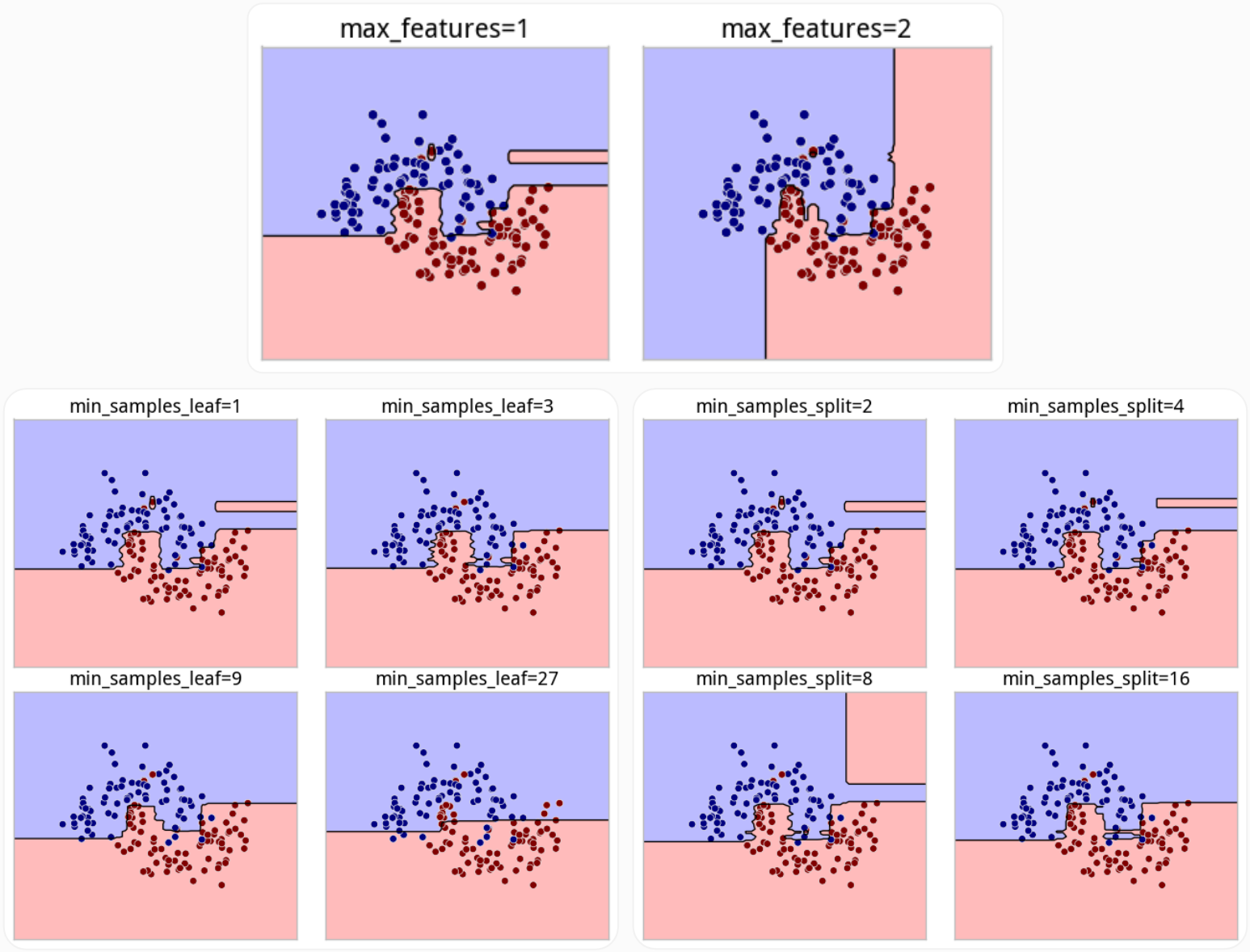
- max_depth — максимальная глубина, больше которой дерево не строится
- min_samples_leaf — минимальное число объектов, которое должно находиться в листе
- min_samples_split — минимальное число объектов, которое должно находиться в вершине, чтобы её можно было дальше разбивать
- max_features — максимальное количество признаков, из которых на каждом шаге выбирается оптимальное разбиение вершины

max_depth



Изображение: Max_depth

Другие гиперпараметры



Изображение: Другие гиперпараметры

Решение №2: стрижка

Можно сначала построить дерево без ограничений —
оно получится громоздким и максимально переобученным,
а уже затем его регуляризовать. Этот подход называется
стрижкой дерева (или pruning).

Автор: Елена Кантонистова

Решение №2: стрижка

Алгоритм стрижки:

- строится дерево без ограничений на гиперпараметры
- производится оптимизация его структуры с целью уменьшения переобучения

Оптимизируется регуляризованный функционал:

$$Q_\alpha(T) = Q(T) + \alpha|T|$$

Вы узнали

- Что в зависимости от данных задачи, иногда лучшее качество показывают решающие деревья, а иногда — линейные модели
- Как снизить переобучение у решающего дерева (или заданием гиперпараметров, или стрижкой)