START ML

KARPOV.COURSES

DECISION TREES 

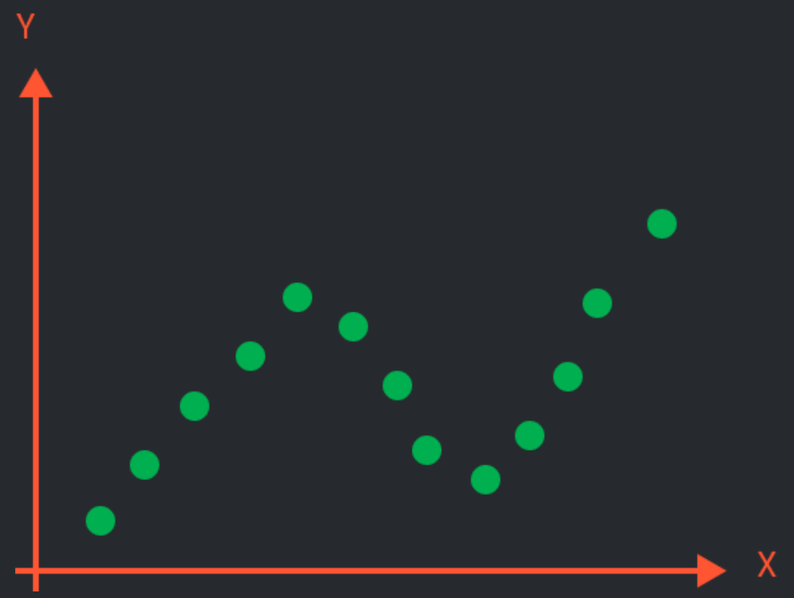
СИЛА РЕШАЮЩИХ ДЕРЕВЬЕВ

- Решающие деревья – невероятно мощные алгоритмы!
- Даже обычных бинарных деревьев может хватить, чтобы получить нулевую ошибку на трейне даже при сложных зависимостях!
- Продемонстрируем эту идею.

DECISION TREES

СИЛА РЕШАЮЩИХ ДЕРЕВЬЕВ

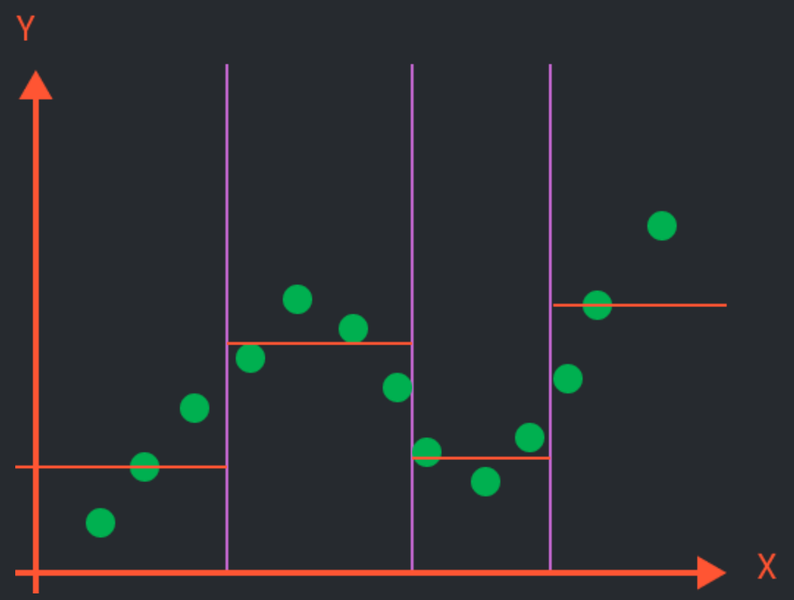
- Пусть имеем выборку, где у каждого объекта есть всего один признак, и зависимость от таргета нелинейная
- Предикаты для бинарного дерева на такой выборке будут вида $[X \leq t]$
- Если в тренировочной выборке m точек, то максимум m таких предикатов $\{[X \leq t_i]\}_i^m$ хватит для идеального прогноза!



DECISION TREES

СИЛА РЕШАЮЩИХ ДЕРЕВЬЕВ

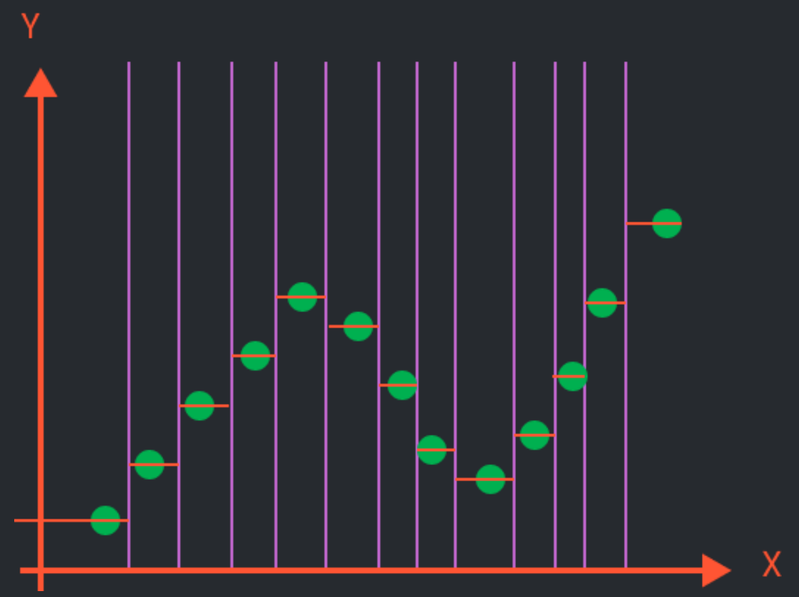
- Пусть имеем выборку, где у каждого объекта есть всего один признак, и зависимость от таргета нелинейная
- Предикаты для бинарного дерева на такой выборке будут вида $[X \leq t]$
- Если в тренировочной выборке m точек, то максимум m таких предикатов $\{[X \leq t_i]\}_i^m$ хватит для идеального прогноза!



DECISION TREES

СИЛА РЕШАЮЩИХ ДЕРЕВЬЕВ

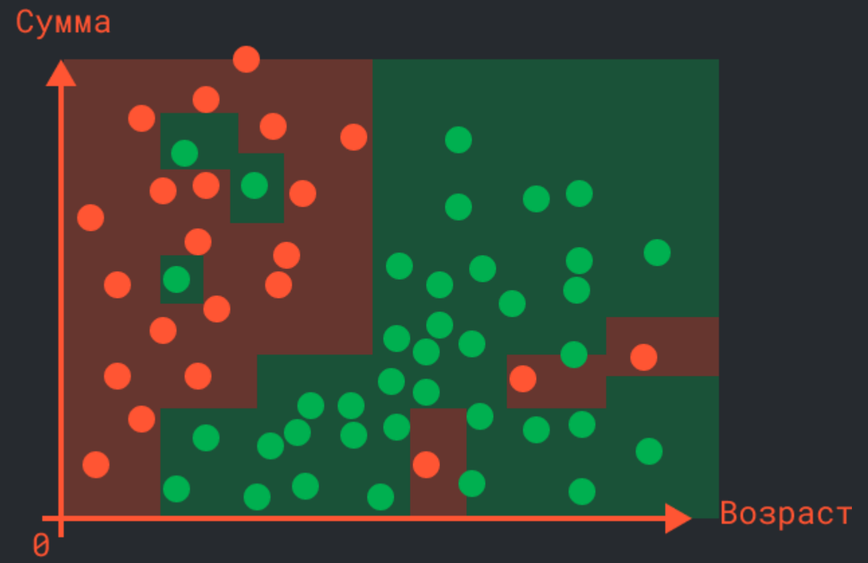
- Пусть имеем выборку, где у каждого объекта есть всего один признак, и зависимость от таргета нелинейная
- Предикаты для бинарного дерева на такой выборке будут вида $[X \leq t]$
- Если в тренировочной выборке m точек, то максимум m таких предикатов $\{[X \leq t_i]\}_i^m$ хватит для идеального прогноза!

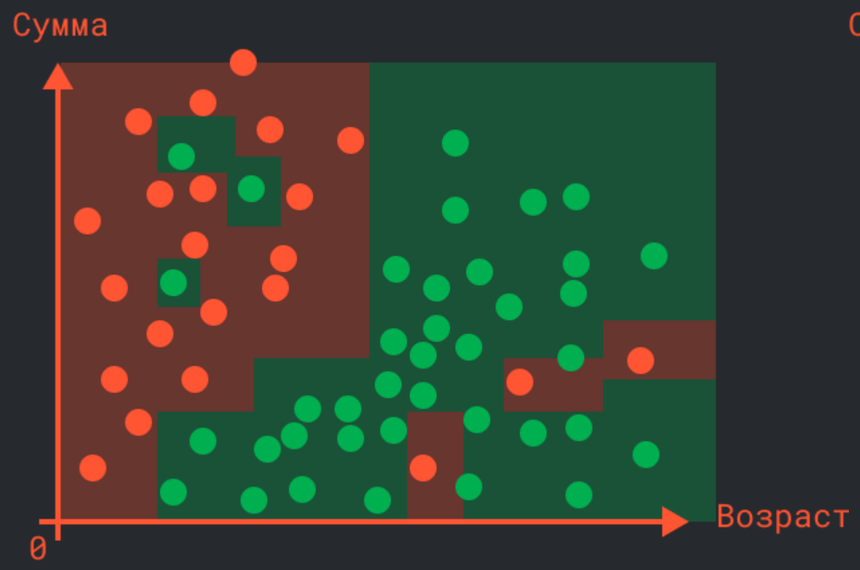


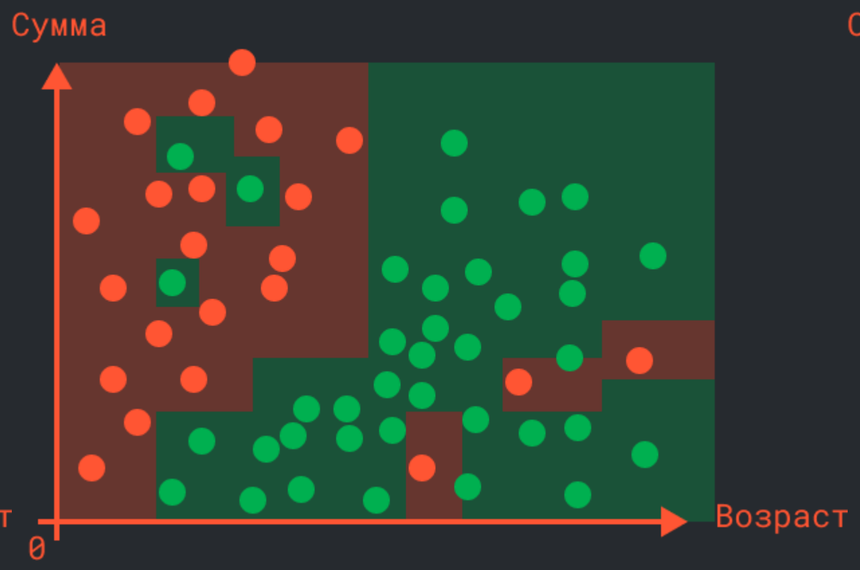
DECISION TREES

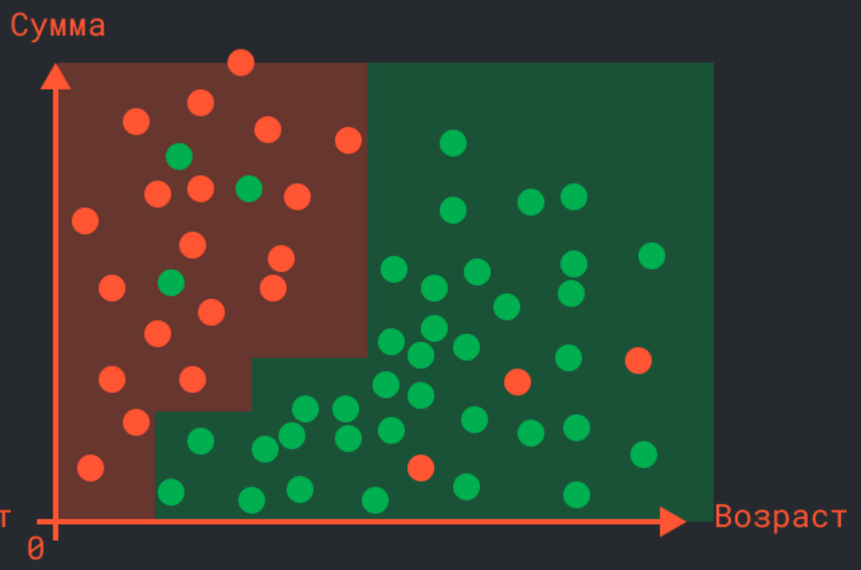
СИЛА РЕШАЮЩИХ ДЕРЕВЬЕВ

- Пусть имеем выборку, где у каждого объекта есть всего один признак, и зависимость от таргета нелинейная
- Предикаты для бинарного дерева на такой выборке будут вида $[X \leq t]$
- Если в тренировочной выборке m точек, то максимум m таких предикатов $\{[X \leq t_i]\}_i^m$ хватит для идеального прогноза!




Сумма
Возраст
0


Сумма
Возраст
0


Сумма
Возраст
0

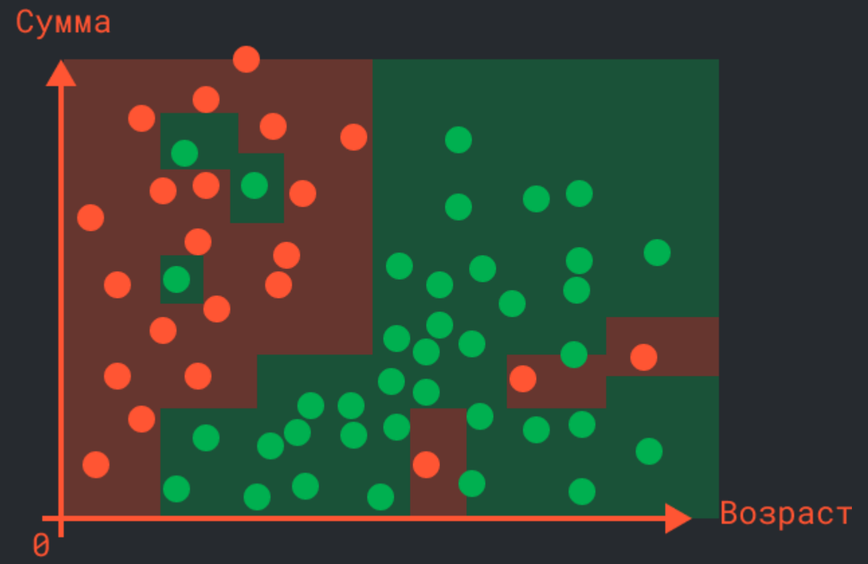
DECISION TREES

СИЛА РЕШАЮЩИХ ДЕРЕВЬЕВ

— При попытке достичь идеального качества получаем очень сложные разбиения

— Они подстраиваются даже под супер-шумовые объекты

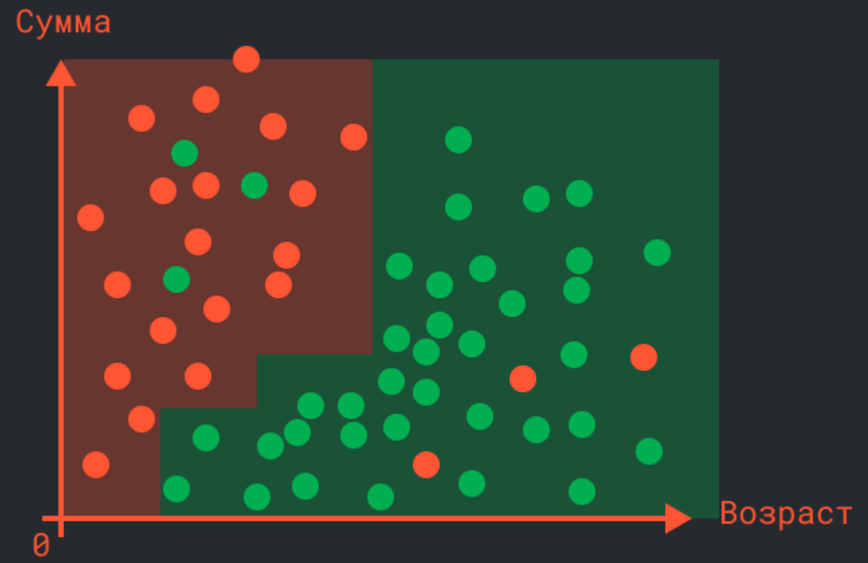
— Хотя на самом деле зависимости могут быть куда проще, и лучше именно их находить



DECISION TREES

СИЛА РЕШАЮЩИХ ДЕРЕВЬЕВ

- При попытке достичь идеального качества получаем очень сложные разбиения
- Они подстраиваются даже под супер-шумовые объекты
- Хотя на самом деле зависимости могут быть куда проще, и лучше именно их находить



РЕЗЮМЕ

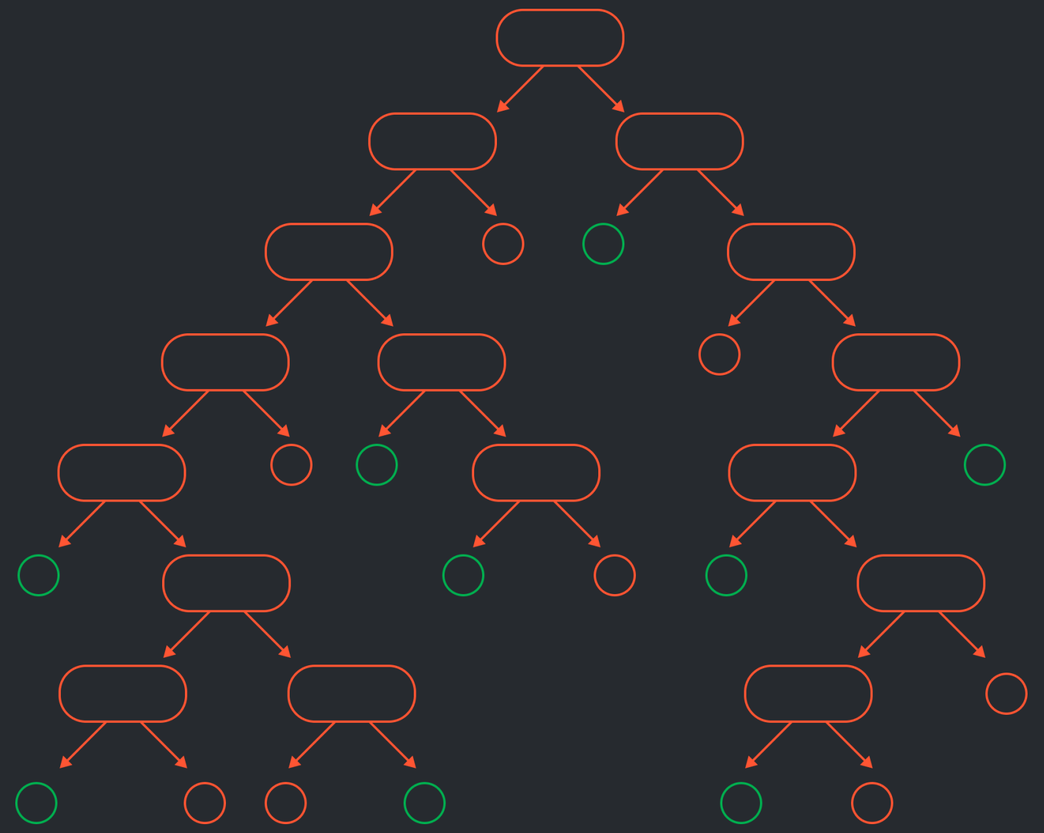
- Убедились в том, что решающие деревья – мощные модели, способные добиваться идеального качества на тренировочной выборке
- Тем не менее, даже интуитивно понятно, что они оказываются в том числе и неустойчивыми
- Склонными к сильному переобучению

- Нужно научиться их как-то “сдерживать”, “регулировать”

DECISION TREES

КРИТЕРИИ ОСТАНОВА

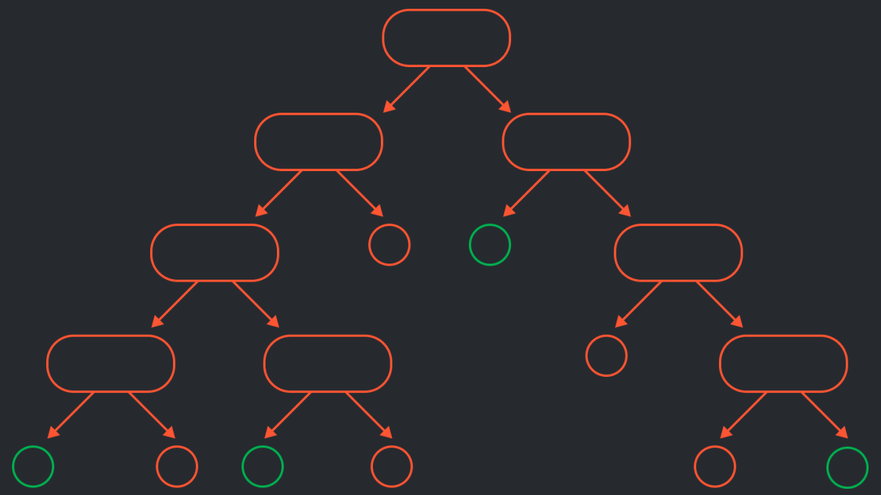
- Деревья крайне склонные к переобучению
- В попытке построить лучшее разбиение на трейне, достигается идеальное качество и/или гигантская глубина и/или малое чисто объектов в листах…
- Давайте устанавливать их как гиперпараметры и делать менее жесткими!
- Гиперпараметр 1: глубина дерева
- Глубина 1000 – переобучение, установим

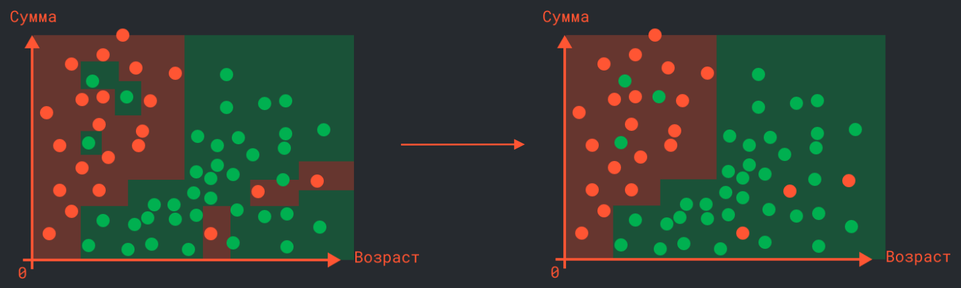


DECISION TREES

КРИТЕРИИ ОСТАНОВА

- Деревья крайне склонные к переобучению
- В попытке построить лучшее разбиение на трейне, достигается идеальное качество и/или гигантская глубина и/или малое чисто объектов в листах…
- Давайте устанавливать их как гиперпараметры и делать менее жесткими!
- Гиперпараметр 1: глубина дерева
- Глубина 1000 – переобучение, установим

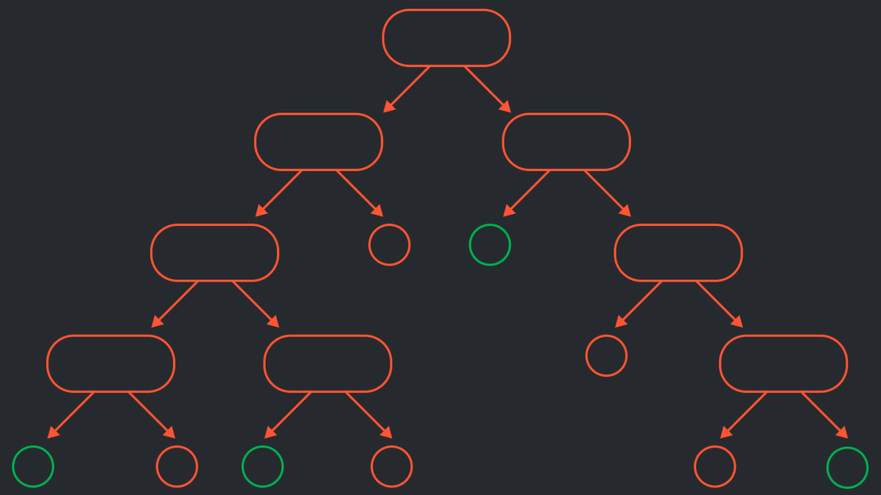


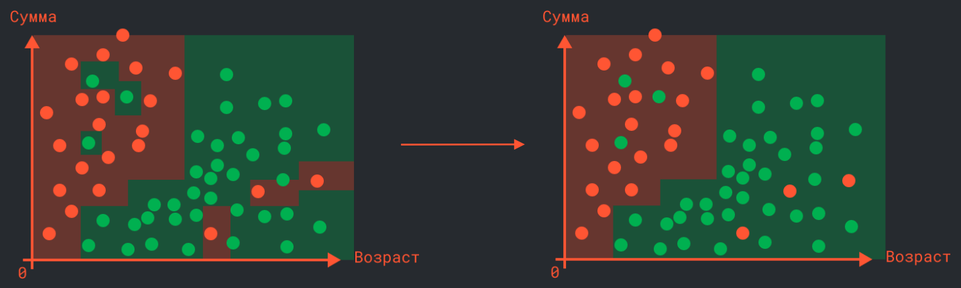


DECISION TREES

КРИТЕРИИ ОСТАНОВА

- Деревья крайне склонные к переобучению
- В попытке построить лучшее разбиение на трейне, достигается идеальное качество и/или гигантская глубина и/или малое чисто объектов в листах…
- Давайте устанавливать их как гиперпараметры и делать менее жесткими!
- Гиперпараметр 2: объекты в внутренней вершине

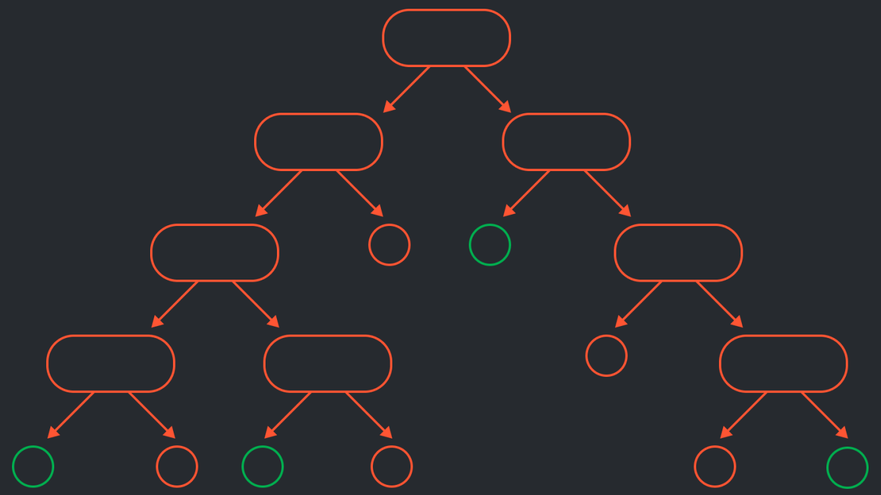


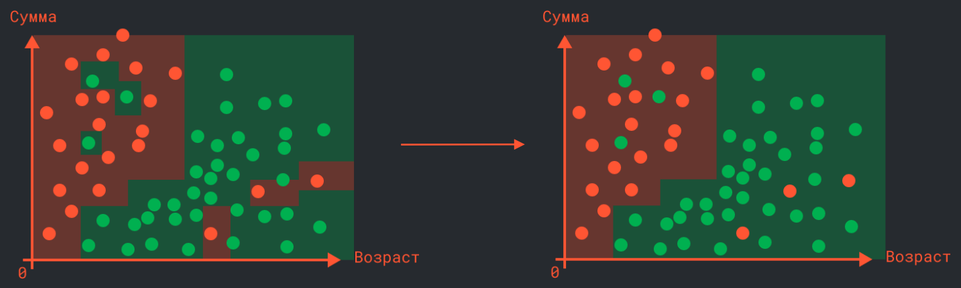


DECISION TREES

КРИТЕРИИ ОСТАНОВА

- Деревья крайне склонные к переобучению
- В попытке построить лучшее разбиение на трейне, достигается идеальное качество и/или гигантская глубина и/или малое чисто объектов в листах…
- Давайте устанавливать их как гиперпараметры и делать менее жесткими!
- Гиперпараметр 3: объекты в листовой вершине

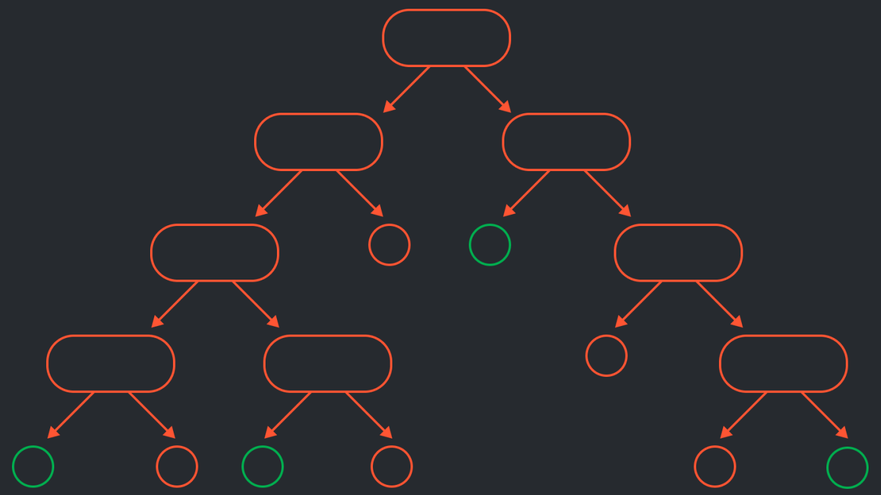


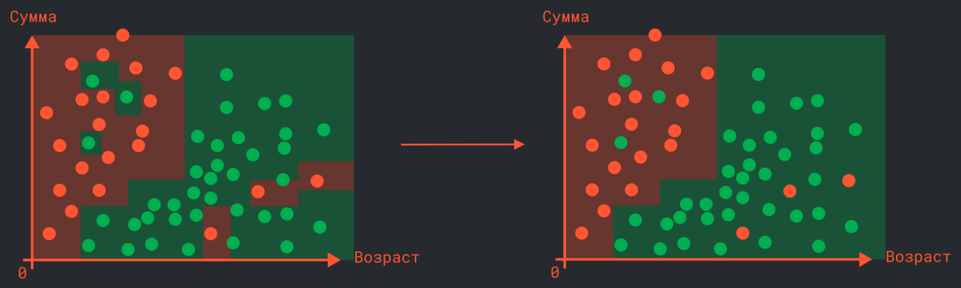


DECISION TREES

КРИТЕРИИ ОСТАНОВА

- Деревья крайне склонные к переобучению
- В попытке построить лучшее разбиение на трейне, достигается идеальное качество и/или гигантская глубина и/или малое чисто объектов в листах…
- Давайте устанавливать их как гиперпараметры и делать менее жесткими!
- Гиперпараметр 4: максимальное число листов

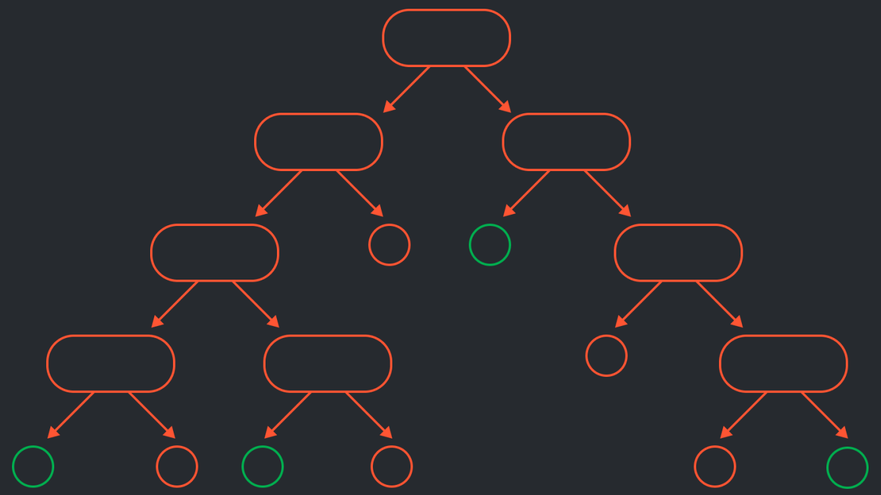


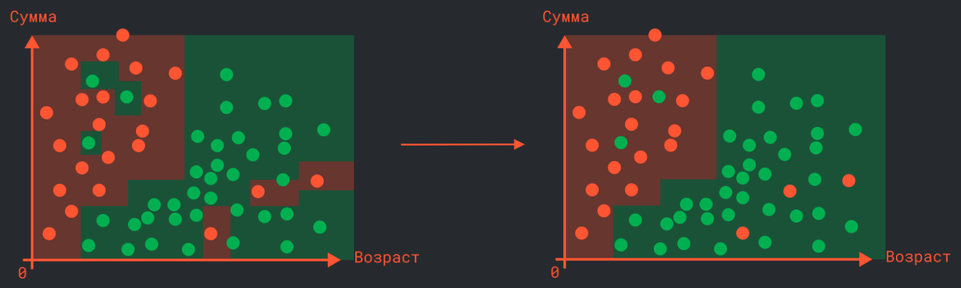


DECISION TREES

КРИТЕРИИ ОСТАНОВА

- Деревья крайне склонные к переобучению
- В попытке построить лучшее разбиение на трейне, достигается идеальное качество и/или гигантская глубина и/или малое чисто объектов в листах…
- Давайте устанавливать их как гиперпараметры и делать менее жесткими!
- Гиперпараметр 5: минимальный прирост качества

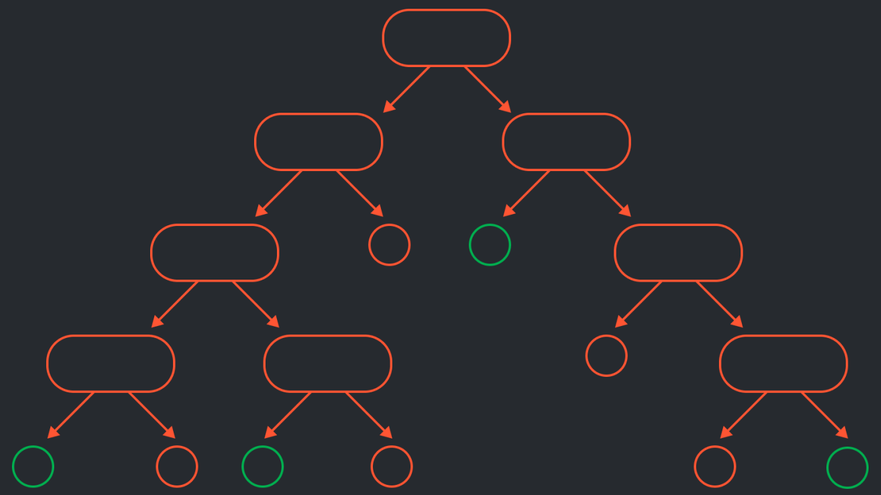


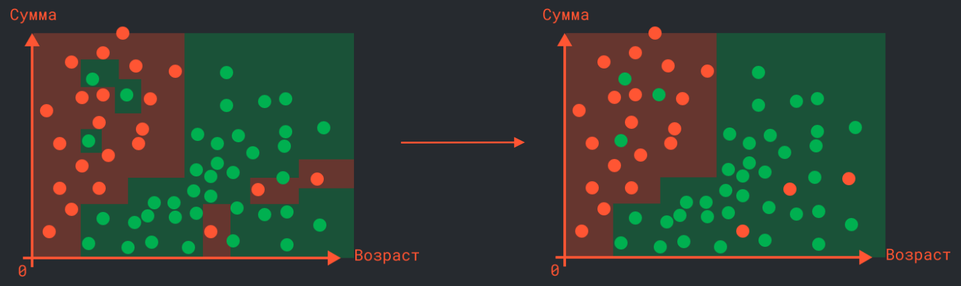


DECISION TREES

КРИТЕРИИ ОСТАНОВА

- Гиперпараметр 1: глубина дерева
- Гиперпараметр 2: объекты в внутренней вершине
- Гиперпараметр 3: объекты в листовой вершине
- Гиперпараметр 4: максимальное число листов
- Гиперпараметр 5: минимальный прирост качества

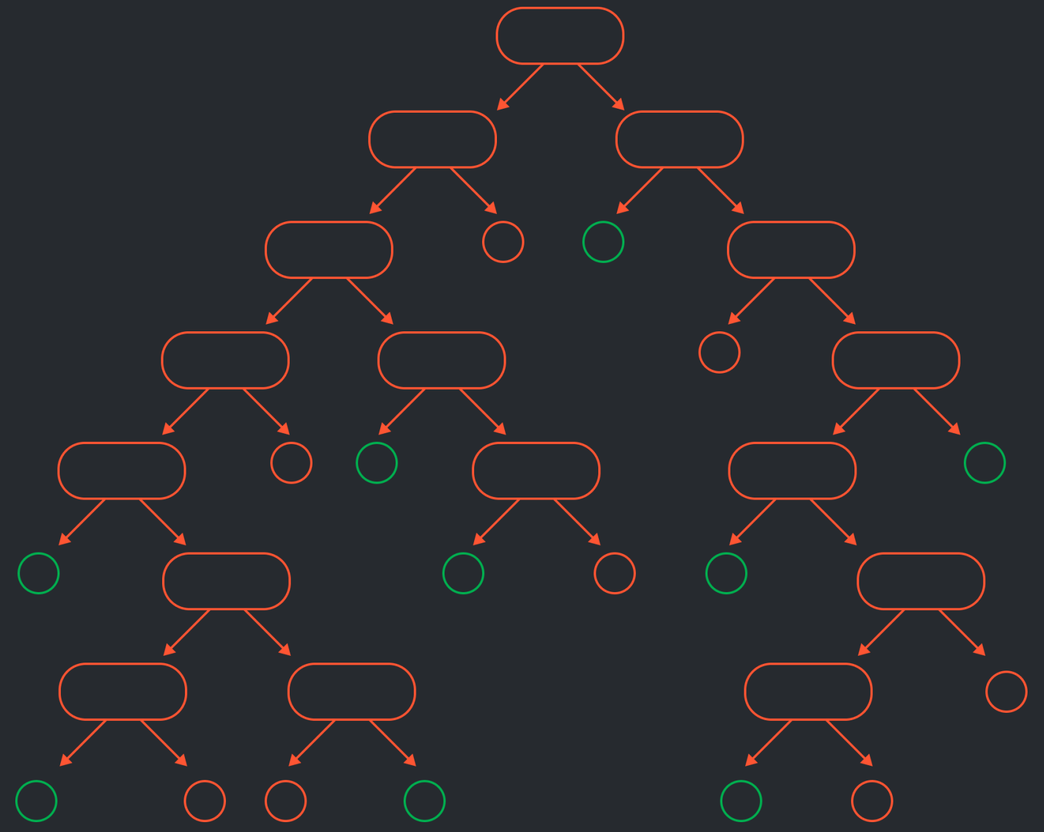




DECISION TREES

СТРИЖКА ДЕРЕВЬЕВ

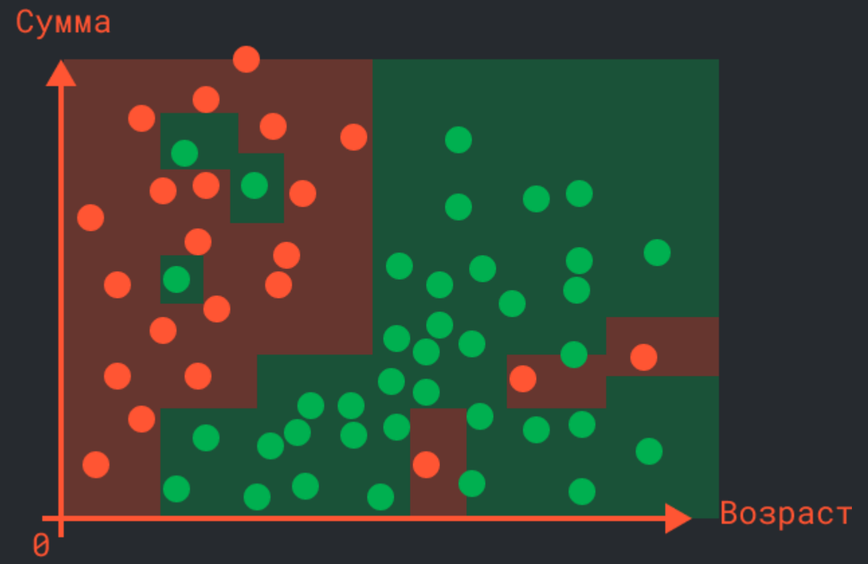
- Есть еще другой подход
- В начале построим глубокое переобученное дерево
- А потом возьмем в руки ножницы и пострижем его!
- Достаточно эвристические механизмы
- Один из подходов cost-complexity pruning
- Не очень популярный метод, мало в каких библиотеках можно найти реализацию



DECISION TREES

СВЯЗЬ С ЛИНЕЙНЫМИ МОДЕЛЯМИ

- Повторим: как деревья дают прогноз?
- Делят признаковое пространство на множество областей $D_1, D_2, \dots D_k$
- Каждой области соответствует прогноз w_j, как усредненное значение или доли классов на трейне
- $a(x) = w_1 \cdot [x \in D_1] + w_2 \cdot [x \in D_2] + \cdots + w_k \cdot [x \in D_k]$
- Это линейная модель над комбинациями бинарных признаков, созданных на основе базовых!



РЕЗЮМЕ

— Рассмотрели ряд гиперпараметров дерева сквозь призму критериев останова

— Кажется, научились контролировать обобщающую способность деревьев

— Узнали, как решающие деревья связаны с линейными моделями

— Пора к практике!

СПАСИБО

ТАБАКАЕВ НИКИТА