ТИНЬКОФФ

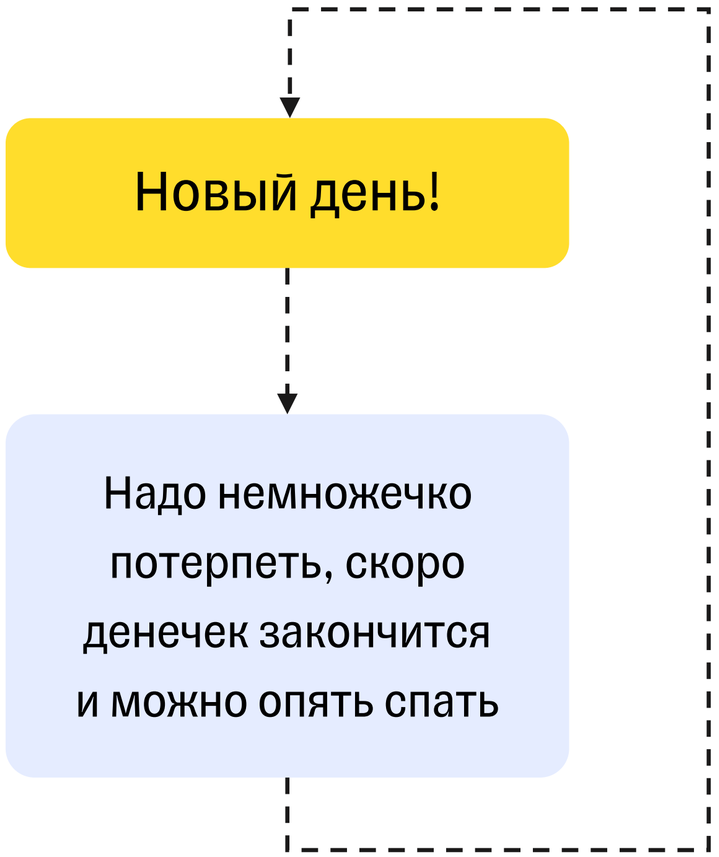
Лекция 9

Деревья решений

Как мы принимаем решения

Что строим в голове
для принятия решения?

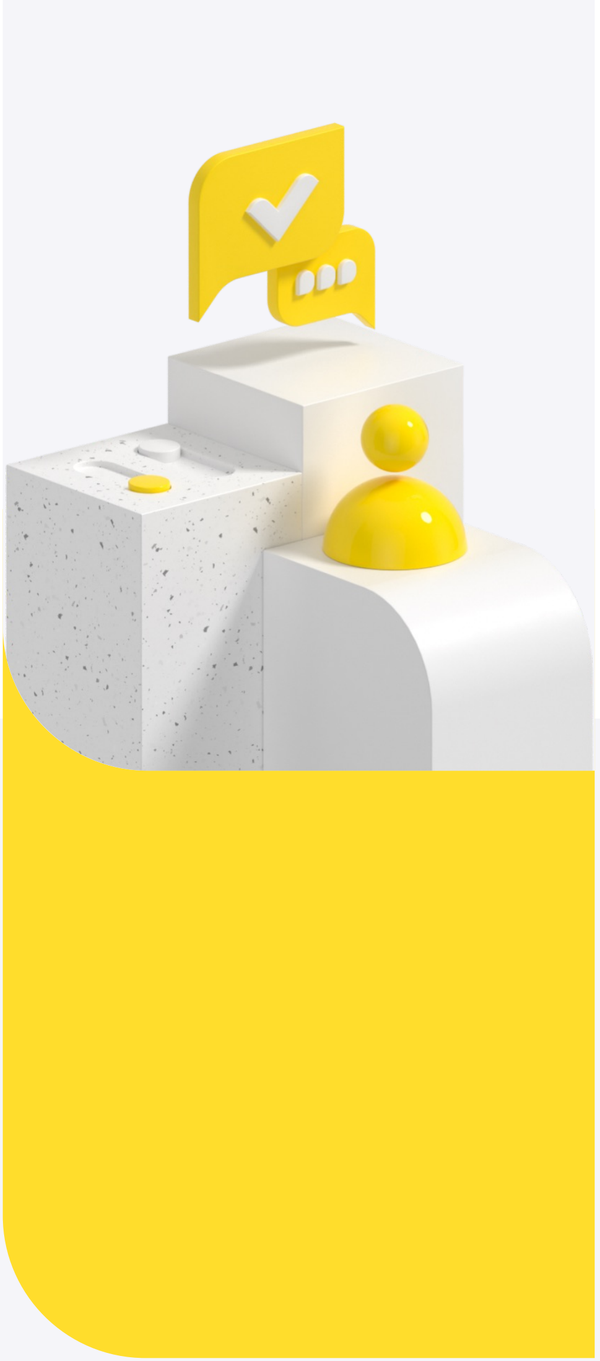
- Что-то похожее на дерево или на граф
- Блок-схемы и прочие UML-диаграммы растут отсюда же
- Хотелось бы, чтобы такие схемы собирались сами



Как работают деревья решений

Играем в «данетки»

- Ведущий загадывает объект
- Игрок задает закрытый вопрос (ответ либо да, либо нет)



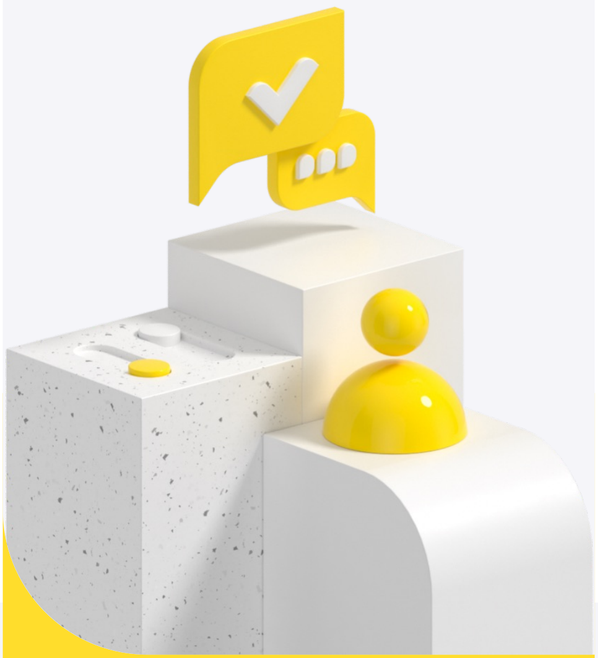
Угадывание удобно описывать деревом

- Придумали «хороший» вопрос
- Для каждого из вариантов («да» или «нет») повторили
- Остановились, когда все объекты стали различимы

Как работают деревья решений

В случае формальных задач смысл сохраняется

- Даны N объектов, для каждого объекта задана категория
- Всего K категорий



- В вершинах дерева поместим вопросы к признакам объектов
- Глядя на ответ, пускаем объект по одной из ветвей вершины
- Делаем так, пока не придём в лист дерева

Как понять, что мы задали хороший вопрос?

Как понять, что итоговое дерево хорошее?

Оптимальность деревьев

Каким свойством должны обладать вопросы?

Хотим такое дерево, чтобы его точность была как можно выше

При этом слишком глубокие деревья нам не подойдут

N может быть очень большим

Для каждого объекта хотим путь примерно одинаковой длины

Оптимальность деревьев

Как формализовать понятие однородности?

Хотим повышать однородность объектов в вершинах дерева

Делим правильно – получаем однородную выборку в листе

Понятие энтропии

Проведём аналогию с предсказуемостью системы

01

В коробке 4096 шаров
уникальных цветов
(мы знаем это заранее)

- Достаем 1 случайный шар
- С какой вероятностью
  мы угадаем его цвет?
- Насколько хаотична
  эта система?

02

В коробке 4096 шаров, все
шары черные
(мы знаем это заранее)

- Достаем 1 случайный шар
- С какой вероятностью мы
  угадаем его цвет?
- Насколько хаотична
  эта система?

О поведении какой
из систем мы имеем
больше информации?

Понятие энтропии

Знакомо понятие энтропии из школьного курса физики

Чем больше энергии в системе, тем выше её энтропия

Чем более упорядочена система, тем меньше энтропия

Чем меньше разных классов, тем меньше энтропия выборки

В идеале хотим нулевую энтропию

?

Может ли энтропия повыситься при непра-вильном разбиении?

Понятие энтропии

Формула информационной энтропии:

$$H(X) = -\sum_{c \in \{c_1, c_2, \ldots, c_N\}} p(x \in c) \log_2 p(x \in c)$$

Похоже на матожидание

Откуда логарифм?

Хотим обратную связь между p_i и величиной энтропии $H(X)$

Отрицательный логарифм дает эту связь

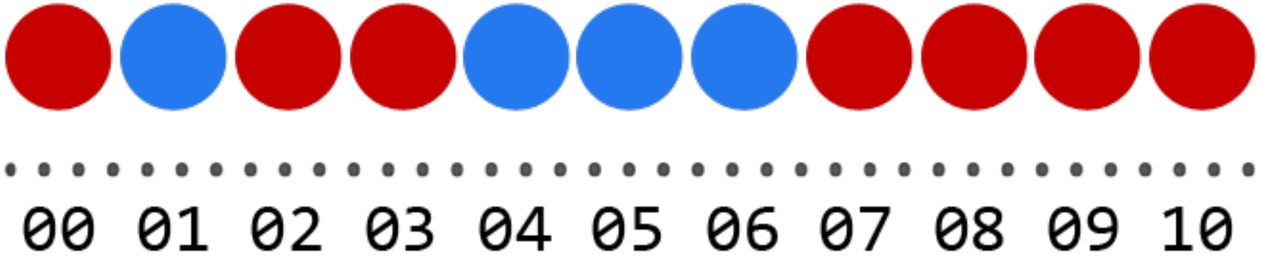
- $H(X) \to 0$ при $p_i \to 1$
- $H(X) \to \infty$ при белом шуме (много равновероятных событий)

Прирост информации

Information Gain

Как понять, что энтропия выборки уменьшилась?

В выборке было 7 красных и 4 синих шара

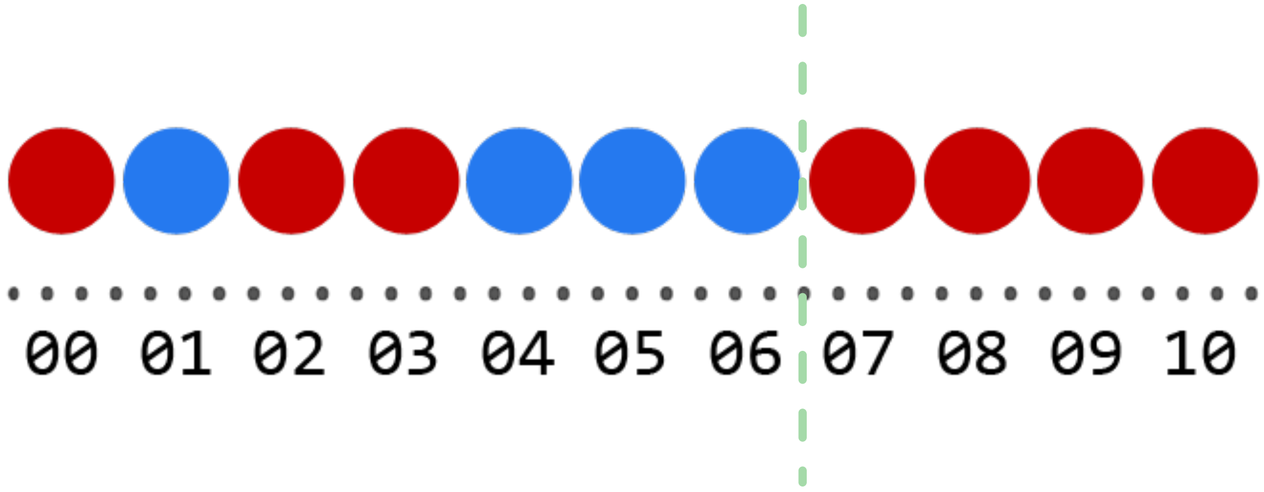


Прирост информации

Information Gain

Разбили выборку вершиной дерева

- Для каждого шара задали вопрос про его признаки
- Слева оказалось 3 красных и 4 синих шара
- Справа оказалось 4 красных шара



Уменьшилась ли общая энтропия?

Прирост информации

Information Gain

Слева энтропия увеличилась, справа стала равной нулю

Как поменялась общая энтропия?

Посчитаем исходную энтропию:

$$H(X) = -\frac{4}{11}\log\frac{4}{11} - \frac{7}{11}\log\frac{7}{11} \approx 0.946$$

- Энтропия слева после разбиения:

$$H_l(X) = -\frac{3}{7}\log\frac{3}{7} - \frac{4}{7}\log\frac{4}{7} \approx 0.985$$

- Энтропия справа:

$$H_r(X) = -\frac{0}{4}\log\frac{0}{4} - \frac{4}{4}\log\frac{4}{4} \approx 0$$

Давайте посчитаем взвешенное среднее

$$H^{l,r}(X) = H_l(X)\times 0.636 + H_r(X)\times 0.364 \approx 0.627$$

Оптимизация дерева

Как построить хорошее дерево?

- Найдём разбиение, дающее максимальный прирост информации
- Для каждой из подвыборок поступим так же

Как будем перебирать разбиения?

- Будем перебирать все признаки
- Для каждого признака переберём т.н. границу разбиения

Перебор непрерывных признаков

01

Выберем диапазон перебора (по умолчанию от min до max)

02

Выберем шаг (или степень дискретизации)

03

Переберём с этим шагом значения признака

04

Для каждого перебираемого значения разобьем выборку

05

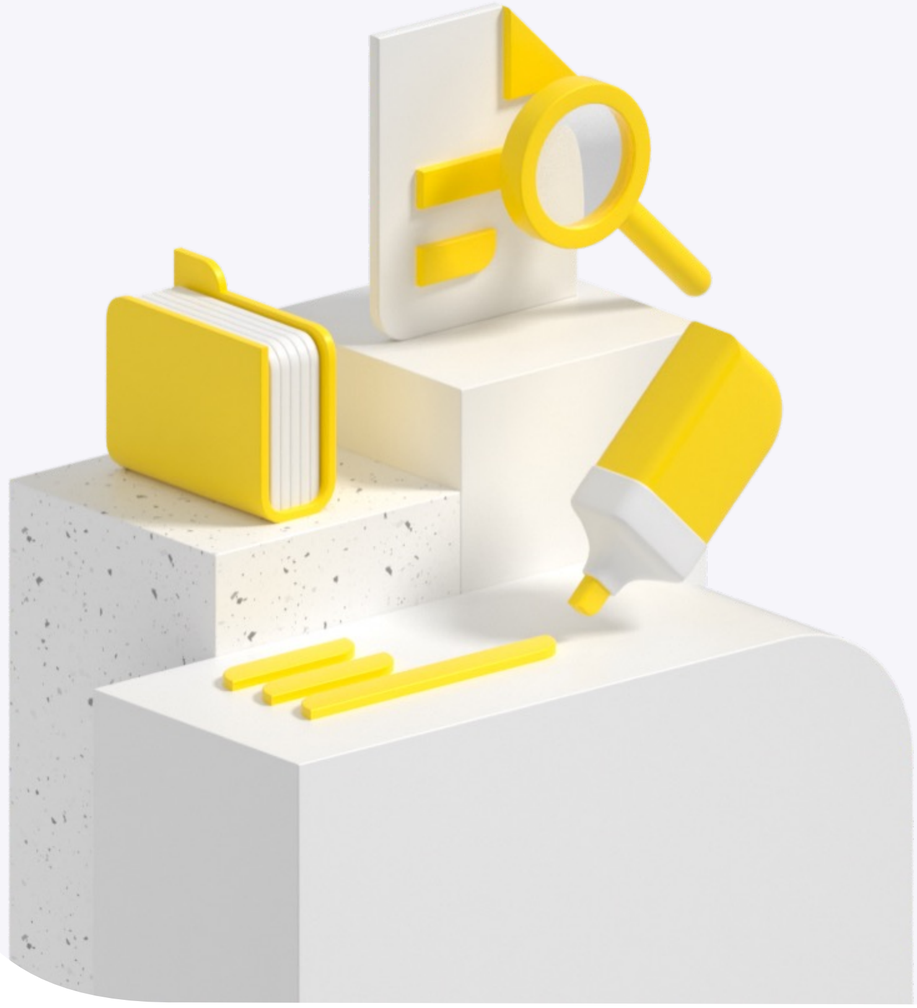
Посчитаем прирост информации

06

Выберем границу по максимальному приросту информации

Перебор категориальных признаков

- Если признак ординальный, можно разбивать аналогично непрерывным признакам
- В этом случае шаг будет равен 1
- Иначе закодируем этот признак через one-hot
- Для каждой one-hot колонки сделаем аналогичный перебор



Какие есть алгоритмы?

Польза от знания названий – можно быстро ответить на собесе

ID3

Алгоритм
для вещественных
признаков

C4.5

Алгоритм
для вещественных
и категориальных

CART

То же самое,
но ещё и для задачи
регрессии

Как еще можно оптимизировать?

Идеи похожи на энтропию

Gini Impurity

По-русски **НЕЛЬЗЯ** называть критерием Джини

Штрафуем, если в выборке нет доминантного класса

$$G = 1 - \sum_k (p_k)^2$$

Misclassification error

Почти то же, что в Gini Impurity

$$E = 1 - \max_k p_k$$

Регрессионные деревья

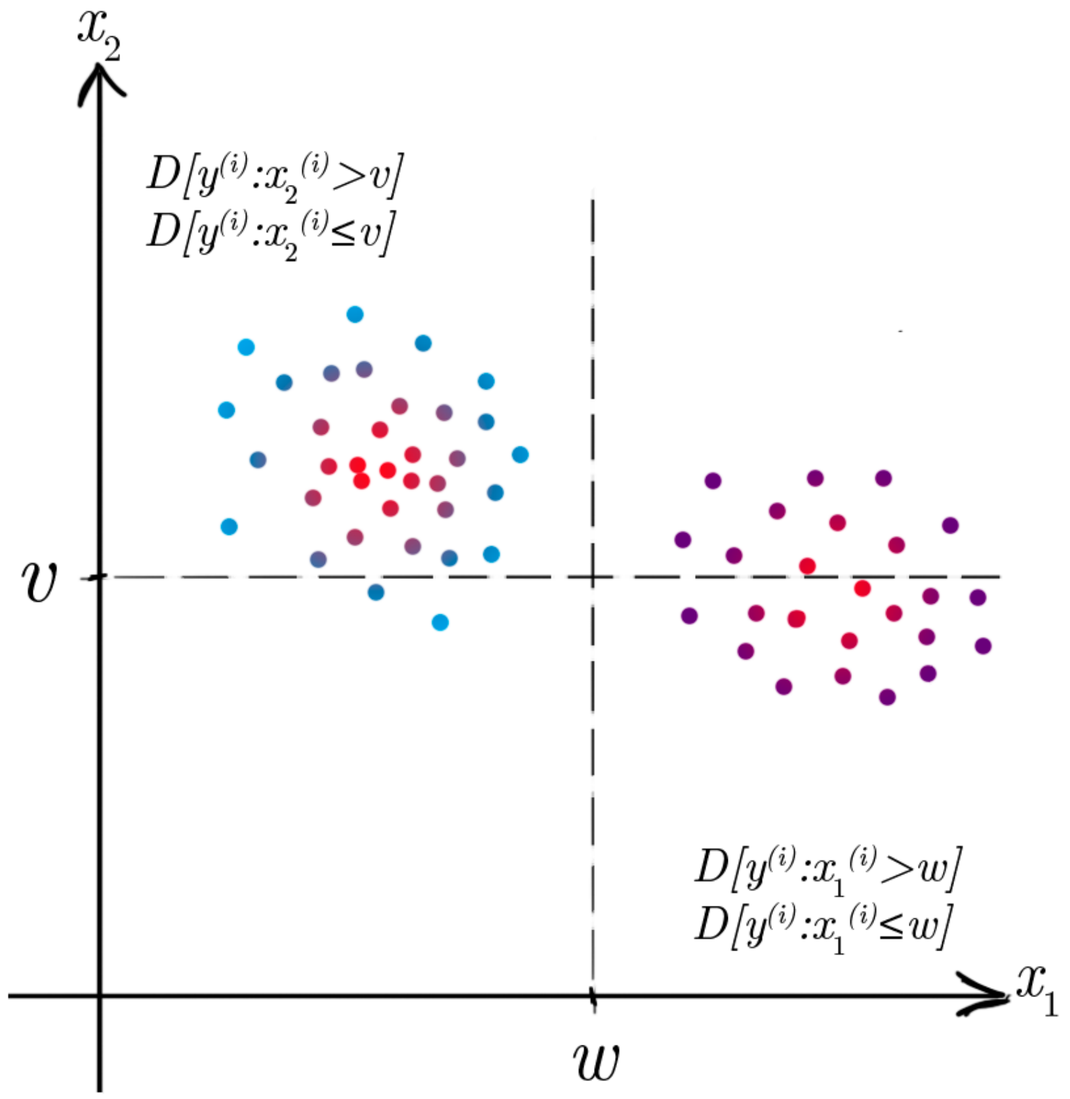
- Как находим класс в случае с классифицирующими деревьями?
- Чему эквивалентны примеси других классов в листе?
- Чему эквивалентен главный класс в листе?

Регрессионные деревья

- Хотим уменьшить дисперсию в листе
- Если много разнородных примеров, дисперсия увеличится
- Чем больше примеров при меньшей дисперсии, тем лучше
- В листе предскажем матожидание

Регрессионные деревья

- Разбили по признаку x_1 и границе w
- Посчитали дисперсию слева и справа
- Разбили по признаку x_2 и границе v
- Ещё раз посчитали дисперсию
- Выбираем вариант, где дисперсия уменьшается сильнее



Проблемы деревьев

01

Решение
глобально
не оптимальное

02

Оптимальное
решение требует
полный перебор
за экспоненту

03

Склонны
к переобучению,
чувствительны
к шумам

Что можно сделать,
чтобы деревья
не переобучались?

04

Разделяющая
граница имеет
ограничения

Как думаете,
какие?

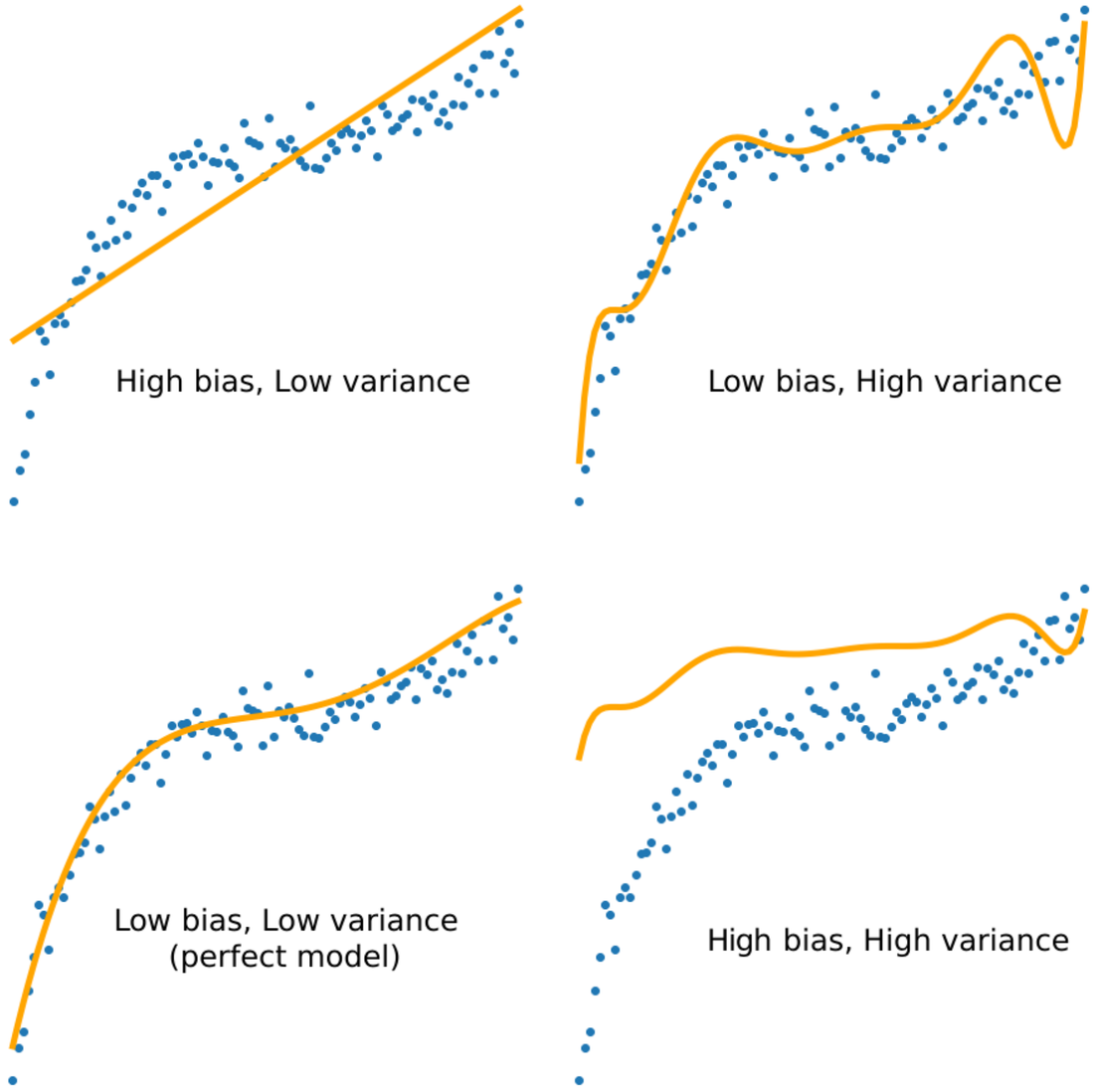
05

Деревья не могут
экстраполировать

Удобно использовать
в регрессии
на ограниченном
множестве

Свойства деревьев

- Разбирали bias-variance
- К чему относится дерево?
- Можно ли улучшить результат, взяв несколько деревьев?



ТИНЬКОФФ

Вопросы?