

Решающие деревья

В основе - слайды Евгения
Соколова

НИУ ВШЭ, 2021

Интерпретация моделей

- Можно интерпретировать веса признаков, если признаки масштабированы
- В остальном все сложно

Предсказание цены квартиры

- Признаки: площадь, этаж, число комнат

$$a(x) = 10 * \text{площадь} + 1.1 * \text{этаж} + 20 * \text{число комнат}$$

- Квадратичные признаки: будут работать лучше, как интерпретировать совсем непонятно

$$a(x) = 10 * \text{площадь} + 1.1 * \text{этаж} + 20 * \text{число комнат} - 0.2 * \text{этаж}^2 + 0.5 * \text{площадь} * \text{число комнат} + \dots$$

Логические правила

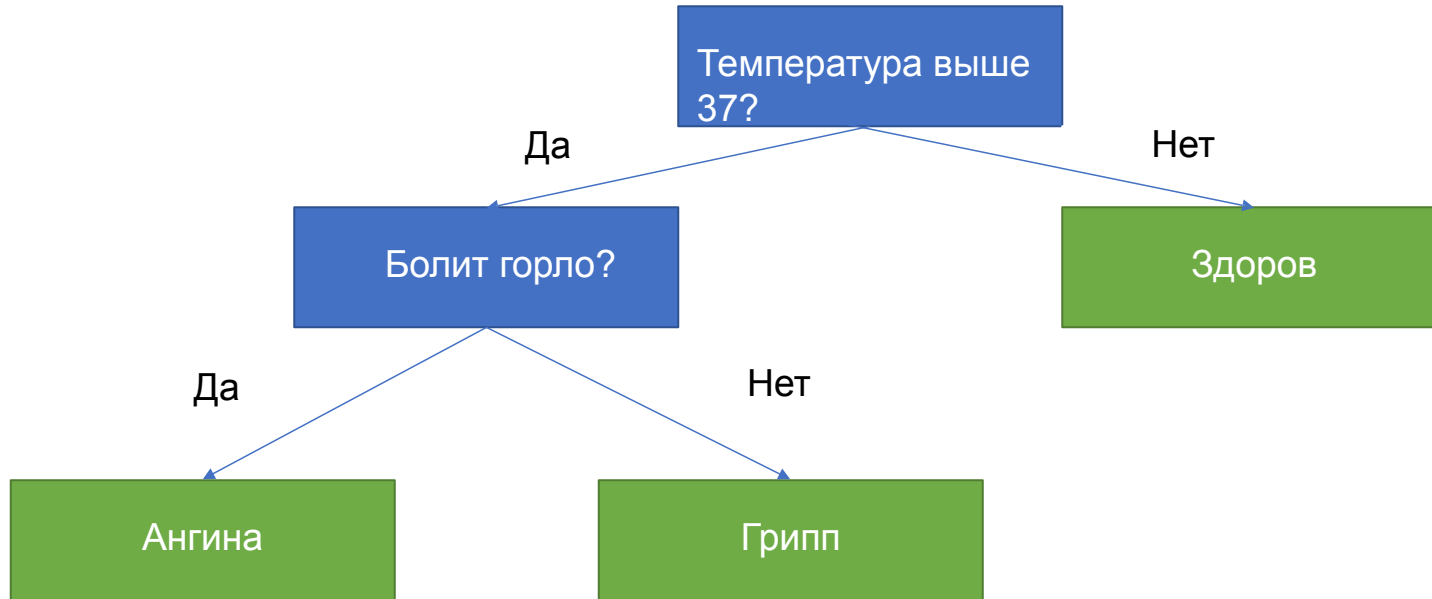
[этаж > 3][площадь < 40][число комнат < 3]

- Легко объяснить заказчику (если ≤ 5 условий)
- Позволяют извлекать знания из данных
- Не факт, что оптимальны с точки зрения качества

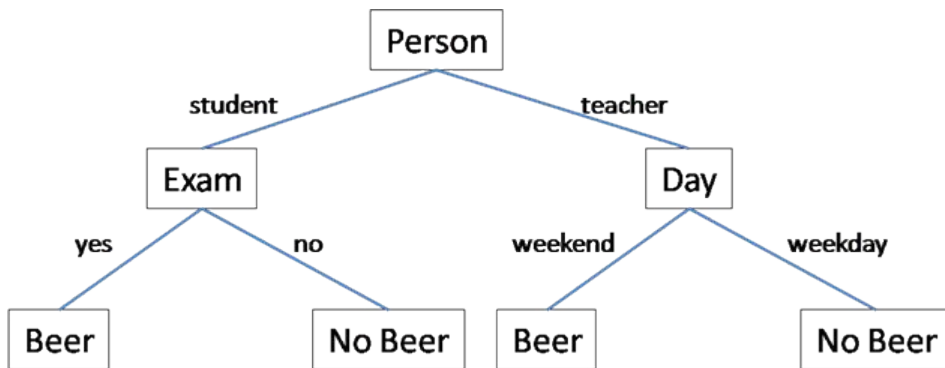
Логические правила

- Как строить?
- Линейные модели
- Перебор, жадное наращивание
- Решающие деревья

Медицинская диагностика

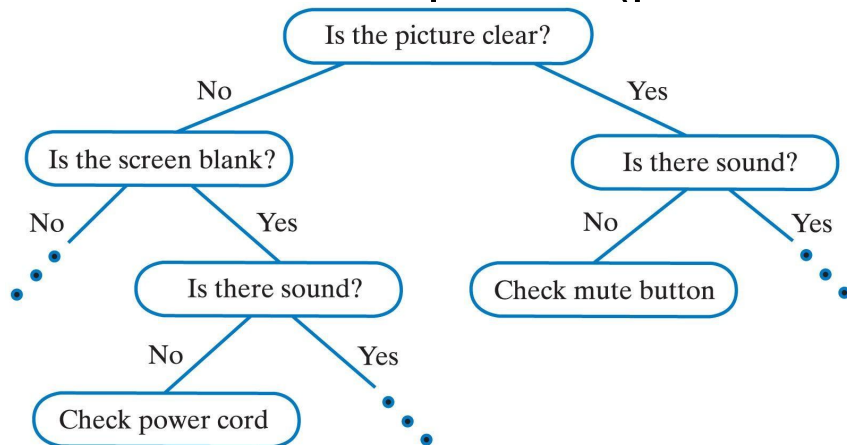


Принятие решений



Решающее дерево

- Бинарное дерево (**ровно 2** дочерних узла)
- В каждой внутренней вершине записано условие
- В каждом листе записан прогноз (решение)



Условия

- Самые популярные варианты:

$$[x^j \leq t] \text{ и } [x^j = t]$$

Примеры:

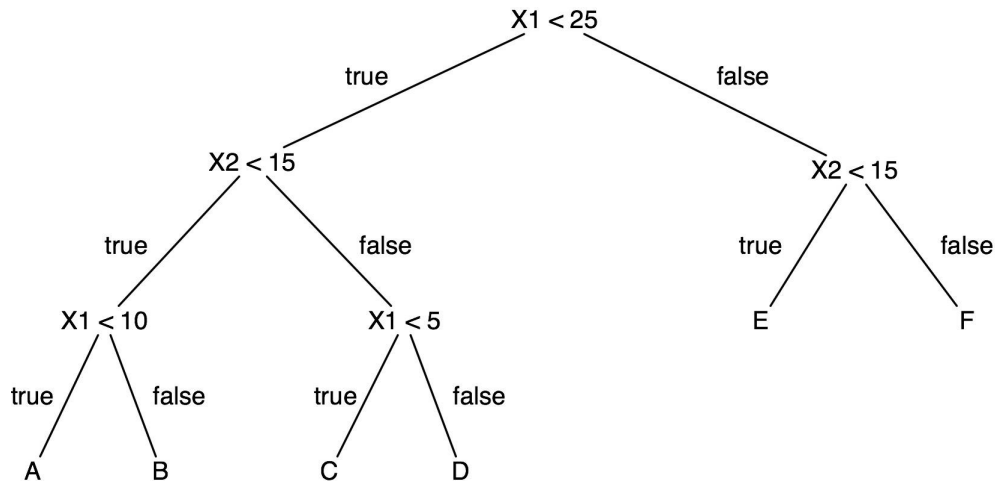
- [этаж = 5]
- [площадь \leq 30]

Прогноз в листе

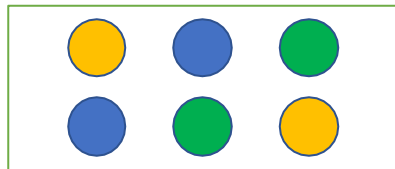
- Регрессия:
 - Вещественное число
- Классификация:
 - Класс
 - Вероятности классов

Жадное построение

- Как правило используется жадный алгоритм построения
- Растим дерево от корня к листьям

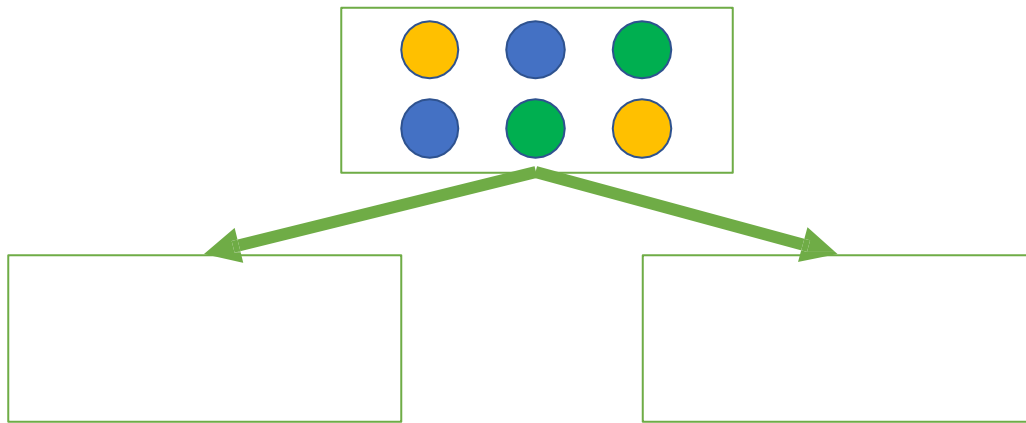


Жадное построение

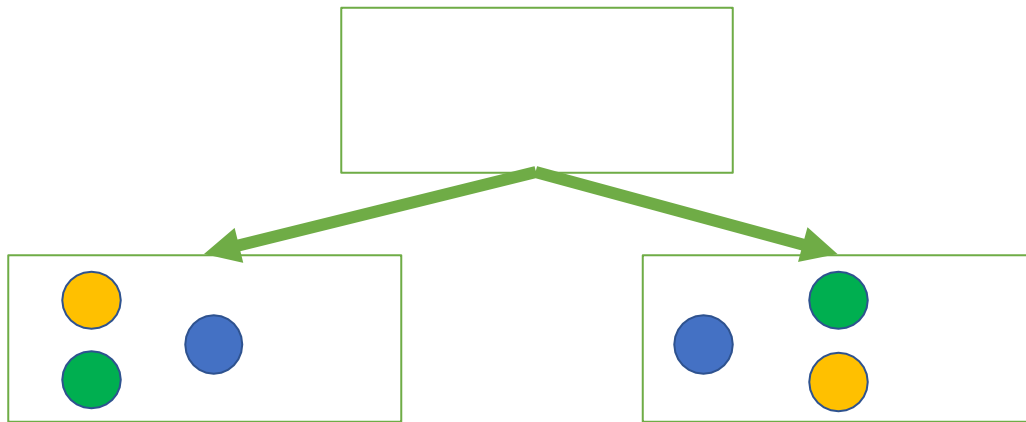


- Как разбить вершину?

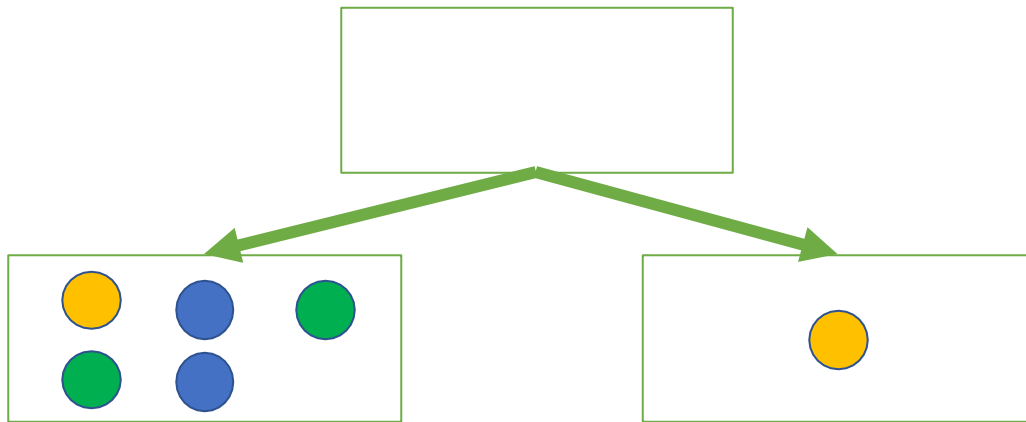
Жадное построение



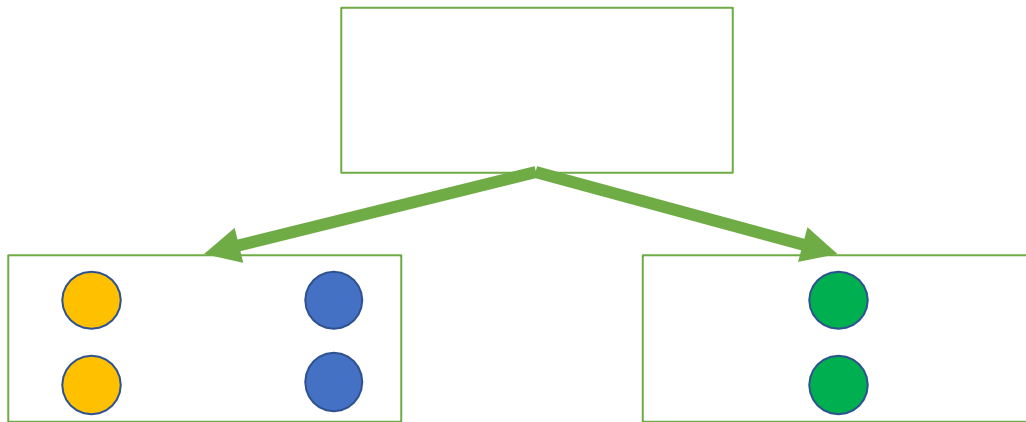
Жадное построение



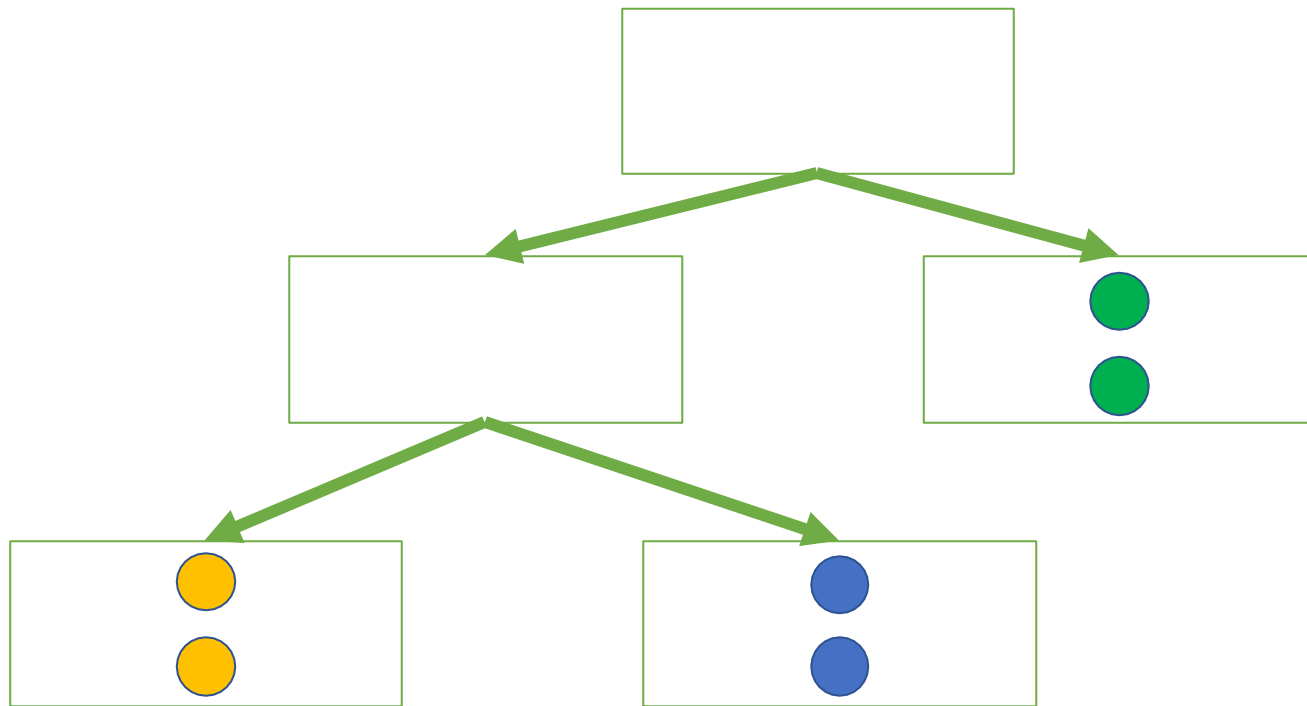
Жадное построение



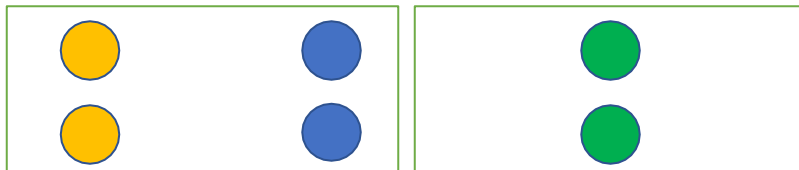
Жадное построение



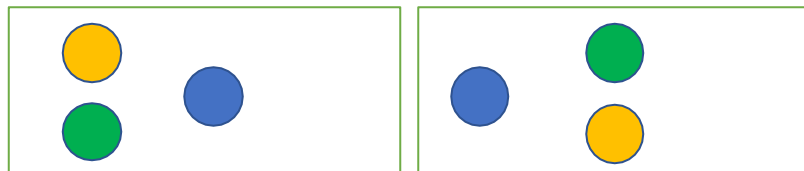
Жадное построение



Как сравнить разбиения?

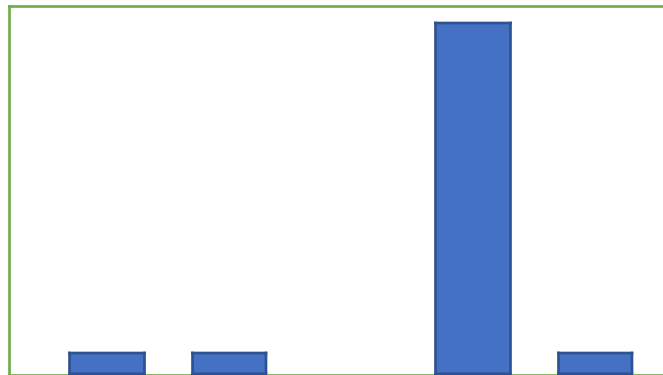
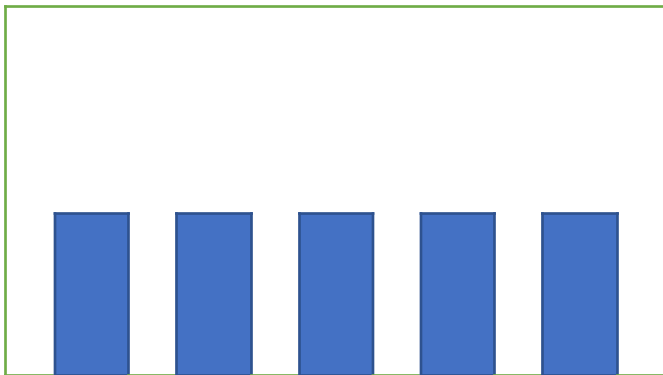


или



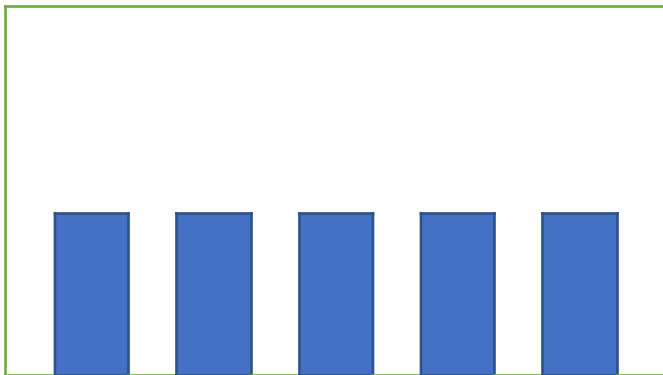
Энтропия

- Мера неопределённости распределения

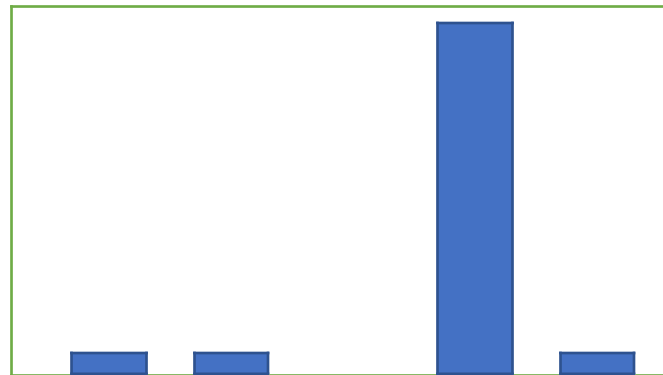


Энтропия

- Мера неопределённости распределения



Высокая энтропия



Низкая энтропия

Энтропия

- Дискретное распределение
- Принимает n значений с вероятностями p_1, \dots, p_n

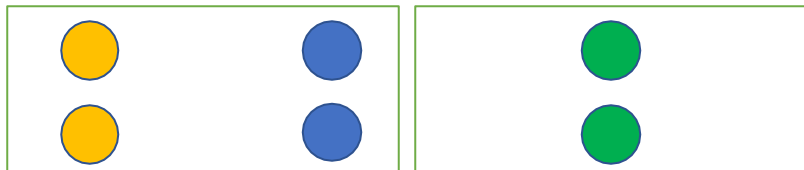
- Энтропия:

$$H(p_1, \dots, p_n) = - \sum_{i=1}^n p_i \log p_i$$

Энтропия

- 5 значений, с вероятностями $(p_1, p_2, p_3, p_4, p_5)$
- $(0.2, 0.2, 0.2, 0.2, 0.2)$ • $(0.9, 0.05, 0.05, 0, 0)$ • $(0, 0, 0, 1, 0)$
- $H = 1.60944 \dots$ • $H = 0.394398 \dots$ • $H = 0$

Как сравнить разбиения?

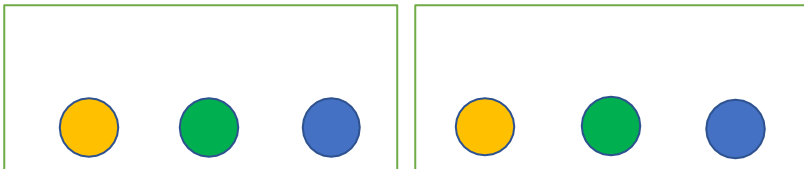


0.693

0

1.09

1.09



- $(0.5, 0.5, 0)$ и $(0, 0, 1)$

- $H = 0.693 + 0 = 0.693$

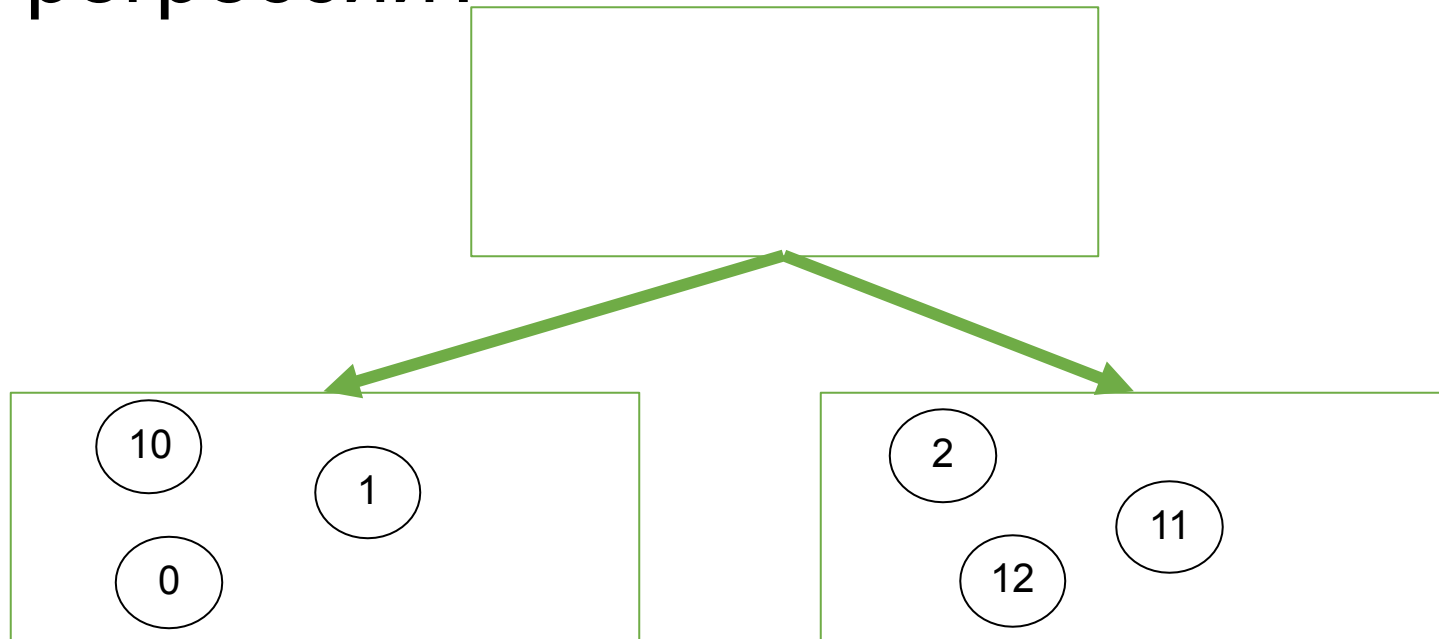
- $(0.33, 0.33, 0.33)$ и $(0.33, 0.33, 0.33)$

- $H = 1.09 + 1.09 = 2.18$

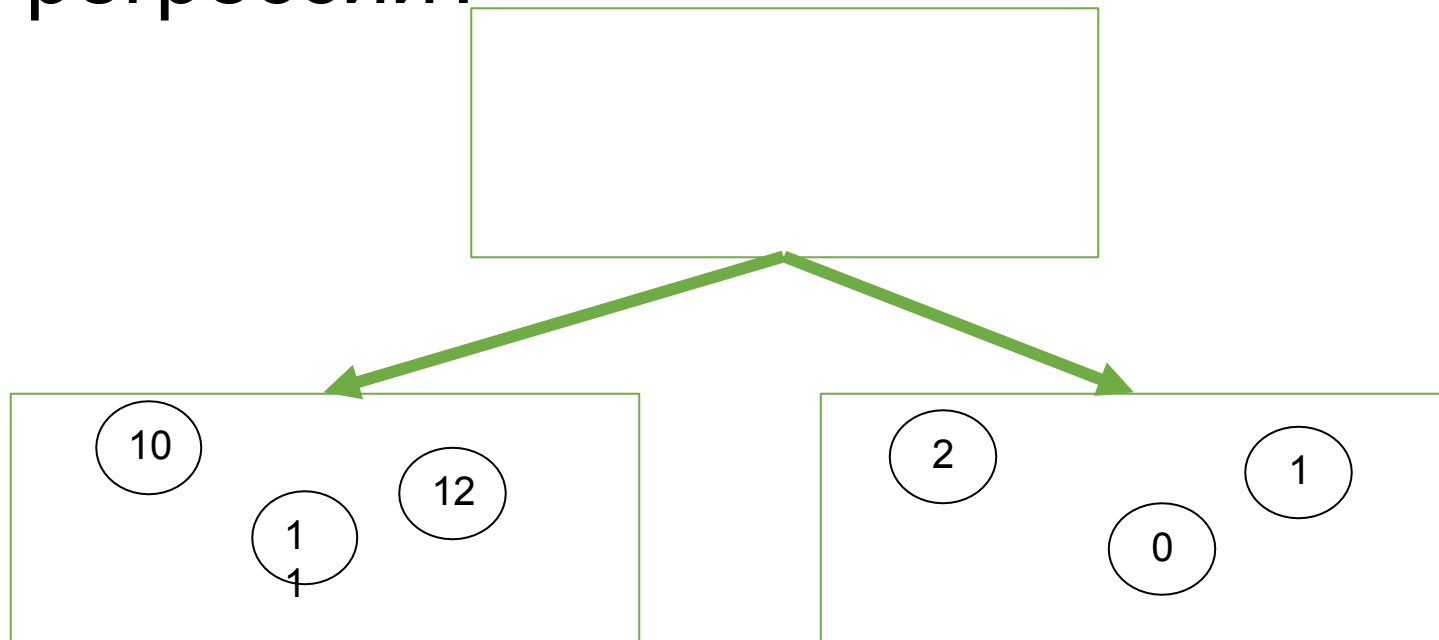
А для
регрессии?

| | | |
|----|---|---|
| 10 | 1 | 2 |
| 12 | 0 | 1 |

А для
регрессии?



А для
регрессии?



А для регрессии?

- Выбираем разбиение с наименьшей суммарной дисперсией
- Чем меньше дисперсия, тем меньше неопределённости

Поиск разбиения

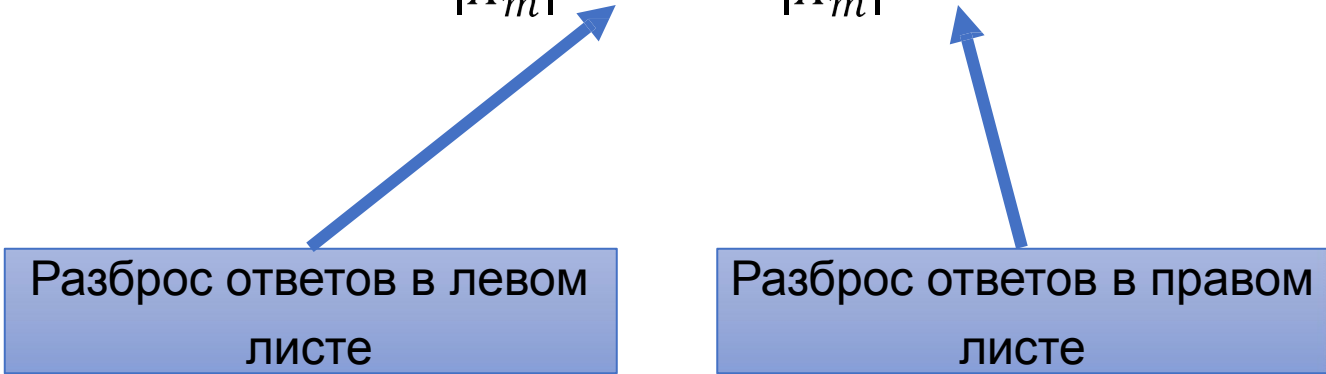
- Пусть в вершине m оказалась выборка X_m
- $Q(X_m, j, t)$ — критерий качества условия $[x^j \leq t]$
- Ищем лучшие параметры j и t перебором:

$$Q(X_m, j, t) \rightarrow \max_{j, t}$$

Критерий качества

$$Q(X_m, j, t) = H(X_m) - \frac{|X_l|}{|X_m|} H(X_l) - \frac{|X_r|}{|X_m|} H(X_r)$$

Разброс ответов в левом
листе



Разброс ответов в правом
листе

Критерий информативности

- $H(X)$
- Зависит от ответов на выборке X
- Чем меньше разброс ответов, тем меньше значение $H(X)$

Критерий информативности

| Impurity | Task | Formula | Description |
|--|----------------|--|--|
| Gini impurity | Classification | $\sum_{i=1}^C f_i(1 - f_i)$ | f_i is the frequency of label i at a node and C is the number of unique labels. |
| Entropy | Classification | $\sum_{i=1}^C -f_i \log(f_i)$ | f_i is the frequency of label i at a node and C is the number of unique labels. |
| Variance / Mean Square Error (MSE) | Regression | $\frac{1}{N} \sum_{i=1}^N (y_i - \mu)^2$ | y_i is label for an instance, N is the number of instances and μ is the mean given by $\frac{1}{N} \sum_{i=1}^N y_i$ |
| Variance / Mean Absolute Error (MAE) (Scikit-learn only) | Regression | $\frac{1}{N} \sum_{i=1}^N y_i - \mu $ | y_i is label for an instance, N is the number of instances and μ is the mean given by $\frac{1}{N} \sum_{i=1}^N y_i$ |

Поиск разбиения

- После того, как разбиение найдено:
- Разбиваем X_m на две части:

$$X_l = \{x \in X_m \mid [x^j \leq t]\}$$

$$X_r = \{x \in X_m \mid [x^j > t]\}$$

- Повторяем процедуру для дочерних вершин

Критерий останова

- В какой момент прекращать разбиение вершин?
- В вершине один объект?
- В вершине объекты одного класса?
- Глубина превысила порог?

Ответ в листе

- Допустим, решили сделать вершину m листом

- Какой прогноз выбрать?

- Регрессия:

$$a_m = \frac{1}{|X_m|} \sum_{i \in X_m} y_i$$

- среднее арифметическое

- Классификация:

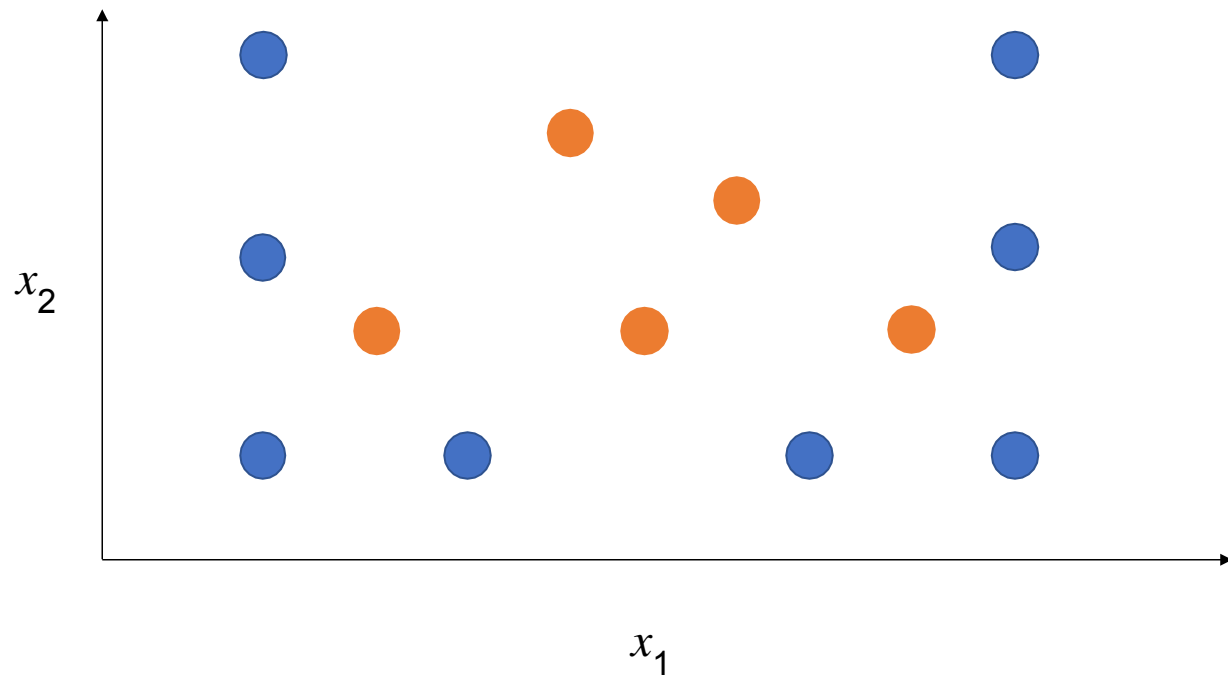
$$a_m = \arg \max_{y \in \mathbb{Y}} \sum_{i \in X_m} [y_i = y]$$

- наиболее частый класс

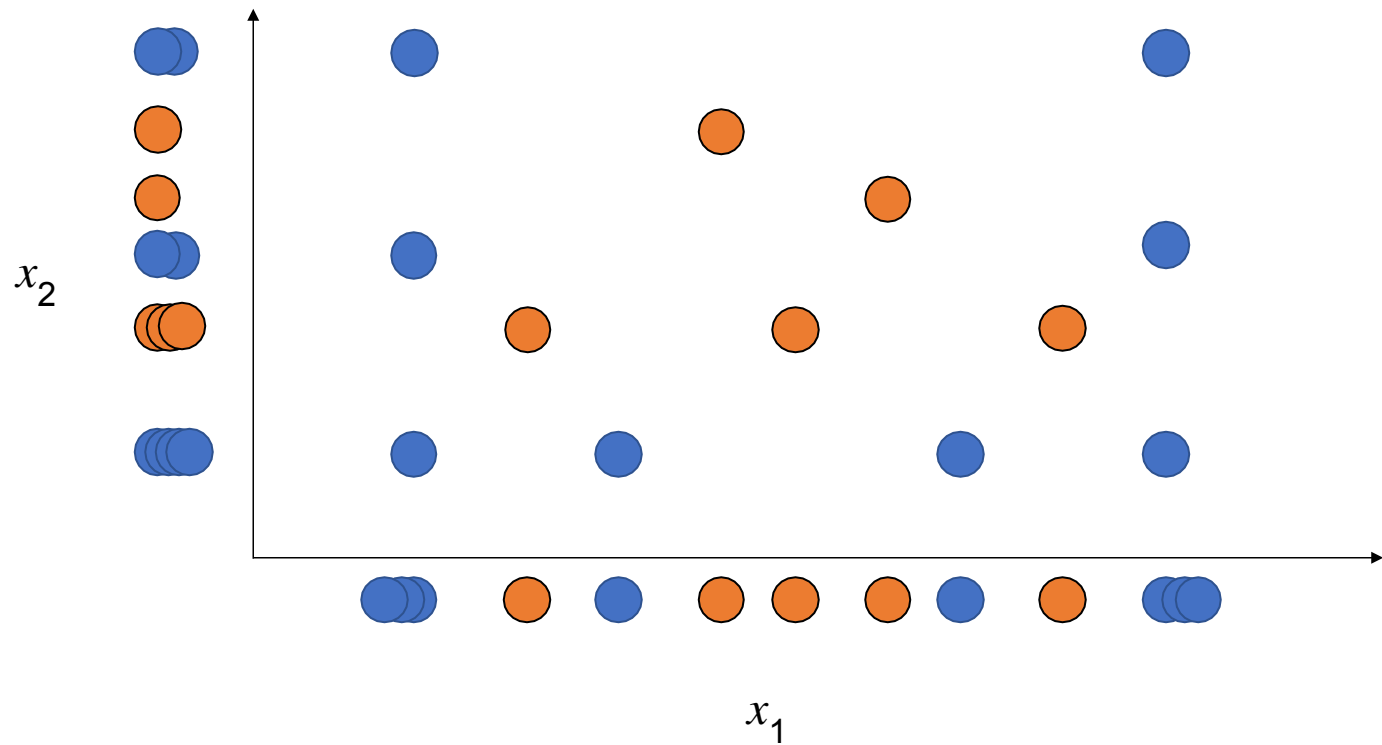
Жадный алгоритм построения дерева

1. Поместить в корень всю выборку: $X_1 = X$
2. Начать построение с корня: $m = 1$
3. Если выполнен критерий останова для вершины m , то выход
4. Найти лучшее разбиение $[x^j \leq t]$ для вершины m
5. Разбить вершину m на дочерние вершины l и r
6. Повторить шаги 3-6 для дочерних вершин l и r

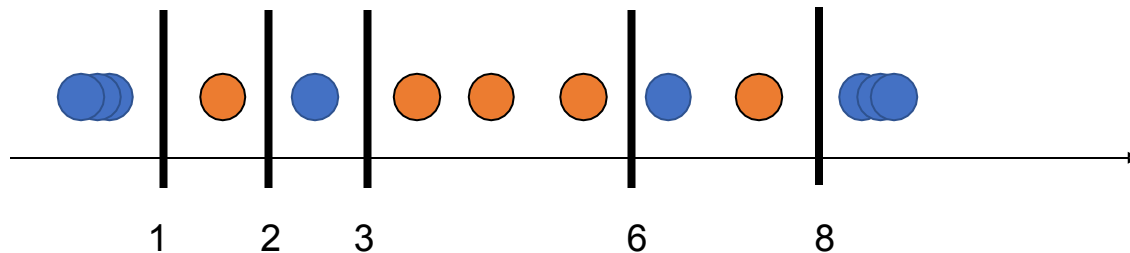
Обучение деревьев



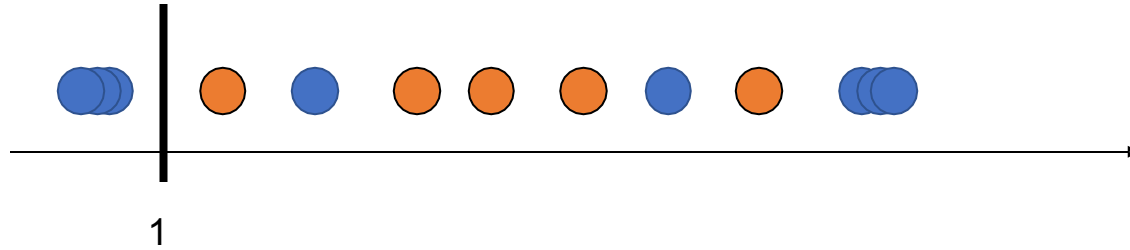
Признаки



Разбиения по признаку 1



Разбиения по признаку 1

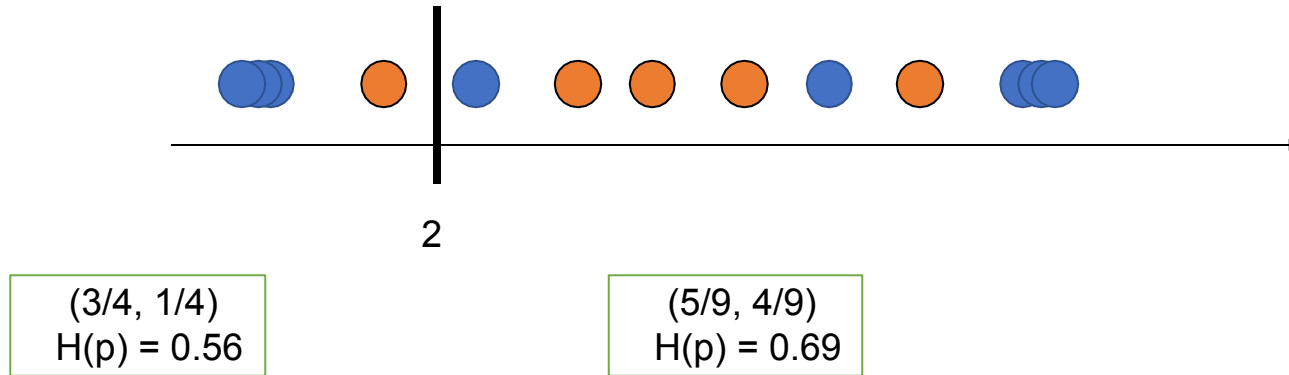


$(1, 0)$
 $H(p) = 0$

$(1/2, 1/2)$
 $H(p) = 0.69$

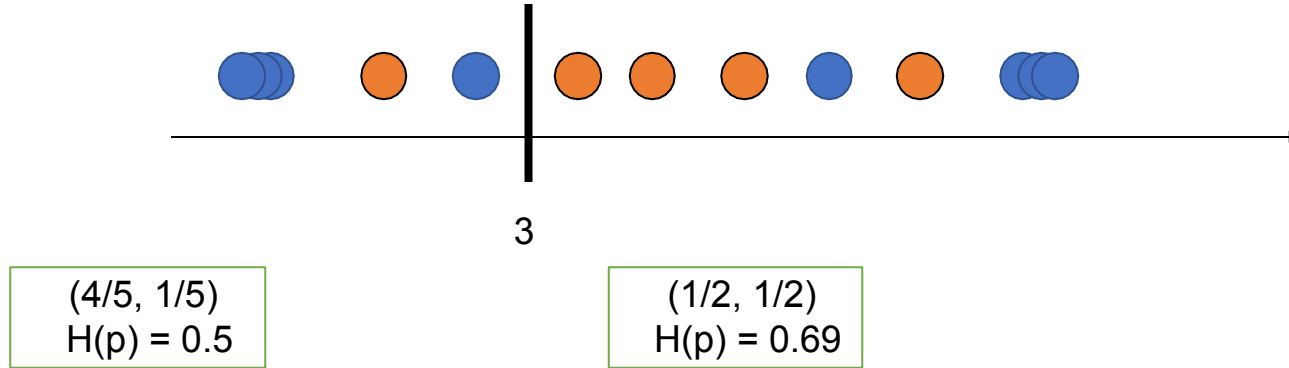
$$\frac{3}{13}H(p_l) + \frac{10}{13}H(p_r) = 0.53$$

Разбиения по признаку 1



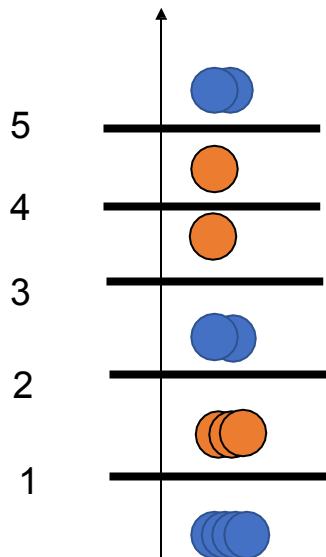
$$\frac{4}{13} H(p_l) + \frac{9}{13} H(p_r) = 0.65$$

Разбиения по признаку 1

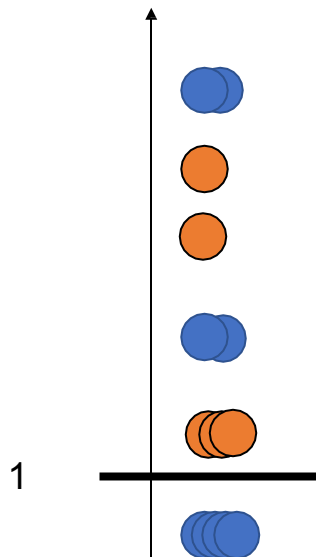


$$\frac{5}{13}H(p_l) + \frac{8}{13}H(p_r) = 0.62$$

Разбиения по признаку 2



Разбиения по признаку 2

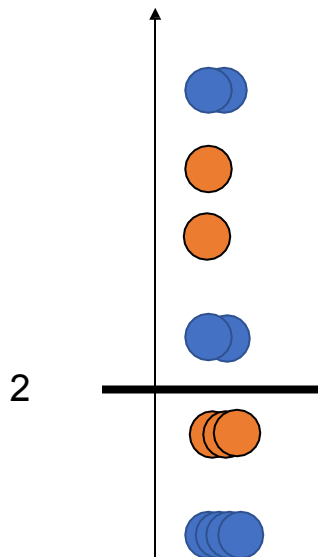


$(4/9, 5/9)$
 $H(p) = 0.69$

$(1, 0)$
 $H(p) = 0$

$$\frac{4}{13}H(p_l) + \frac{9}{13}H(p_r) = 0.47$$

Разбиения по признаку 2

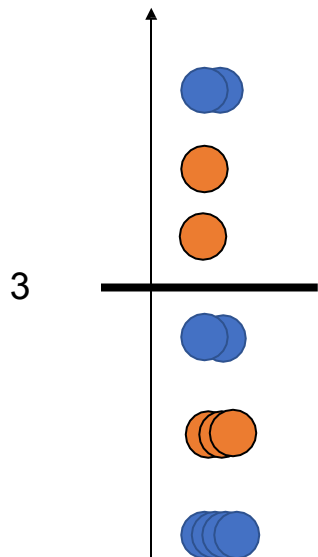


$(4/6, 2/6)$
 $H(p) = 0.64$

$$\frac{7}{13}H(p_l) + \frac{6}{13}H(p_r) = 0.66$$

$(4/7, 3/7)$
 $H(p) = 0.68$

Разбиения по признаку 2

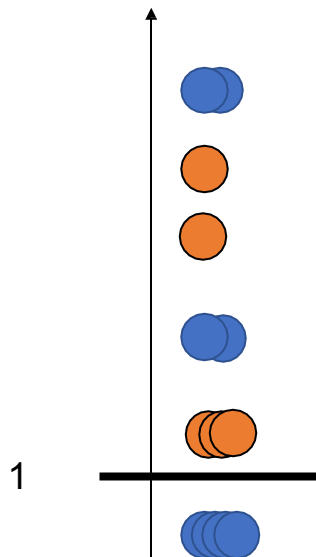


$(1/2, 1/2)$
 $H(p) = 0.69$

$(6/9, 3/9)$
 $H(p) = 0.46$

$$\frac{9}{13}H(p_l) + \frac{4}{13}H(p_r) = 0.53$$

Разбиения по признаку 2



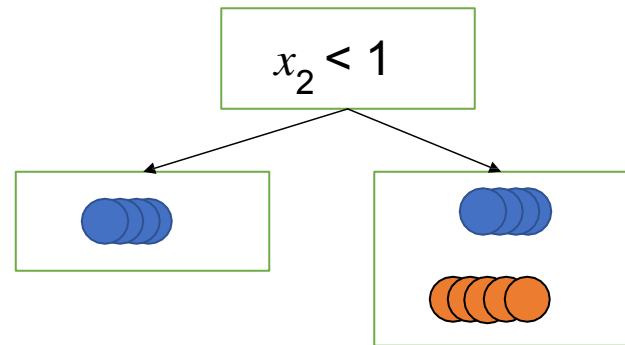
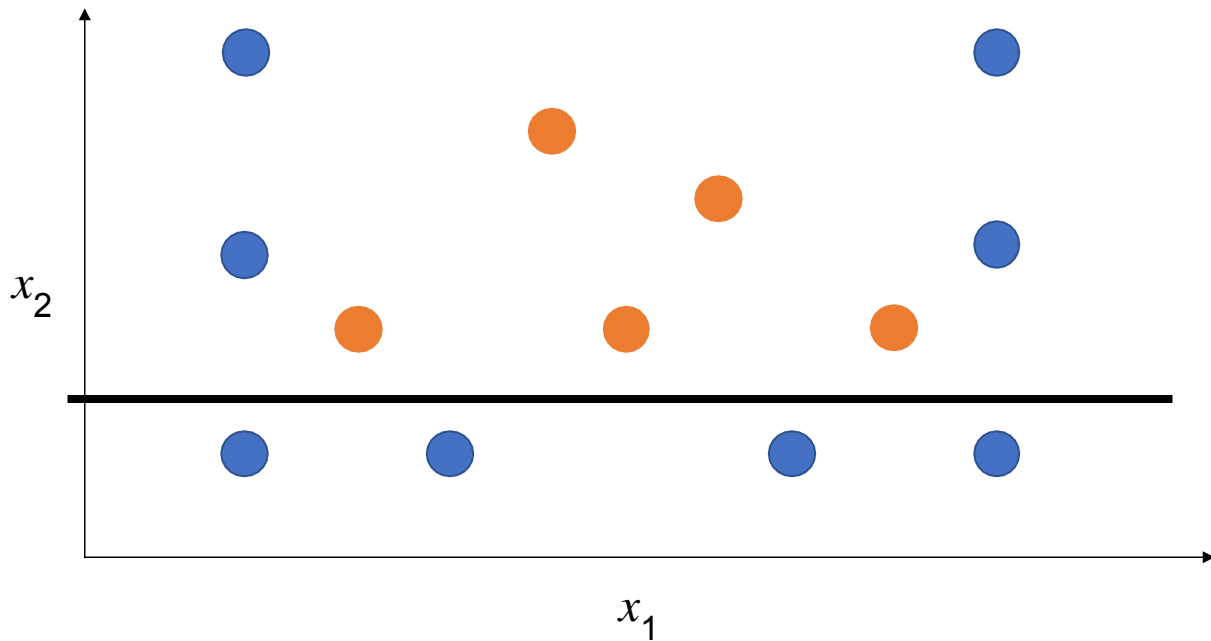
$(4/9, 5/9)$
 $H(p) = 0.69$

$(1, 0)$
 $H(p) = 0$

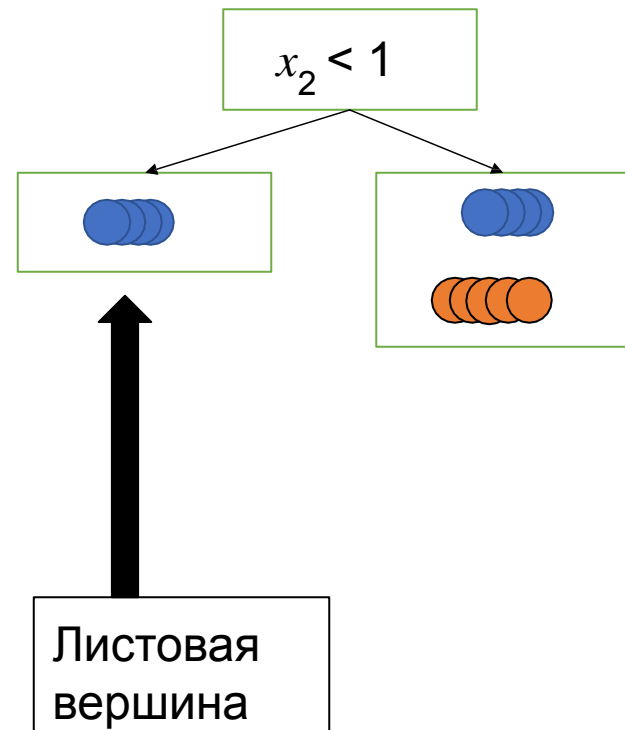
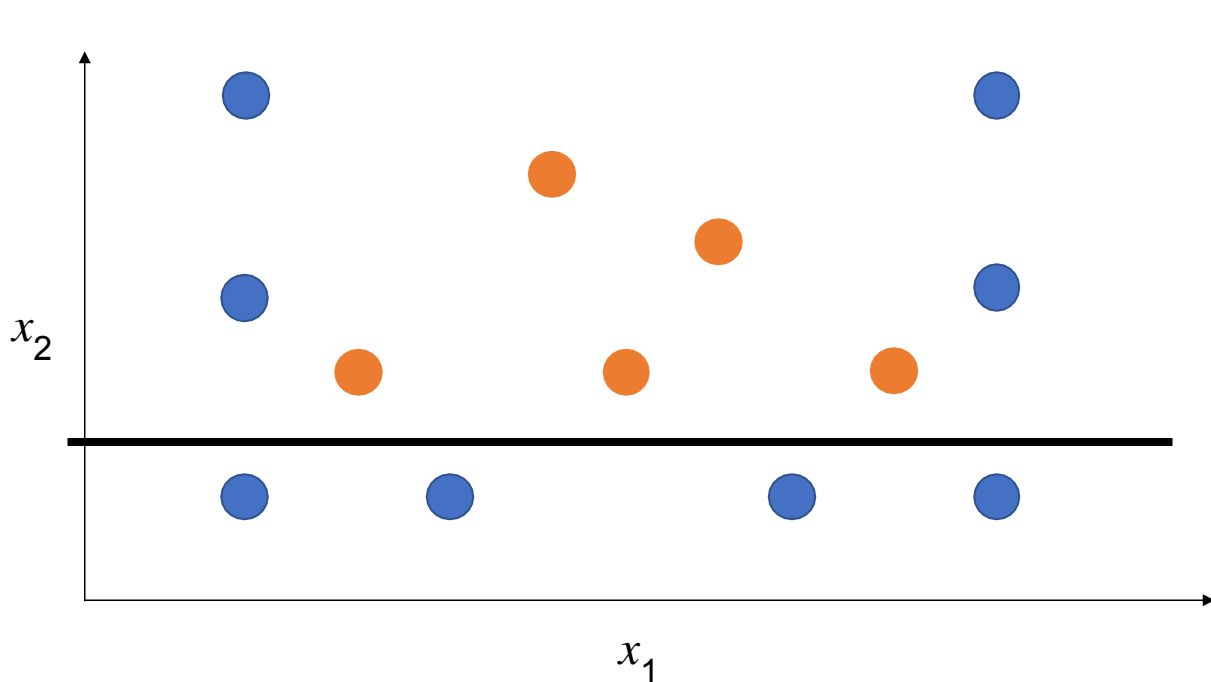
$$\frac{4}{13}H(p_l) + \frac{9}{13}H(p_r) = 0.47$$

Лучшее
разбиение!

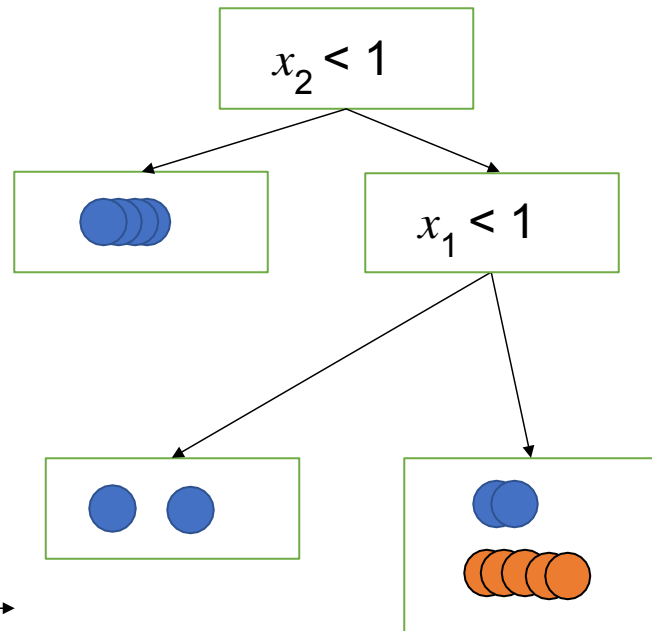
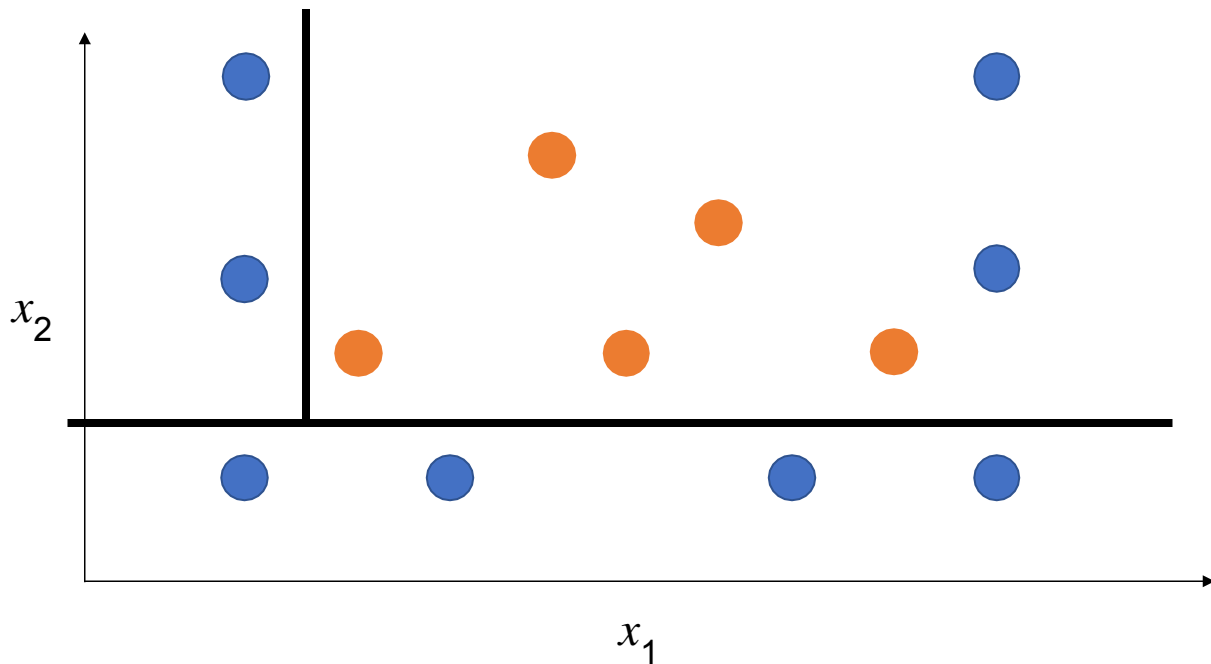
Обучение деревьев



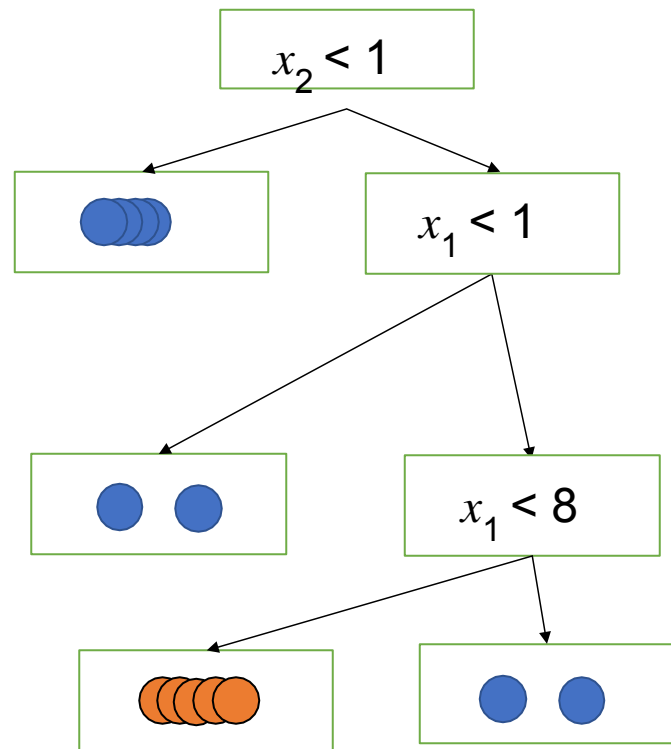
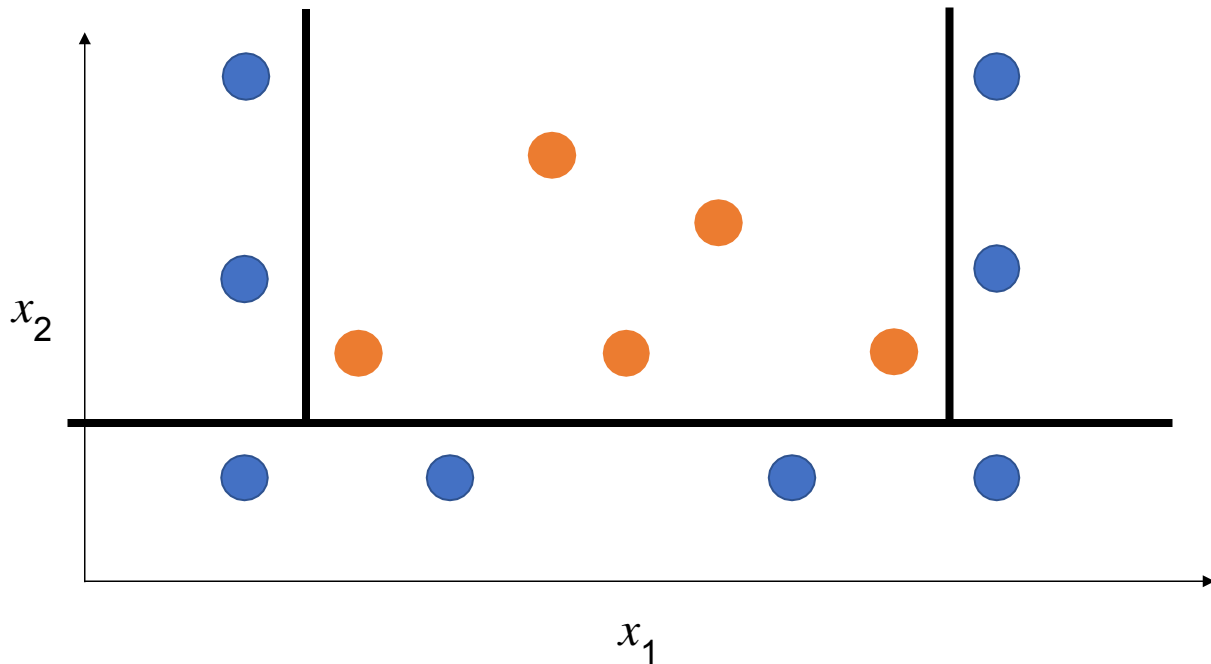
Обучение деревьев



Обучение деревьев

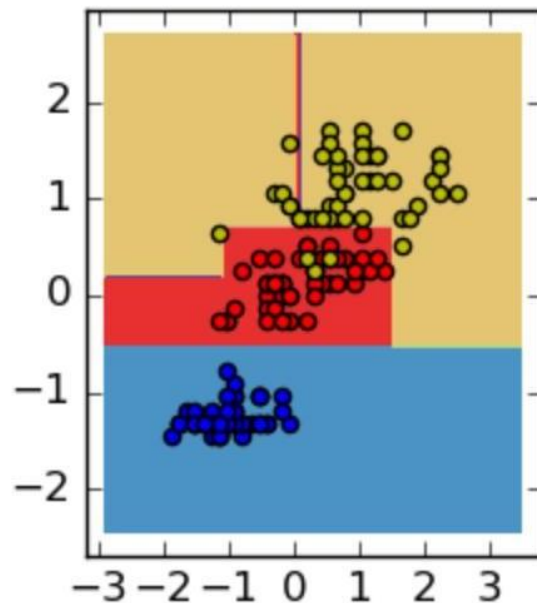


Обучение деревьев

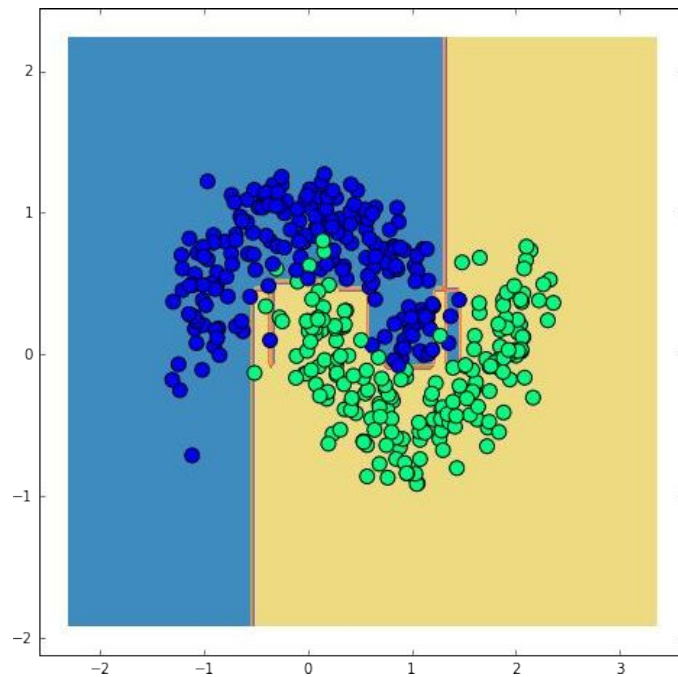


Переобучение деревьев и
борьба с ним

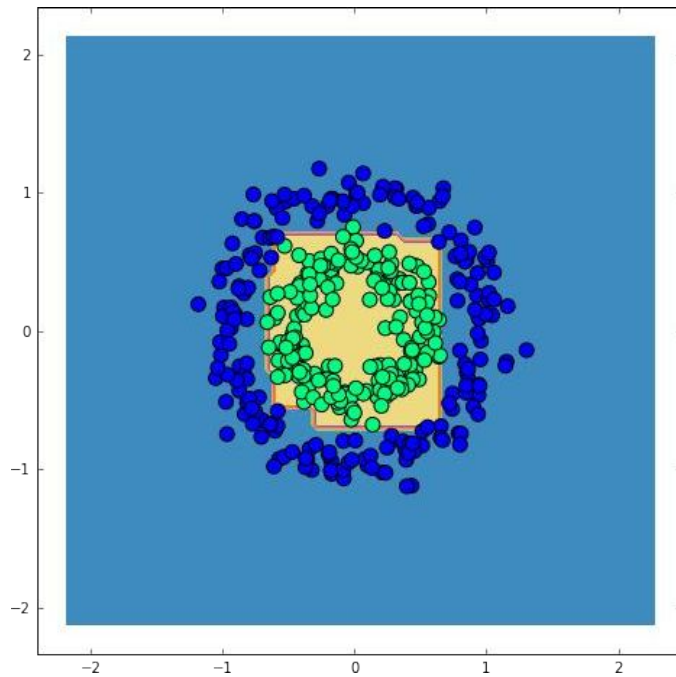
Классификация



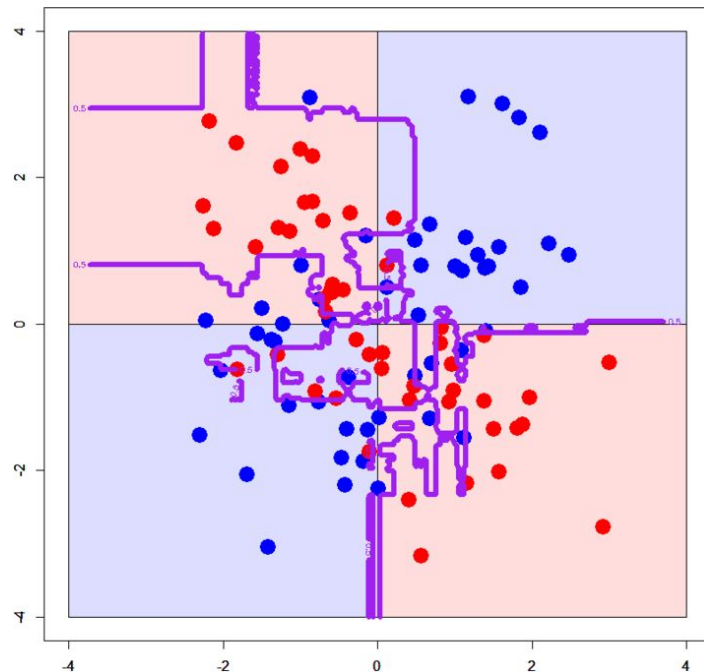
Классификация



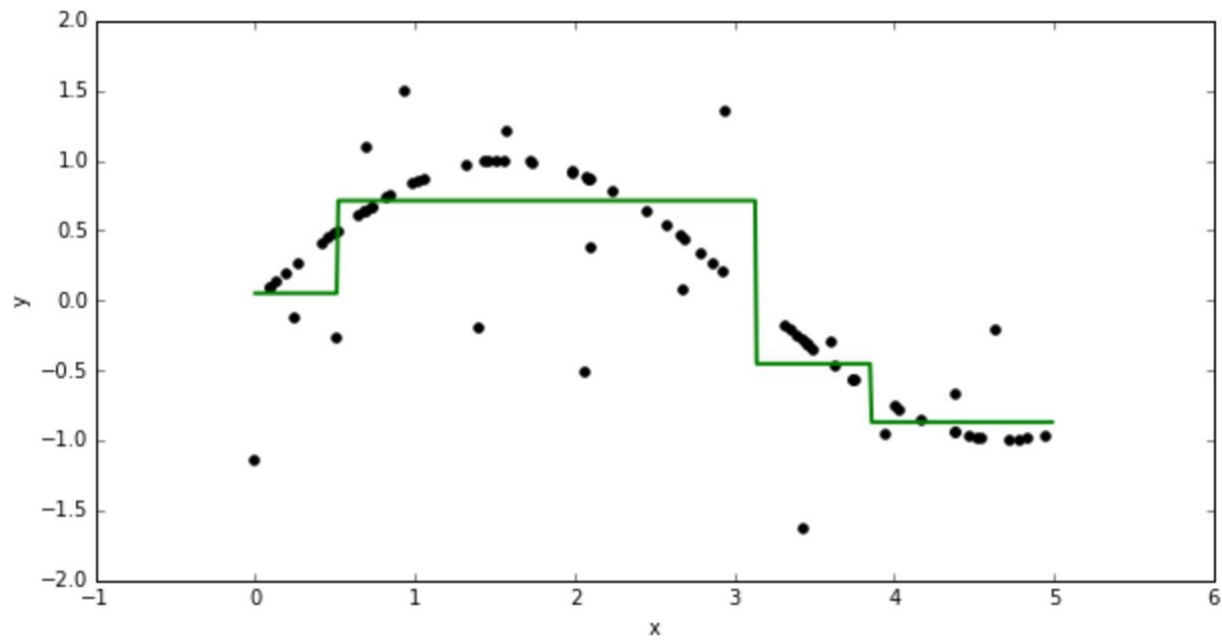
Классификация



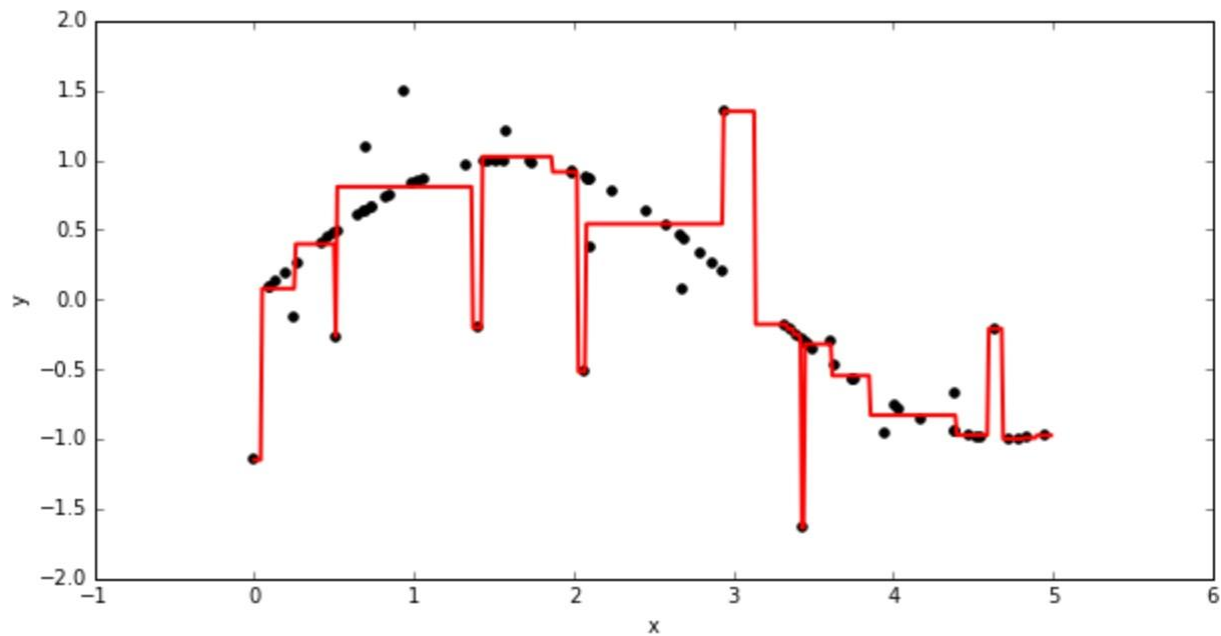
Классификация



Регрессия



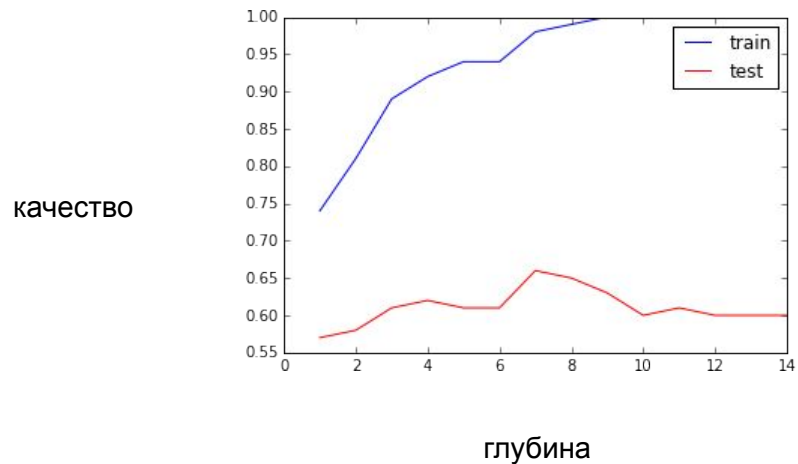
Регрессия



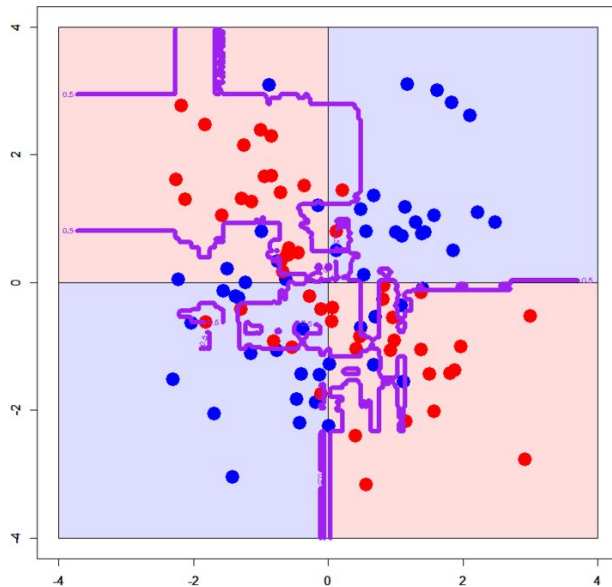
Решающие деревья

- Восстанавливают сложные закономерности
- Могут построить сколь угодно сложную поверхность
- Чем больше глубина — тем сложнее поверхность
- Склонны к переобучению

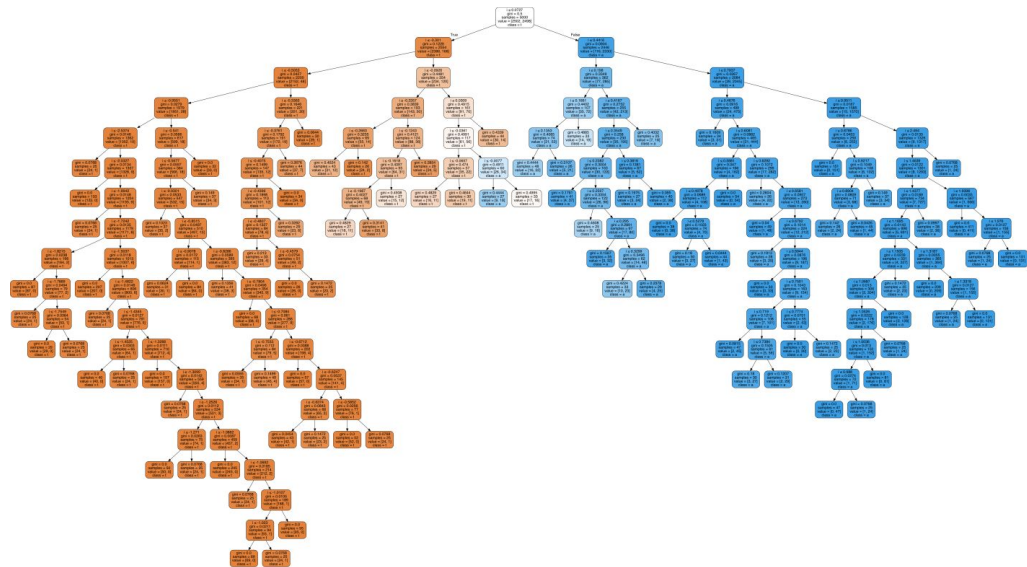
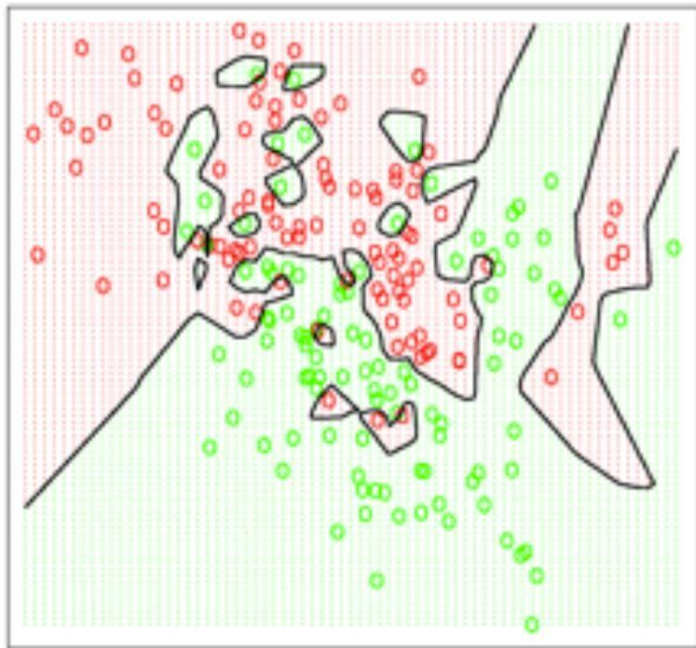
Глубина деревьев



Переобучение деревьев



Переобучение деревьев



Переобучение деревьев

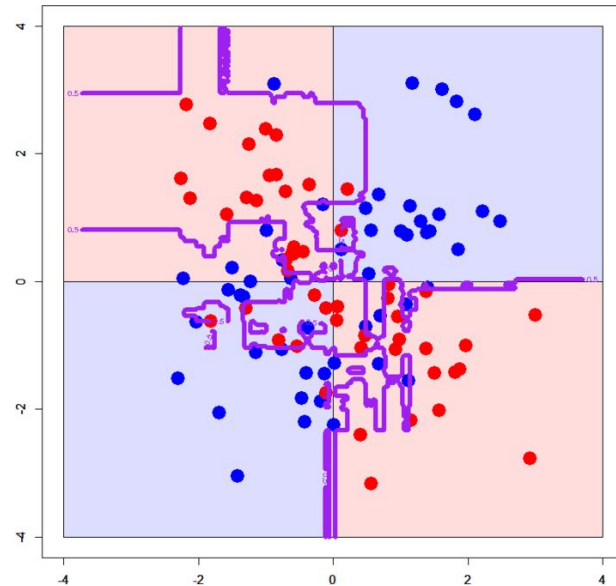
- Дерево может достичь нулевой ошибки на любой выборке
- Как правило, такое дерево окажется переобученным
- Выход — ограничивать глубину или число объектов в листе

Критерий останова

- Как понять, разбивать вершину или делать листовой?
- Способ борьбы с переобучением

Критерий останова

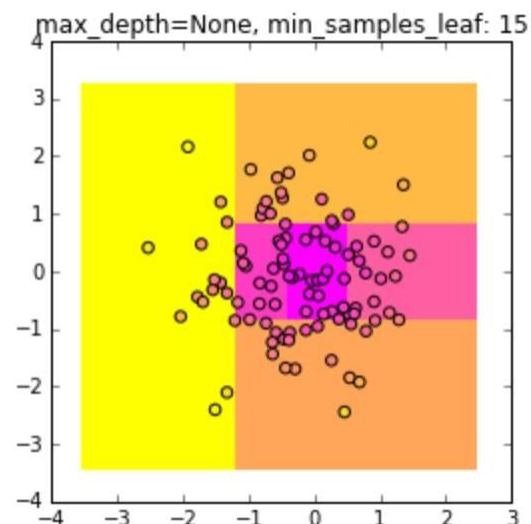
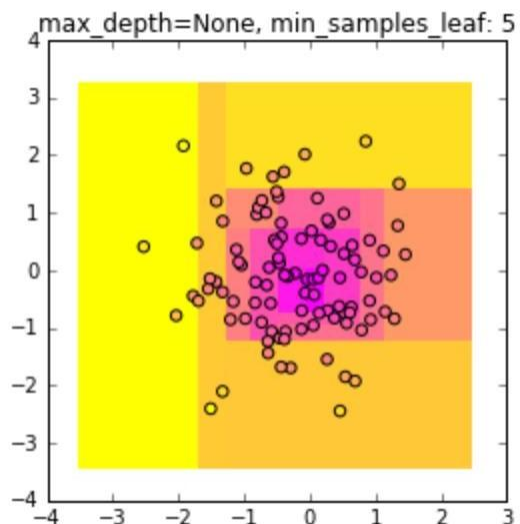
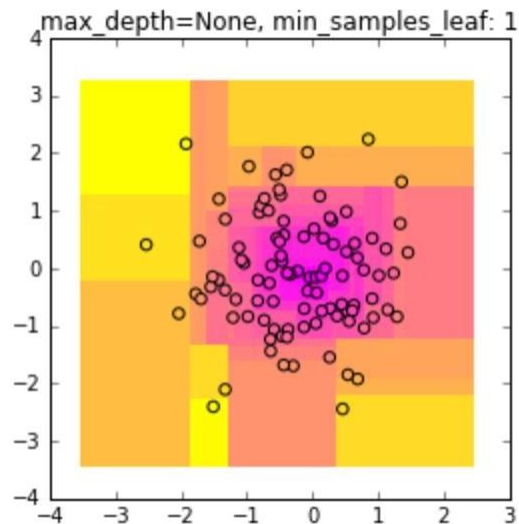
- Все объекты в вершине относятся к одному классу
- Простое условие
- Но приводит к переобучению



Число объектов в листе

- В вершину попало $\leq n$ объектов
- При $n = 1$ получаем максимально переобученные деревья
- n должно быть достаточно, чтобы построить надёжный прогноз
- Рекомендация: $n = 5$

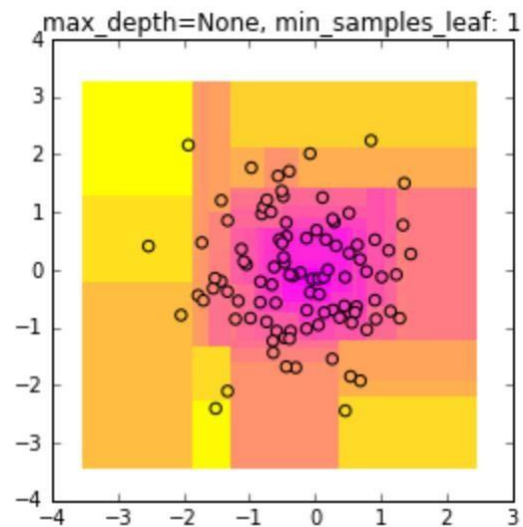
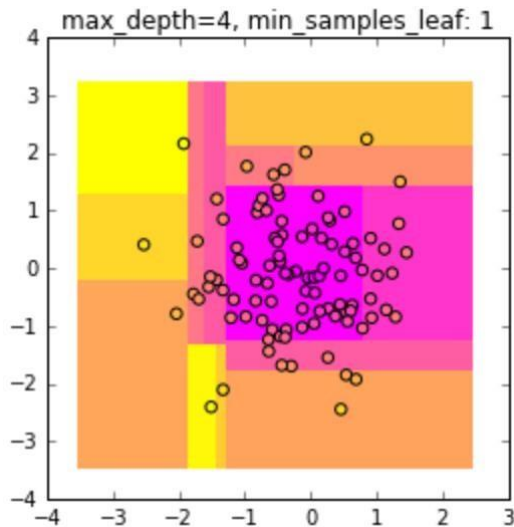
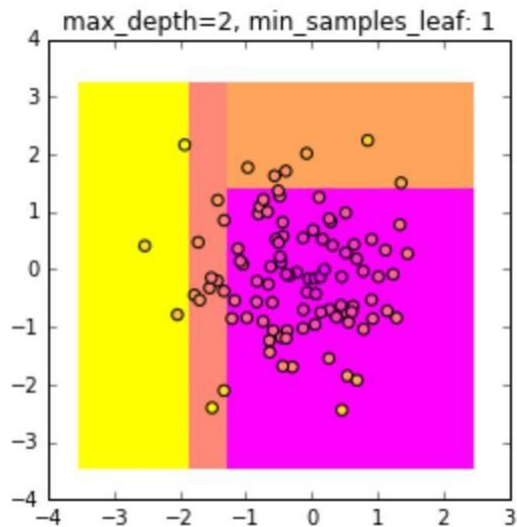
Число объектов в листе



Глубина дерева

- Ограничение на глубину
- Достаточно грубый критерий

Глубина дерева



Резюме

- Решающее дерево — очень мощная модель
- Много тонкостей с переобучением
- Обычно используется в композициях