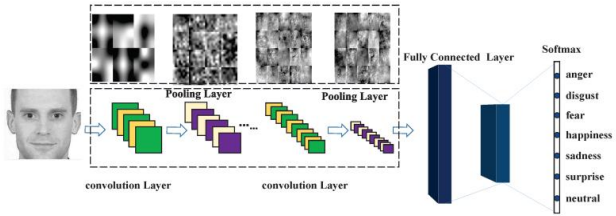
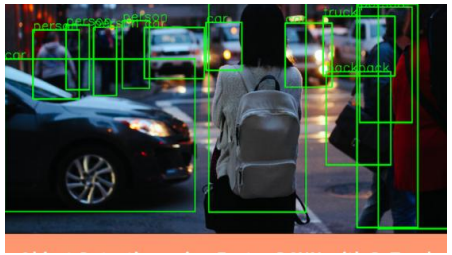
# Введение в Машинное обучение













**Лекция**

Обучение с учителем: Решающие деревья

к.ф.-м.н., доцент УИ
Корнаева Е.П.

# Обучение с учителем: Решающие деревья

**Мотивация:** Как моделировать нелинейные зависимости и при этом легко интерпретировать прогноз?

## Пример: оценка стоимости квартиры



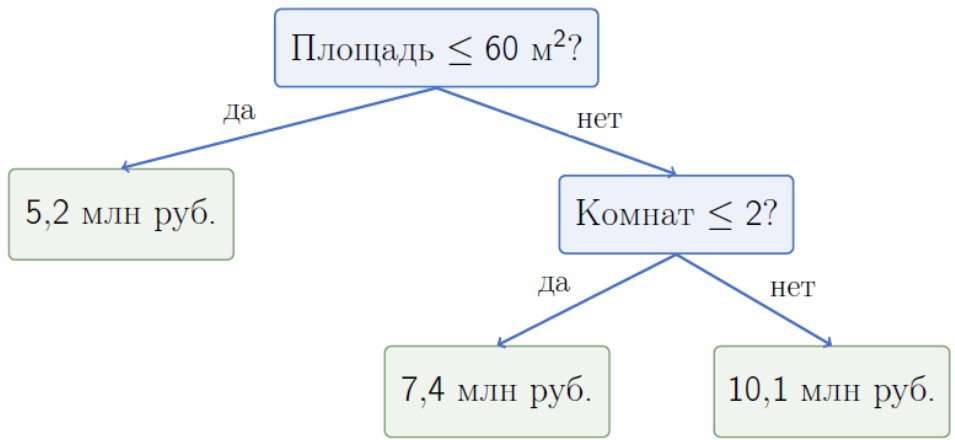

### Модель линейной регрессии

Целевая переменная: рыночная стоимость квартиры (Price);
Признаки: Room, Size, Floor и т.д.
Пусть модель линейной регрессии от полиномиальных признаков имеет вид:

$$Price(\boldsymbol{X}, \boldsymbol{\Theta}) = \Theta_0 + \Theta_1 Room + \cdots \Theta_{m+1} Room^2 + \\ + \Theta_{m+2} Size^2 + \Theta_{m+3} SizeRoom + \cdots$$

Чем больше таких слагаемых, как $Room^2$ , $SizeRoom$ ...тем сложнее объяснить, почему получена именно эта цена.

### Альтернатива: простые вопросы



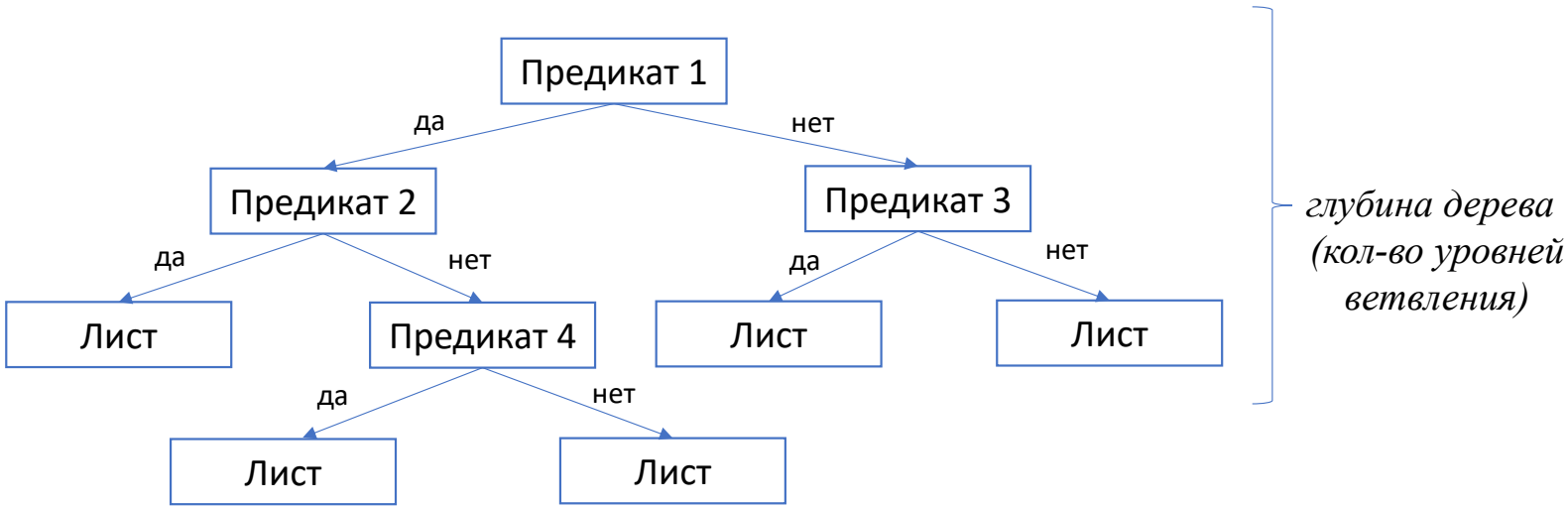

Учебная схема: пороги и цены условные.
При обучении разбиения и прогнозы определяются по данным.

Дерево: нелинейная модель и понятный путь от признаков к прогнозу.

# Обучение с учителем: Решающие деревья

## Деревья классификации и регрессии (Classification and regression trees (CART))

**Решающие деревья (decision trees) -** модели, которые определяются путем рекурсивного разделения факторного пространства и определения локальной модели в каждой результирующей области входного пространства [1]




*Вершины (узлы) дерева* содержат предикаты вида $X_j \leq t_m$

*Листья дерева* содержат прогнозы:

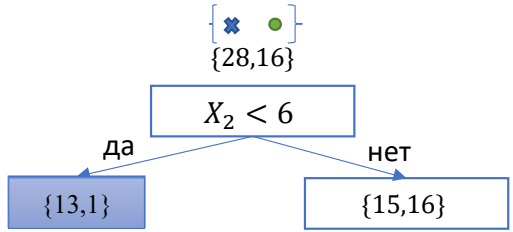
- для классификации – класс или вероятность (относительная частота);
- для регрессии – значение целевой переменной.

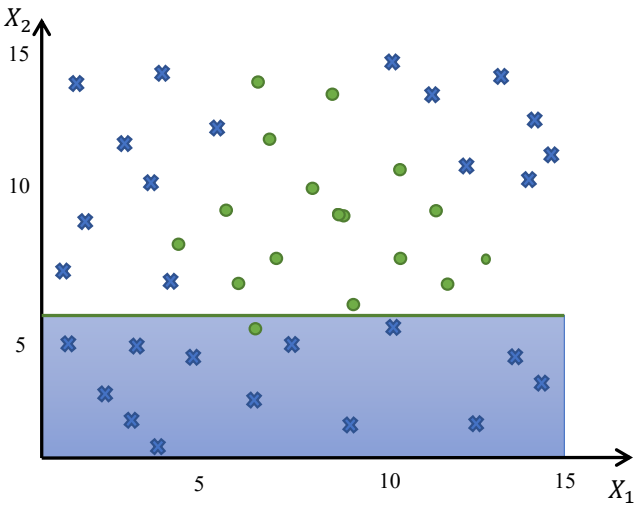
[1] Murphy, Kevin P. Machine learning : a probabilistic perspective / Kevin P. Murphy. – 2012. - 1098 p.

# Обучение с учителем: Решающие деревья

## Деревья классификации и регрессии (Classification and regression trees (CART))

Пример для задачи классификации



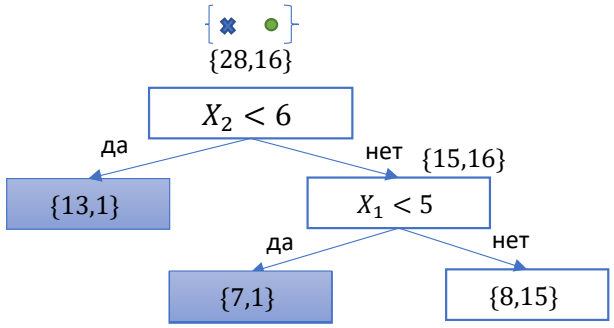



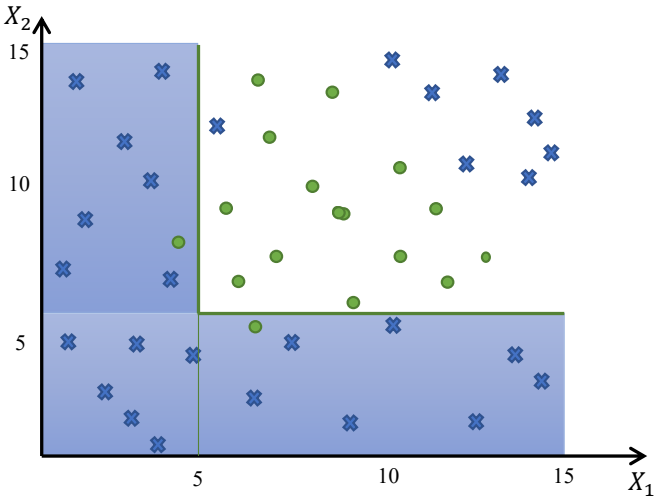

$$y_{pred} = argmax(p_k)$$

# Обучение с учителем: Решающие деревья

## Деревья классификации и регрессии (Classification and regression trees (CART))

Пример для задачи классификации

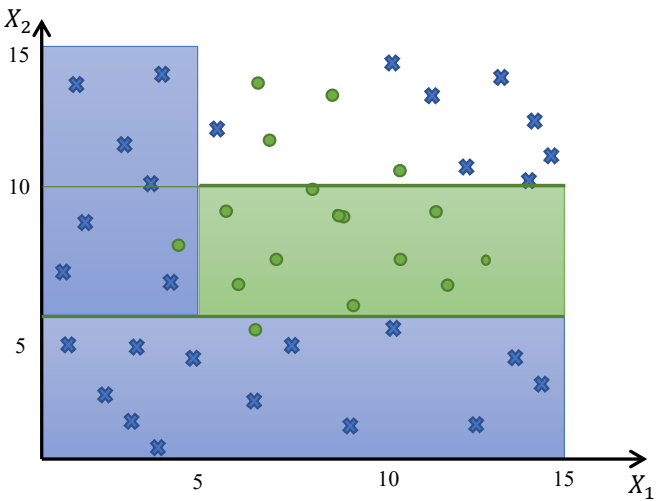


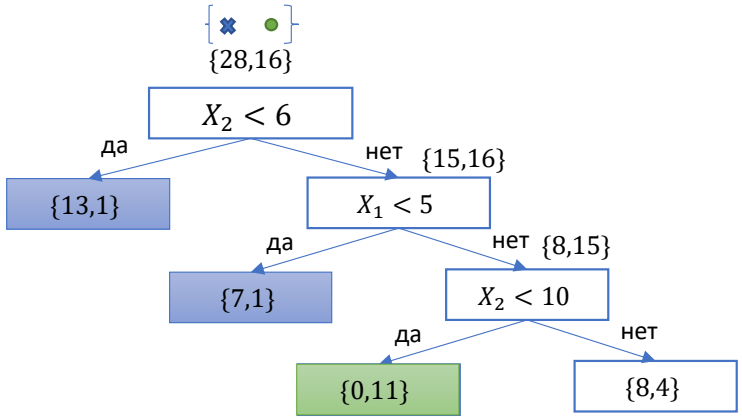



# Обучение с учителем: Решающие деревья

## Деревья классификации и регрессии (Classification and regression trees (CART))

Пример для задачи классификации

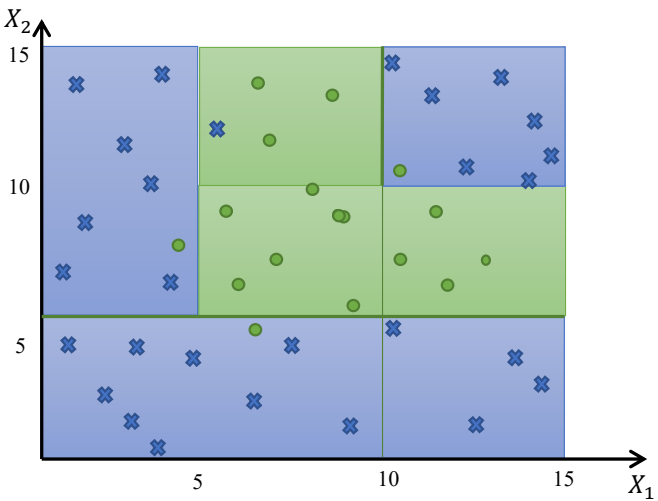


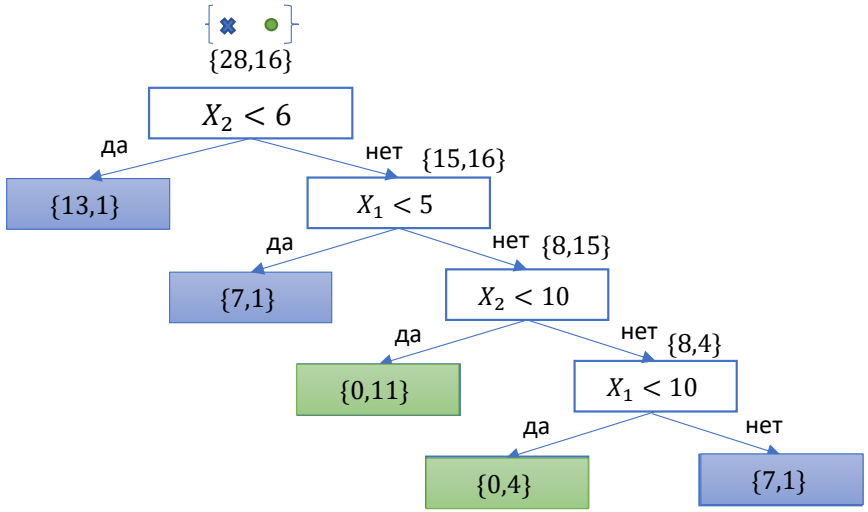



# Обучение с учителем: Решающие деревья

## Деревья классификации и регрессии (Classification and regression trees (CART))

Пример для задачи классификации

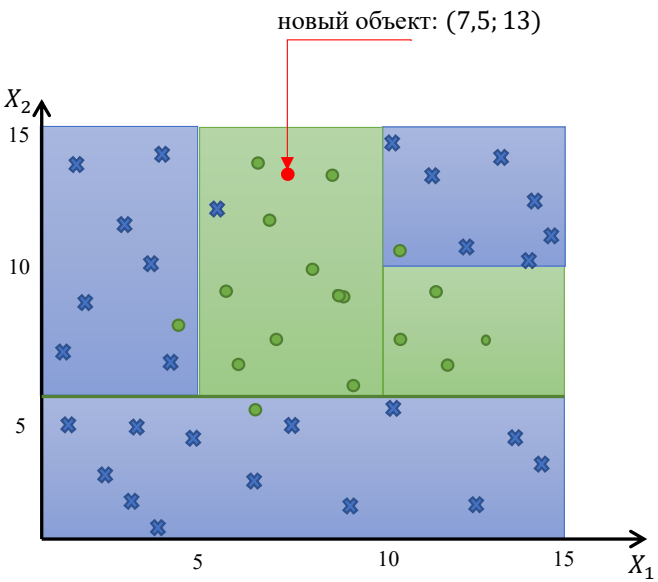


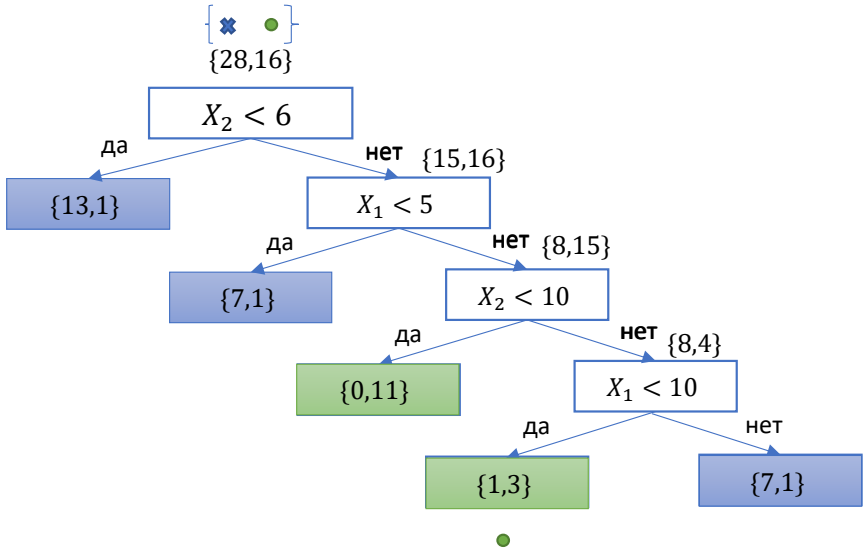



# Обучение с учителем: Решающие деревья

## Деревья классификации и регрессии (Classification and regression trees (CART))

Пример для задачи классификации



# Обучение с учителем: Решающие деревья

## Деревья классификации и регрессии (Classification and regression trees (CART))

### Выбор предикатов

| Вид предиката | Название предиката |
|---|---|
| $[x_j < t_i]$ | Порог на признак |
| $[XW < t_i]$ | Предикат с линейной моделью |
| $[\rho(X, X_0) < t_i]$ | Предикат с метрикой |
| | ... |

*Для деревьев можно выбирать разные предикаты, но обычно хватает самых простых!

# Обучение с учителем: Решающие деревья

## Деревья классификации и регрессии (Classification and regression trees (CART))

### Выбор предикатов

Критерий качества разбиения вершины дерева: $Q(R_i, \Theta) \to min$

Обозначения:

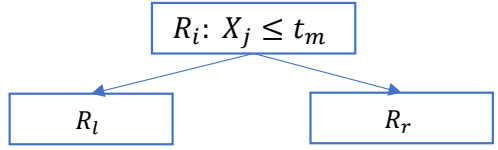
$R_i$ – текущая вершина;

$\Theta = (j, t_m)$ – номер признака $X_j$ и порог для предиката $X_j \leq t_m$ в вершине $m$;

$R_l(\Theta) = \{(X, y)|x_j \leq t_i\}$ – объекты, попавшие в левую вершину;

$R_r(\Theta) = R_i \backslash \mathrm{R}_l(\Theta)$ - объекты, попавшие в правую вершину;

$N_l, N_r$ – количество объектов в левой и правой вершине.

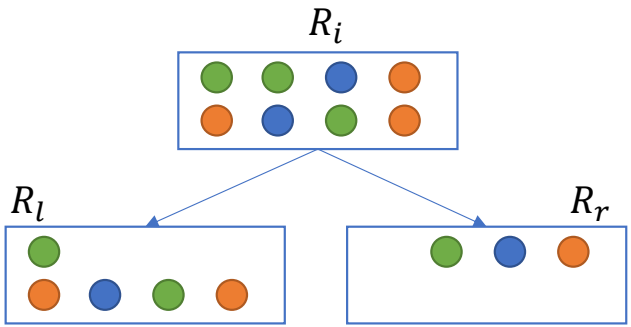


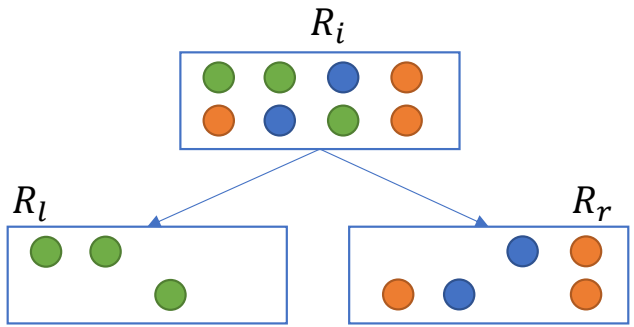
# Обучение с учителем: Решающие деревья

## Деревья классификации и регрессии (Classification and regression trees (CART))

### Выбор предикатов

Пример для классификации. Какое разбиение лучше?



## Обучение с учителем: Решающие деревья

### Деревья классификации и регрессии (Classification and regression trees (CART))

**Выбор предикатов**

*Мера неопределённости* $H(R_i)$ – оценивает меру неоднородности целевых переменных в вершине $R_i$.

- ***Критерий Джини*** (Gini index) [1] как мера неопределенности в $i^{\text{ой}}$ вершине:

$$H(R_i) = \sum_{k=1}^{K} p_{ik}(1 - p_{ik}) = 1 - \sum_{k=1}^{K} p_{ik}^2$$

- ***Энтропия*** (Entropy, or deviance) [1]:

$$H(R_i) = -\sum_{i=1}^{K} p_{ik} \log p_{ik}$$

- ***Ошибка классификации*** (Misclassification rate) [1]:

$$H(R_i) = 1 - \min_k p_{ik}$$

где $p_{ik} = \frac{1}{N_i}\sum_{y \in R_i}[y == k]$ – доля объектов $k^{\text{го}}$ класса, попавшие в $i^{\text{ю}}$ вершину.

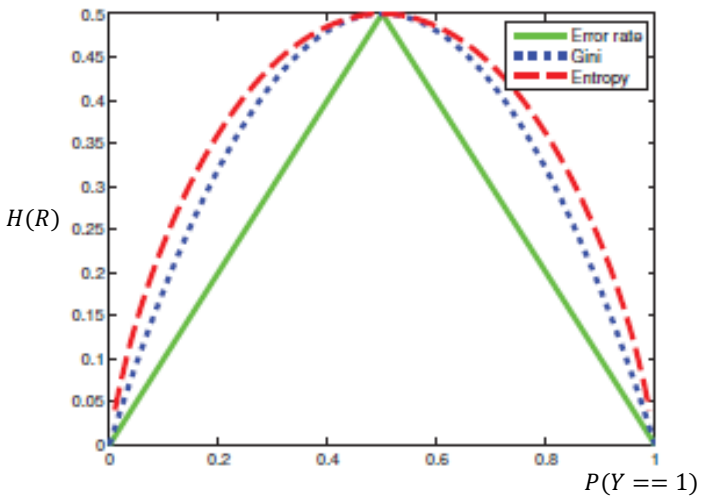



Рис. 2. Критерий информативности для бинарной классификации [1]. ф. энтропии масштабирована.

[1] Sarah Guido, Andreas Muller. Introduction to Machine Learning with Python: A Guide for data scientists. O'Reilly & Associates. - 2017.

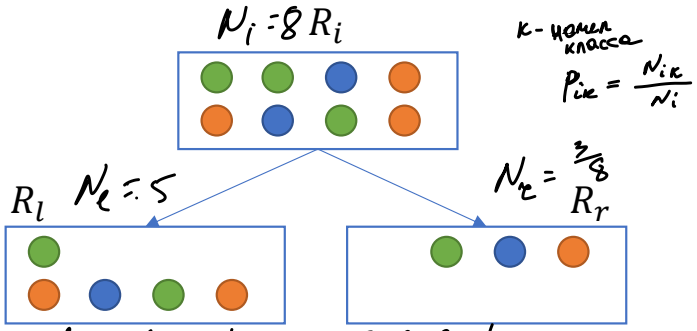
# Обучение с учителем: Решающие деревья

## Деревья классификации и регрессии (Classification and regression trees (CART))

### Выбор предикатов

Рассчитать значения $H(R_l), H(R_r)$ для каждой вершины, используя критерий Джини:

$$H(R_i) = \sum_{k=1}^{K} p_{ik}(1 - p_{ik}), = 1 - \sum_{k=1}^{K} p_{ik}^2$$




k - номер класса

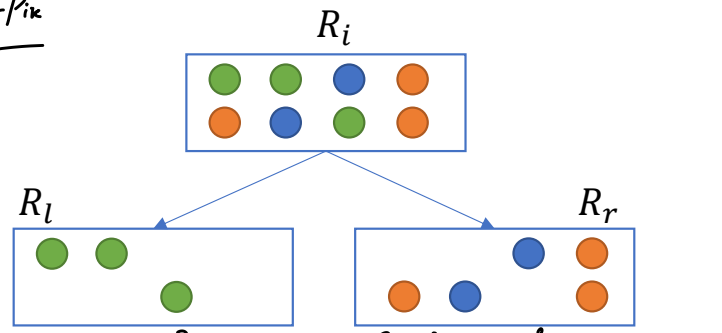
$p_{ik} = \frac{N_{ik}}{N_i}$

$p_g = \frac{2}{5}$, $p_o = \frac{2}{5}$, $p_c = \frac{1}{5}$

$H(R_l) = 1 - 2\left(\frac{2}{5}\right)^2 - \left(\frac{1}{5}\right)^2 = \frac{16}{25}$

$p_g = p_c = p_o = \frac{1}{3}$

$H(R_r) = 1 - 3\cdot\left(\frac{1}{3}\right)^2 = \frac{1}{3}$




$p_g = 1$, $p_o = p_c = 0$

$H(R_l) = 0$

$p_o = \frac{3}{5}$, $p_c = \frac{2}{5}$

$H(R_r) = 1 - \left(\frac{3}{5}\right)^2 - \left(\frac{2}{5}\right)^2 = \frac{12}{25}$

## Обучение с учителем: Решающие деревья

### Деревья классификации и регрессии (Classification and regression trees (CART))

**Выбор предикатов**

*Критерий качества (информативности)* разбиения вершины $R_i$:

$$\tilde{Q}(R_i, \Theta) = H(R_i) - \frac{N_l}{N_i} H\big(R_l(\Theta)\big) - \frac{N_r}{N_i} H\big(R_r(\Theta)\big) \underset{\Theta}{\rightarrow} max,$$

или $$Q(R_i, \Theta) = \frac{N_l}{N_i} H\big(R_l(\Theta)\big) + \frac{N_r}{N_i} H\big(R_r(\Theta)\big) \underset{\Theta}{\rightarrow} min,$$

где $H(R_l), H(R_r)$ – меры неопределенности в левой и правой вершинах;

$N_l, N_r$ – количество объектов в левой и правой вершине.

Таким образом:

- предикат можно выбирать так, чтобы он как можно сильнее уменьшал «хаотичность» вершин;
- качество предиката измеряется с помощью критерия информативности.

[1] Sarah Guido, Andreas Muller. Introduction to Machine Learning with Python: A Guide for data scientists. O'Reilly & Associates. - 2017.

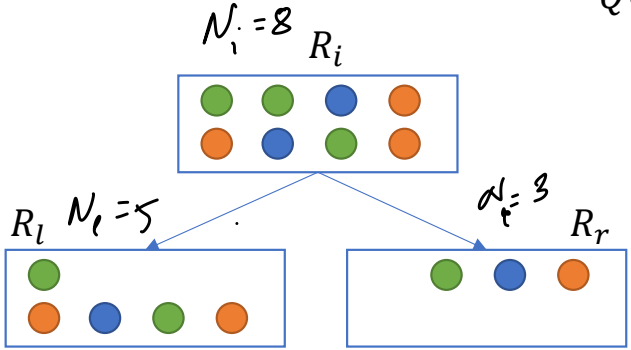
# Обучение с учителем: Решающие деревья

## Деревья классификации и регрессии (Classification and regression trees (CART))

**Выбор предикатов.**

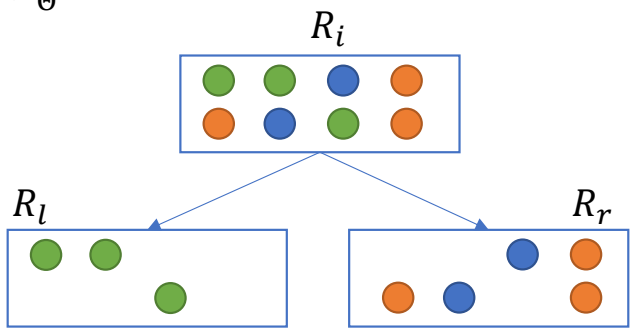
Рассчитать критерий качества разбиения вершины $R_i$ для предыдущей задачи. Какое разбиение лучше?

$$Q(R_i, \Theta) = \frac{N_l}{N_i} H\big(R_l(\Theta)\big) + \frac{N_r}{N_i} H\big(R_r(\Theta)\big) \underset{\Theta}{\rightarrow} min,$$




$H(R_l) = \frac{16}{25}$ $\quad H(R_r) = \frac{2}{3}$

(1) $Q(R_i) = \frac{5}{8} \cdot \frac{16}{25} + \frac{3}{8} \cdot \frac{2}{3} = \frac{13}{20}$




$H(R_l) = 0$ $\quad H(R_r) = \frac{12}{25}$

(2) $Q(R_i) = \frac{3}{8} \cdot 0 + \frac{5}{8} \cdot \frac{12}{25} = \frac{3}{10}$

$Q^{(2)} < Q^{(1)}$

**Обучение с учителем: Решающие деревья**

**Деревья классификации и регрессии (Classification and regression trees (CART))**

**Выбор предикатов**. Пример разбиения по критерию Джини

В вершине $R_i$ семь объектов трёх классов.
Какой предикат лучше: $x_1 < 6$ или $x_3 < 5$?

Исходная выборка: четыре признака

| $x_1$ | $x_2$ | $x_3$ | $x_4$ | Класс |
|---|---|---|---|---|
| 5 | 8 | 8 | 2,5 | 2 |
| 3 | 3 | 7 | 7,7 | 1 |
| 6 | 7,7 | 1 | 1,1 | 3 |
| 3,3 | 1 | 2 | 1,2 | 1 |
| 24 | 3,9 | 5 | 3,9 | 1 |
| 12 | 10 | 10,1 | 8 | 2 |
| 1 | 2 | 2 | 9,1 | 1 |

$$H(Y) = 1 - \sum_{k=1}^{3} p_k^2, \qquad Q = \frac{N_\ell}{7} H(Y_\ell) + \frac{N_r}{7} H(Y_r)$$

1. Предикат x1 < 6

2. Предикат x3 < 5

Рассчитайте Q1 и Q2. Выберите предикат с меньшим значением критерия.

**Обучение с учителем: Решающие деревья**

**Деревья классификации и регрессии (Classification and regression trees (CART))**

**Выбор предикатов**. Пример разбиения по критерию Джини

В вершине $R_i$ семь объектов трёх классов.
Какой предикат лучше: $x_1 < 6$ или $x_3 < 5$?

Исходная выборка: четыре признака

| $x_1$ | $x_2$ | $x_3$ | $x_4$ | Класс |
|---|---|---|---|---|
| 5 | 8 | 8 | 2,5 | 2 |
| 3 | 3 | 7 | 7,7 | 1 |
| 6 | 7,7 | 1 | 1,1 | 3 |
| 3,3 | 1 | 2 | 1,2 | 1 |
| 24 | 3,9 | 5 | 3,9 | 1 |
| 12 | 10 | 10,1 | 8 | 2 |
| 1 | 2 | 2 | 9,1 | 1 |

$$H(Y) = 1 - \sum_{k=1}^{3} p_k^2, \qquad Q = \frac{N_\ell}{7} H(Y_\ell) + \frac{N_r}{7} H(Y_r)$$

1. Предикат x1 < 6

$$Y_\ell = (2,1,1,1), \quad Y_r = (3,1,2)$$
$$H(Y_\ell) = 1 - \left(\tfrac{3}{4}\right)^2 - \left(\tfrac{1}{4}\right)^2 = \tfrac{3}{8}$$
$$H(Y_r) = 1 - 3\left(\tfrac{1}{3}\right)^2 = \tfrac{2}{3}$$
$$Q_1 = \tfrac{4}{7} \cdot \tfrac{3}{8} + \tfrac{3}{7} \cdot \tfrac{2}{3} = \tfrac{1}{2} = 0{,}5$$

2. Предикат x3 < 5

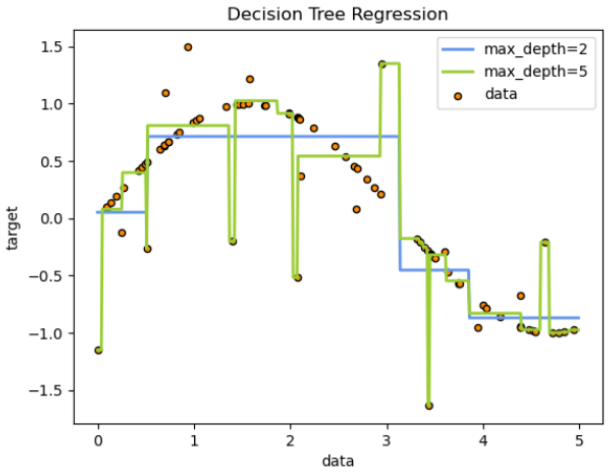
$$Y_\ell = (3,1,1), \quad Y_r = (2,1,1,2)$$
$$H(Y_\ell) = 1 - \left(\tfrac{2}{3}\right)^2 - \left(\tfrac{1}{3}\right)^2 = \tfrac{4}{9}, \quad H(Y_r) = \tfrac{1}{2}$$
$$Q_2 = \tfrac{3}{7} \cdot \tfrac{4}{9} + \tfrac{4}{7} \cdot \tfrac{1}{2} = \tfrac{10}{21} \approx 0{,}4762$$

Рассчитайте Q1 и Q2. Выберите предикат с меньшим значением критерия.

# Обучение с учителем: Решающие деревья

## Деревья классификации и регрессии (Classification and regression trees (CART))

### Для задачи регрессии




Предсказание значения в i[ом] листе:

$$y_{pred}(x) = \frac{1}{N_i}\sum_{k=1}^{N_i} y_k(x)$$

*Мера неопределённости* - это дисперсия целевой переменной (для объектов, попавших в этот лист):

$$H(R_i) = \frac{1}{N_i}\sum_{k=1}^{N_i} (y_k - \bar{y})^2$$

*Критерий качества* разбиения вершины $R_i$:

$$Q(R_i) = \frac{N_l}{N_i} H\big(R_l(\Theta)\big) + \frac{N_r}{N_i} H\big(R_r(\Theta)\big) \underset{\Theta}{\rightarrow} min,$$

**Обучение с учителем: Решающие деревья**

**Деревья классификации и регрессии (Classification and regression trees (CART))**

**Прогноз** по готовому дереву в $k^{\text{ом}}$ листе.

- Классификация:

$$y_{pred}(x) = \arg\max_{y \in \mathbb{Y}} \sum_{k=1}^{N_i} [y_k(x) = y]$$

- Регрессия:

$$y_{pred}(x) = \frac{1}{N_i} \sum_{k=1}^{N_i} y_k(x)$$

$\tilde{y}_k(x)$ - значение отклика для объекта, попавшего в $k$-й лист

## Обучение с учителем: Решающие деревья

### Деревья классификации и регрессии (Classification and regression trees (CART))

Как строить дерево?



: «перебрать все возможные деревья, выбрать самое маленькое среди безошибочных?»



: «слишком доооолго…»



: «будем строить жадно!»

**Обучение с учителем: Решающие деревья**

**Деревья классификации и регрессии (Classification and regression trees (CART))**

**Жадный алгоритм**

**Жадный алгоритм построения дерева** - алгоритм, который на каждом шаге рекурсивного разбиения факторного пространства выбирает **локально оптимальный предикат** (пару признак $X_j$ и порог $t_m$), минимизирующий заданный критерий качества разбиения $Q(R_i, \Theta)$ в текущей вершине $R_i$ , без учёта влияния этого выбора на последующие разбиения.

**Ключевые свойства жадного алгоритма:**

- **Локальность выбора.** На каждом шаге оптимизируется только «здесь и сейчас» - качество текущего разбиения. Будущие разбиения в дочерних вершинах не учитываются.
- **Рекурсивность.** Алгоритм спускается сверху вниз (top-down), от корня к листьям.
- **Необратимость.** После того как разбиение сделано, оно не пересматривается.
- **Пошаговый перебор.** Для каждого признака перебираются все возможные пороги и выбирается лучший.

# Обучение с учителем: Решающие деревья

## Деревья классификации и регрессии (Classification and regression trees (CART))

Пример для задачи классификации

Какое дерево выбрать (рис.1 или рис.2 ?)

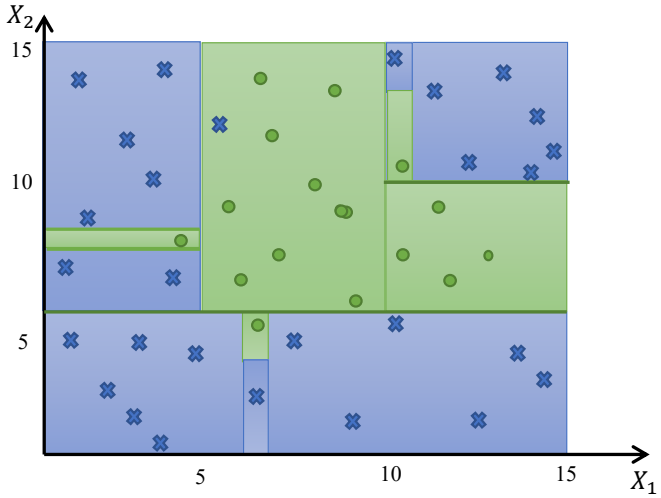



Рис.1

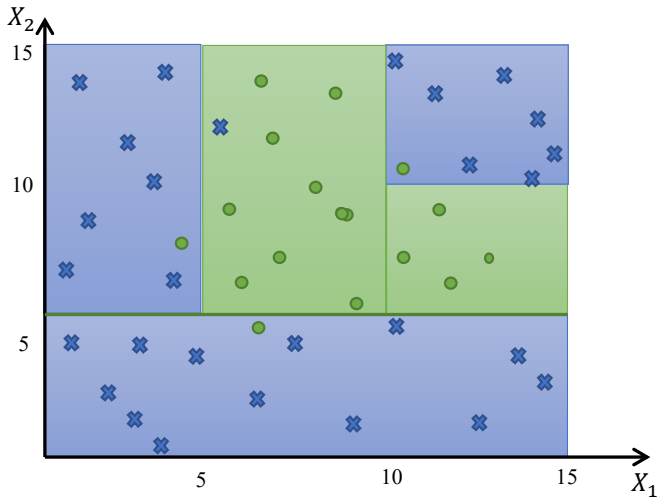



Рис.2

# Обучение с учителем: Решающие деревья

## Деревья классификации и регрессии (Classification and regression trees (CART))

Пример для задачи классификации

Какое дерево выбрать (рис.1 или рис.2 ?)

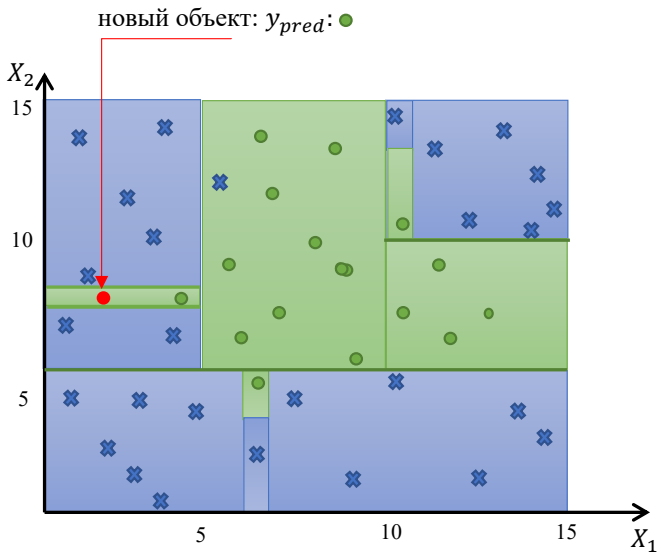



Рис.1

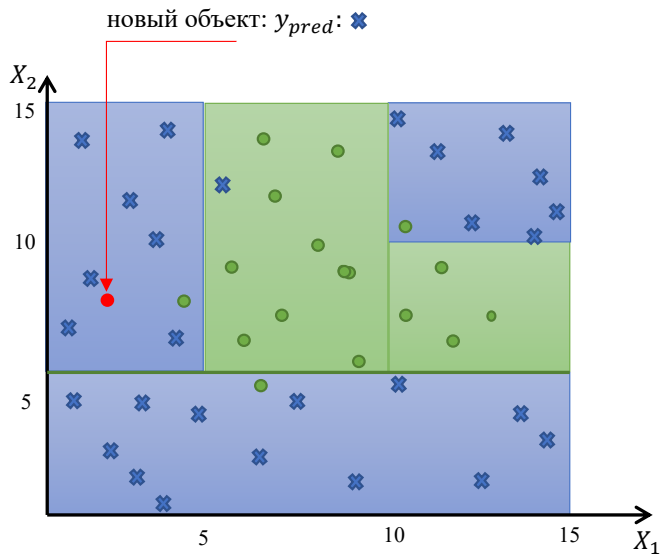



Рис.2

**Замечание**. Дерево можно построить так, чтобы точность на обучающей выборке была 100%, т.е. деревья легко переобучаются (например, как на рис.1). Если только нет объектов с одинаковыми признаками, но разными ответами.

# Обучение с учителем: Решающие деревья

## Деревья классификации и регрессии (Classification and regression trees (CART))

Чтобы дерево не переобучалось:
- глубина дерева должна быть ограничена;
- нужны условия на продолжение ветвления.

**Как остановить процесс построения дерева?**

- Ограничить глубину дерева;
- Ограничить количество листьев;
- Задать минимальное количество объектов в вершине;
- Задать минимальное уменьшение хаотичности при разбиении (порог ниже которого разбиение не проводится);
- и т.д.

https://scikit-learn.org/stable/auto_examples/tree/plot_iris_dtc.html#sphx-glr-auto-examples-tree-plot-iris-dtc-py

# Обучение с учителем: Решающие деревья

## Резюме

Преимущества:

- Четкие правила классификации (легко интерпретировать);
- Легко визуализируются;
- Быстро выдают прогноз;
- Малое кол-во параметров;
- Не требуется нормировка данных (в подавляющем большинстве моделей);
- Модель не чувствительна к коррелированным признакам (в плане предсказательной силы) как в регрессии.

Недостатки:

- Чувствительны к шумам в данных (модель сильно меняется при небольшом изменении обучающей выборки);
- Разделяющая граница имеет свои ограничения (состоит из гиперплоскостей, параллельных осям признаков);
- Необходимость борьбы с переобучением (стрижка или какой-либо из критериев останова);
- Проблема поиска оптимального дерева (на практике используется жадное построение дерева): не гарантирует глобальный оптимум.