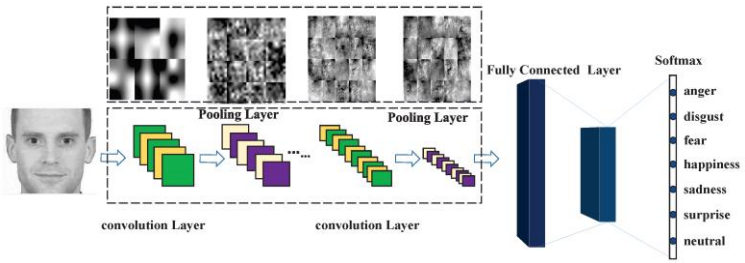
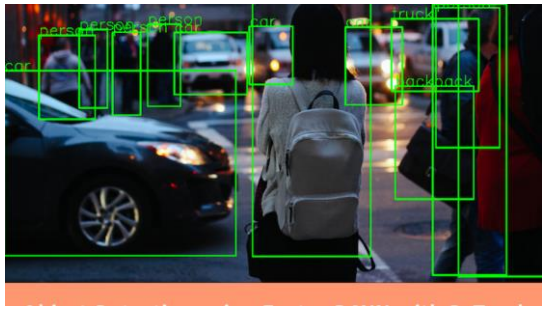
# Машинное обучение







## Приглашенная лекция 10

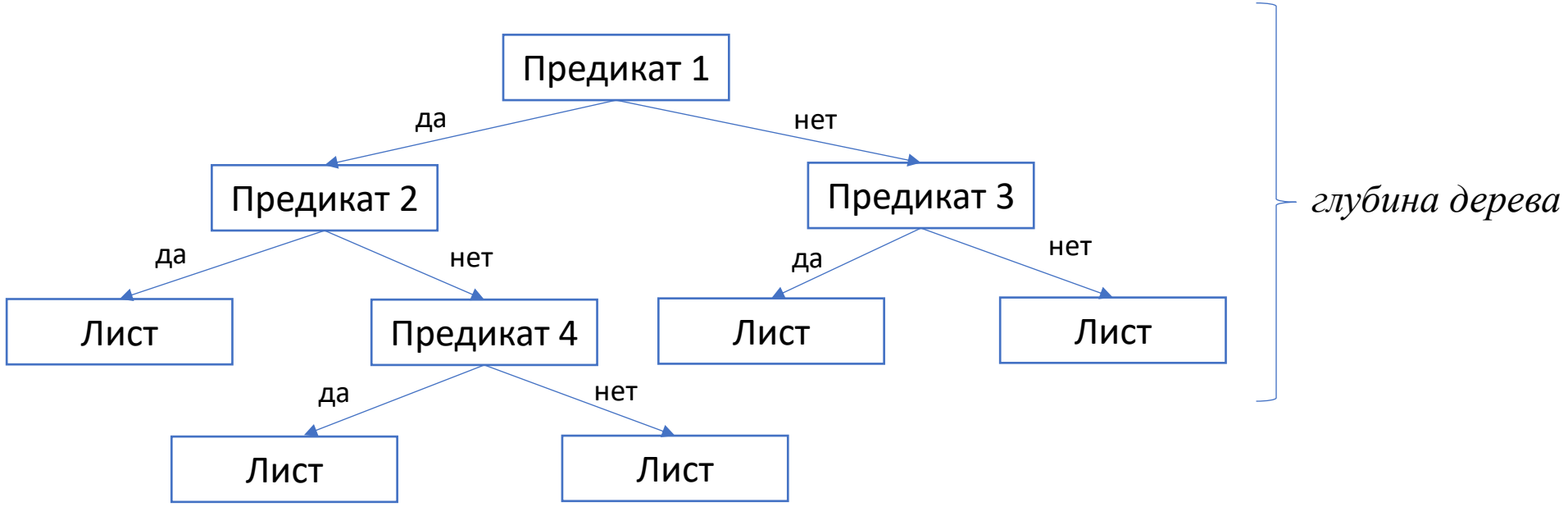
Обучение с учителем: Решающие деревья

к.ф.-м.н., доцент кафедры ИСиЦТ ОГУ
Корнаева Е.П.

# 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 дерева* содержат предикаты вида $X_j \leq t_m$

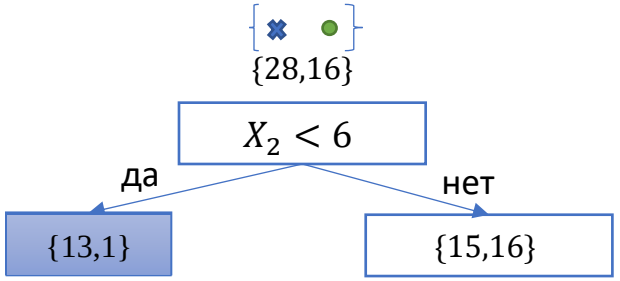
*Листья дерева* содержат прогнозы (для классификации – класс или вероятность, для регрессии – значение целевой переменной)

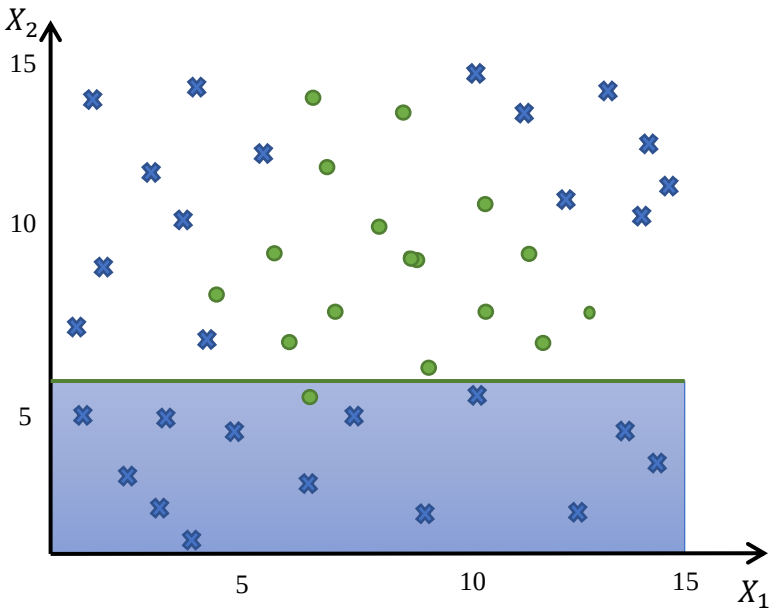
[1] Murphy, Kevin P. Machine learning : a probabilistic perspective / Kevin P. Murphy. – 2012. - 1098 p.

# Обучение с учителем: Решающие деревья

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

Пример для задачи классификации



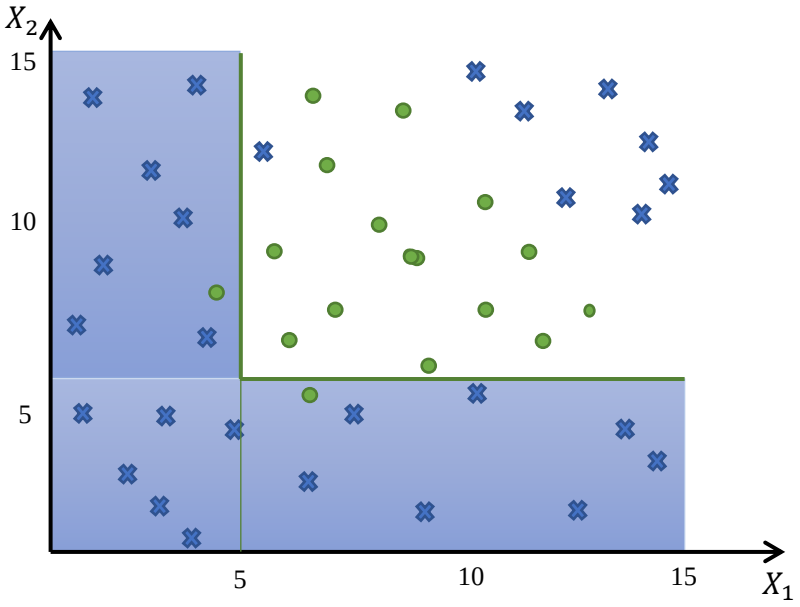



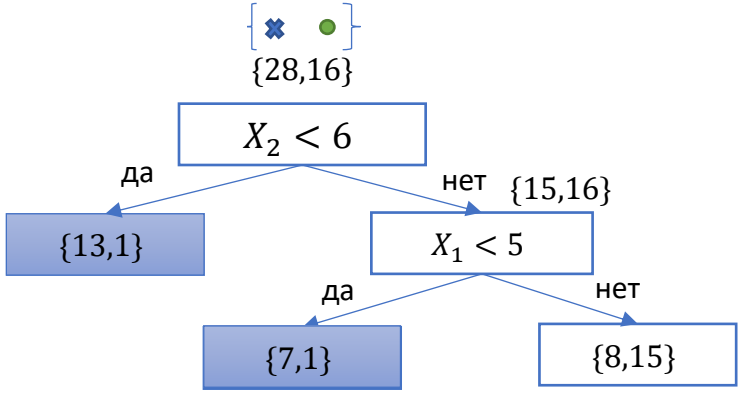

$$y_{pred} = argmax(p_k)$$

# Обучение с учителем: Решающие деревья

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

Пример для задачи классификации

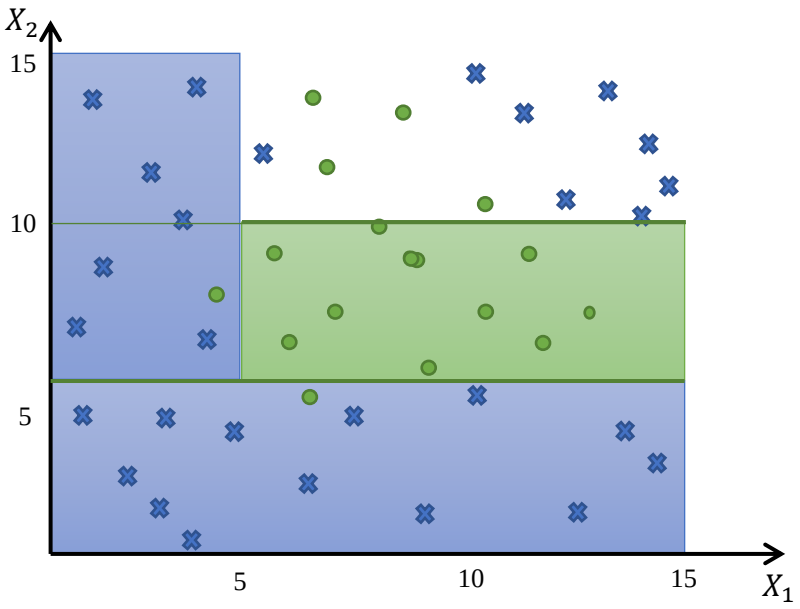


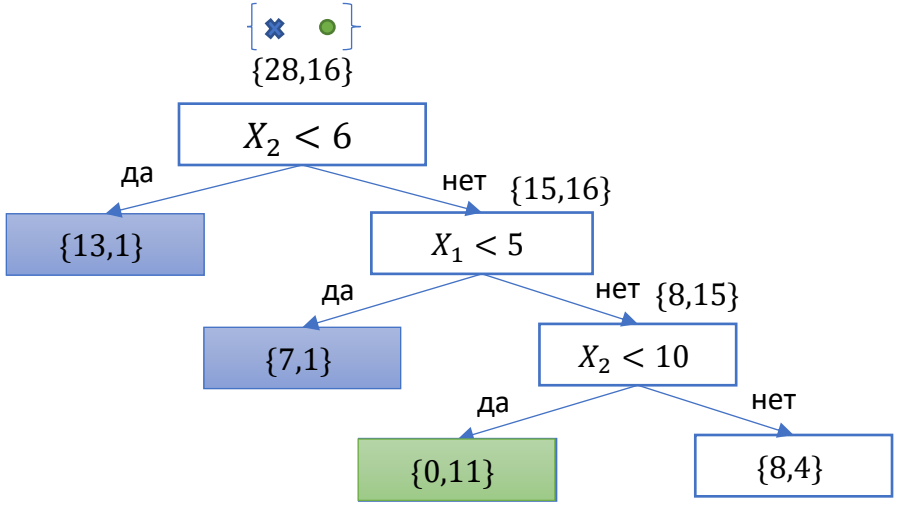



# Обучение с учителем: Решающие деревья

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

Пример для задачи классификации

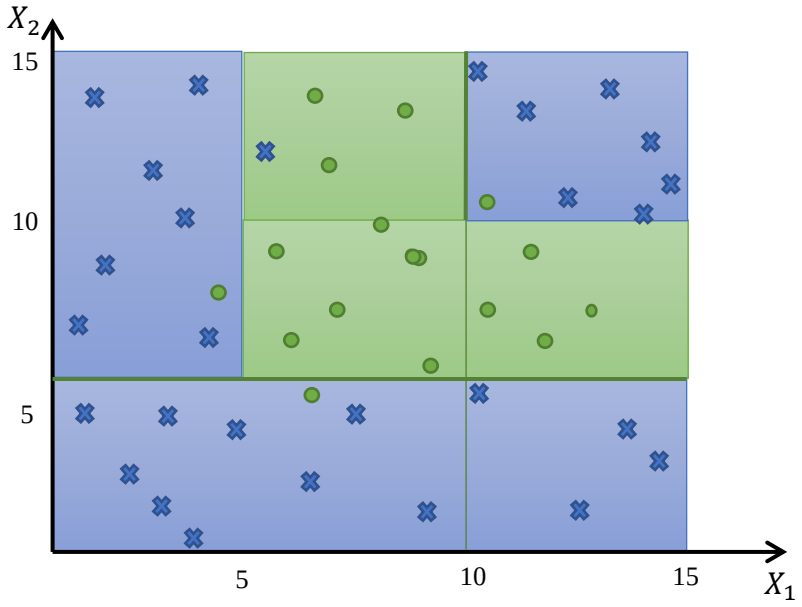


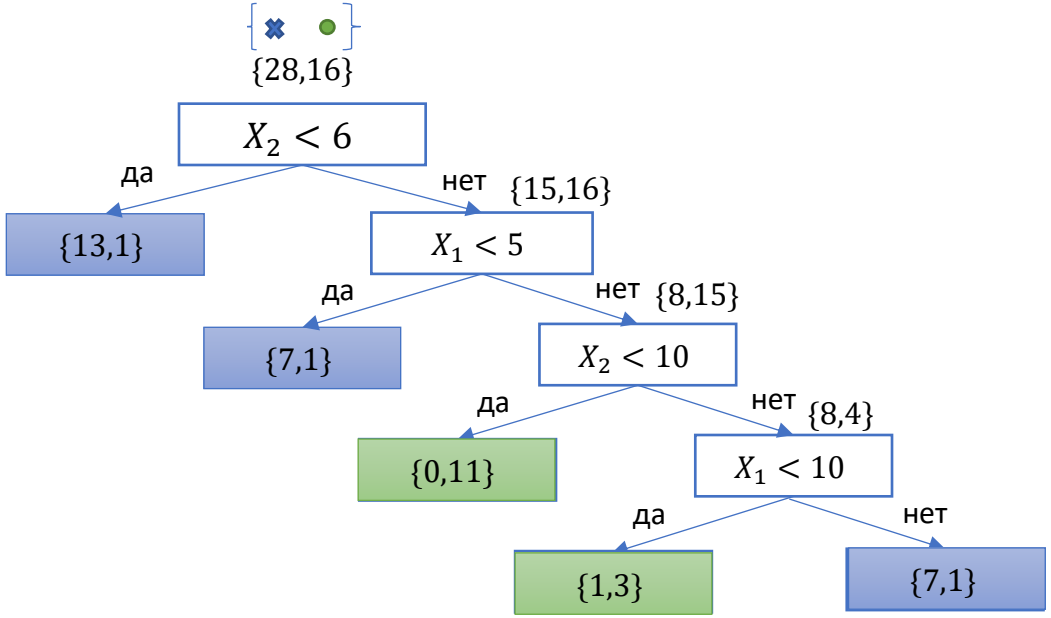



# Обучение с учителем: Решающие деревья

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

Пример для задачи классификации

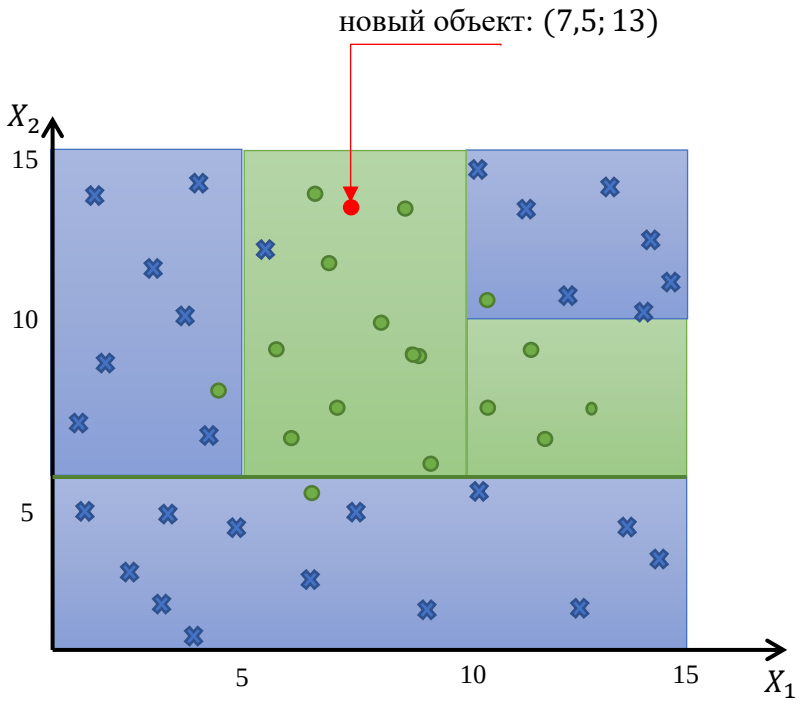


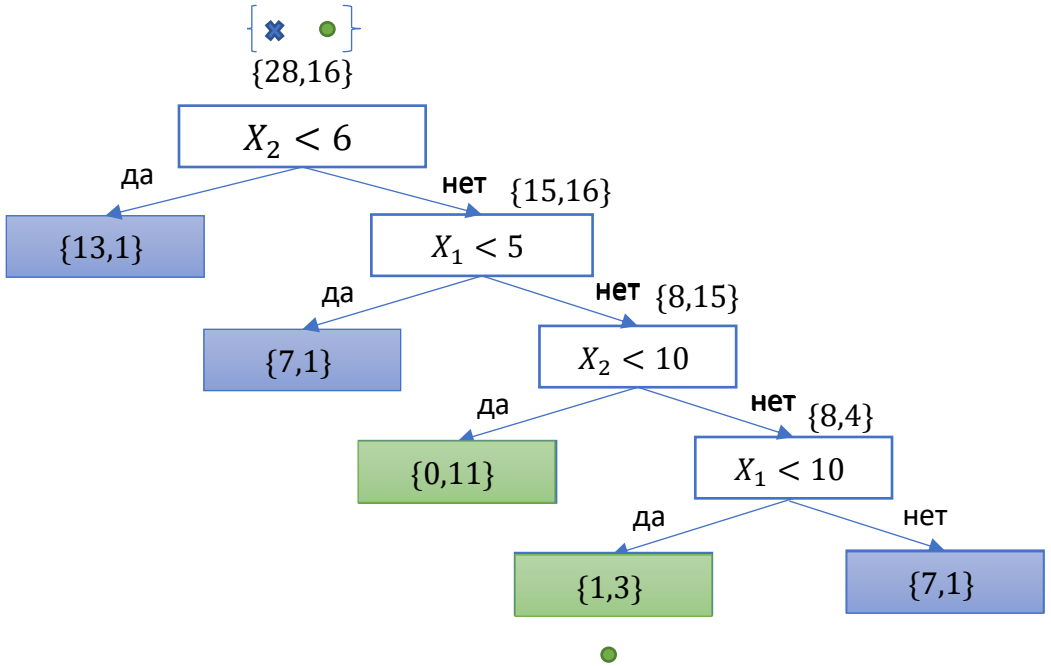



# Обучение с учителем: Решающие деревья

## 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))

Пример для задачи классификации

Дерево можно построить так, чтобы точность на обучающей выборке была 100%,
т.е. деревья легко переобучаются (например, как на рис.1)

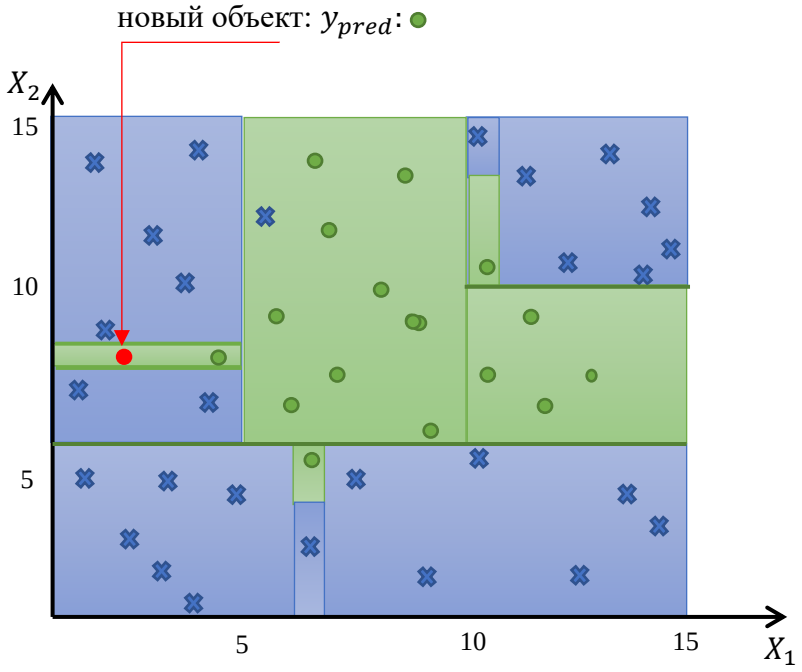



Рис.1

Чтобы дерево не переобучалось:

- глубина дерева должна быть ограничена;
- нужны условия на продолжение ветвления.

# 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 |
| $[X\Theta < t_i]$ | Предикат с линейной моделью |
| $[\rho(X, X_0) < t_i]$ | Предикат с метрикой |
| | … |

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

# Обучение с учителем: Решающие деревья

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

## Выбор предикатов. Жадный алгоритм построения

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

Обозначения:

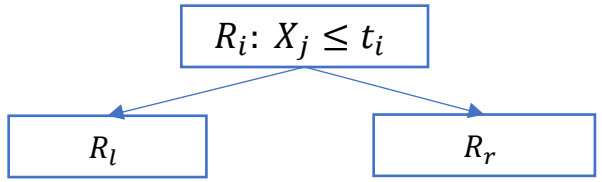
$R_i$ – текущая вершина;

$\Theta = (j, t_m)$ – 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.

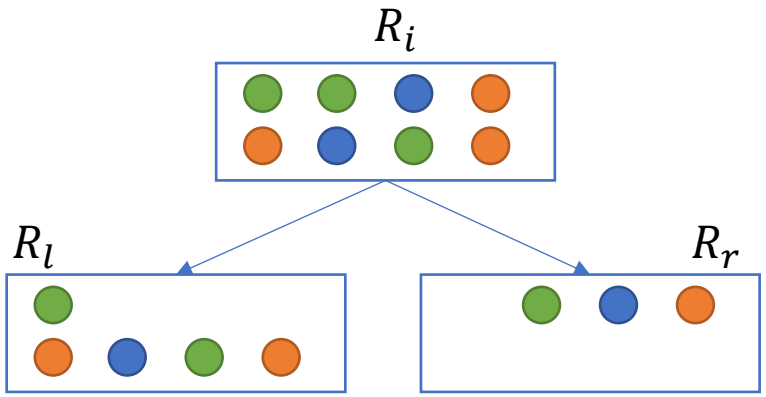


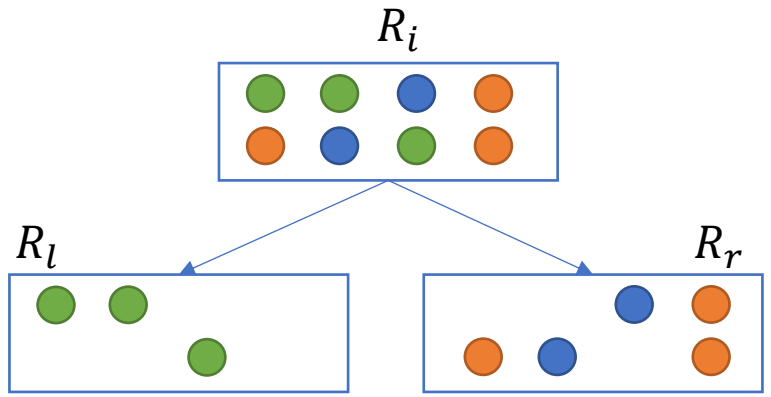
# Обучение с учителем: Решающие деревья

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

**Выбор предикатов. Жадный алгоритм построения**

Пример для классификации. Какое разбиение лучше?



# Обучение с учителем: Решающие деревья

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

### Выбор предикатов. Жадный алгоритм построения

*Критерий информативности* $H(R_i)$ – 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.

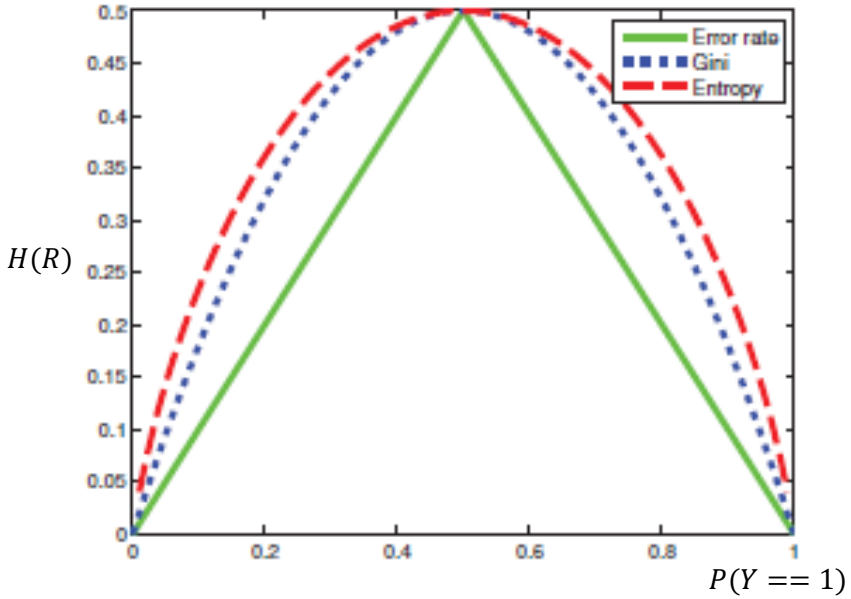



Рис. 2. Критерий информативности для бинарной классификации [1]. Ф. Энтропии масштабирована.

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

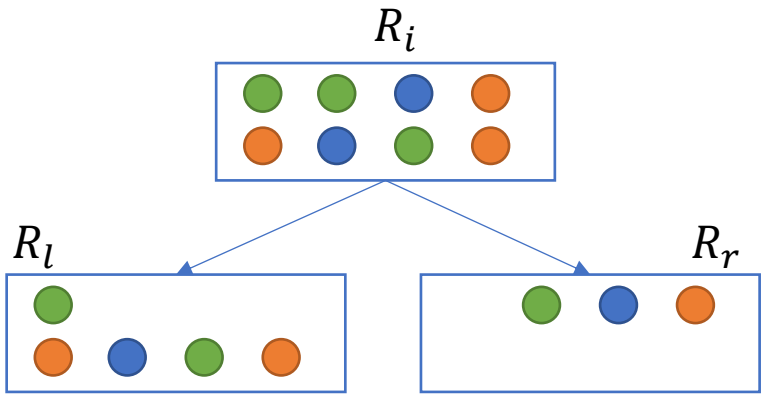
# Обучение с учителем: Решающие деревья

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

## Выбор предикатов. Жадный алгоритм построения

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

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

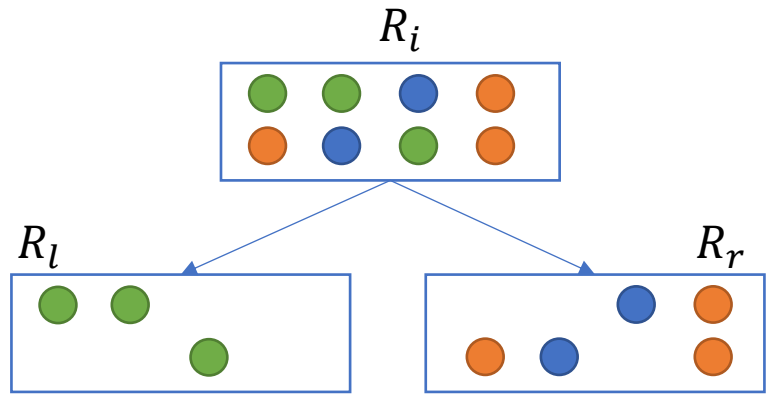



$H(R_l) =$

$H(R_r) =$

$$H(R_l) = \begin{bmatrix} p_з = \frac{2}{5} \\ p_с = \frac{1}{5} \\ p_о = \frac{2}{5} \end{bmatrix} = \frac{2}{5}\left(1-\frac{2}{5}\right) + \frac{1}{5}\left(1-\frac{1}{5}\right) + \frac{2}{5}\left(1-\frac{2}{5}\right) = \frac{16}{25}$$

$$H(R_r) = \left[p_з = p_с = p_о = \frac{1}{3}\right] = 3 \cdot \frac{1}{3}\left(1-\frac{1}{3}\right) = \frac{2}{3}$$




$H(R_l) =$

$H(R_r) =$

$$H(R_l) = \left[p_з = 1,\ p_с = p_о = 0\right]$$

$$= 1 \cdot (1-1) + 0 = 0$$

$$H(R_r) = \begin{bmatrix} p_з = 0 \\ p_с = \frac{2}{5} \\ p_о = \frac{3}{5} \end{bmatrix} =$$

$$= 0 + \frac{1}{5}\left(1-\frac{2}{5}\right) + \frac{3}{5}\left(1-\frac{3}{5}\right) = \frac{12}{25}$$

# Обучение с учителем: Решающие деревья

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

### Выбор предикатов. Жадный алгоритм построения

*Критерий качества* разбиения вершины $R_i$:

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

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.

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

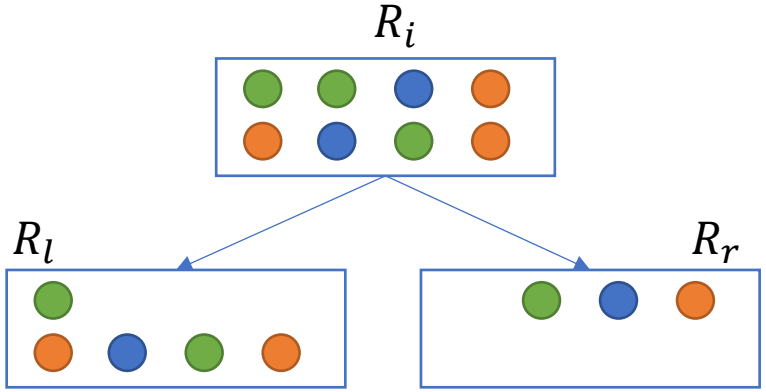
# Обучение с учителем: Решающие деревья

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

**Выбор предикатов. Жадный алгоритм построения**

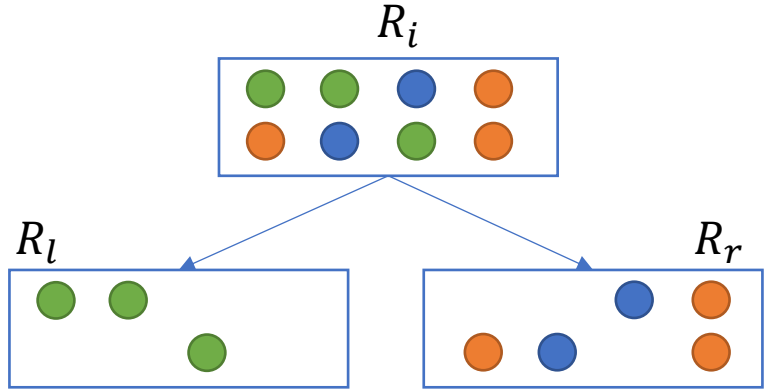
Рассчитать критерий качества разбиения вершины $R_i$ для предыдущей задачи. Какое разбиение лучше?

$$Q(R_i) = \frac{N_l}{N_m} H\big(R_l(\Theta)\big) + \frac{N_r}{N_m} H\big(R_r(\Theta)\big)$$




$H(R_l) = \frac{16}{25}$ $H(R_r) = \frac{1}{3}$

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




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

$Q(R_i) = \frac{3}{8} \cdot 0 + \frac{5}{8} \cdot \frac{12^3}{25^2} = \frac{6}{20}$

# Обучение с учителем: Решающие деревья

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

**Как остановить процесс построения дерева?**

Задать:

- Ограничение на глубину дерева;
- Ограничение количества листьев;
- Минимальное количество объектов в вершине;
- Минимальное уменьшение хаотичности при разбиении;
- и т.д.

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

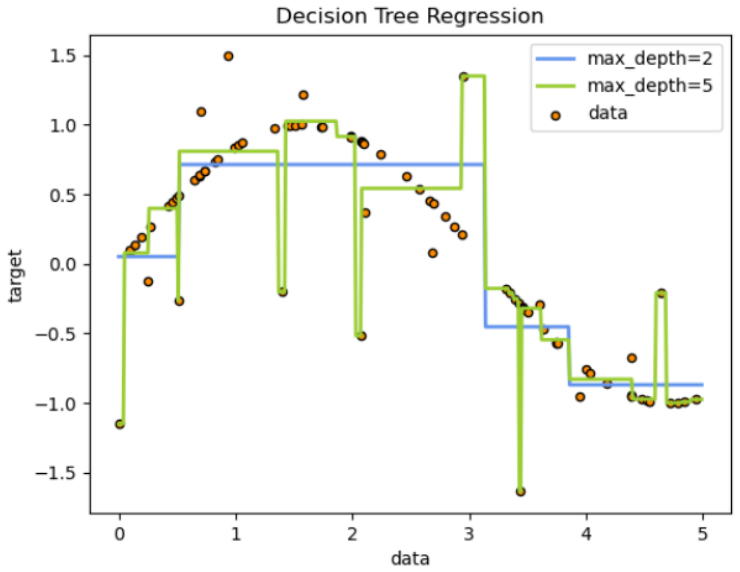
# Обучение с учителем: Решающие деревья

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

### Для задачи регрессии

Предсказание значения в i[ом] листе:

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




*Критерий информативности* - 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 $R_i$:

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

# Обучение с учителем: Решающие деревья

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

## Резюме

| Преимущества | Недостатки |
| --- | --- |
| Четкие правила классификации (легко интерпретировать) | Чувствительны к шумам в данных (модель сильно меняется при небольшом изменении обучающей выборки) |
| Легко визуализируются | Разделяющая граница имеет свои ограничения (состоит из гиперплоскостей) |
| Быстро обучаются и выдают прогноз | Необходимость борьбы с переобучением (стрижка или какой-либо из критериев останова) |
| Малое кол-во параметров | Проблема поиска оптимального дерева (на практике используется жадное построение дерева) |