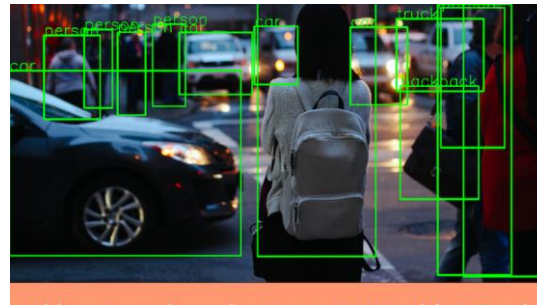
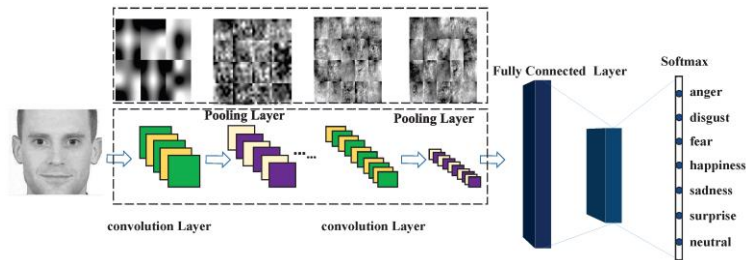


Машинное обучение



Приглашенная лекция 04

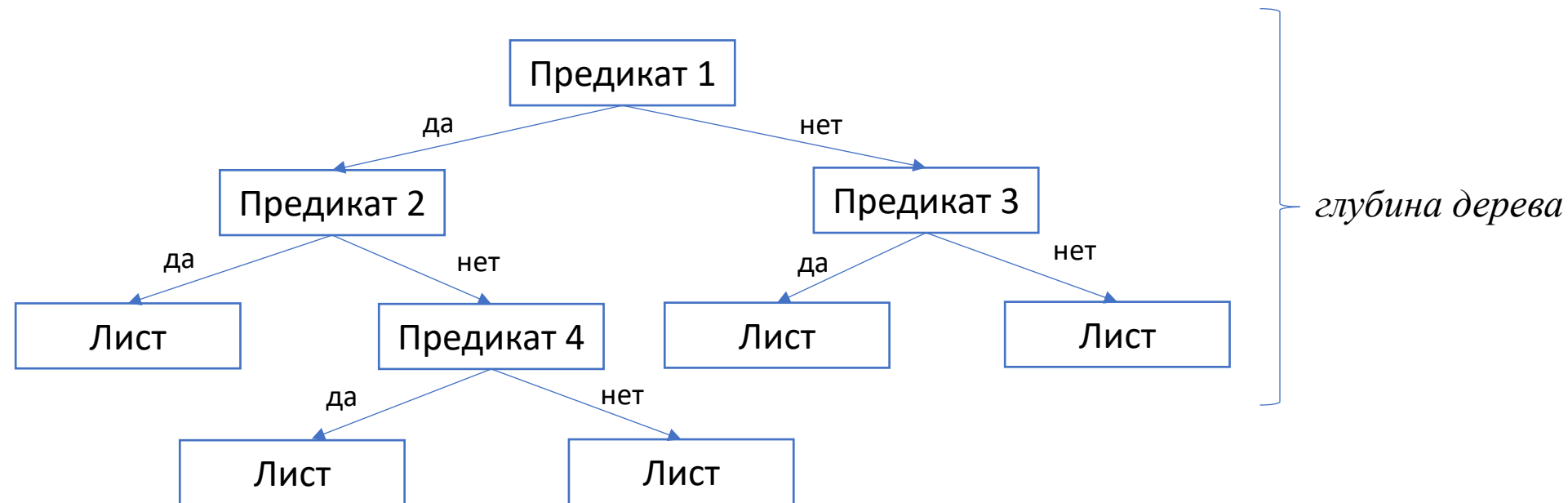
Обучение с учителем: Решающие деревья и случайный лес

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]



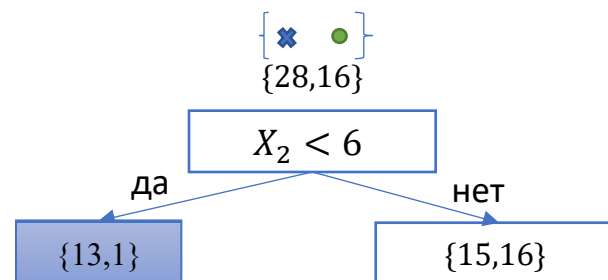
Вершины дерева содержат предикаты вида $X_j \leq t_m$

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

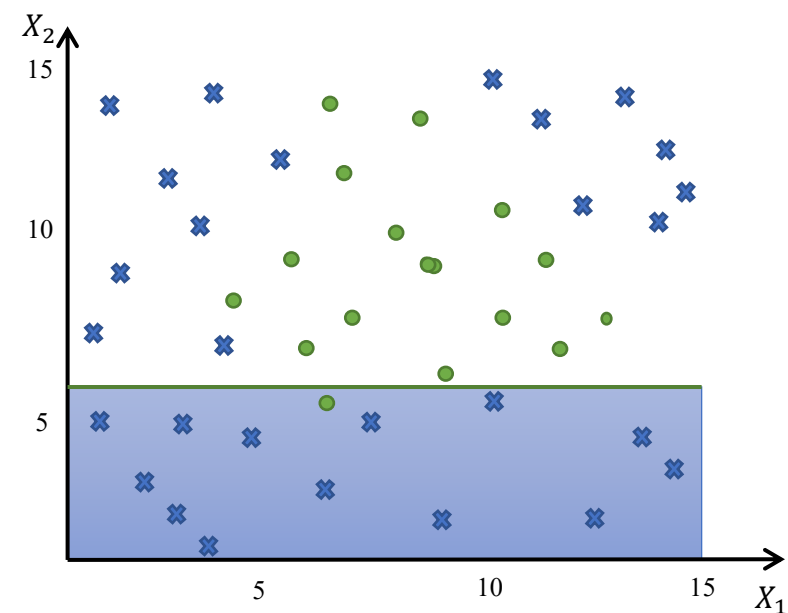
Обучение с учителем: Решающие деревья

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

Пример для задачи классификации



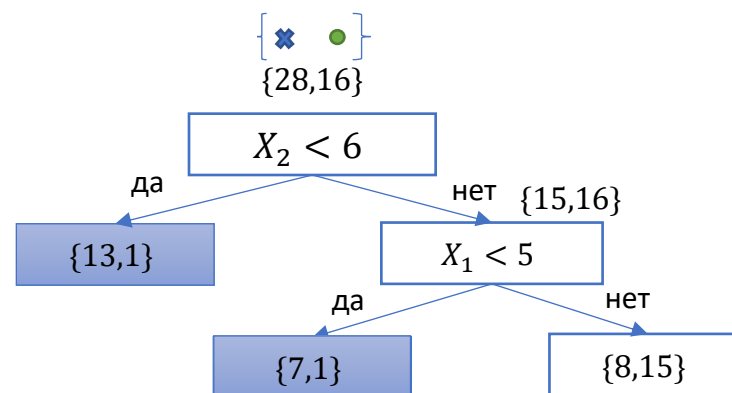
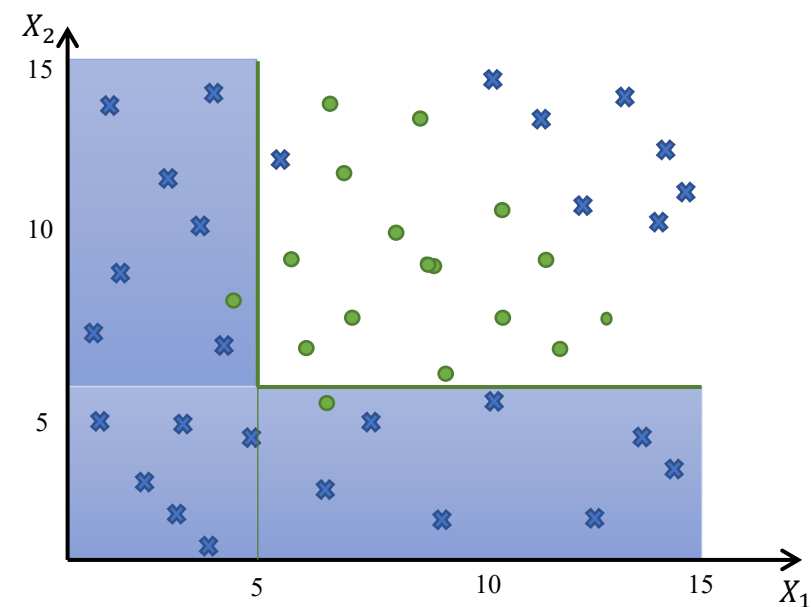
$$y_{pred} = \operatorname{argmax}(p_k)$$



Обучение с учителем: Решающие деревья

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

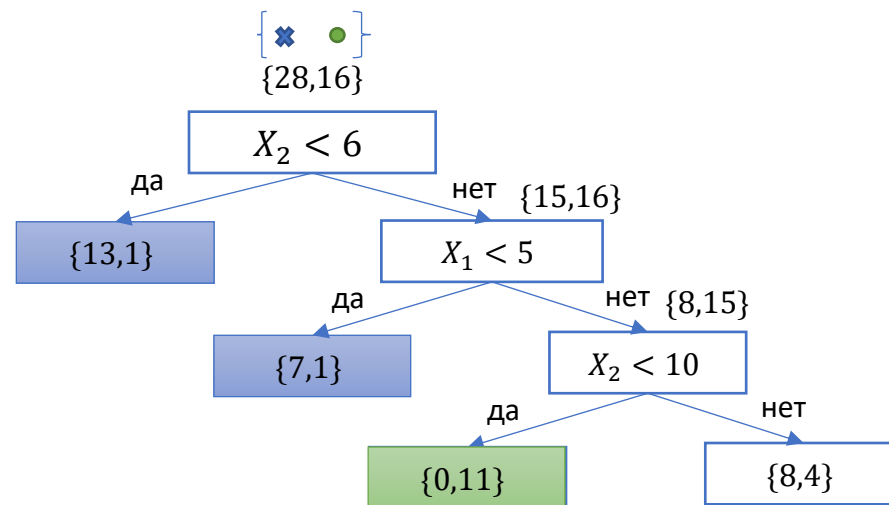
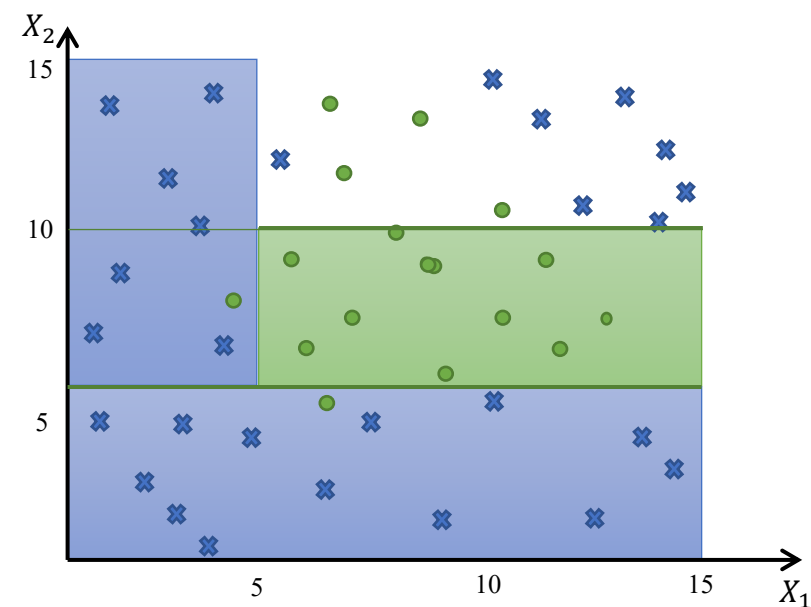
Пример для задачи классификации



Обучение с учителем: Решающие деревья

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

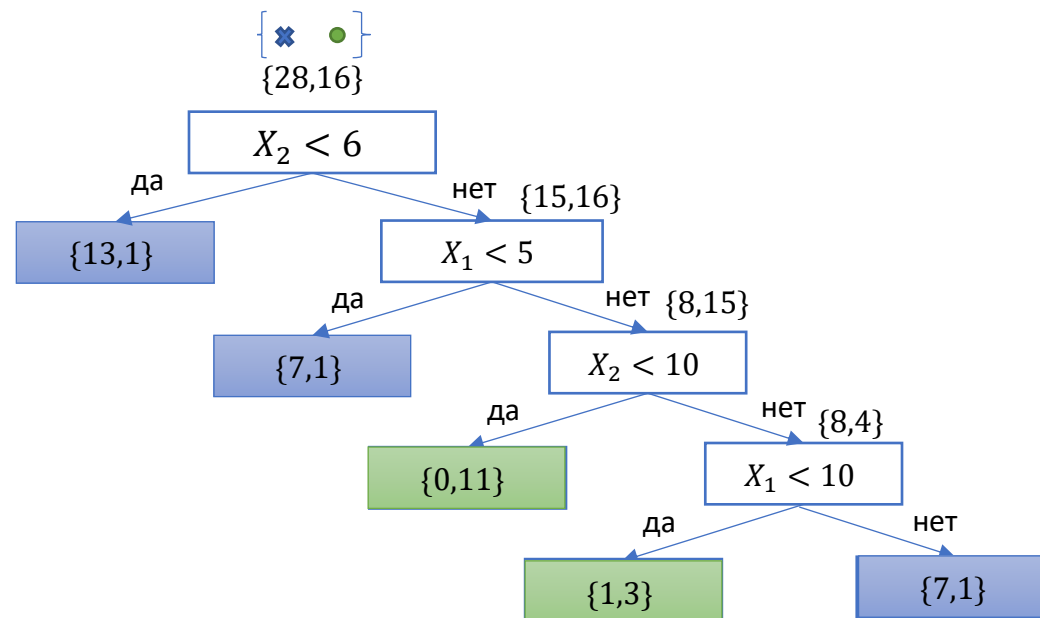
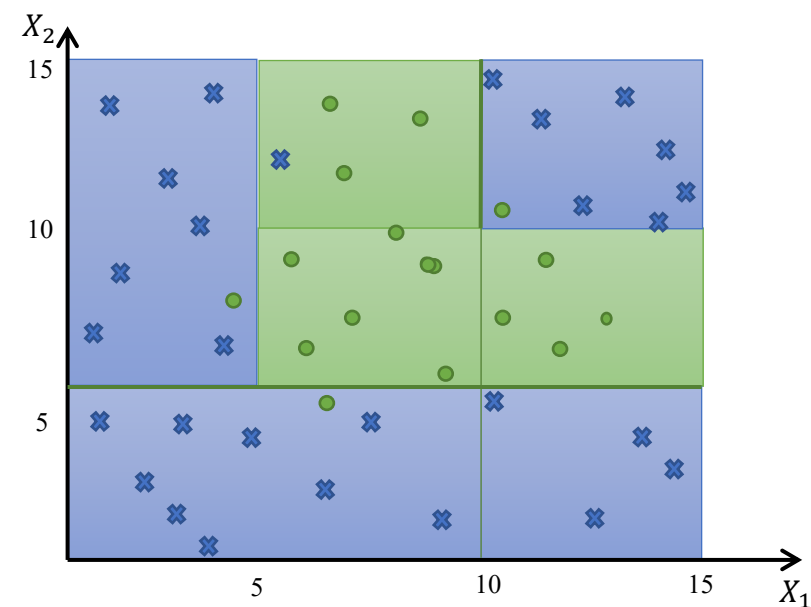
Пример для задачи классификации



Обучение с учителем: Решающие деревья

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

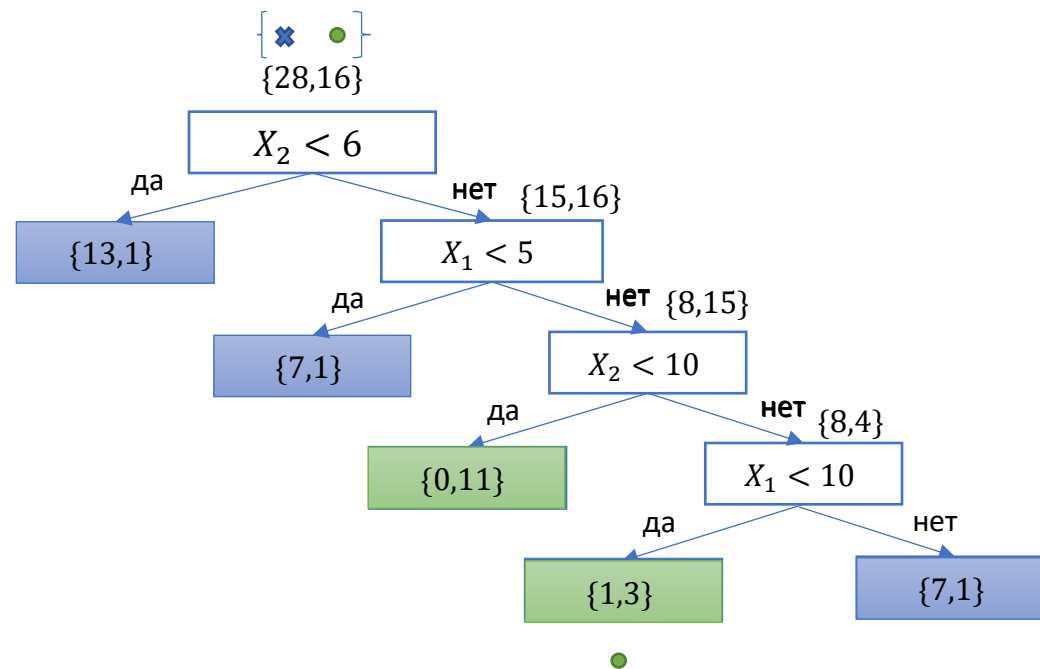
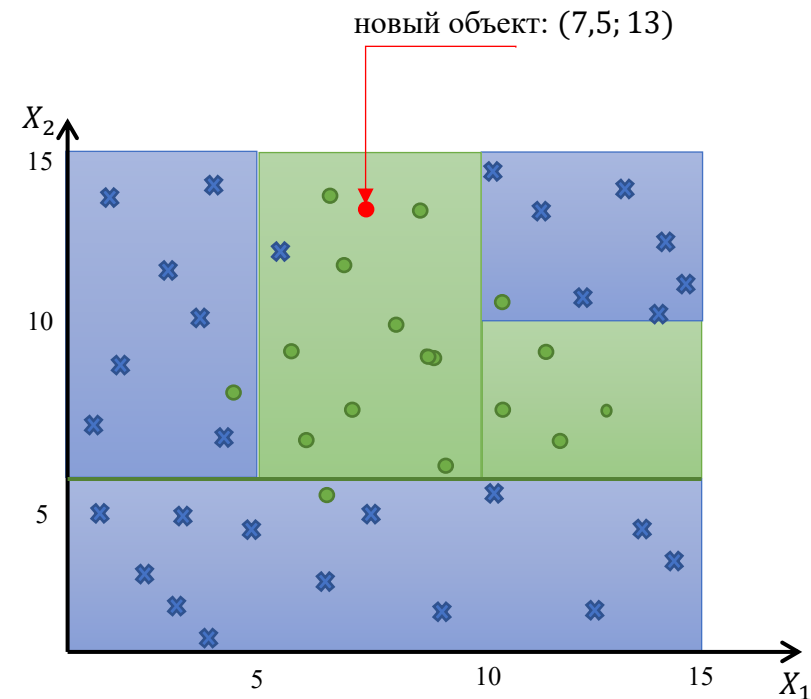
Пример для задачи классификации



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)

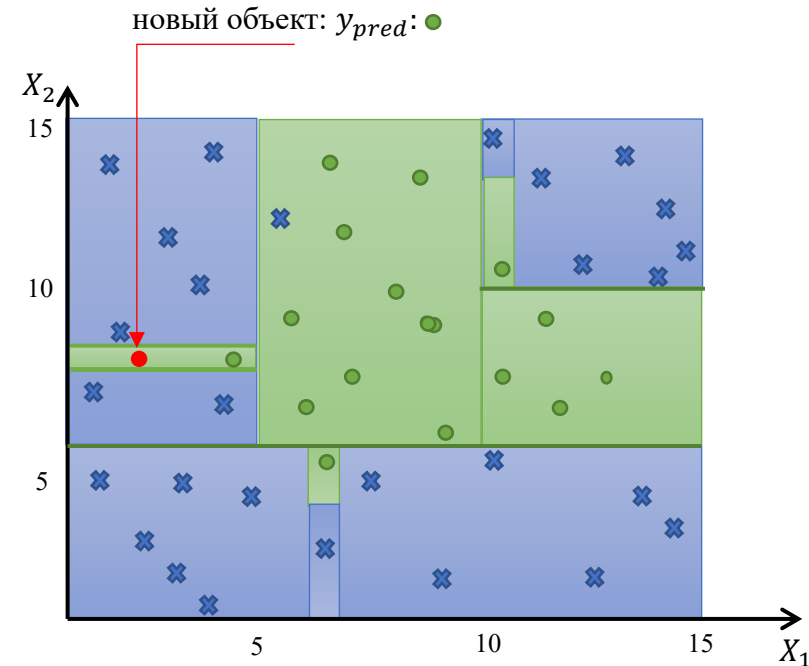


Рис.1

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: $Q(R_i, \Theta) \rightarrow \min$

Обозначения:

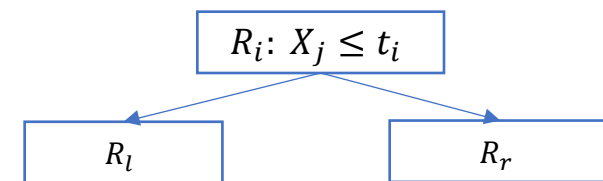
R_i – текущая вершина;

$\Theta = (j, t_m)$ – 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;

$R_r(\Theta) = R_i \setminus R_l(\Theta)$ – объекты, попавшие в правую вершину;

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

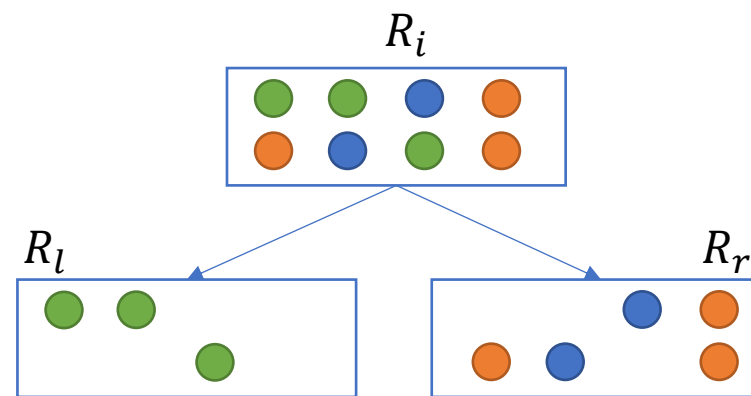
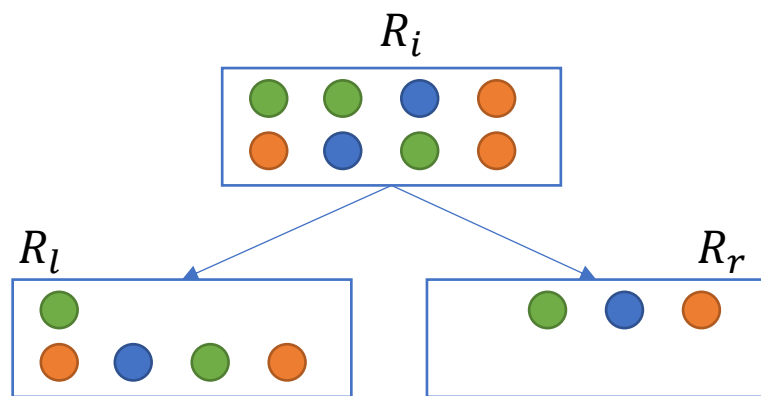


Обучение с учителем: Решающие деревья

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.

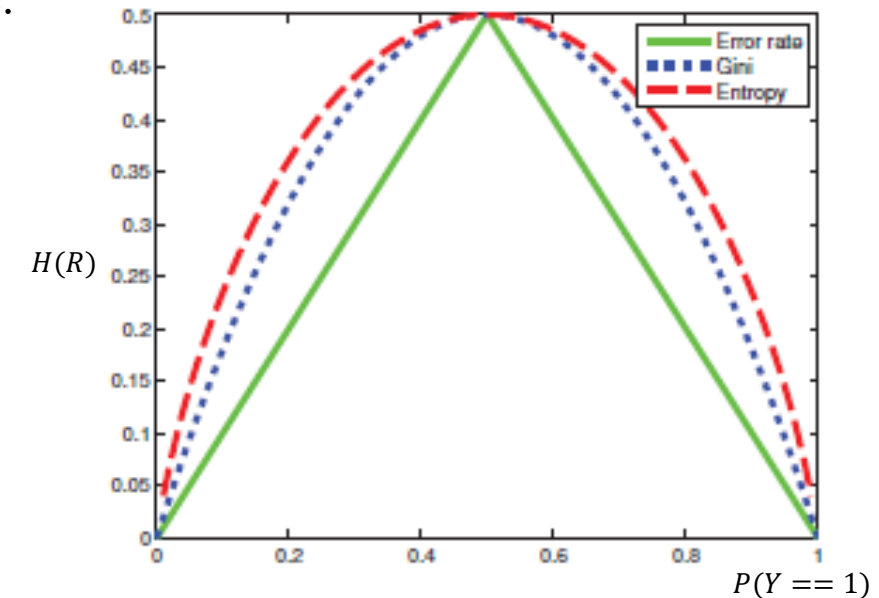


Рис. 2. Критерий информативности для бинарной классификации [1]. Ф. Энтропии масштабирована.

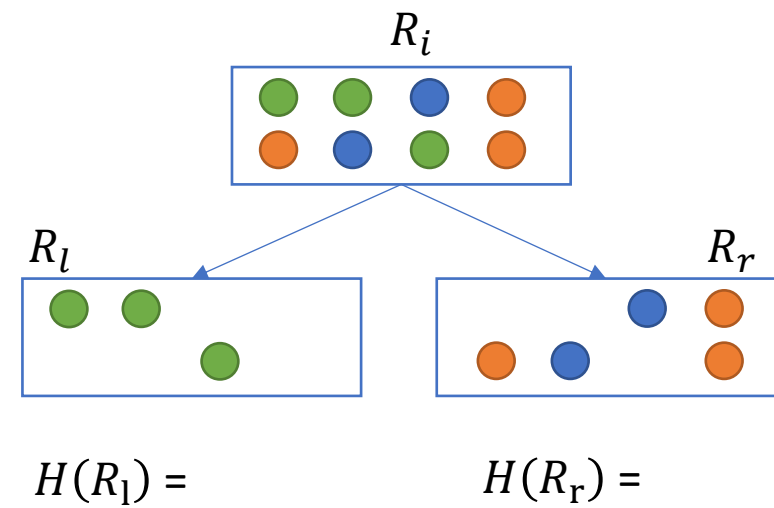
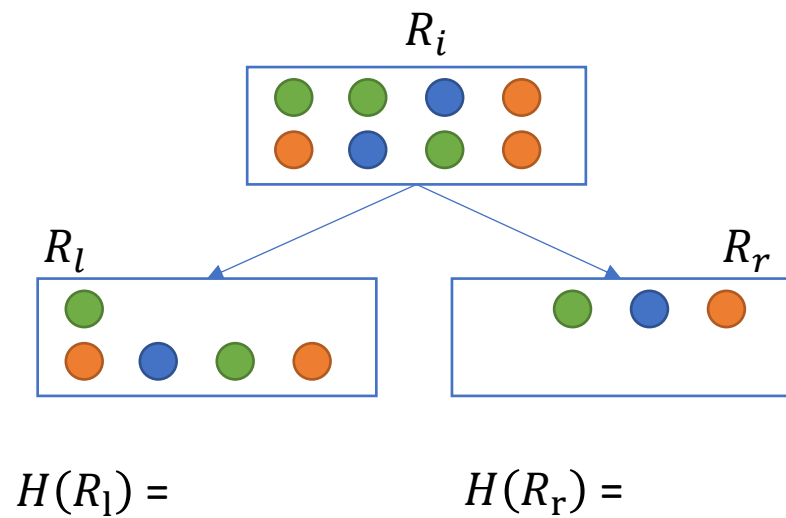
Обучение с учителем: Решающие деревья

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:

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



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 :

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

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.

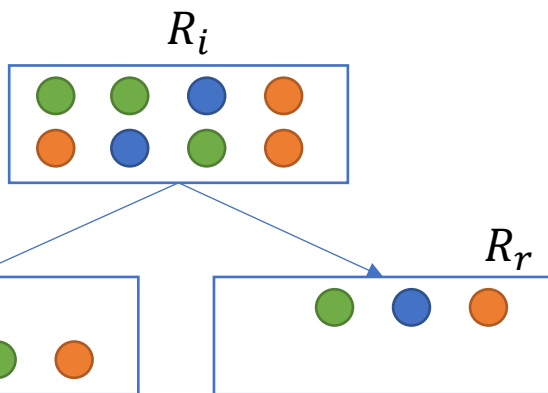
Обучение с учителем: Решающие деревья

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?

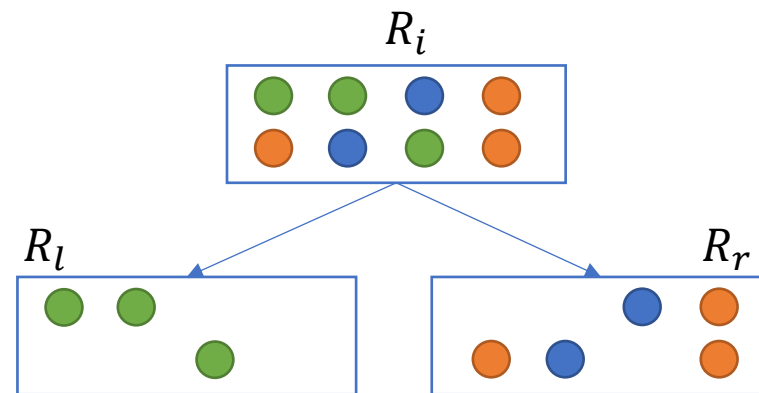
$$Q(R_i) = \frac{N_l}{N_m} H(R_l(\Theta)) + \frac{N_r}{N_m} H(R_r(\Theta))$$



$H(R_l) =$

$H(R_r) =$

$Q(R_i) =$



$H(R_l) =$

$H(R_r) =$

$Q(R_i) =$

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.

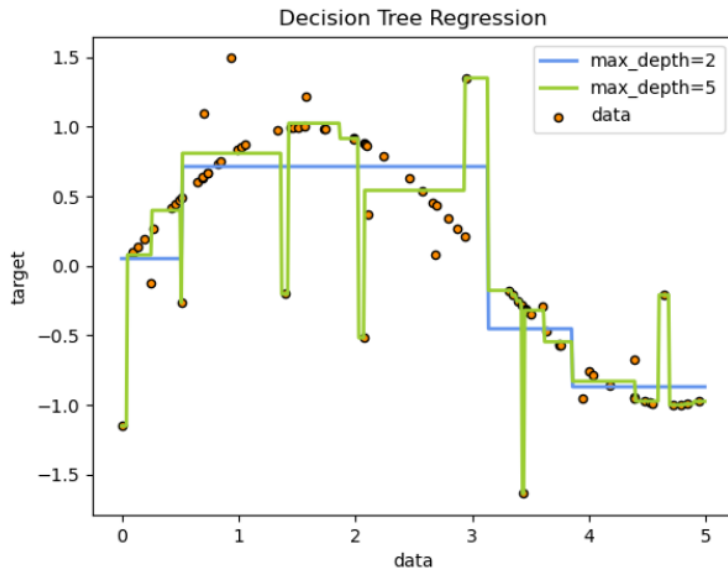
Обучение с учителем: Решающие деревья

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

Для задачи регрессии

Предсказание значения в $i^{\text{ом}}$ 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 :

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



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)