

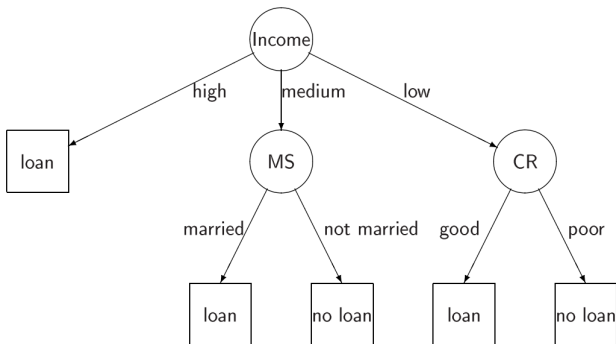
# Решающие деревья

Виктор Владимирович Китов

# Содержание

- 1 Определение решающего дерева
- 2 Правила перехода
- 3 Выбор правил перехода
- 4 Назначение прогнозов листовым вершинам
- 5 Критерий остановки

# Пример решающего дерева



## Определение решающего дерева

- Прогнозирование осуществляется деревом  $T$ :
  - направленный граф
  - без циклов
  - с одной корневой вершиной

## Определение решающего дерева

- каждому узлу  $t$  соответствуют функции от признаков  $Q_t(x)$
- допустим, из узел  $t$  содержит  $K(t)$  дочерних узлов  $t_1, \dots, t_{K(t)}$
- переход в узел  $t_i$  осуществляется, если  $Q_t(x) \in S_i(t)$
- Множества  $S_1(t), \dots, S_{K(t)}(t)$  таковы, что:
  - $\bigcup_k S_k(t)$  - все множество значений  $Q_t(x)$
  - $S_i(t) \cap S_j(t) = \emptyset \quad \forall i \neq j$

## Процесс прогнозирования

- множество вершин делится на:
  - внутренние вершины  $int(T)$ , каждая из которых имеет  $\geq 2$  дочерних вершины
  - внешние вершины  $terminal(T)$ , которые не содержат потомков, но которым сопоставлены прогнозные значения.

## Процесс прогнозирования

- множество вершин делится на:
  - внутренние вершины  $int(T)$ , каждая из которых имеет  $\geq 2$  дочерних вершины
  - внешние вершины  $terminal(T)$ , которые не содержат потомков, но которым сопоставлены прогнозные значения.
- Процесс прогнозирования для дерева  $T$ :
  - $t = root(T)$
  - пока  $t$  не является листом:
    - рассчитать  $Q_t(x)$
    - определить номер множества  $j$ , среди  $S_1(t), \dots, S_{K(t)}(t)$ , куда попадет  $Q_t(x)$ :  $Q_t(x) \in S_j(t)$
    - перейти по ребру  $r_j(t)$  к  $j$ -му потомку  $t$ :  
 $t = (j\text{-й потомок } t)$ .
  - вернуть прогноз, сопоставленный листу  $t$ .

## Спецификация решающего дерева

- Определить решающее правило:  $Q_t(x)$ ,  $K(t)$  и  $S_1(t), \dots, S_{K(t)}(t)$ .
- Определить критерий останова (когда делаем вершину листом дерева).
- Определить сопоставление прогнозов каждому листу.



# Содержание

- 1 Определение решающего дерева
- 2 Правила перехода**
- 3 Выбор правил перехода
- 4 Назначение прогнозов листовым вершинам
- 5 Критерий остановки

## Правила перехода в алгоритме CART

- Рассматривается единственный признак:

$$Q_t(x) = x^{i(t)}$$

- На основе него делается бинарное разбиение:

$$K(t) = 2$$

- Признак сравнивается с порогом:

$$S_1 = \{x^{i(t)} \leq \text{threshold}(t)\}, S_2 = \{x^{i(t)} > \text{threshold}(t)\}$$

- $\text{threshold}(t) \in \{x_1^{i(t)}, x_2^{i(t)}, \dots, x_N^{i(t)}\}$

- применимо только к вещественным, упорядоченным и бинарным признакам
- дискретные неупорядоченные признаки:

## Правила перехода в алгоритме CART

- Рассматривается единственный признак:

$$Q_t(x) = x^{i(t)}$$

- На основе него делается бинарное разбиение:

$$K(t) = 2$$

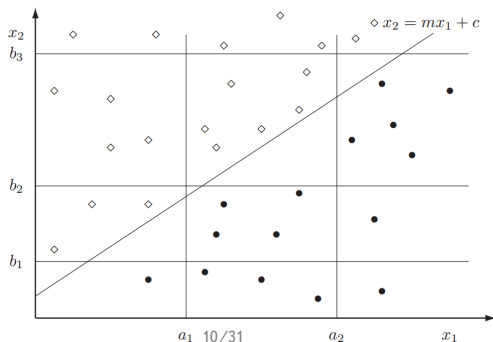
- Признак сравнивается с порогом:

$$S_1 = \{x^{i(t)} \leq \text{threshold}(t)\}, S_2 = \{x^{i(t)} > \text{threshold}(t)\}$$

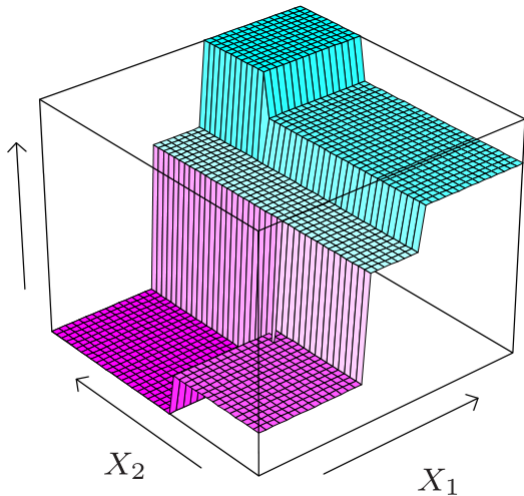
- $\text{threshold}(t) \in \{x_1^{i(t)}, x_2^{i(t)}, \dots, x_N^{i(t)}\}$ 
  - применимо только к вещественным, упорядоченным и бинарным признакам
  - дискретные неупорядоченные признаки: можно применить one-hot encoding.

# Анализ правила перехода CART

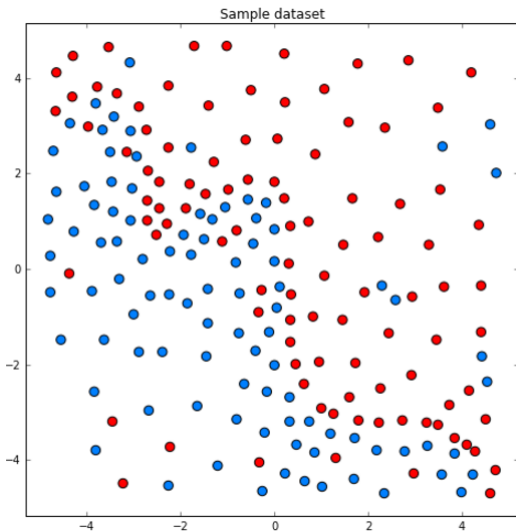
- Преимущества:
  - простота
  - интерпретируемость
- Недостатки
  - много узлов может потребоваться, если разделяющая кривая не параллельна осям координат:



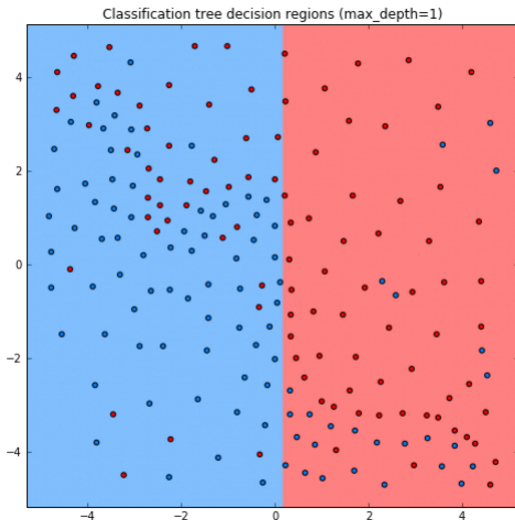
# Кусочно-постоянное решение



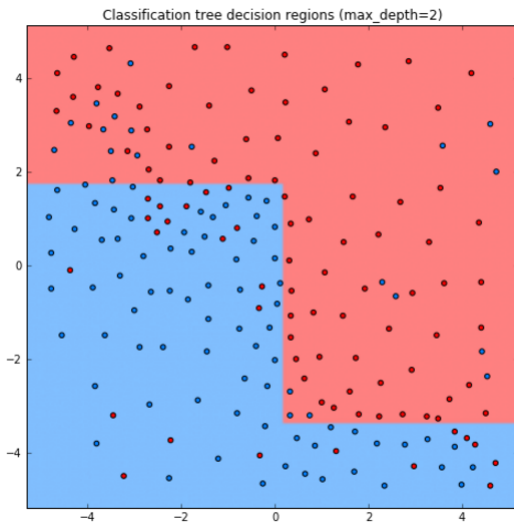
# Пример: обучающая выборка



# Пример: классификация решающим деревом

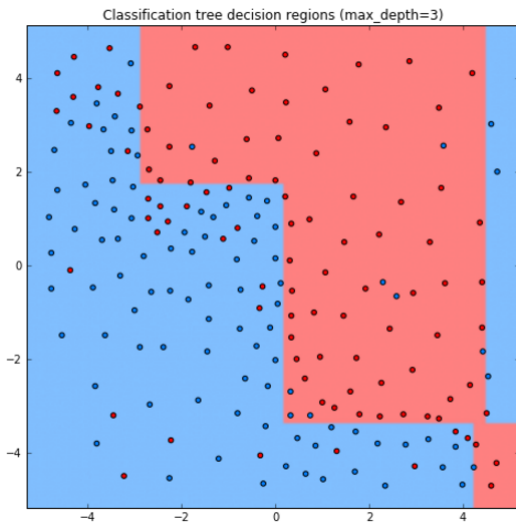


# Пример: классификация решающим деревом

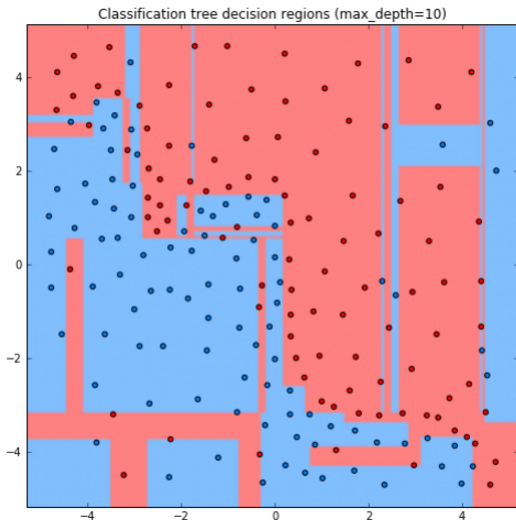




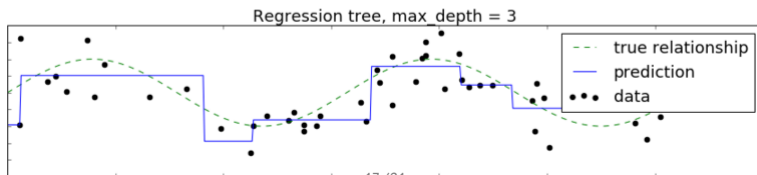
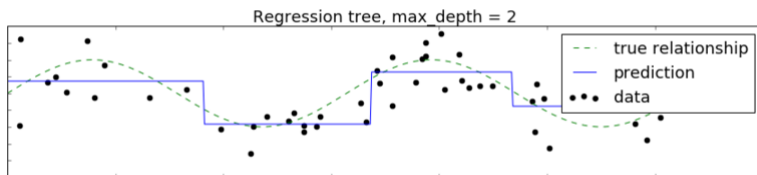
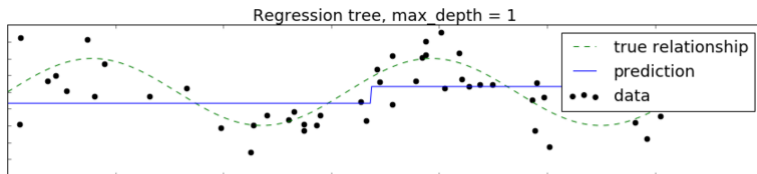
# Пример: классификация решающим деревом



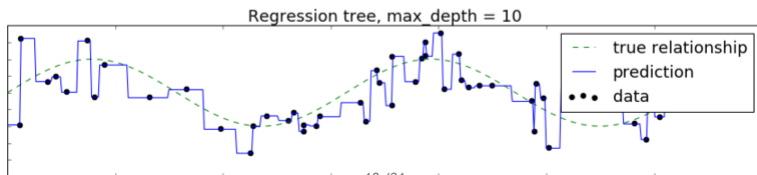
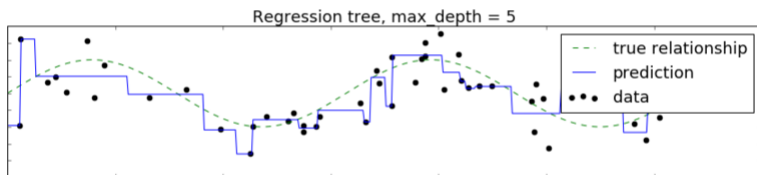
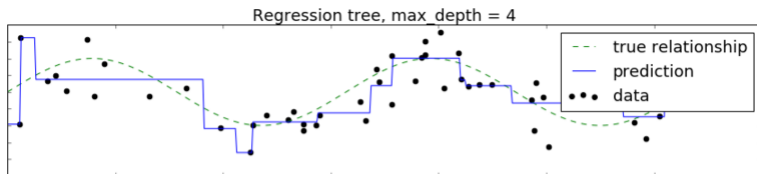
# Пример: классификация решающим деревом



# Пример: регрессия решающим деревом



# Пример: регрессия решающим деревом



# Содержание

- 1 Определение решающего дерева
- 2 Правила перехода
- 3 Выбор правил перехода**
- 4 Назначение прогнозов листовым вершинам
- 5 Критерий остановки

## Функция смешанности классов

- Пусть  $u(t)$ -область точек, которые попадают в узел  $t$ . Вероятности классов в узле  $t$ :

$$p(\omega_j | x \in u(t)) = p(\omega_j | t) \approx \frac{N_j(t)}{N(t)}$$

- Функция смешанности (impurity function):

$$I(t) = \phi(p(\omega_1 | t), \dots, p(\omega_C | t))$$

- $\phi(q_1, q_2, \dots, q_C)$  определена для  $q_j \geq 0$  и  $\sum_j q_j = 1$ .
- $\phi$  достигает максимума при  $q_j = 1/C$  для всех  $j$ .
- $\phi$  достигает минимума, когда  $\exists j : q_j = 1, q_i = 0$  для всех  $i \neq j$ .
- $\phi$  симметричная функция от  $q_1, q_2, \dots, q_C$ .

## Типичные функции смешанности классов

- Критерий Джини

- вероятность сделать ошибку, сопоставляя класс случайно с вероятностями  $[p(\omega_1|t), \dots, p(\omega_C|t)]$

$$I(t) = \sum_i p(\omega_i|t)(1 - p(\omega_i|t)) = 1 - \sum_i [p(\omega_i|t)]^2$$

- Энтропия

- мера неопределенности дискретной случайной величины

$$I(t) = - \sum_i p(\omega_i|t) \ln p(\omega_i|t)$$

- Ошибка классификации

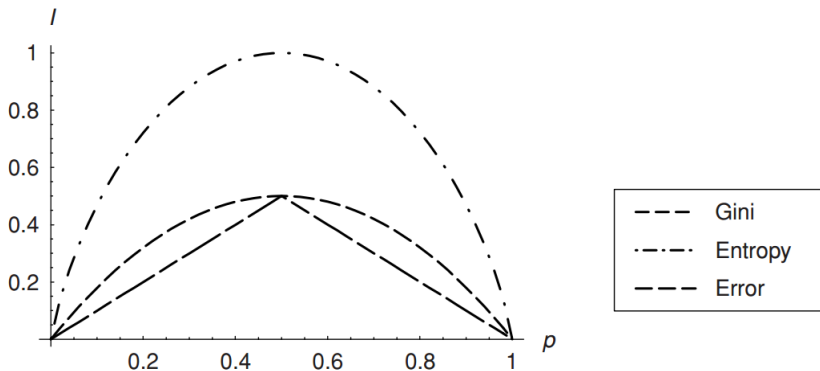
- показывает частоту ошибок при классификации самым частым классом

$$I(t) = 1 - \max_i p(\omega_i|t)$$

## Типичные функции смешанности классов

Функции смешанности в случае 2х классов:

$$p = p(\omega_1|t), \quad 1 - p = p(\omega_2|t).$$





## Улучшение функции смешанности

- Выбрать правило ветвления, максимизирующее снижение смешанности классов:

$$\Delta I(t) = I(t) - \sum_{i=1}^S I(t_i) \frac{N(t_i)}{N(t)}$$

где  $S$  - число ветвей,  $t_1, \dots, t_S$  дочерние вершины для  $t$ ,  $N(t)$  - число наблюдений в  $t$ , а  $N_j(t)$  - число наблюдений класса  $\omega_j$  в  $t$ .

- Если  $I(t)$  - энтропия, то  $\Delta I(t)$  называется *information gain*.

## Улучшение функции смешанности

- Выбрать правило ветвления, максимизирующее снижение смешанности классов:

$$\Delta I(t) = I(t) - \sum_{i=1}^S I(t_i) \frac{N(t_i)}{N(t)}$$

где  $S$  - число ветвей,  $t_1, \dots, t_S$  дочерние вершины для  $t$ ,  $N(t)$  - число наблюдений в  $t$ , а  $N_j(t)$  - число наблюдений класса  $\omega_j$  в  $t$ .

- Если  $I(t)$  - энтропия, то  $\Delta I(t)$  называется *information gain*.

## CART: выбор правила для узла

- CART: выбрать признак  $k(t)$  и порог  $h(t)$ , которые максимизируют  $\Delta I(t)$ :

$$k(t), h(t) = \arg \max_{f,h} \Delta I(t)$$

- CART: из узла  $t$  перейти в дочерний узел:

$$\begin{cases} t_1, & \text{если } x^{k(t)} \geq h(t) \\ t_2, & \text{если } x^{k(t)} < h(t) \end{cases}$$

# Содержание

- 1 Определение решающего дерева
- 2 Правила перехода
- 3 Выбор правил перехода
- 4 Назначение прогнозов листовым вершинам**
- 5 Критерий остановки

## Назначение прогнозов листьям

- Регрессия:

$$\hat{y} = \arg \min_{\mu} \sum_{i: x_i \in u(t)} (y_i - \mu)^2 = \text{mean}\{y_i\},$$

$$\hat{y} = \arg \min_{\mu} \sum_{i: x_i \in u(t)} |y_i - \mu| = \text{median}\{y_i\},$$

где  $N = |\{x_i : x_i \in u(t)\}|$

- могут использоваться и др. функции потерь.
- Классификация:
  - Самый частый класс
  - Сопоставление класса с минимальной совокупной ценой

# Содержание

- 1 Определение решающего дерева
- 2 Правила перехода
- 3 Выбор правил перехода
- 4 Назначение прогнозов листовым вершинам
- 5 Критерий остановки**
  - **Правиловые подходы**

## Критерий останова

- Противоречие: смещение/дисперсия
  - наиболее разросшиеся деревья - переобучение
  - слишком простые - недообучение
- Подходы к остановке
  - основанные на правиле: сравниваем критерий с порогом
  - строим до самого низа, а потом обрезаем лишнее (pruning)

- 5 Критерий остановки
  - Правиловые подходы



## Правиловый подход к остановке

- Делаем вершину листом, если критерий больше или меньше порога.
- Варианты критерия:
  - глубина дерева
  - количество наблюдений в вершине
  - минимальное количество наблюдений в одной из дочерних вершин
  - делимость классов
  - изменение делимости классов после дробления

# Анализ решающих деревьев в целом

- Преимущества решающих деревьев:
  - простота
  - интерпретируемость
  - встроенный отбор признаков
  - работает одновременно с дискретными и непрерывными признаками
  - прогноз инвариантен к монотонным преобразованиям признаков для  $Q_t(x) = x^{i(t)}$ 
    - не нужна нормализация
- Недостатки решающих деревьев:
  - если разделяющая кривая не параллельна осям координат, то может потребоваться много вершин
  - субоптимальность (пример XOR)
  - для новых наблюдений требуется полная перестройка всего дерева.