Основы машинного обучения

Тема: Решающие деревья

Выполнили: Губанов Алексей, Шевченко Глеб

СПбПУ 2024

Мотивация

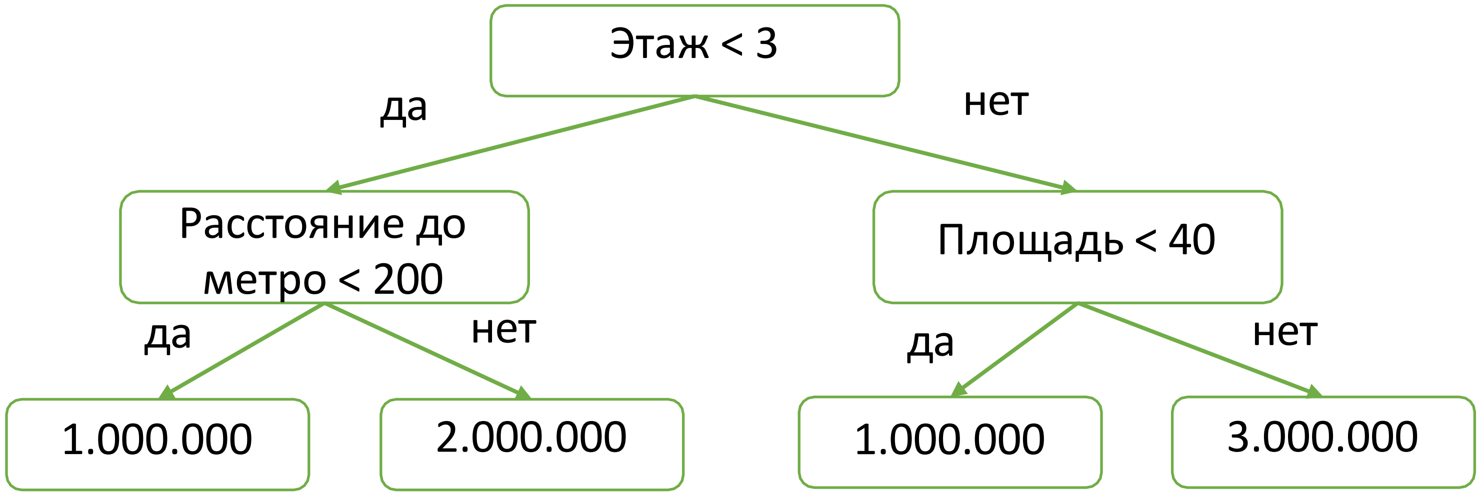
- Легко интерпретировать данные
- Находят нелинейные закономерности

Для этого нужно

- как-то искать хорошие логические правила
- Уметь составлять модели из логических правил

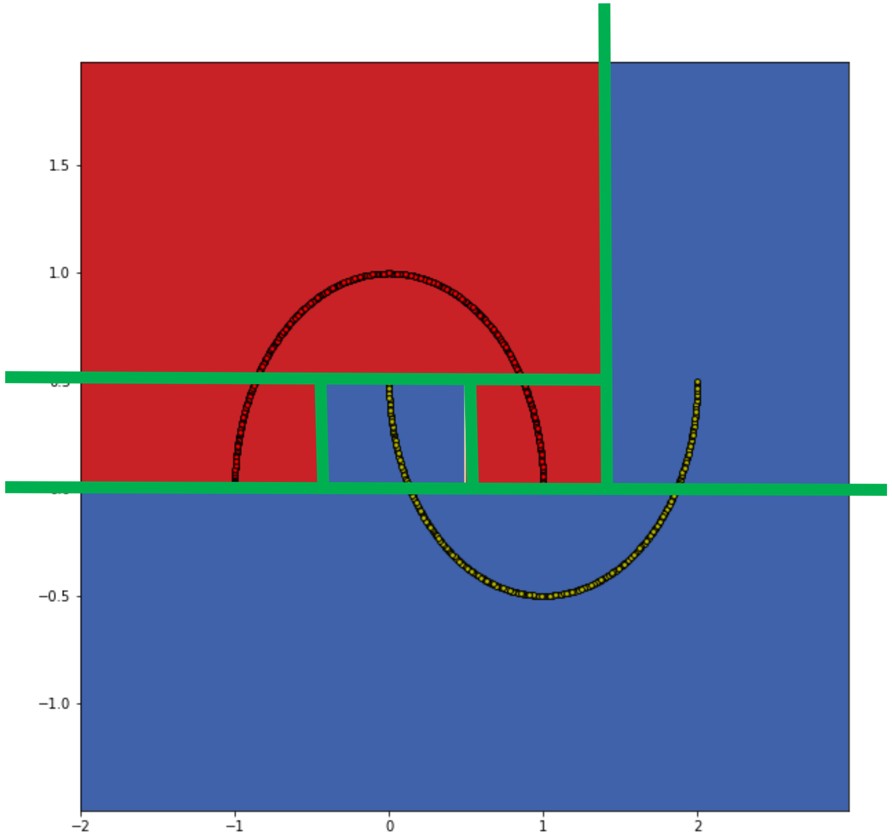
Примеры

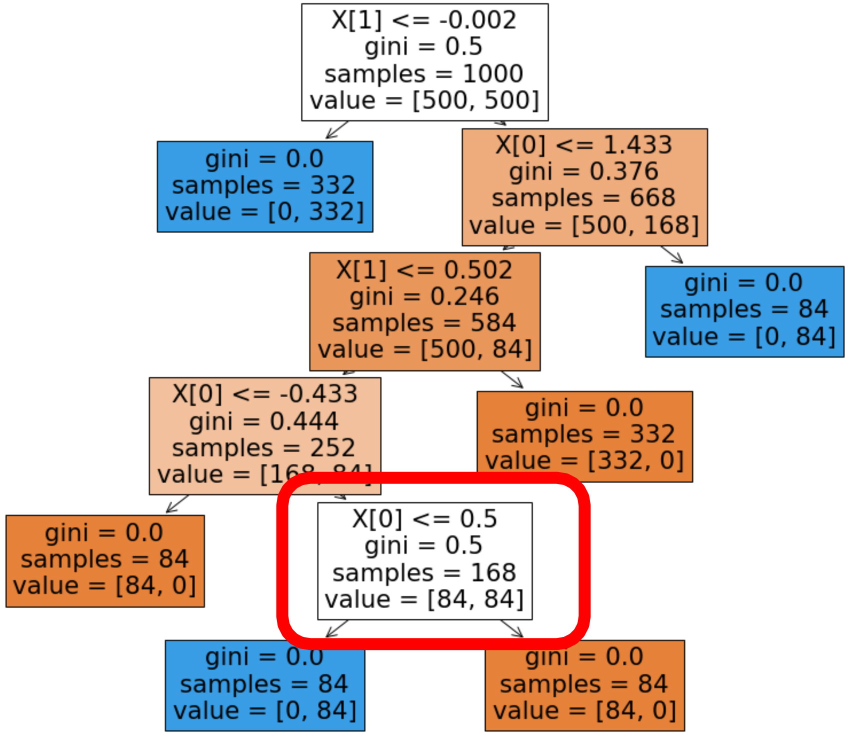
Решающее дерево



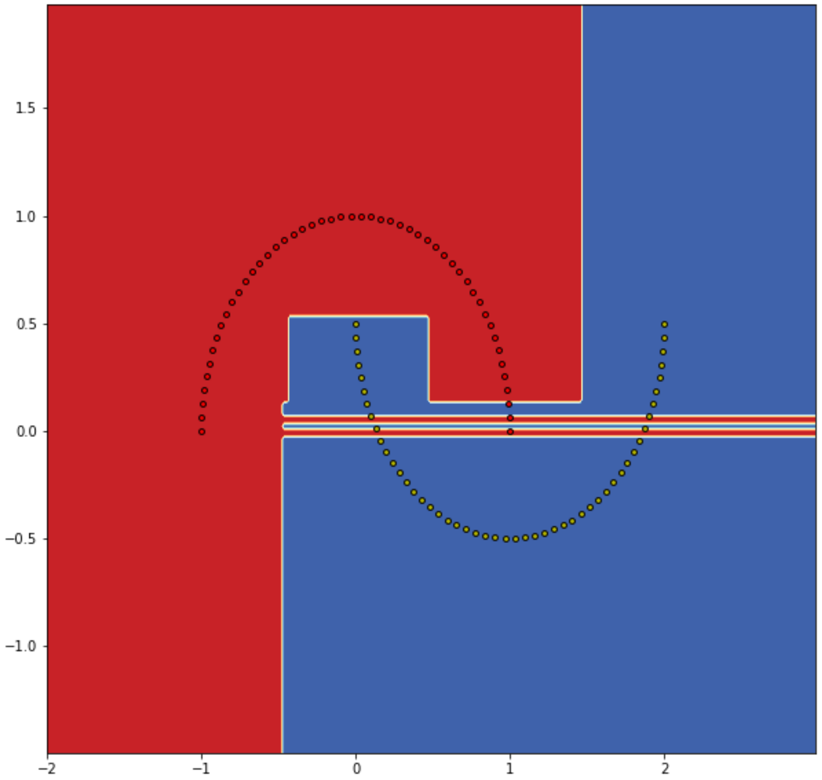
- Внутренние вершины: предикаты $[x_j < t]$
- Листья: прогнозы $c \in \mathbb{Y}$

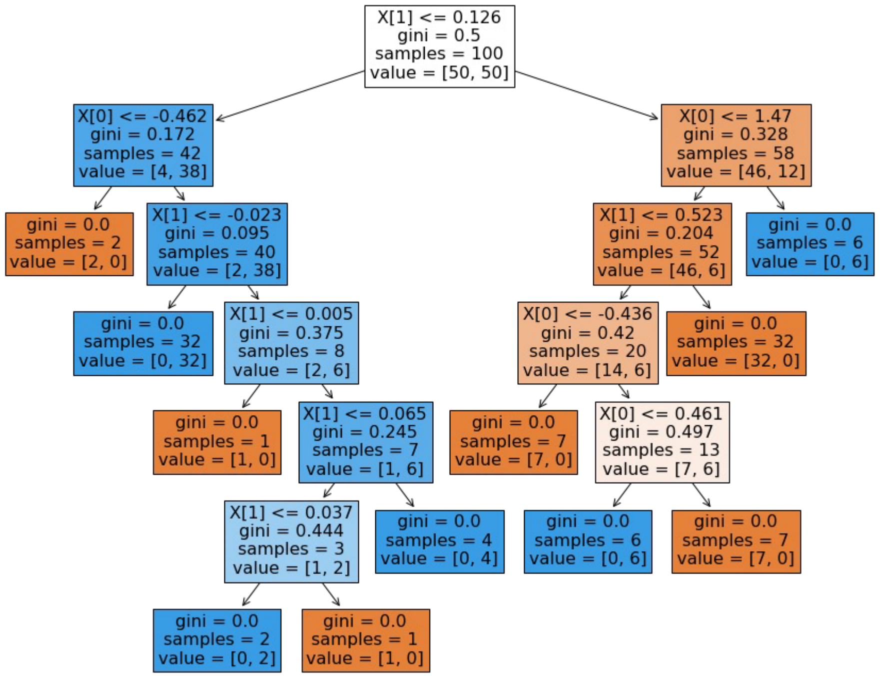
Решающее дерево





Решающее дерево

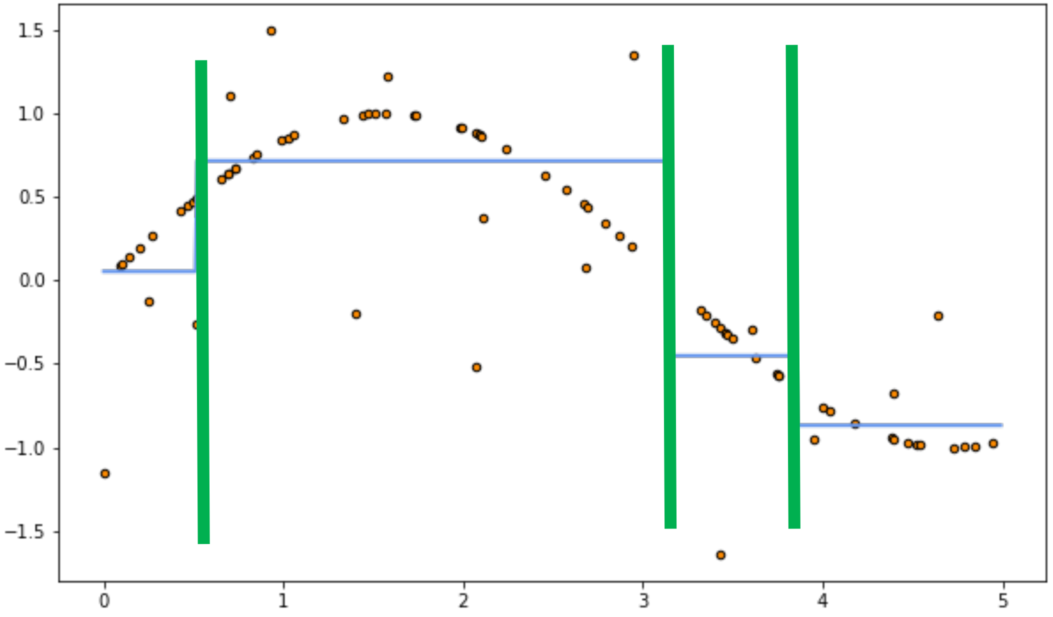


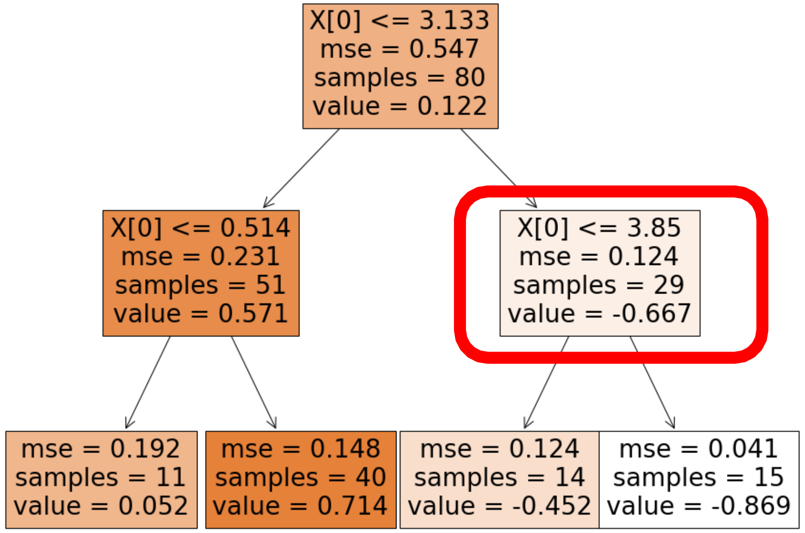


Сложность дерева

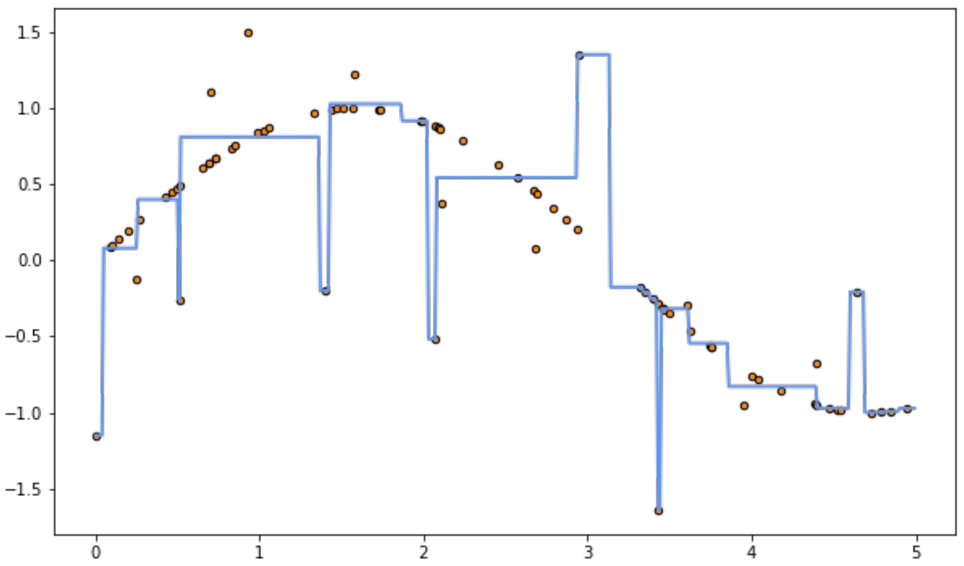
- Решающее дерево можно строить до тех пор, пока каждый лист не будет соответствовать ровно одному объекту
- Деревом можно идеально разделить любую выборку! Если только нет объектов с одинаковыми признаками, но разными ответами

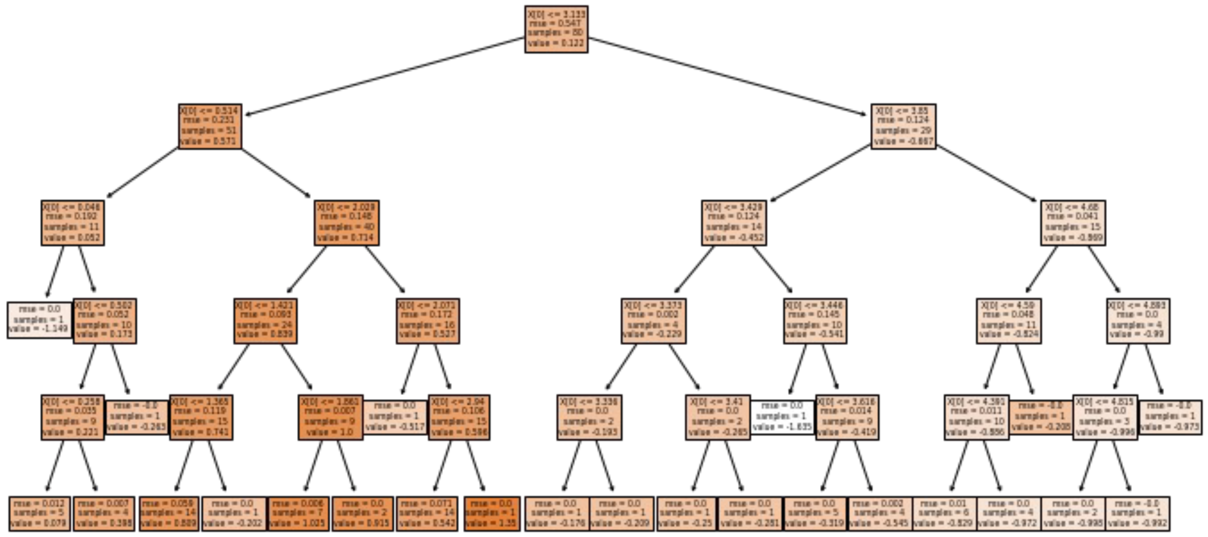
Решающее дерево для регрессии





Решающее дерево для регрессии





Предикаты

- Порог на признак $[x_j < t]$ — не единственный вариант
- Предикат с линейной моделью: $[\langle w, x\rangle < t]$
- Предикат с метрикой: $[\rho(x, x_0) < t]$
- И много других вариантов
- Но даже с простейшим предикатом можно строить очень сложные модели

Прогнозы в листьях

- Наш выбор: константные прогнозы $c_v \in \mathbb{Y}$
- Регрессия:

$$c_v = \frac{1}{|R_v|} \sum_{(x_i, y_i) \in R_v} y_i$$

- Классификация:

$$c_v = \arg\max_{k \in \mathbb{Y}} \sum_{(x_i, y_i) \in R_v} [y_i = k]$$

Прогнозы в листьях

- Наш выбор: константные прогнозы $c_v \in \mathbb{Y}$
- Классификация и вероятности классов:

$$c_{vk} = \frac{1}{|R_v|} \sum_{(x_i, y_i) \in R_v} [y_i = k]$$

Как выбирать предикаты

Энтропия

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

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

Критерий Джини

$$H(p_1, \dots, p_K) = \sum_{i=1}^{K} p_i\,(1 - p_i)$$

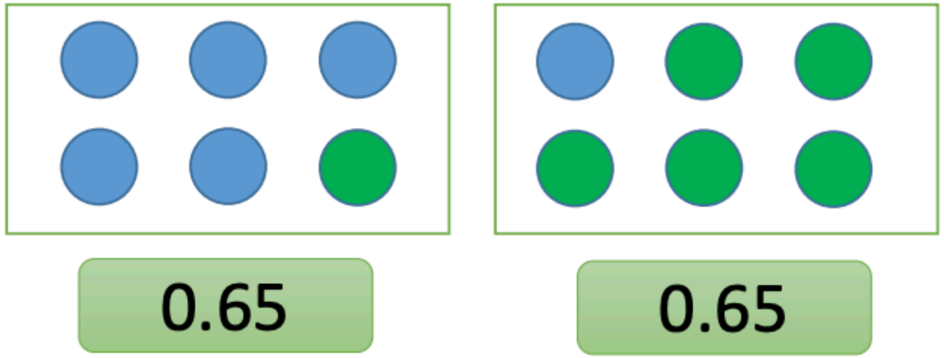
- Вероятность ошибки случайного классификатора, который выдаёт класс k с вероятностью p_k
- Примерно пропорционально количеству пар объектов, относящихся к разным классам

Критерий информативности

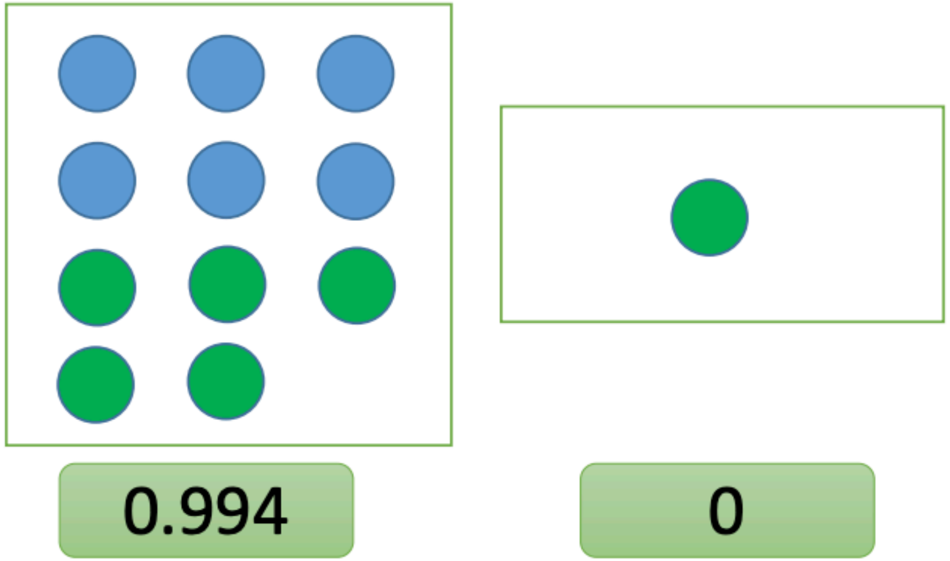
$$Q(R, j, t) = H(R) - \frac{|R_\ell|}{|R|} H(R_\ell) - \frac{|R_r|}{|R|} H(R_r) \to \max_{j,t}$$

- Или так:

$$Q(R, j, t) = \frac{|R_\ell|}{|R|} H(R_\ell) + \frac{|R_r|}{|R|} H(R_r) \to \min_{j,t}$$



- $(5/6, 1/6)$ и $(1/6, 5/6)$
- $0.65 + 0.65 = 1.3$



- $(6/11, 5/11)$ и $(0, 1)$
- $0.994 + 0 = 0.994$

Задача регрессии

$$H(R) = \frac{1}{|R|} \sum_{(x_i, y_i) \in R} (y_i - y_R)^2$$

$$y_R = \frac{1}{|R|} \sum_{(x_i, y_i) \in R} y_i$$

- То есть «хаотичность» вершины можно измерять дисперсией ответов в ней

1. Классификация и регрессия с помощью деревьев" (Classification and Regression Trees) Л. Бреймана, Дж. Фридмана, Р. Олшена и Ч. Стоуна (1984)

2. "Решающие деревья и случайные леса" (Decision Trees and Random Forests) Т. Хасти, Р. Тибширани и Дж. Фридмана (2009)

3. "Решающие деревья и случайные леса" (Decision Trees and Random Forests) Т. Хасти, Р. Тибширани и Дж. Фридмана (2009)