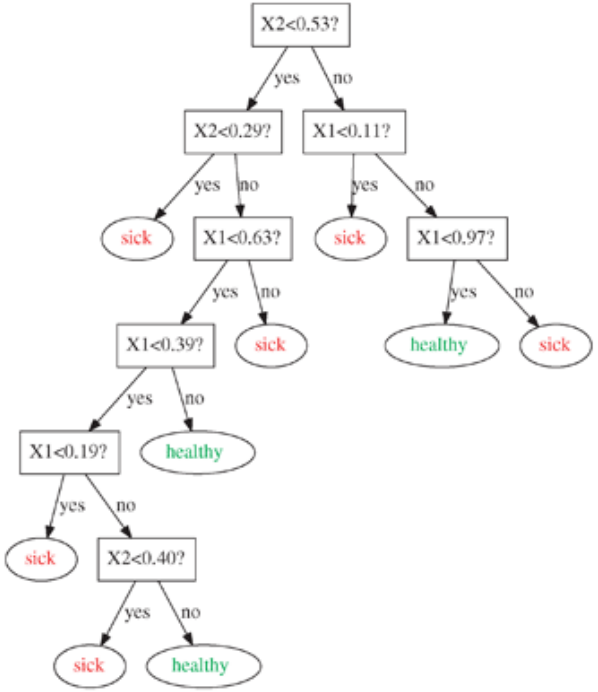
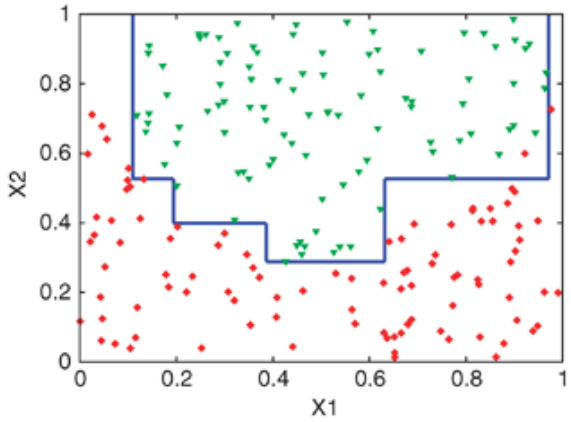




Анализ данных на практике

Решающие деревья

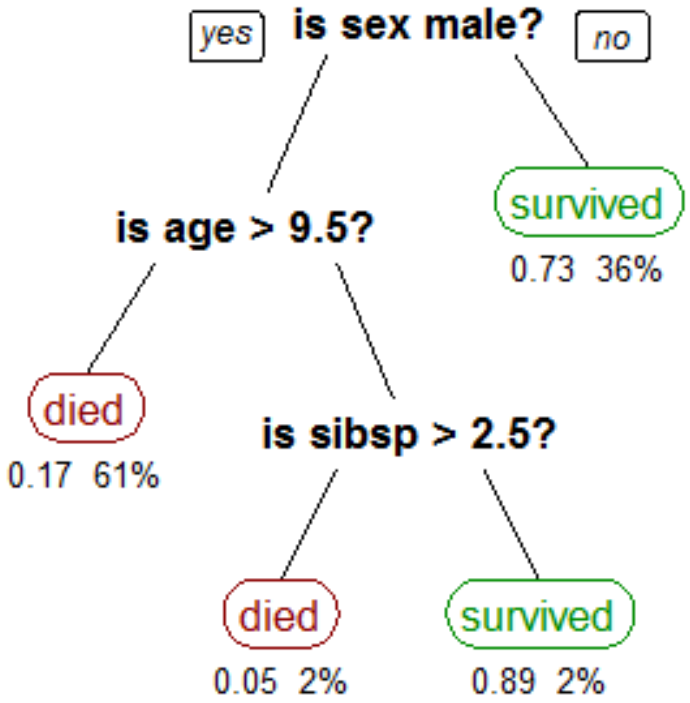
Виктор Кантор

Решающие деревья

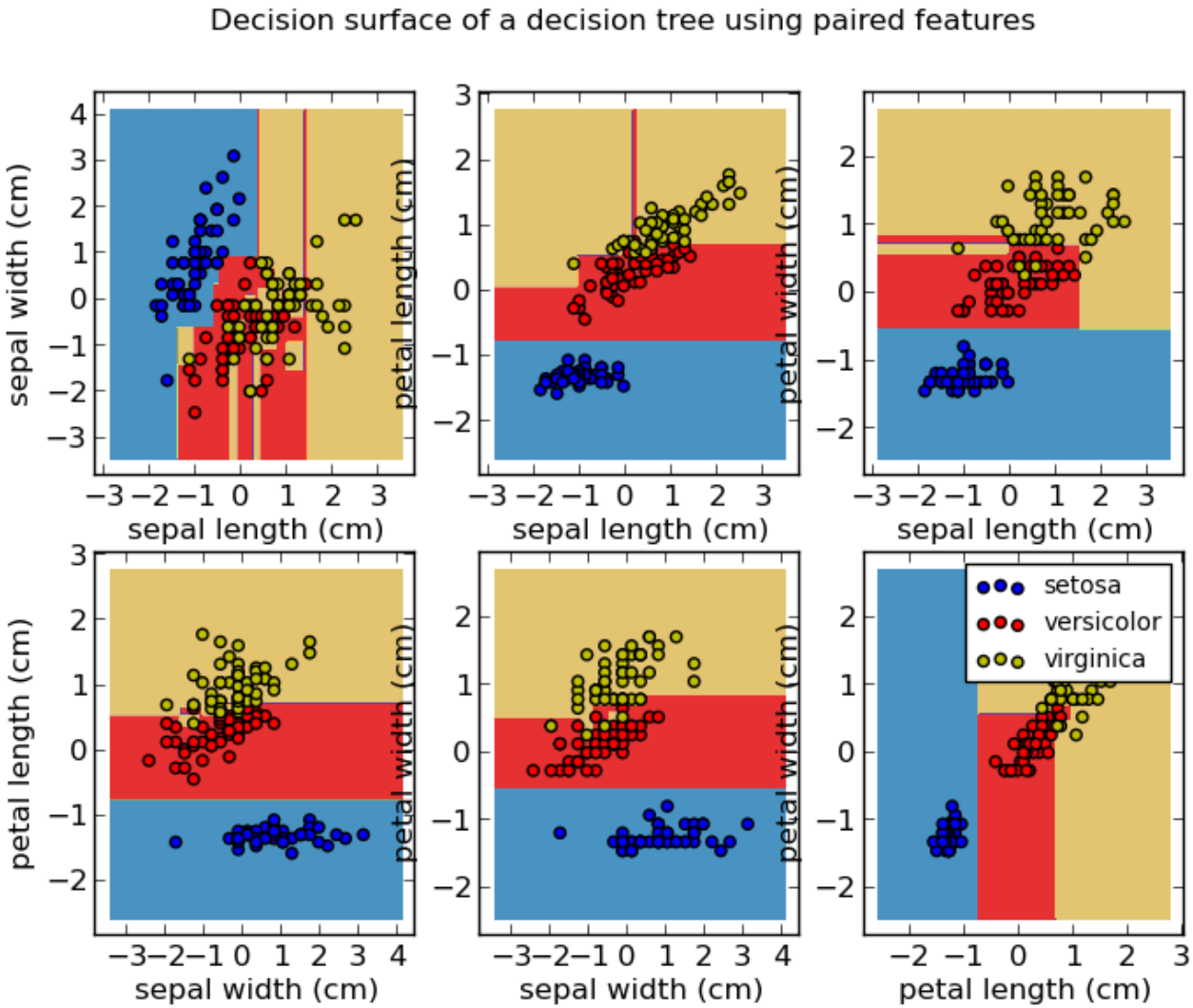
- Как работают
- Критерии информативности закономерностей
- ID3
- Идея регрессионных деревьев
- Random Forest

Как работают решающие деревья

Решающее дерево для Titanic dataset:



Как работают решающие деревья



Точный тест Фишера и Information Gain

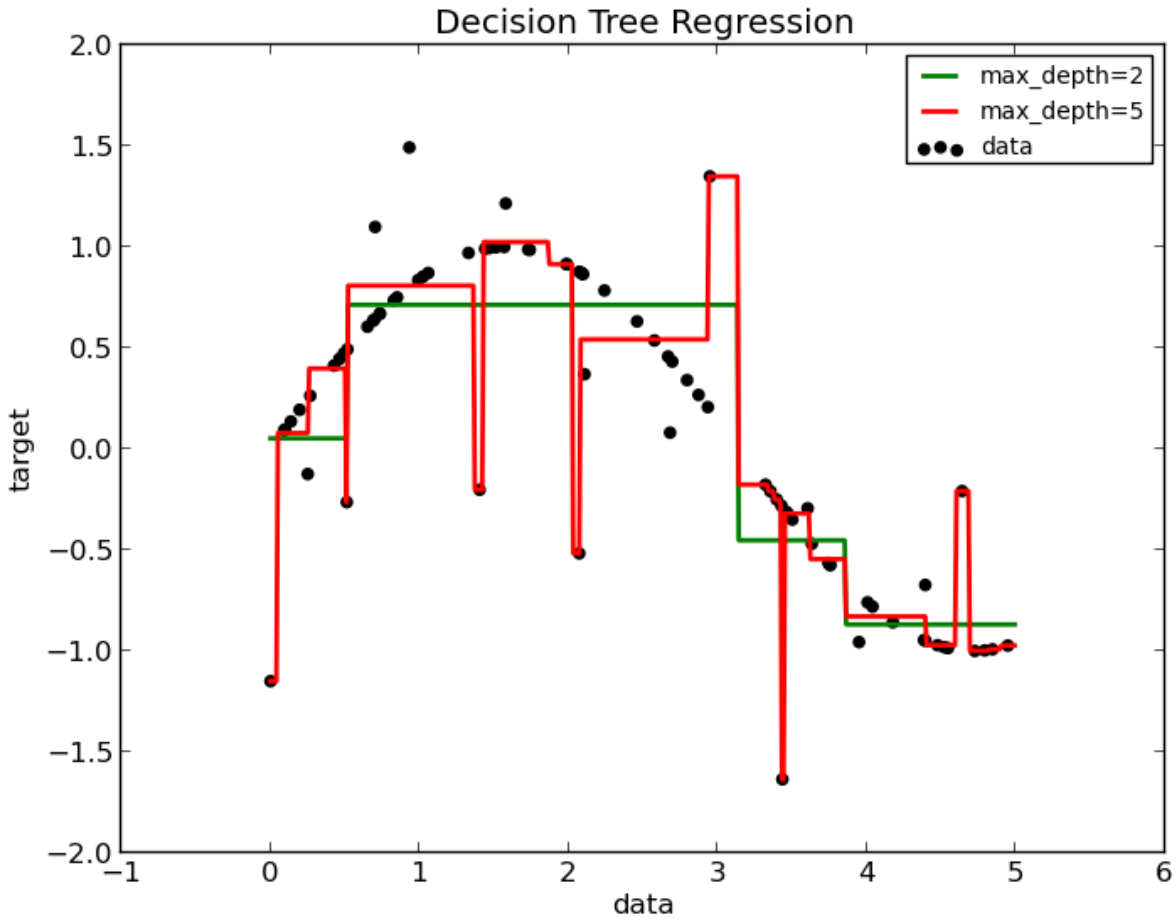
$$\mathrm{IStat}(p,n) = -\log_2 C_P^p C_N^n / C_{P+N}^{p+n}$$

$$\mathrm{IGain}(p,n) = h\left(\frac{P}{\ell}\right) - \frac{p+n}{\ell} h\left(\frac{p}{p+n}\right) - \frac{\ell-p-n}{\ell} h\left(\frac{P-p}{\ell-p-n}\right)$$

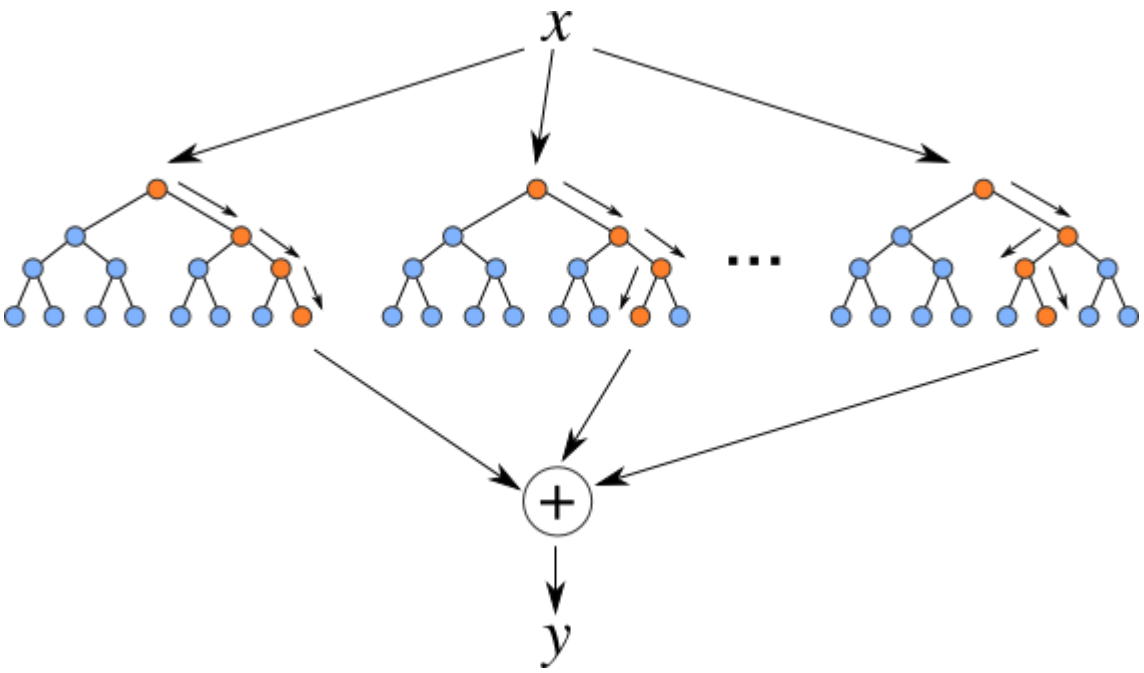
ID3

- Выбираем наиболее информативное разбиение
- Если в одном из поддеревьев меньше αl примеров или все примеры из одного класса: сделать вершину терминальной
- Иначе: повторить процедуру для поддеревьев

Идея регрессионных деревьев



Random Forest



1. Бэггинг над деревьями
2. Рандомизированные разбиения в деревьях: выбираем k случайных признаков и ищем наиболее информативное разбиение по ним

Дополнительные темы

- C4.5
- CART

Решающие деревья

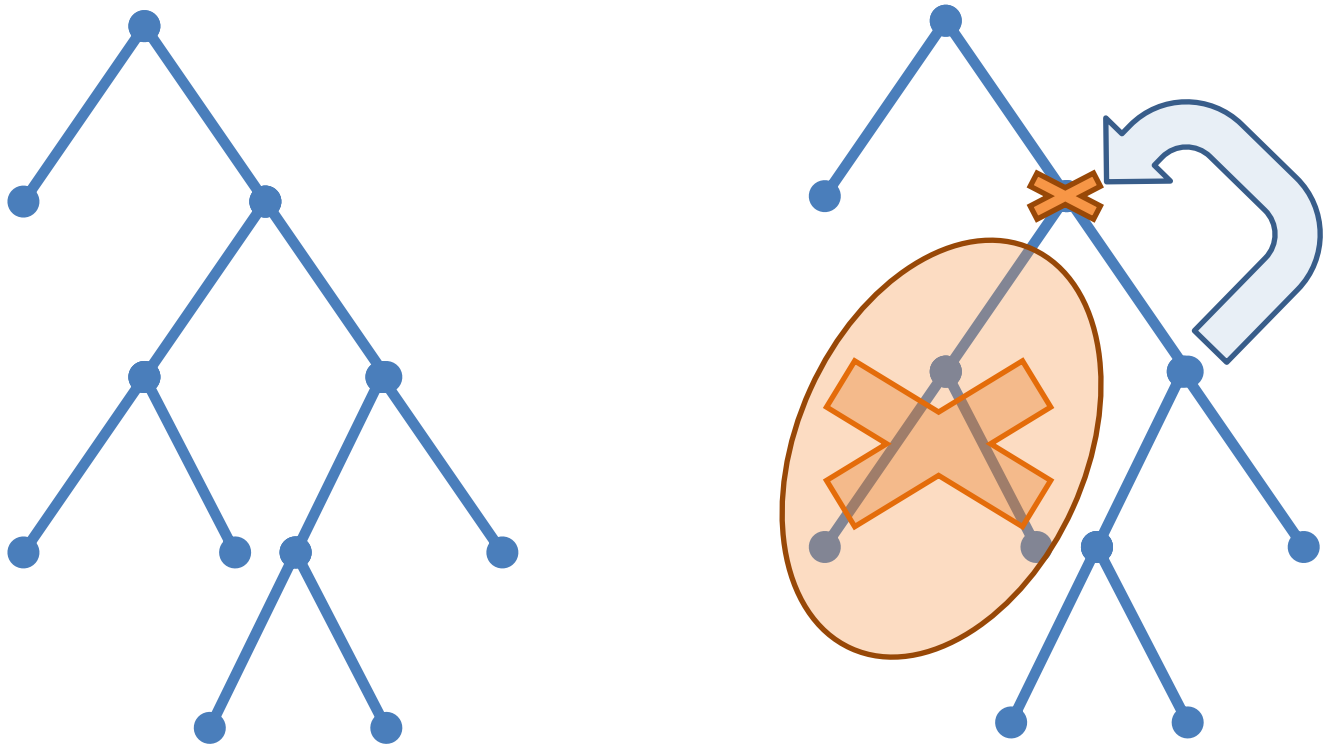
- Как работают
- Критерии информативности закономерностей
- ID3
- Идея регрессионных деревьев
- Random Forest

Дополнительные слайды

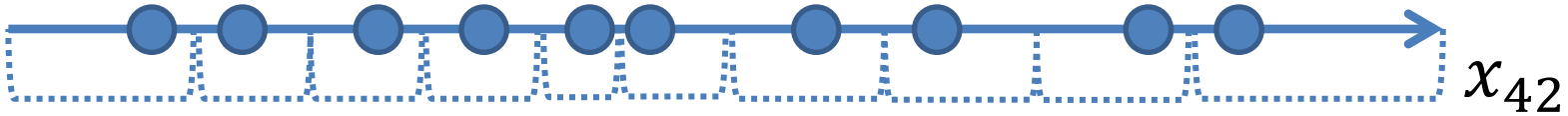
Prunning

- Pre-prunning:
  - Ограничиваем рост дерева до того как оно построено
  - Если в какой-то момент информативность признаков в разбиении меньше порога – не разбиваем вершину
- Post-prunning:
  - Упрощаем дерево после того как дерево построено

Post-prunning



Бинаризация



$$\varphi(x) = [a_k < x \leq a_{k+1}]$$