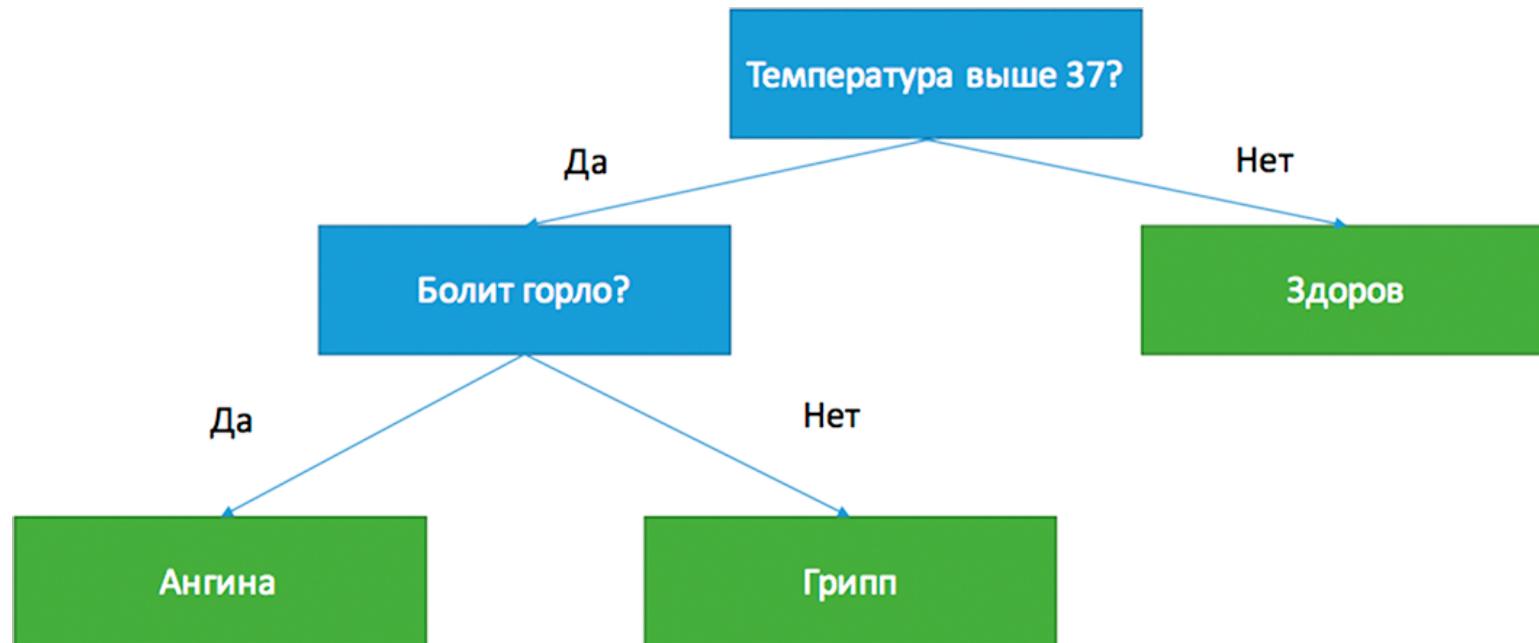


РЕШАЮЩИЕ ДЕРЕВЬЯ

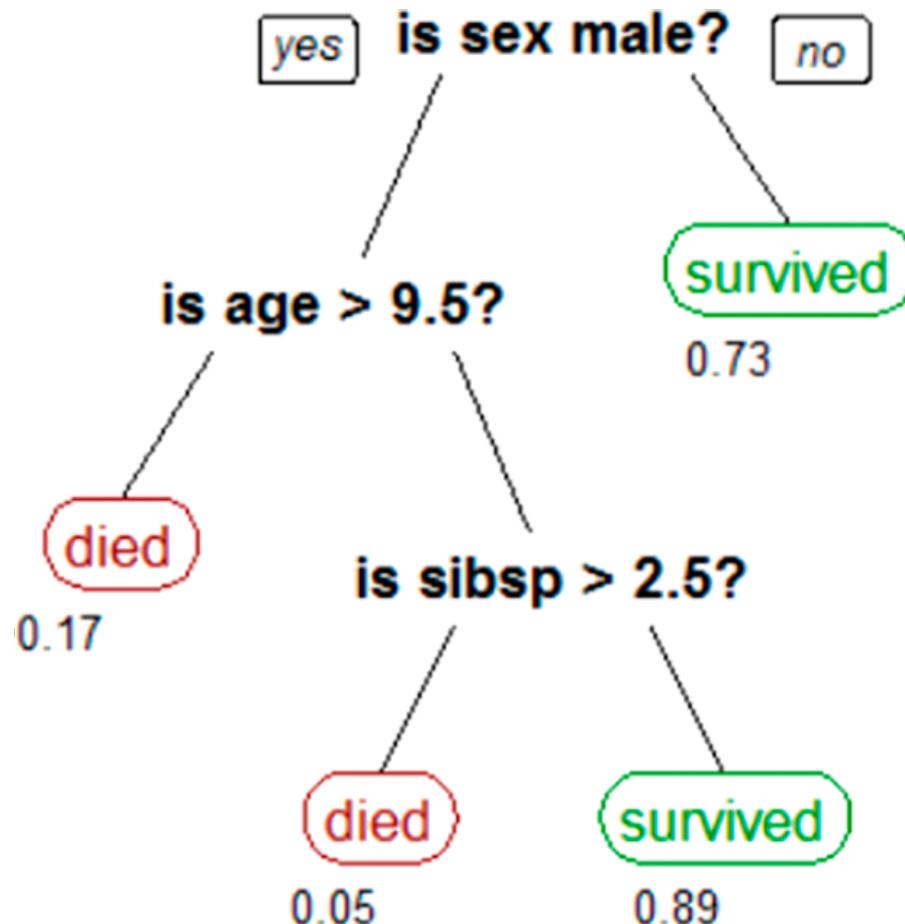
ЛИНЕЙНЫЕ МОДЕЛИ

- › Легко обучаются
- › Восстанавливают только простые зависимости
- › Усложнение — через спрямляющие пространства

МЕДИЦИНСКАЯ ДИАГНОСТИКА



ПАССАЖИРЫ ТИТАНИКА



РЕШАЮЩИЕ ДЕРЕВЬЯ

- › Бинарное дерево (не обязательно)
- › В каждой внутренней вершине записано условие
- › В каждом листе записан прогноз

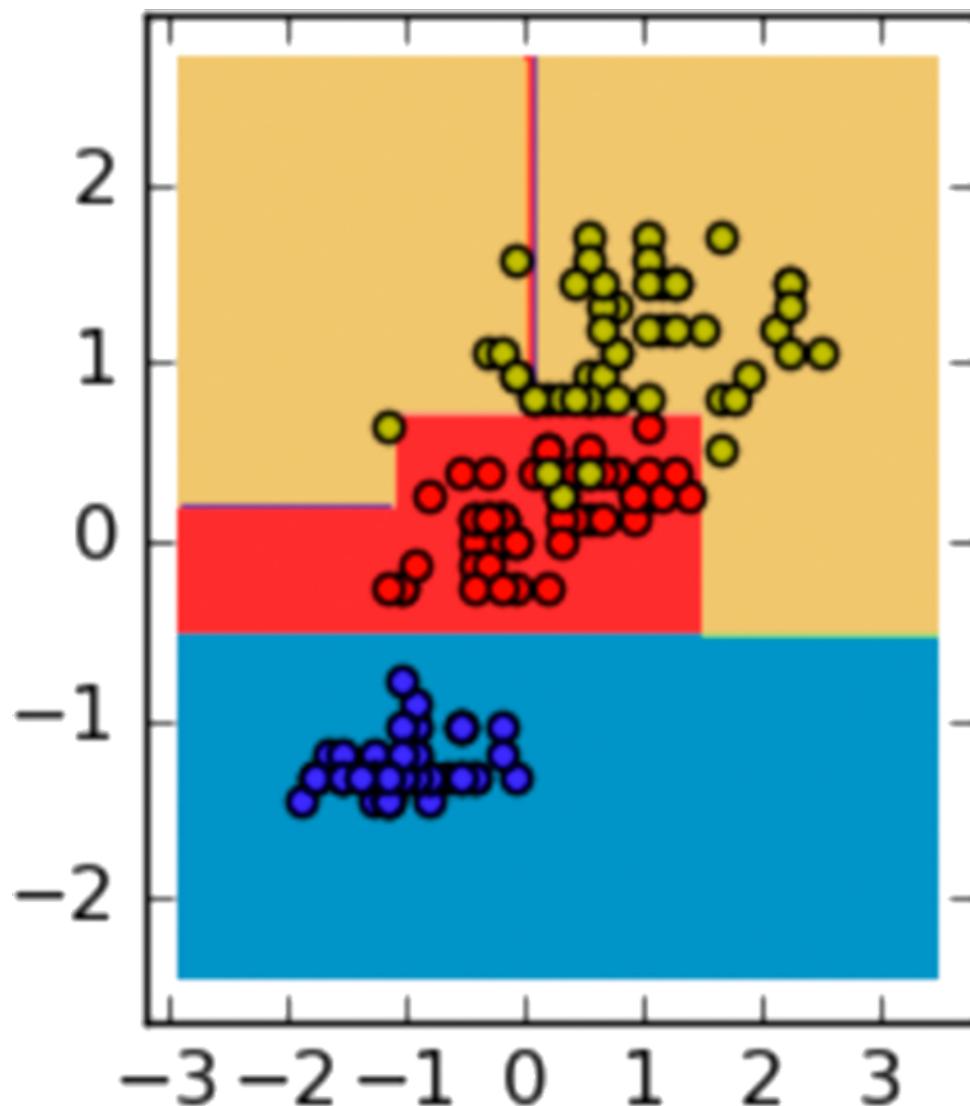
УСЛОВИЯ

- Самый популярный вариант: $[x^j \leq t]$

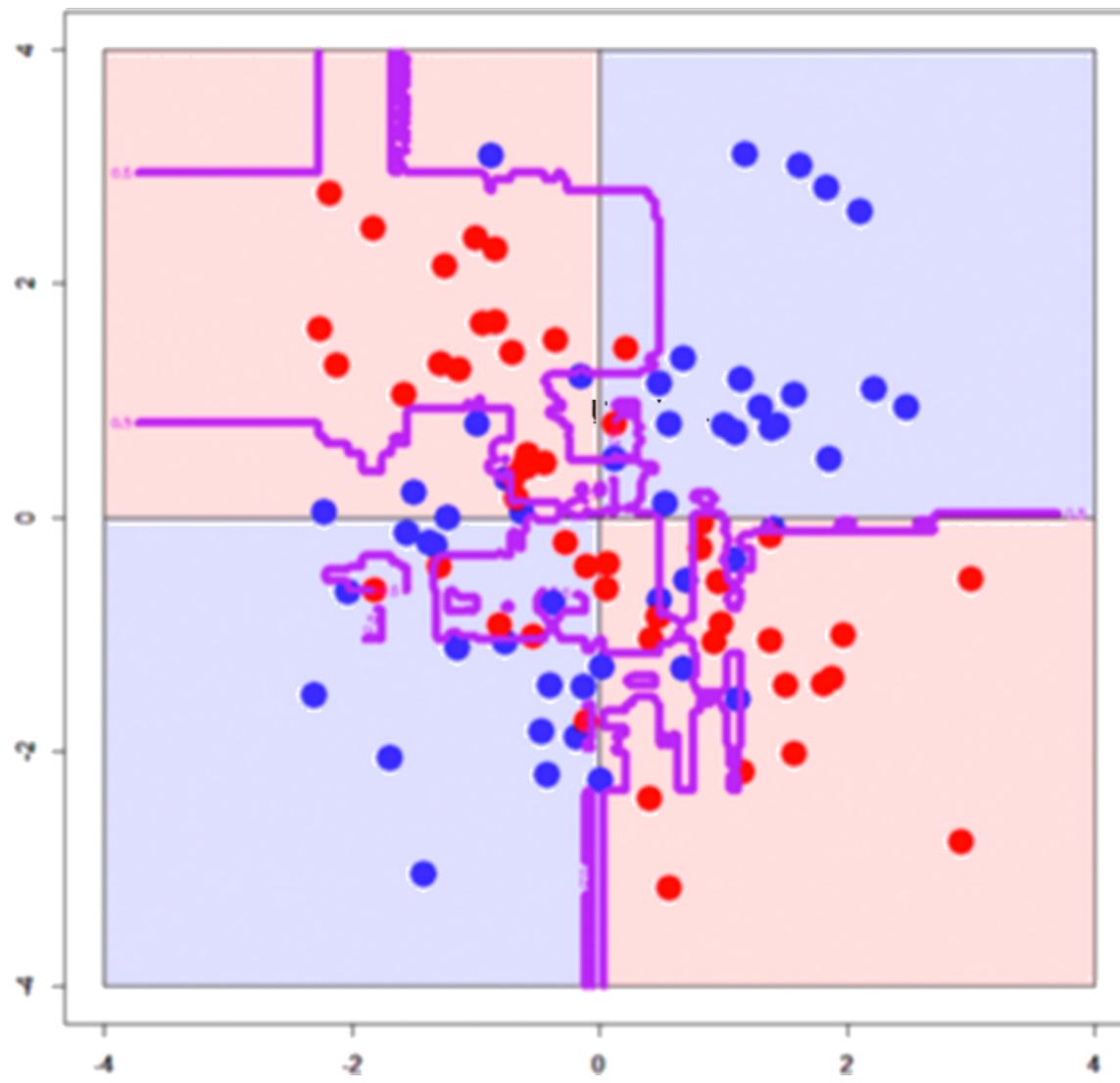
ПРОГНОЗ В ЛИСТЕ

- › Регрессия:
 - ▶ Вещественное число
- › Классификация:
 - ▶ Класс
 - ▶ Вероятности классов

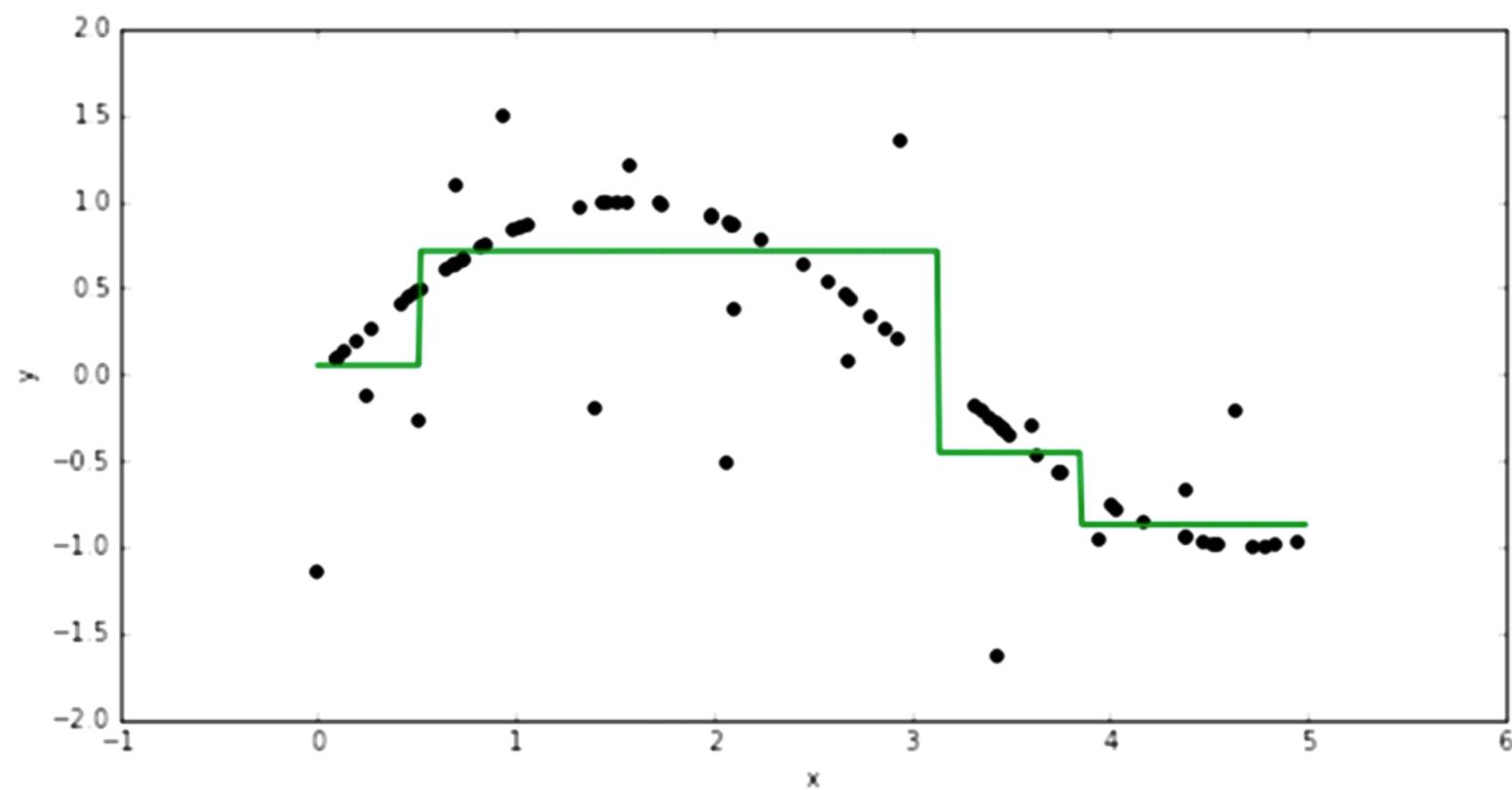
КЛАССИФИКАЦИЯ



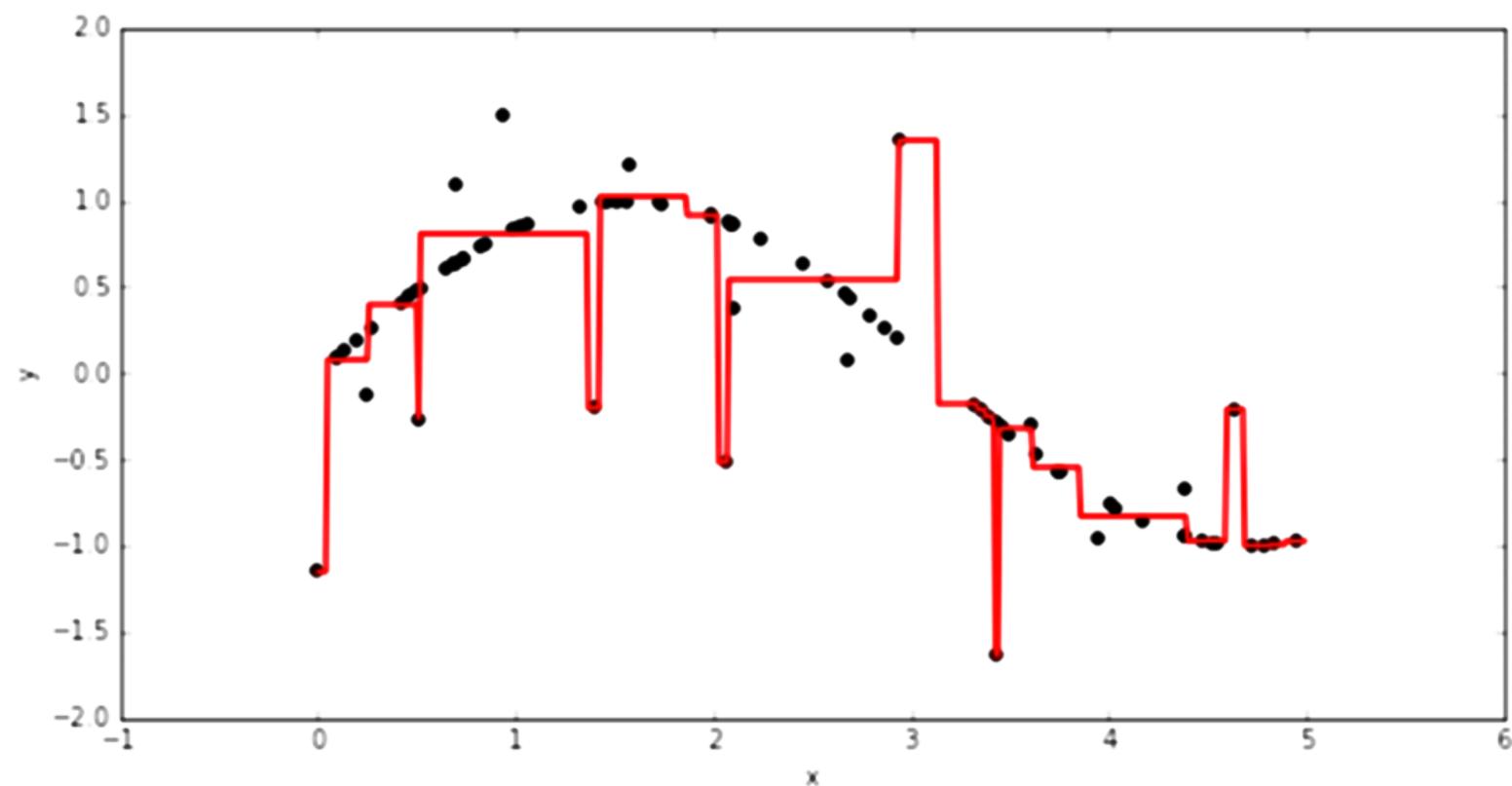
КЛАССИФИКАЦИЯ



РЕГРЕССИЯ



РЕГРЕССИЯ

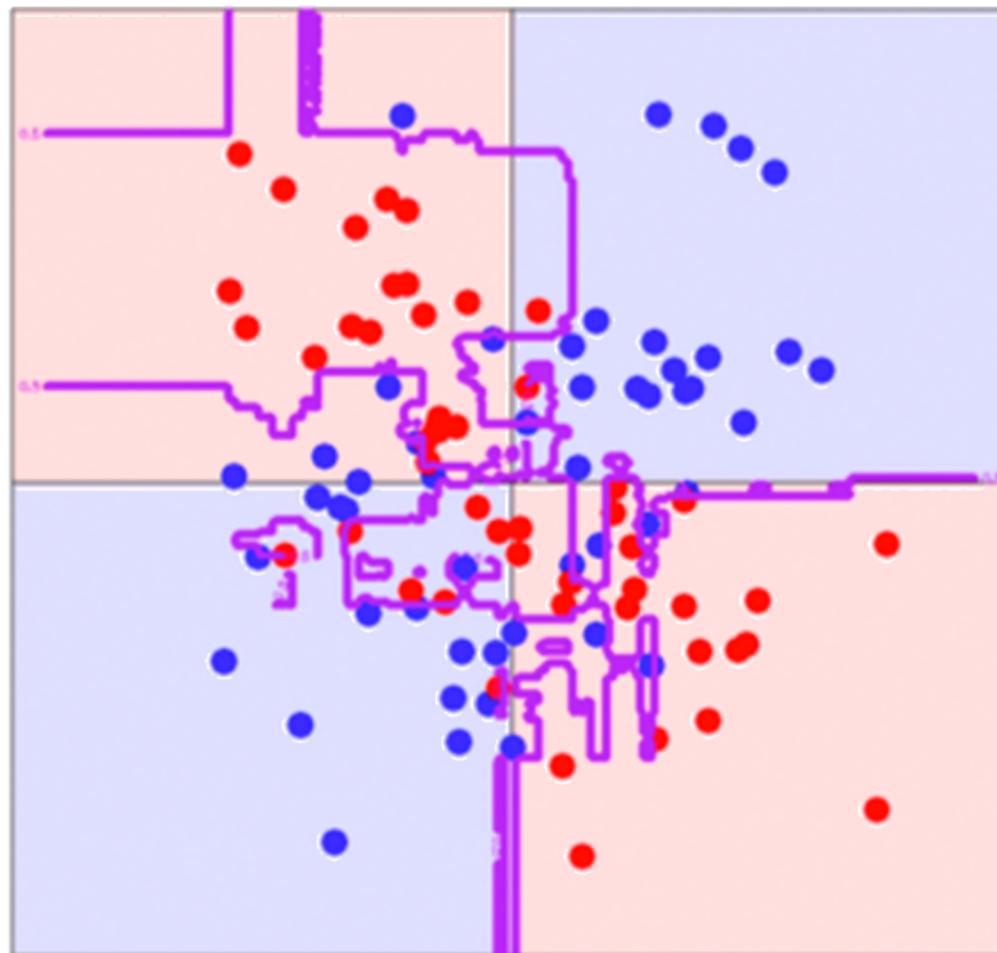


ПРОГНОЗ В ЛИСТЕ

- › Решающие деревья последовательно проверяют простые условия
- › Интерпретируемые
- › Позволяют восстанавливать нелинейные зависимости
- › Легко переобучаются

ОБУЧЕНИЕ РЕШАЮЩИХ ДЕРЕВЬЕВ

ПЕРЕОБУЧЕНИЕ ДЕРЕВЬЕВ

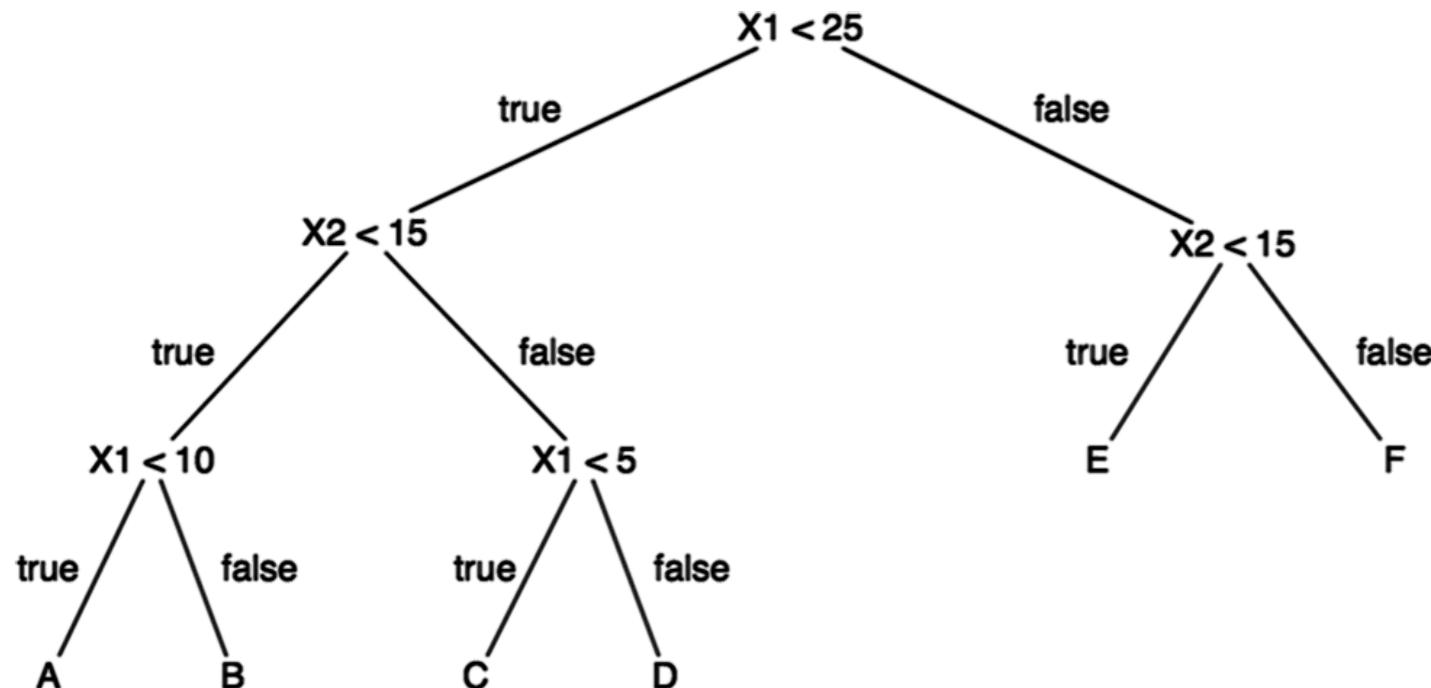


ПЕРЕОБУЧЕНИЕ ДЕРЕВЬЕВ

- › Дерево может достичь нулевой ошибки на любой выборке
- › Борьба с переобучением: минимальное дерево среди всех с нулевой ошибкой
- › NP-полная задача

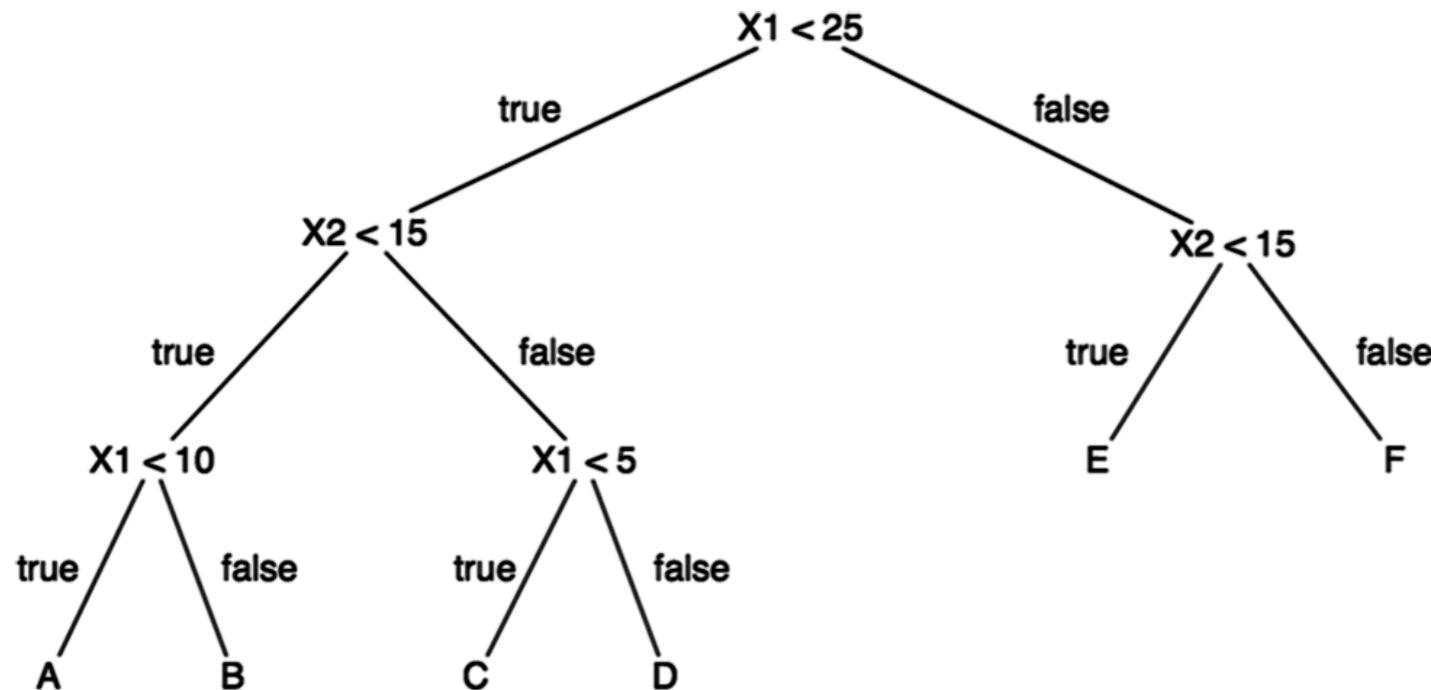
ЖАДНОЕ ПОСТРОЕНИЕ

› Растим дерево от корня к листьям



ЖАДНОЕ ПОСТРОЕНИЕ

› Как разбить вершину на две?



ПОИСК РАЗБИЕНИЯ

- › Пусть в вершине m оказалась выборка X_m
- › $Q(X_m, j, t)$ — критерий ошибки условия $[x^j \leq t]$
- › Ищем лучшие параметры j и t перебором:

$$Q(X_m, j, t) \rightarrow \min_{j,t}$$

ПОИСК РАЗБИЕНИЯ

- › После того, как разбиение найдено
- › Разбиваем X_m на две части:
$$X_\ell = \{x \in X_m | [x^j \leq t]\}$$
$$X_r = \{x \in X_m | [x^j > t]\}$$
- › Повторяем процедуру для дочерних вершин

КРИТЕРИЙ ОСТАНОВА

- › В какой момент прекращать разбиение вершин?
- › В вершине один объекты?
- › В вершине объекты одного класса?
- › Глубина превысила порог?

ОТВЕТ В ЛИСТЕ

- › Допустим, решили сделать вершину листом
- › Какой прогноз выбрать?
- › Регрессия: $a_m = \frac{1}{|X_m|} \sum_{i \in X_m} y_i$
- › Классификация:

$$a_m = \operatorname{argmax}_{y \in \mathbb{Y}} \sum_{i \in X_m} [y_i = y]$$

ОТВЕТ В ЛИСТЕ

- › Допустим, решили сделать вершину листом
- › Какой прогноз выбрать?
- › Вероятности классов:

$$a_{mk} = \frac{1}{|X_m|} \sum_{i \in X_m} [y_i = k]$$

ВОПРОСЫ

- › Критерий ошибки разбиения?
- › Критерий останова?

РЕЗЮМЕ

- › Решающие деревья строятся жадно — от корня к листьям
- › Разбиение выбирается, исходя из критерия ошибки
- › Критерий останова
- › Прогнозы в листе

КРИТЕРИИ ИНФОРМАТИВНОСТИ

ПОИСК РАЗБИЕНИЯ

- › Пусть в вершине m оказалась выборка X_m
- › $Q(X_m, j, t)$ — критерий ошибки условия $[x^j \leq t]$
- › Ищем лучшие параметры j и t перебором:

$$Q(X_m, j, t) \rightarrow \min_{j,t}$$

ПОИСК РАЗБИЕНИЯ

» Разбиваем X_m на две части:

$$X_\ell = \{x \in X_m | [x^j \leq t]\}$$

$$X_r = \{x \in X_m | [x^j > t]\}$$

КРИТЕРИЙ ОШИБКИ

$$Q(X_m, j, t) = \frac{|X_\ell|}{|X_m|} H(X_\ell) + \frac{|X_r|}{|X_m|} H(X_r)$$

КРИТЕРИЙ ОШИБКИ

$$Q(X_m, j, t) = \frac{|X_\ell|}{|X_m|} H(X_\ell) + \frac{|X_r|}{|X_m|} H(X_r)$$



Разброс ответов в левом листе

КРИТЕРИЙ ОШИБКИ

$$Q(X_m, j, t) = \frac{|X_\ell|}{|X_m|} H(X_\ell) + \frac{|X_r|}{|X_m|} H(X_r)$$



Разброс ответов в правом листе

КРИТЕРИЙ ОШИБКИ

$$Q(X_m, j, t) = \frac{|X_\ell|}{|X_m|} H(X_\ell) + \frac{|X_r|}{|X_m|} H(X_r)$$


Доля объектов в листьях

КРИТЕРИЙ ИНФОРМАТИВНОСТИ

- › $H(X)$
- › Зависит от ответов на выборке X
- › Чем меньше разброс ответов, тем меньше значение $H(X)$

РЕГРЕССИЯ

$$\bar{y}(X) = \frac{1}{|X|} \sum_{i \in X} y_i$$

$$H(X) = \frac{1}{|X|} \sum_{i \in X} (y_i - \bar{y}(X))^2$$

КЛАССИФИКАЦИЯ

› Доля объектов класса k в выборке X :

$$p_k = \frac{1}{|X|} \sum_{i \in X} [y_i = k]$$

КРИТЕРИЙ ДЖИНИ

$$H(X) = \sum_{k=1}^K p_k(1 - p_k)$$

- › Если $p_1 = 1, p_2 = \dots = p_k = 0$, то $H(X) = 0$
- › Вероятность ошибки классификатора, который выдаёт ответы пропорционально p_k

ЭНТРОПИЙНЫЙ КРИТЕРИЙ

$$H(X) = - \sum_{k=1}^K p_k \ln p_k$$

- › Считаем, что $0 \ln 0 = 0$
- › Если $p_1 = 1, p_2 = \dots = p_k = 0$, то $H(X) = 0$
- › Мера отличия распределения классов от вырожденного

РЕЗЮМЕ

- › Критерий ошибки измеряет разброс ответов после разбиения
- › Выражается через критерий информативности
- › В регрессии: **MSE**
- › В классификации: критерий Джини и энтропийный критерий

КРИТЕРИИ ОСТАНОВА И СТРИЖКА ДЕРЕВЬЕВ

КРИТЕРИЙ ОСТАНОВА

- › Как понять, разбивать вершину или делать листовой?
- › Способ борьбы с переобучением

КРИТЕРИЙ ОСТАНОВА

- › Все объекты в вершине относятся к одному классу

КРИТЕРИЙ ОСТАНОВА

- › В вершину попало $\leq n$ объектов
- › При $n = 1$ получаем максимально переобученные деревья
- › n должно быть достаточно, чтобы построить надёжный прогноз
- › Рекомендация: $n = 5$

КРИТЕРИЙ ОСТАНОВА

- › Ограничение на глубину
- › Грубый критерий
- › Неплохо работает в композициях

СТРИЖКА ДЕРЕВЬЕВ

- › Строим максимально переобученное дерево
- › Удаляем листья по некоторому критерию
- › Пример: удаляем, пока улучшается ошибка на валидации
- › Считается, что работает лучше критериев останова

СТРИЖКА ДЕРЕВЬЕВ

- › Трудоёмкая процедура
- › Имеет смысл только при использовании одного дерева
- › В композициях деревьев достаточно простых критериев останова

РЕЗЮМЕ

- › Критерии останова: глубина, число объектов в листе
- › Стрижка деревьев

РЕШАЮЩИЕ ДЕРЕВЬЯ И КАТЕГОРИАЛЬНЫЕ ПРИЗНАКИ

УСЛОВИЕ РАЗБИЕНИЯ

$$[x^j \leq t]$$

- › Только для вещественных и бинарных признаков!

N-АРНЫЕ ДЕРЕВЬЯ

- › Нужно сделать разбиение вершины m
- › Для бинарных или вещественных признаков:

$$[x^j \leq t]$$

N-АРНЫЕ ДЕРЕВЬЯ

- › Нужно сделать разбиение вершины m
- › Для категориального признака x^j с n значениями $\{c_1, \dots, c_n\}$
- › Пробуем разбить на n вершин
- › В i -ю дочернюю вершину идут объекты с $x^j = c_i$

N-АРНЫЕ ДЕРЕВЬЯ

- » Разбили X_m на n частей: X_1, \dots, X_n
- » Критерий ошибки:

$$Q(X_m, j) = \sum_{i=1}^n \frac{|X_i|}{|X_m|} H(X_i) \rightarrow \min_j$$

N-АРНЫЕ ДЕРЕВЬЯ

- › Выбираем из всех разбиений то, на котором меньше всего $Q(X_m, j)$ или $Q(X_m, j, t)$
- › Будем очень часто выбирать категориальные признаки с большим n
- › Высокий риск переобучения
- › Подходит для очень больших выборок

БИНАРНЫЕ ДЕРЕВЬЯ

- › Нужно сделать разбиение вершины m
- › Для категориального признака $f(x^j)$ с n значениями $C = \{c_1, \dots, c_n\}$
- › Разобьём множество значений на две части:
$$C = C_1 \cup C_2$$
- › Разбиение: $[x^j \in C_1]$

БИНАРНЫЕ ДЕРЕВЬЯ

- › Как разбить C ?
- › Отсортируем значения
- › Заменим $c_{(1)}, \dots, c_{(n)}$ на $1, \dots, n$
- › Будем работать как с вещественным признаком

СОРТИРОВКА ЗНАЧЕНИЙ

› Для бинарной классификации:

$$\frac{\sum_{i \in X_m} [x_i^j = c_{(1)}] [y_i = +1]}{\sum_{i \in X_m} [x_i^j = c_{(1)}]} \leq \dots \leq \frac{\sum_{i \in X_m} [x_i^j = c_{(n)}] [y_i = +1]}{\sum_{i \in X_m} [x_i^j = c_{(n)}]}$$

СОРТИРОВКА ЗНАЧЕНИЙ

› Для регрессии:

$$\frac{\sum_{i \in X_m} [x_i^j = c_{(1)}] y_i}{\sum_{i \in X_m} [x_i^j = c_{(1)}]} \leq \dots \leq \frac{\sum_{i \in X_m} [x_i^j = c_{(n)}] y_i}{\sum_{i \in X_m} [x_i^j = c_{(n)}]}$$

БИНАРНЫЕ ДЕРЕВЬЯ

- › Аналогично перебору всех разбиений C
- › Выполнено для MSE, критерий Джини, энтропийного критерия

РЕЗЮМЕ

- › n -арные деревья
- › Бинарные деревья с разбиением множества значений