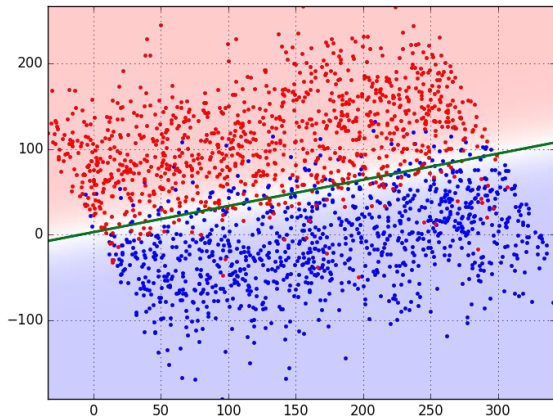
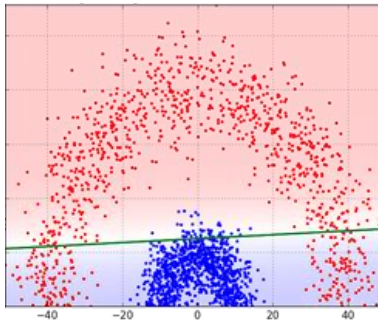


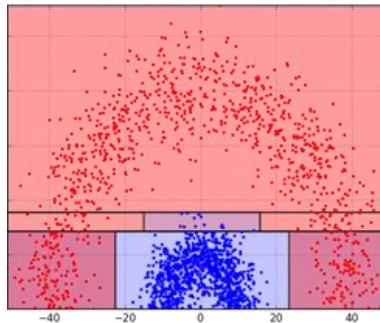
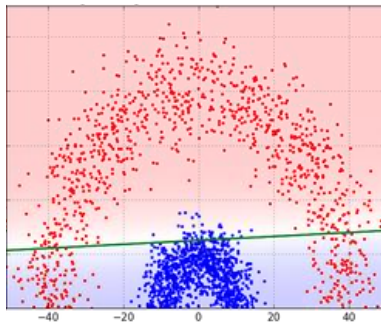
# Регрессия. Всегда ли она нас спасает?



# А вот в таком случае?

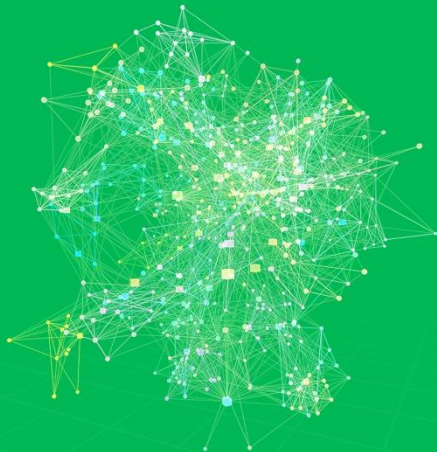


# А вот в таком случае?



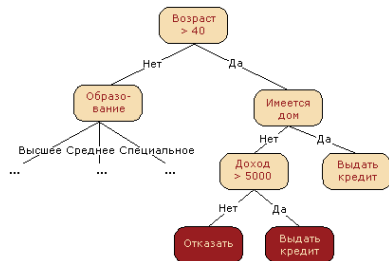


# Деревья решений



# Деревья решений относятся к логическим методам классификации, т.е. ищут в данных логические закономерности

- Температура > 38? **Да** -> Есть кашель? **Да** -> Кашель влажный? **Да** -> **Назначать антибиотики**
- Возраст > 40? **Да** -> Имеется дом? **Нет** -> Доход > 5000? **Да** -> **Выдать кредит**

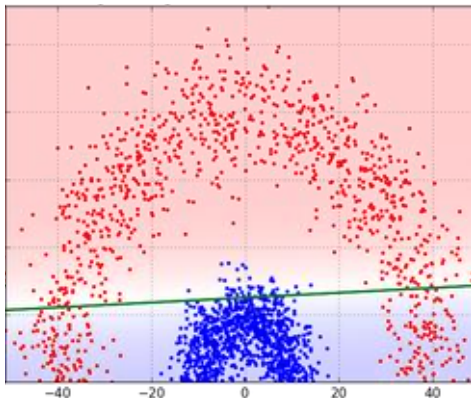


В ходе лекции будем рассматривать задачу бинарной классификации, потом обсудим как можно масштабировать

# Мотивация 1. Пример нелинейного датасета



Logistic Regression, f-measure = 0.854290

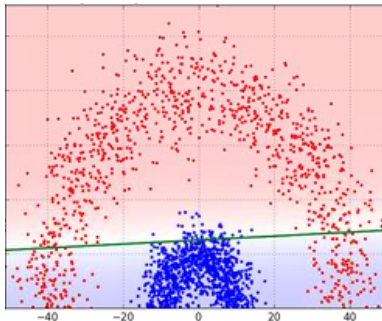


Целевая переменная **нелинейно** зависит от признаков

## Мотивация 2. Лог. рег. и наш пример



Logistic Regression, f-measure = 0.854290

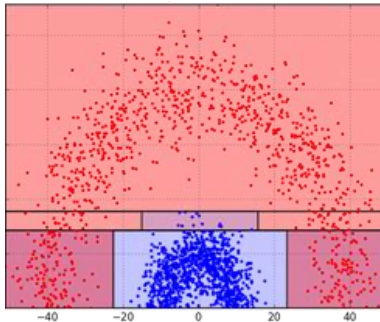


- Лог. рег. хорошо работает при линейной зависимости признаков и целевой переменной
- Экспериментировать с преобразованием признаков и добиться более хорошего качества, но такой подход является эвристиким и вы можете потратить на много времени и при этом не получить желаемый результат

## Мотивация 3. Дерево решений и наш пример



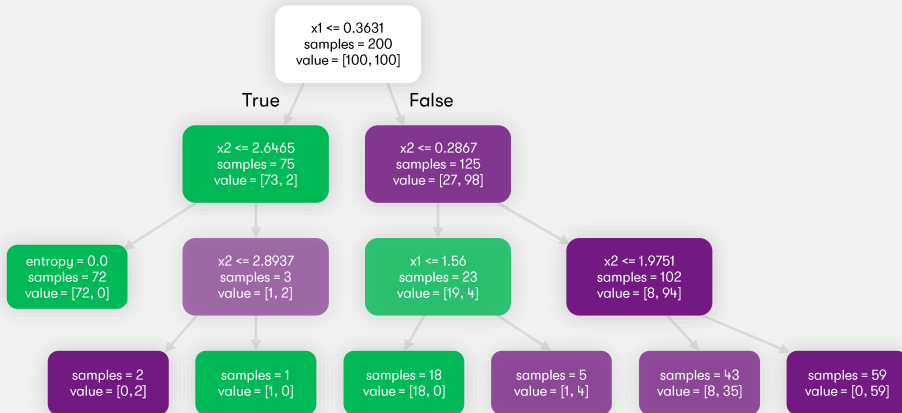
Decision Tree, f-measure = 1.000000



- разделяет пространство на многомерные прямоугольники (подпространства)
- в подпространстве формируется ответ на основе обучающей выборки



# Мотивация 4. Представление дерева



- Последовательность логических правил
- Константа в листьях

# Бинарное дерево решений



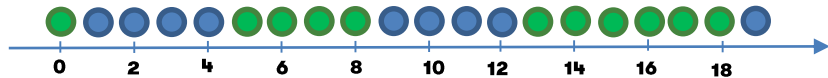
**Вершины** - логические правила

1. Кол-во этажей в доме  $\geq 5$ ?
2. Квартира студия?

**Листья** - предсказания в виде константы



# Как выгоднее всего строить дерево?



# Пример на костях:

Случай 1.

“Я бросил 10 игральных кубиков и получил сумму 30.”



Случай 2.

“Я бросил 10 игральных кубиков и получил сумму 59.”

2 930 455 комбинаций

60 комбинаций

# Как же определить меру беспорядка?

Энтропия Шеннона – мера беспорядка системы:

$$S = - \sum_{i=1}^N p_i \log_2 p_i,$$

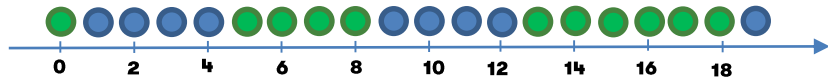
В случае бинарной классификации:

$$S = -p_+ \log_2 p_+ - p_- \log_2 p_- = -p_+ \log_2 p_+ - (1 - p_+) \log_2 (1 - p_+);$$

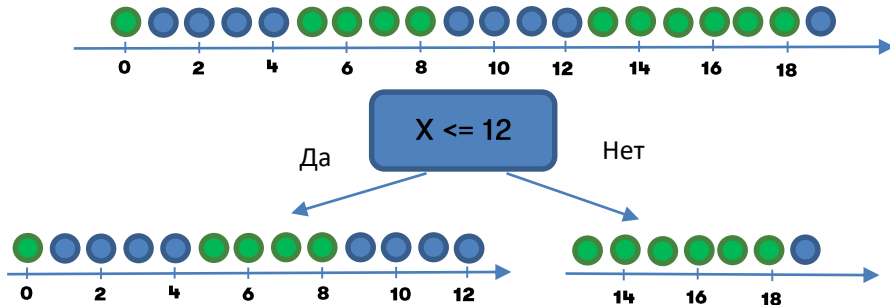
Прирост информации:

$$IG(Q) = S_O - \sum_{i=1}^q \frac{N_i}{N} S_i,$$

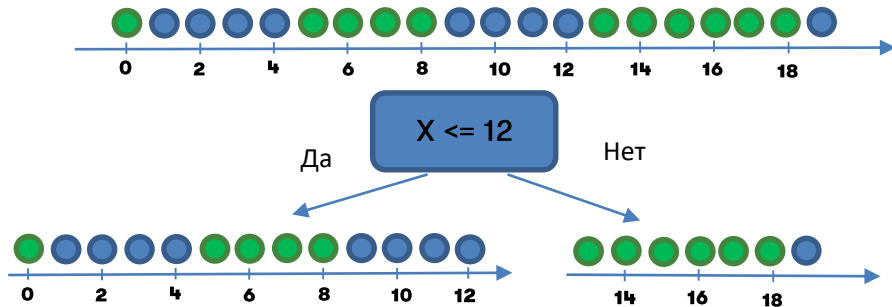
# Как выгоднее всего строить дерево?



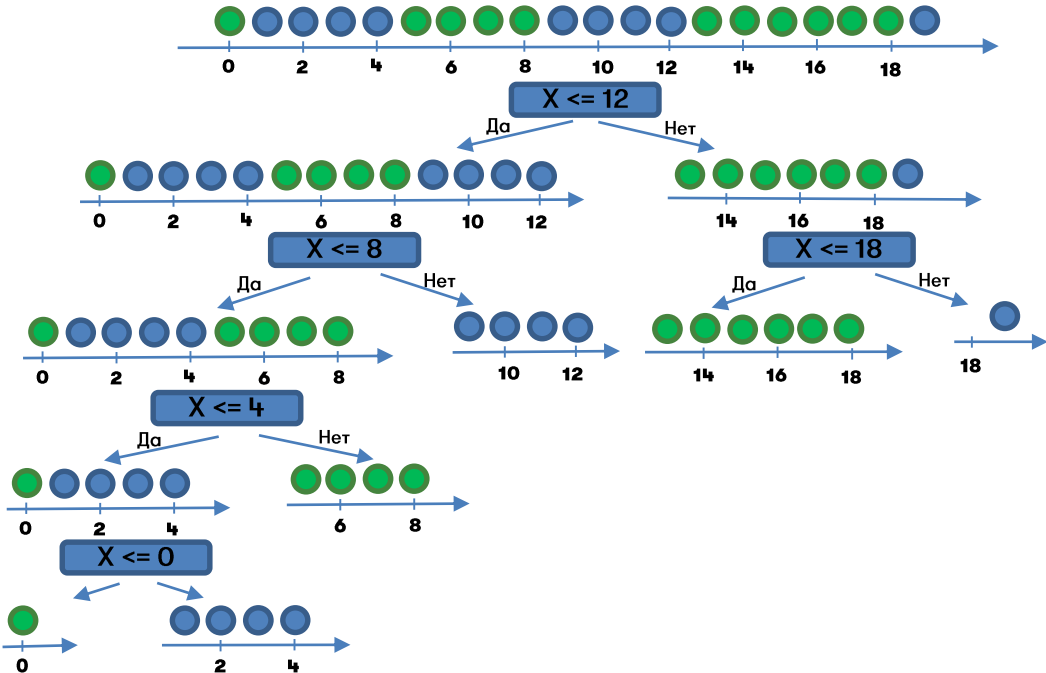
# Как выгоднее всего строить дерево?



# Давайте посчитаем энтропию и IG







# Алгоритм построения дерева

$$S = \{(\mathbf{x}_1, y_1), \dots, (\mathbf{x}_N, y_N)\}$$

$$\mathbf{x} = (x_1, \dots, x_d)$$

$$x_j, y \in \{0, 1\}$$

GROWTREE( $S$ )

**if** ( $y = 0$  for all  $\langle \mathbf{x}, y \rangle \in S$ ) **return** new leaf(0)

**else if** ( $y = 1$  for all  $\langle \mathbf{x}, y \rangle \in S$ ) **return** new leaf(1)

**else**

choose best attribute  $x_j$

$S_0 =$  all  $\langle \mathbf{x}, y \rangle \in S$  with  $x_j = 0$ ;

$S_1 =$  all  $\langle \mathbf{x}, y \rangle \in S$  with  $x_j = 1$ ;

**return** new node( $x_j$ , GROWTREE( $S_0$ ), GROWTREE( $S_1$ ))

# Кроме энтропии:

Джини:

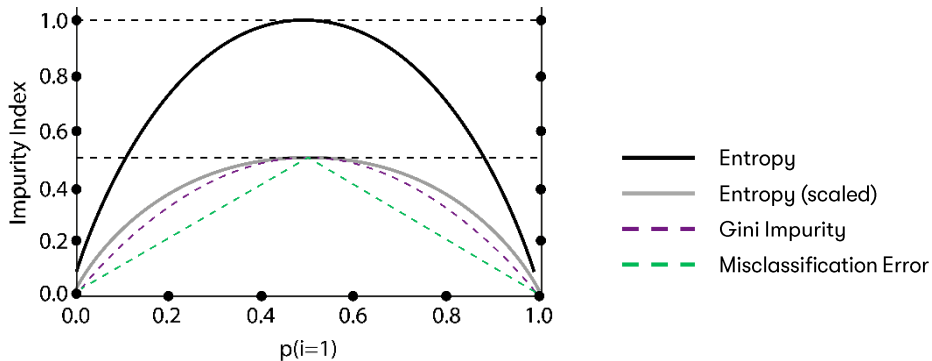
$$G = 1 - \sum_k (p_k)^2$$

$$G = 1 - p_+^2 - p_-^2 = 1 - p_+^2 - (1 - p_+)^2 = 2p_+(1 - p_+).$$

Misclassification error:

$$E = 1 - \max_k p_k$$

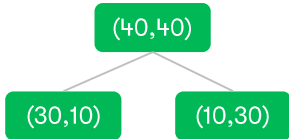
# Сравнение КИ для классификации



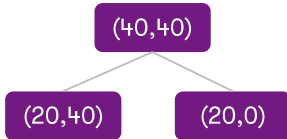
# Пример. Чувствительность КИ для классификации



**A**



**B**



## Пример. Чувствительность MSI



---

$$A : IG_E = 0.5 - \frac{4}{8} \times 0.25 - \frac{4}{8} \times 0.25 = 0.25$$

$$B : I_E (D_{\text{left}}) = 1 - \frac{4}{6} = \frac{1}{3}$$

$$B : I_E (D_{\text{Right}}) = 1 - 1 = 0$$

$$B : IG_E = 0.5 - \frac{6}{8} \times \frac{1}{3} - 0 = 0.25$$

# Пример. Чувствительность Gini



---

$$I_G(D_p) = 1 - (0.5^2 + 0.5^2) = 0.5$$

$$A : I_C(D_{\text{left}}) = 1 - \left( \left( \frac{3}{4} \right)^2 + \left( \frac{1}{4} \right)^2 \right) = \frac{3}{8} = 0.375$$

$$A : I_G(D_{\text{right}}) = 1 - \left( \left( \frac{1}{4} \right)^2 + \left( \frac{3}{4} \right)^2 \right) = \frac{3}{8} = 0.375$$

$$A : IG_G = 0.5 - \frac{4}{8} \times 0.375 - \frac{4}{8} \times 0.375 = 0.125$$

$$B : I_C(D_{\text{left}}) = 1 - \left( \left( \frac{2}{6} \right)^2 + \left( \frac{4}{6} \right)^2 \right) = \frac{4}{9} = 0.4$$

$$B : I_G(D_{\text{right}}) = 1 - (1^2 + 0^2) = 0$$

$$B : IG_G = 0.5 - \frac{6}{9} \times 0.4 - 0 = 0.16$$

## Пример. Чувствительность Entropy



---

$$A : I_H(D_{right}) = - \left( \frac{1}{4} \log_2 \left( \frac{1}{4} \right) + \frac{3}{4} \log_2 \left( \frac{3}{4} \right) \right) = 0.81$$

$$A : IG_H = 1 - \frac{4}{8} \times 0.81 - \frac{4}{8} \times 0.81 = 0.19$$

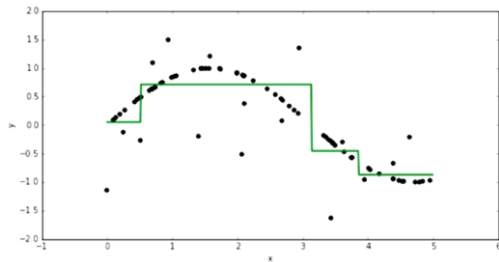
$$B : I_H(D_{left}) = - \left( \frac{2}{6} \log_2 \left( \frac{2}{6} \right) + \frac{4}{6} \log_2 \left( \frac{4}{6} \right) \right) = 0.92$$

$$B : I_N(D_{right}) = 0$$

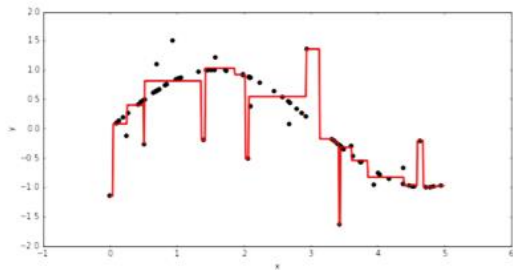
$$B : IG_H = 1 - \frac{6}{8} 0.92 - 0 = 0.31$$



# Регрессия



Использование деревьев для  
решения задачи регрессии



Можно легко переобучиться

# Как определить меру беспорядка в задаче регрессии?

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

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

# Критерии останова



## Возможные критрии останова:

- Ограничение максимальной глубины дерева
- Ограничение минимального числа объектов в листе
- Ограничение максимального количества листьев
- Ограничение на значение предсказания в листе
- Ограничение на дельту улучшения функционала качества

# Вывод. Достоинства



- Учитывает нелинейность данных
- Интерпретируемость и возможность визуализации
- Гибкость можно варьировать множество  $B$  (любые критерии разделения в вершинах)
- Допустимы разнотипные данные и данные с пропусками
- Не бывает отказов от классификации



# Вывод. Недостатки



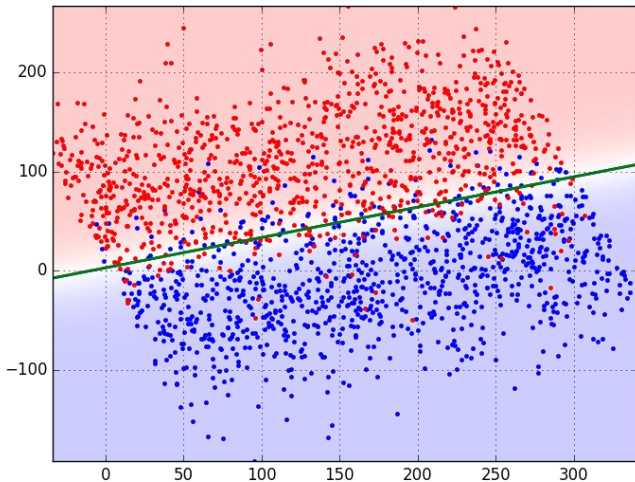
- **Жадный ID3** переусложняет структуру дерева и, как следствие, переобучается - <https://youtu.be/MFS0gKU3ICQ?t=2634>
- **Фрагментация выборки:** чем дальше от корня, тем меньше статистическая надежность
- **Высокая чувствительность** к шуму к составу выборки и КИ



# Недостатки линейных моделей



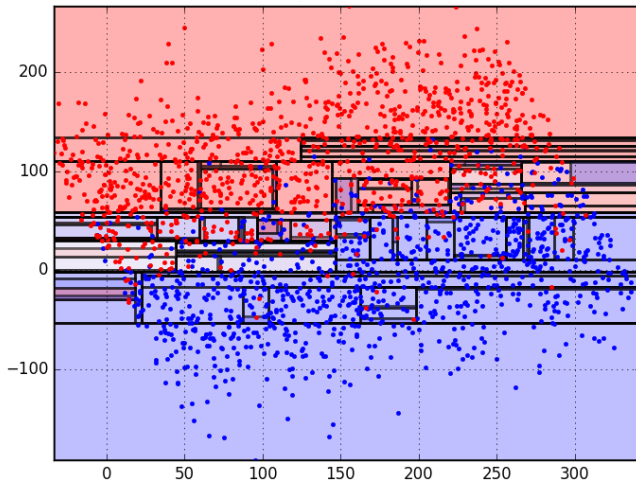
Logistic Regression, f-measure = 0.922420



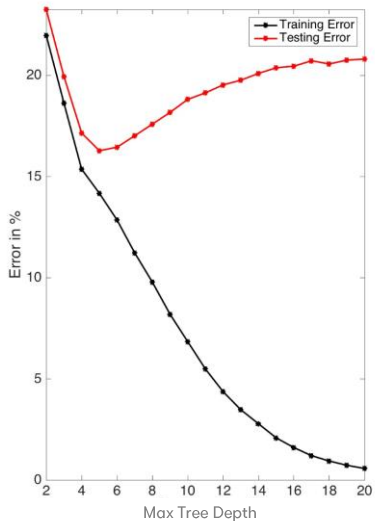
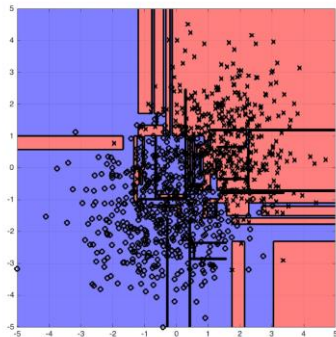
# Недостатки бинарных деревьев решений



Decision Tree, f-measure = 0.889780




# Недостатки бинарных деревьев решений





# Решение: Композиция деревьев



- 
- уменьшается чувствительность к изменению в данных;
  - уменьшается разбор ответом;
  - смещение остается неизменным.