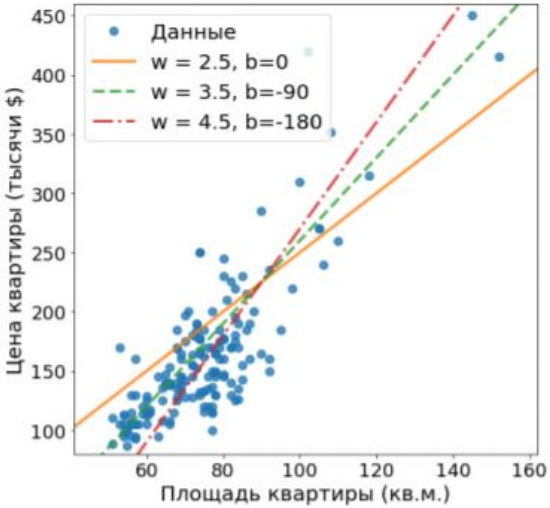
# План занятия

- Линейная регрессия
- Логистическая регрессия
- Решающее дерево
- Линейные модели

# Линейная регрессия

# Линейная регрессия

| x | y | $\hat{y}(x)$, w = 2.5, b = 0 | $\hat{y}(x)$, w = 3.5, b = -90 |
|---|---|---|---|
| 77 | 115 | 192.5 | 179.5 |
| 79 | 175 | 197.5 | 186.5 |
| 84 | 170 | 210.0 | 204.0 |
| 65 | 140 | 162.5 | 137.5 |




- y – настоящая цена квартиры
- ŷ(x) - то, что выдала модель
- x – площадь квартиры
- w, b – коэффициенты (параметры) модели линейной регрессии

$$\hat{y}(x) = w\,x + b$$

# Другие loss’ы

Среднеквадратичная ошибка имеет свои минусы, и иногда не подходит для конкретной задачи

- MSE чувствительна к выбросам
- MAE не гладкая
- Huber их объединяет, но имеет внешний параметр

Не для каждой функции потерь есть аналитическое решение

Градиентный спуск работает всегда, когда функция хотя бы кусочно дифференцируема

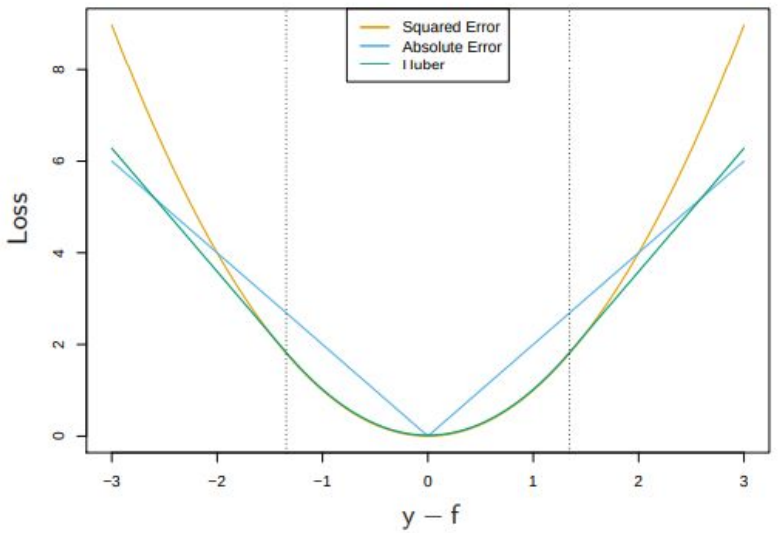



**FIGURE 10.5.** *A comparison of three loss functions for regression, plotted as a function of the margin $y-f$. The Huber loss function combines the good properties of squared-error loss near zero and absolute error loss when $|y-f|$ is large.*

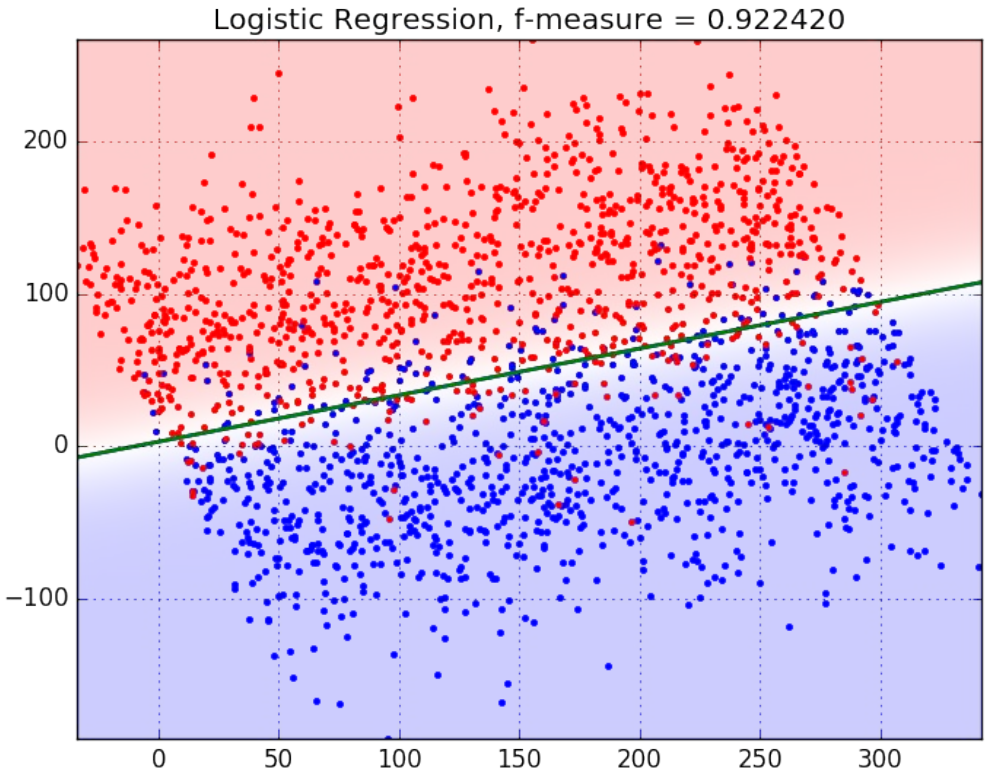
# Логистическая регрессия

# Logistic regression

Логистическая регрессия - линейный метод **классификации**

Название исторически сложилось, т. к. этот метод предсказывает вероятность

Также называют линейным классификатором



# MSE Loss

Попробуем обучить модель с помощью MSE (будем относить объект к положительному классу если прогноз для него положительное число)

$$L(\hat{y}, y) = (y - \hat{y})^2 = (y - \sum_{i}^{n} w_i x_i)^2$$

Плюсы:

- Простота

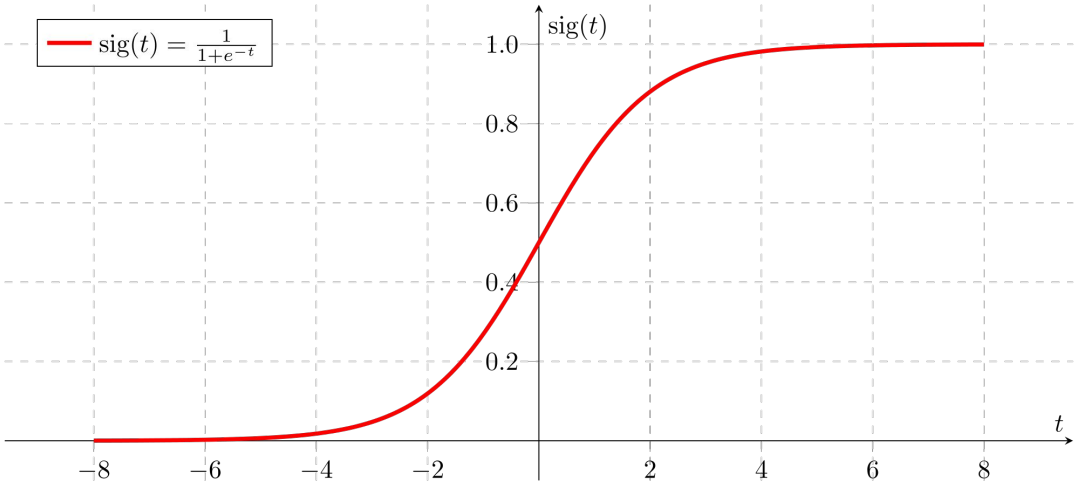
Минусы:

- Нет информации об уверенности прогноза
- Модель получает штраф за большое положительное предсказание

# Решение

Будем предсказывать вероятность, а не метку класса
Для этого будем использовать функцию сигмойды (отображает расстояние в вероятность)

Формула сигмойды $\sigma(x) = \frac{1}{1 + e^{-x}}$

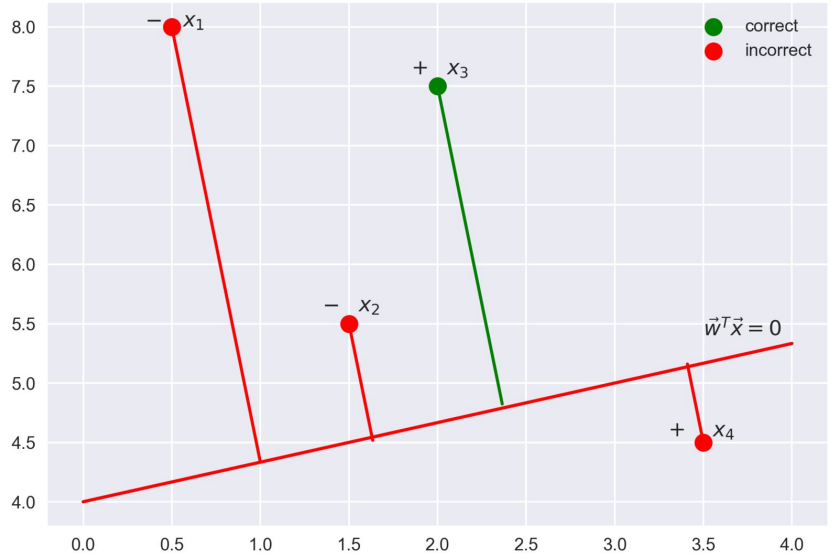


# Логистическая регрессия

$$\hat{y} = \sigma(x_1 w_1 + x_2 w_2 + b)$$

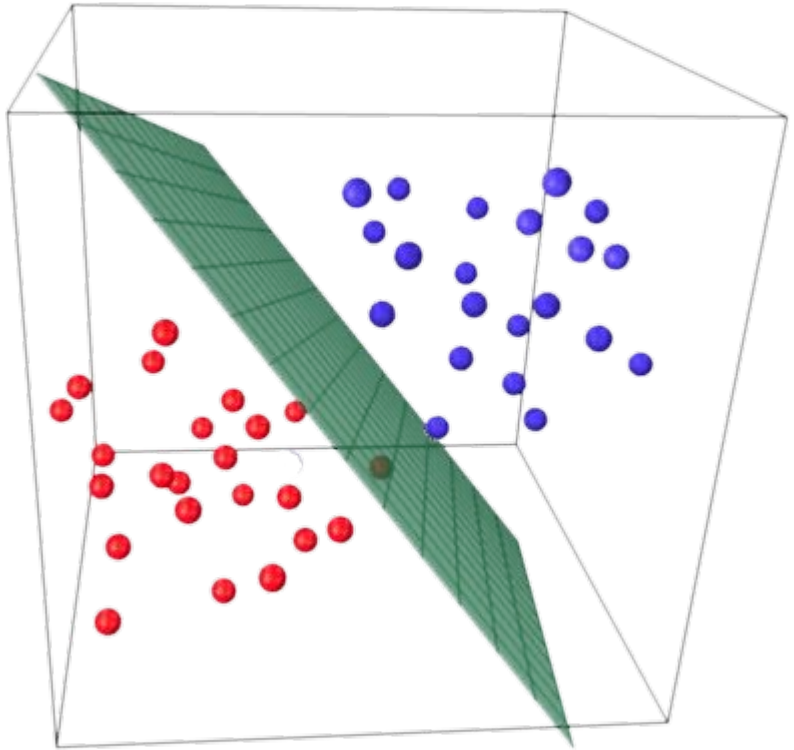
$$S = \frac{|x_1 w_1 + x_2 w_2 + b|}{\sqrt{w_1^2 + w_2^2}}$$

- Выражение внутри сигмоиды — расстояние S от точки до разделяющей прямой (с точностью до нормировки)
- Чем расстояние больше, тем более сеть уверена в своем ответе
- Сигмоида делает из расстояния вероятность
- Метрика качества Accuracy (доля правильных ответов)

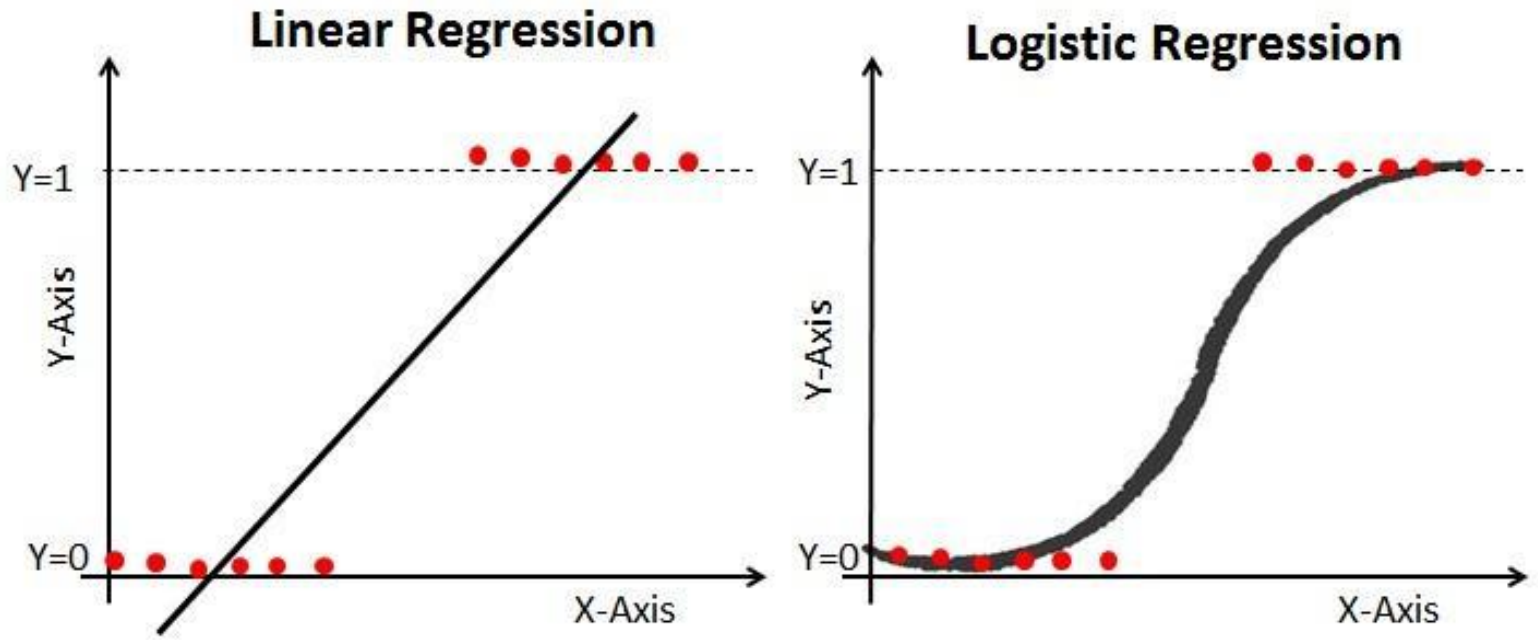


# Пример построенной разделяющей поверхности

- Разделяющая плоскость —всегда линейная
- На рисунке — разделяющая плоскость для датасета с 3 признаками



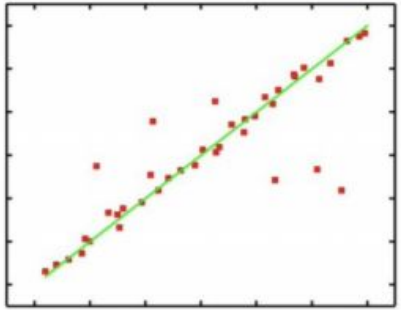
# А почему бы не предсказывать класс как число?



# Линейные модели

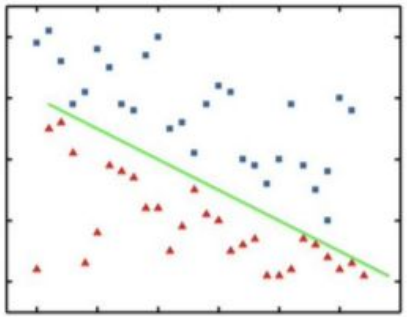
- Линейная регрессия: задача регрессии
  - Метрика качества MSE
  - Формула прогноза:

$$\hat{y} = \sum_{i=1}^{n} x_i w_i$$



- Логистическая регрессия: задача бинарной классификации
  - Метрика качества Accuracy
  - Формула прогноза:

$$\hat{y} = \sigma\left(\sum_{i=1}^{n} x_i w_i\right)$$

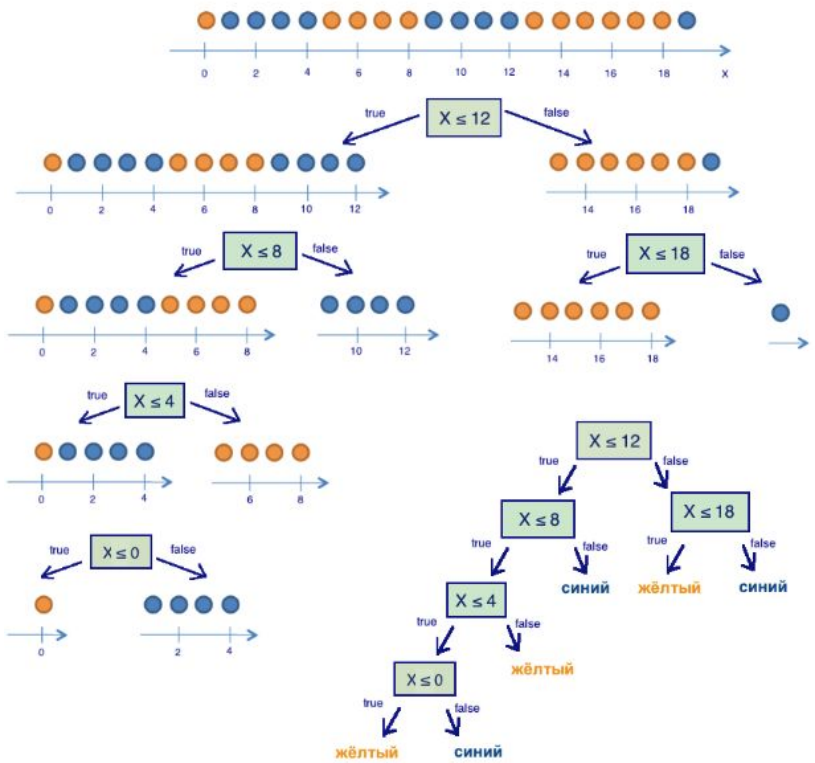


# Решающее дерево

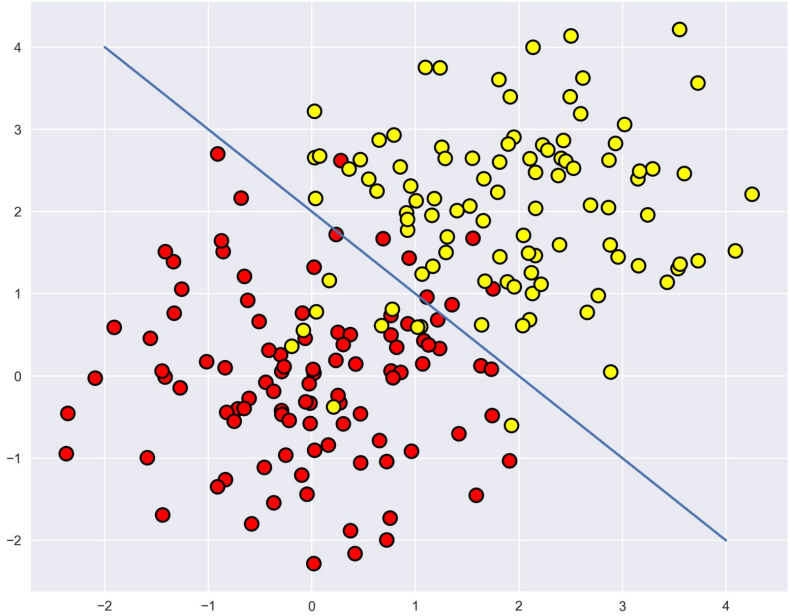
# Как выглядит?

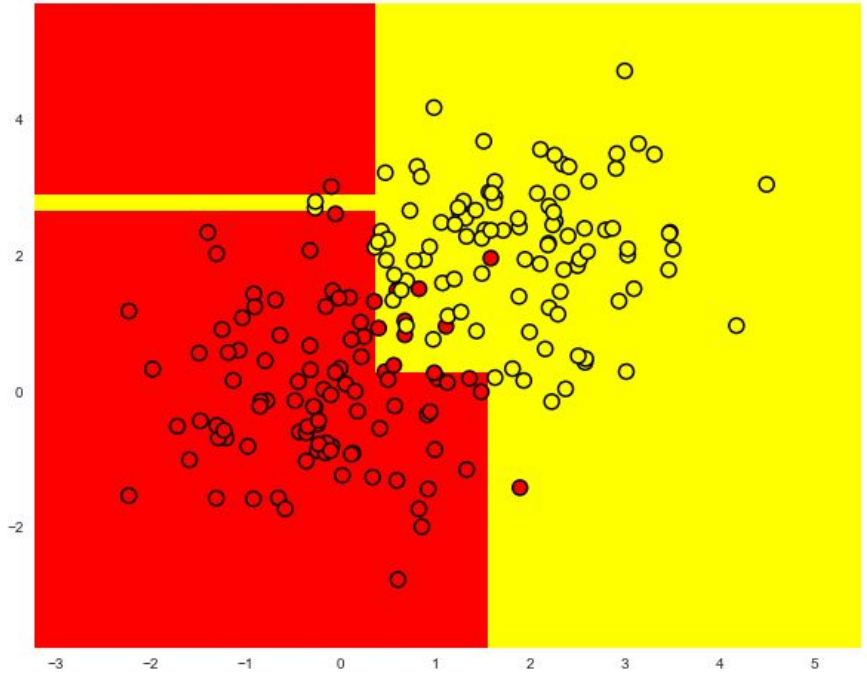
# Как строится?

- Выбираем признак, по которому наблюдается наибольший прирост порядка
- Выбираем порог для разделения
- повторяем процедуру в каждой из двух групп

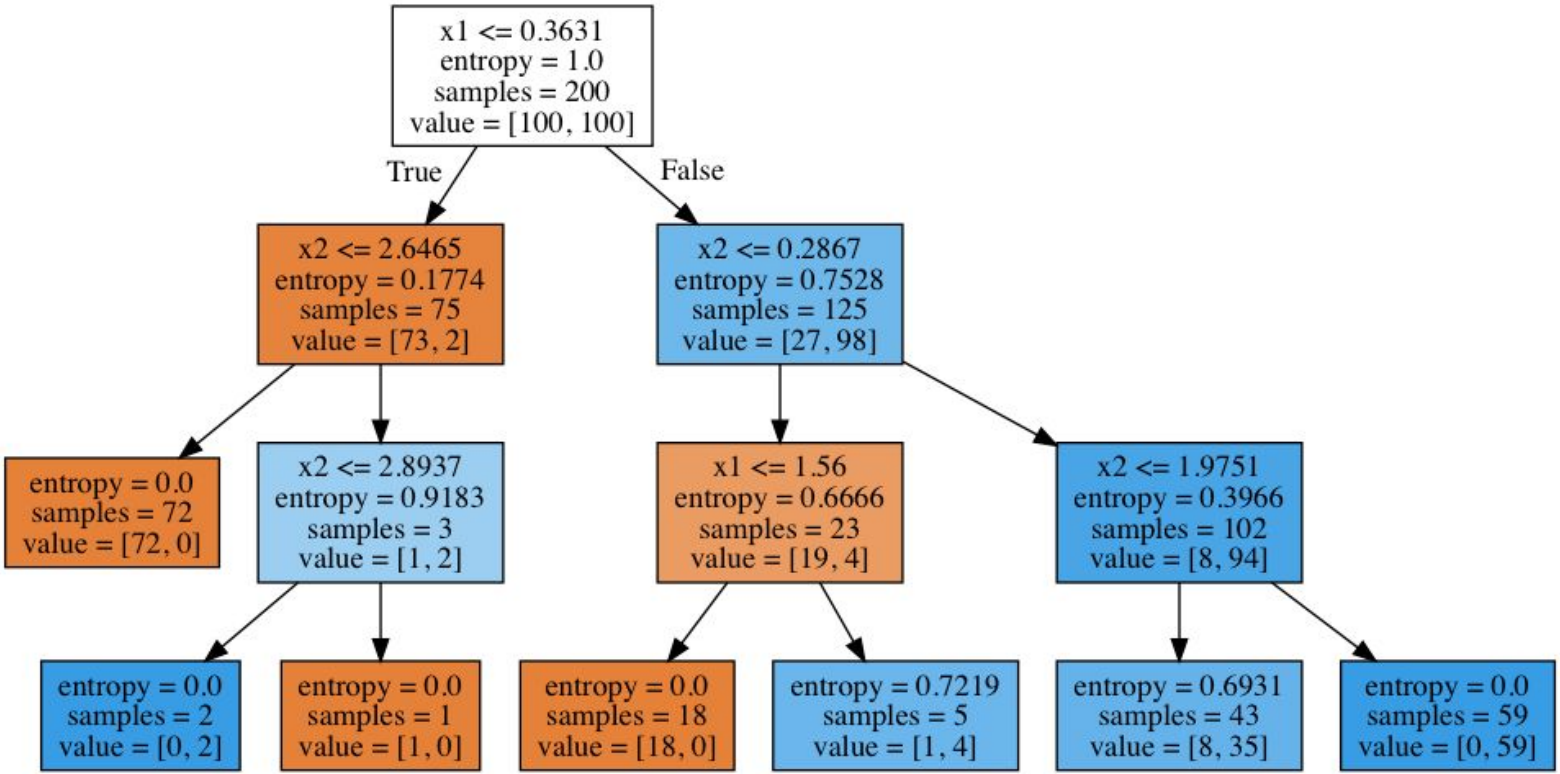


# Как выглядит?



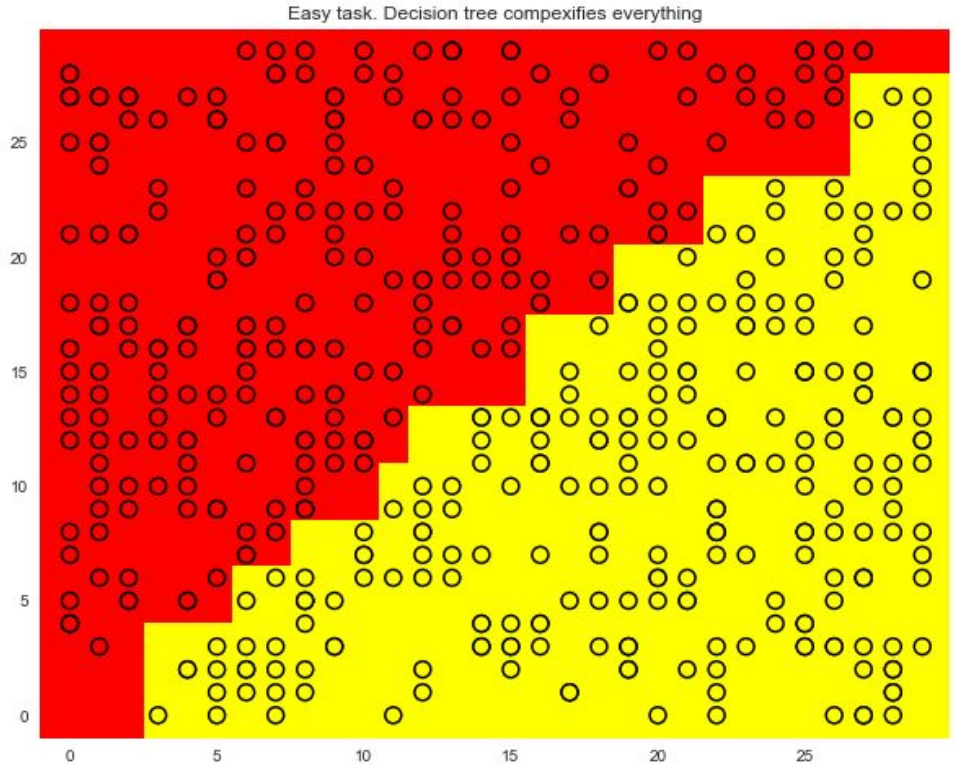
# Принятие решения



# Дерево линейно?

- Если в качестве новых признаков взять индикатор попадания в конкретную область
- Прогноз:

$$\hat{y} = \sum_{i=1}^{n} [x \in A_i] w_i$$



# Решающее дерево

- Плюсы
  - поддается интерпретации
  - обладает большей обобщающей способностью чем простые модели
  - могут работать с пропусками в данных “из коробки”

- Минусы
  - склонно к переобучению
  - неустойчивость к шуму
  - фрагментация (избыточная сложность структуры)

# Итоги

- Поговорили о линейных моделях
- Изучили два типа моделей
  - логистическая регрессия
  - решающие деревья
- Разобрались с метрикой Accuracy