Machine learning: Regression

The problem

Regression

- Let us say I tell you that I have a house I want to sell:
  - It has a square footage of 2500 sqft
  - It has 4 bedrooms
  - It has a pool
  - It needs repairs
- I want to predict how much money can I sell this house for?

What makes this a regression problem?

- **Inputs**:
  - A single input value: e.g. square footage
  - A set of values: e.g. square footage, bedrooms, condition
  - A picture, a video or a sound file
    - Eg. forward looking camera in a self-driving car
- **Outputs**: one or more **numerical** values
  - Price of the house
  - Distance to the car in front

How do I know that it works?

- **Accuracy**: distance of between the predicted value and real value
  - We cannot just subtract them, as negative values do not make sense for accuracy!
  - We can take the **absolute value of the difference** and average them (Mean Absolute Error or L1 loss)
  - We can take the **square of the difference**, and average them (Mean Square Error or L2 loss)

The solution

Building a regressor with knowledge engineering

- Maybe I can create a formula to predict the price of the house:
  - +$100 for every square ft
  - +$30,000 for a pool
  - -$50,000 if it needs repairs

$$price = 100 \times size + 30,000 \times has\_pool - 50,000 \times needs\_repairs$$

- the more complicated the problem, the harder is to come up with the right formula

Building a regressor with learning

- Start with a number of examples: houses for which we know the parameters as well as the price for which they had been sold!
- This is another example of supervised learning, just like in the case of classification
  - The difference is that we are predicting a **number** rather than a **class**

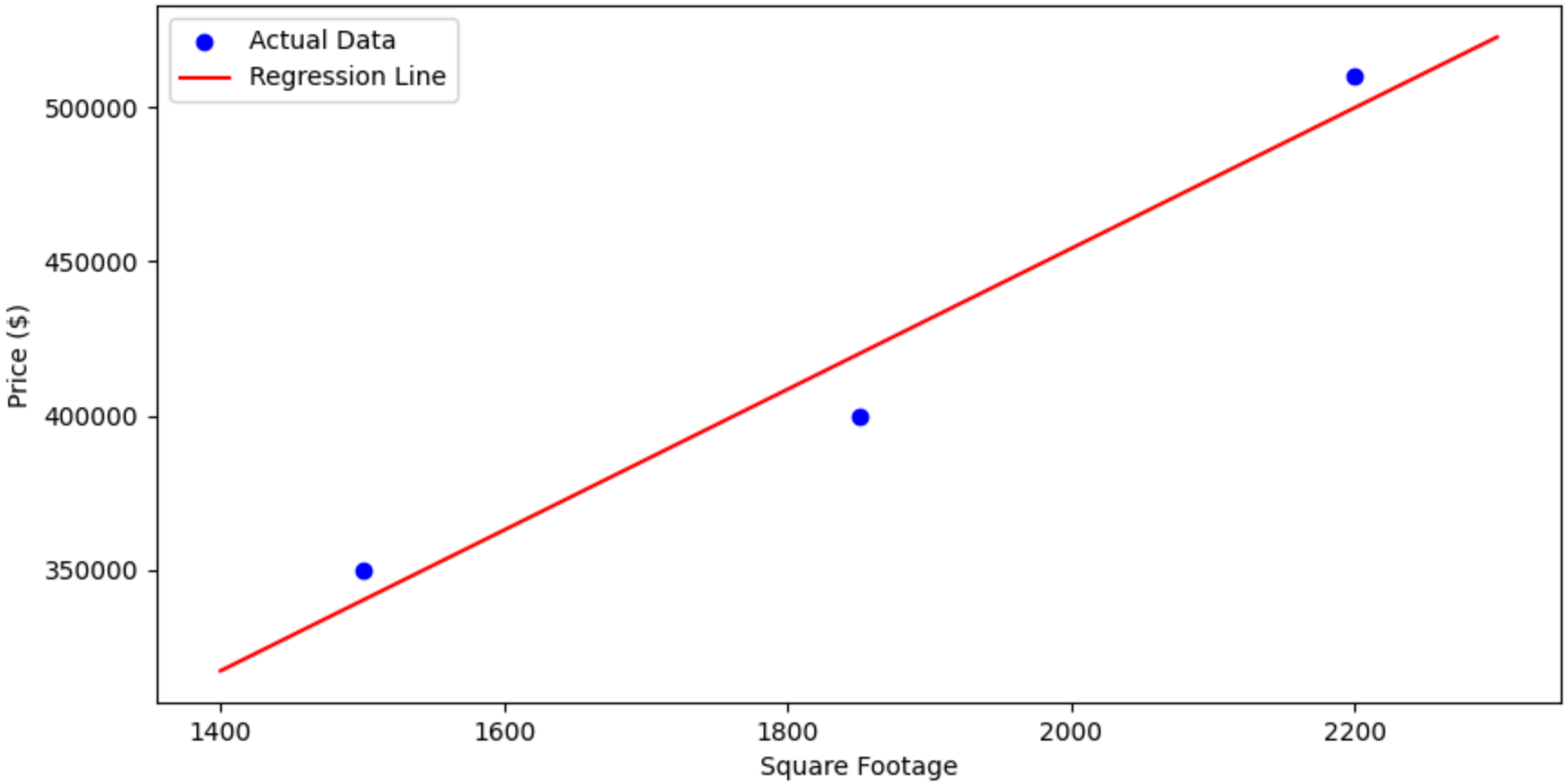
Simplest case: linear regression

- Assume that the inputs are numbers $x_1, x_2, ...$
  -For instance, x_1 is the size of the house in square ft, x_2 is 1 if there is a pool and 0 if not
- The output is a number y
  -For instance, y is the predicted price of the house
- We assume that the output can be described in the following way:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n + \epsilon$$

- This is called a $linear$ expression, because when we plot it, it will be something like a line, a flat plane etc.

Example of linear regression



What does linear regression do?

- Imagine that the training data are **points**
- Linear regression finds the line which is the **closest possible to the points**
  - This is called **fitting a line to the points**
  - It might not be perfect because the points might not be on a line
- How do we use it?
  - If we have have to predict the price for a house, and we know the square footage $x = 1200$
    - We find the position on the x axis
    - Find the corresponding position on the line
    - Read out the y value: that is our prediction!

More complicated cases: multiple inputs

- What if there are multiple input variables?
  - If we have 1 input variable, we fit a 1-dimensional **line** in a 2D plane
  - If we have 2 input variables, we fit a 2-dimensional **plane** in a 3D space
  - If we have 3 input variables, we fit a 3-dimensional **hyperplane** in a 4D space
    - We cannot conveniently visualize this, but computer code can take care of this
    - And the output is still just one number!

More complicated cases: nonlinear problems

- 30ft frontage: $150,000

- 60ft frontage: $800,000

More complicated cases: nonlinear problems

- What if the points just totally cannot fit on a line?
  - For instance: house frontage
- There are several approaches:
  - We can try to fit a curve instead (non-linear regression)
  - We can try to find several similar houses, and return their average price (K-nearest neighbors)
  - We can try to use a neural network (more about this later)

Applications

Applications of regression

- Any time we need to predict a number
  - Predicting the sales of a product
    - function of the price, advertising budget etc
  - Predict the stock market
    - function of company profitability, expenses etc.
  - Predict the price of a house
    - function of its attributes
  - Predict the crop yield
    - function of rain, fertilizer expense, etc
  - Predict a student's GPA
    - function of hours studied, tutoring hours, previous grades etc.

Pitfalls and dangers

- What if I have the following house price data:
  - 2000sqft, $2,300,000
  - 3200sqft, $4,500,000
- What happened?
  - These prices are not from Orlando, but from Palo Alto, California
  - Our predictions are **way off**
- Out of distribution

Try it out + Homework