1.0 - What is Data Science?

Extract information from Data

Example 1

Years of study / salary relation

1. Understand the relation between an "input" and an "output"
2. Find a function that roughly estimates the data points (called *regression*)

Other examples

- Using user rating to generate movie recommendations
- Label an entire dataset from only a few labels
- Classify data
    - Whole dataset is labeled (e.g. Cats and Dogs)
    - Draw a new sample (take a new image) and classify it (as a Dog or a Cat)

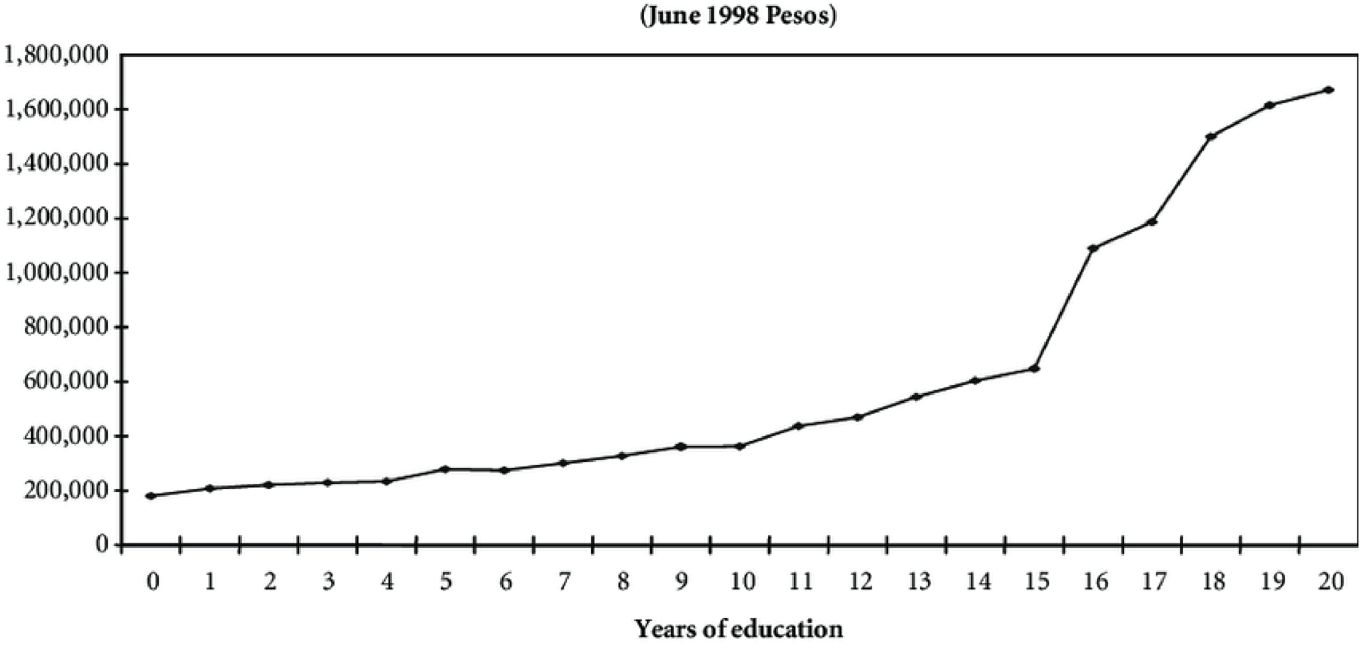
3 Fields

1. Supervised learning
    - classification / regression
2. Unsupervised learning
    - clustering
    - you only have data and no information on it
3. Semi-supervised learning
    - if you have a few labels and you want to infer all the labels

Terminology

step size = learning rate

Example: The income dataset



The data points $(x_i, y_i) \in \mathbb{R}^2$ are supposed to be of the form

$$y_i = f(x_i) + \epsilon \quad (\epsilon \rightarrow \text{noise})$$

Remark: In general, the function f is *unknown*
We want to approximate this function f using the data!
Find an approximation $\hat{f}$ of f using $\{(x_i, y_i)\}_{1 \leq i \leq N}$

Nearest neighbours interpolation

The nearest neighbours interpolation of $\{(x_i, y_i)\}_{1 \leq i \leq N}$ is the function

$$x \Rightarrow \hat{f}(x) = y_{i*}$$

This means for an input x you find the nearest data point x_i (i.e. the one with the smallest absolute distance to x) and assign its corresponding value y_i to $\hat{f}(x)$

when $i_x \in \text{argmin}_{1 \leq i \leq N} |x - x_i|$

- Perfectly fits the data, i.e. $\hat{f}(x_i) = y_i$
- However, its sensible to outlayers
    - **Outlayers**: Some point in the graph is somehow not along the expected function (imagine point (18, 40,000)), which is very far from the function we are seeking