School of Electronic Engineering and Computer Science
Queen Mary University of London

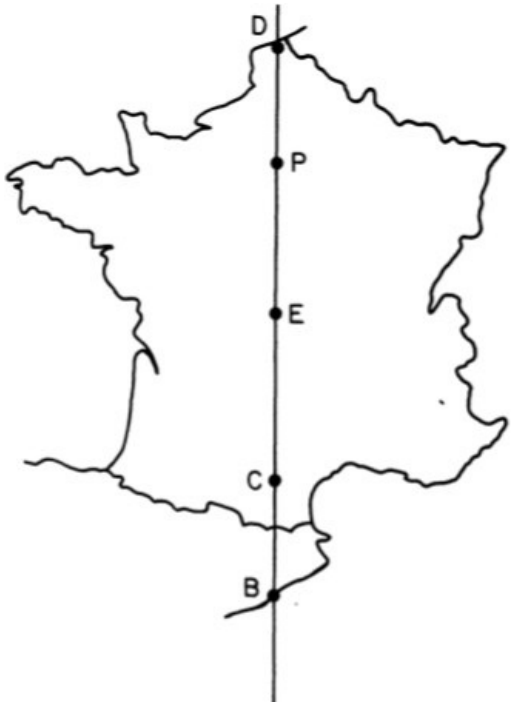
CBU5201 Machine Learning
Supervised learning: Regression

Dr Chao Liu

Sep 2023

Credit to Dr Jesús Requena Carrión

How far is the equator from the north pole?



”By using this method, a sort of equilibrium is established between the **errors** which prevents the extremes from prevailing [...] [getting us closer to the] **truth**.”

Adrien-Marie Legendre, 1805

Embrace the error!

Agenda

Machine learning

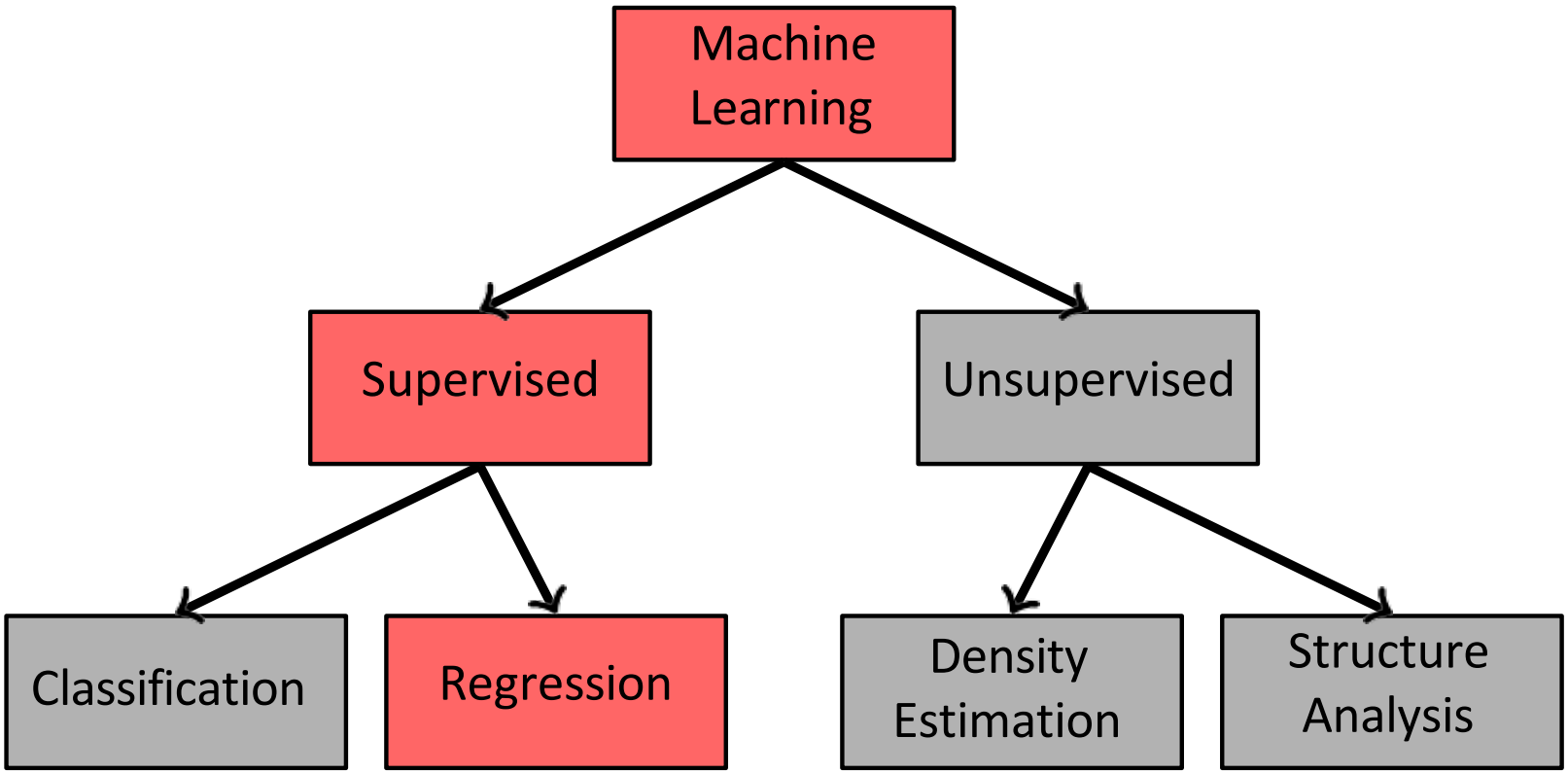
There are two main ways of thinking about ML:

- **Data-first** view: ML is a set of tools for extracting knowledge from data.
- **Deployment-first** (our) view: ML is a set of tools together with a methodology for solving problems using data.

In ML, data is organised as a **dataset** (a collection of **items** described by a set of **attributes**) and knowledge is represented as a **model**.

Machine learning distinguishes between different types of problems, techniques and models, which can be arranged into a **taxonomy**.

Machine learning taxonomy



Agenda

Problem formulation

- Regression is a **supervised** problem: Our goal is to predict the value of one attribute (**label**) using the remaining attributes (**predictors**).
- The label is a **continuous** variable.
- Our job is then to **find the best model** that assigns a unique label to a given set of predictors.
- We use **datasets** consisting of **labelled samples**.

Predictors $\longrightarrow$ Model? $\longrightarrow$ Label

Examples of regression problems

The following are examples of business and scientific problems that can be formulated as a regression problem:

- Predict the energy consumption of a household, given the location of the house, household size, income, intensity of occupation.
- Predict future values of a company stock, given past stock prices.
- Predict distance driven by a vehicle given its speed and journey duration.
- Predict demand given past demand and currency exchange rate.
- Predict tomorrow's temperature given today's temperature and pressure.
- Predict the probability to develop a specific heart condition given BMI, alcohol consumption, diet, number of daily steps.

Can you identify labels and predictors? Do we need data to solve them?

Predictors and labels

	Age	Salary
S_1	18	12000
S_2	37	68000
S_3	66	80000
S_4	25	45000
S_5	26	30000
...	...	...

In this dataset:

(a) *Age* is the predictor, *Salary* is the label

(b) *Salary* is the predictor, *Age* is the label

(c) Both options can be considered

Association and causation

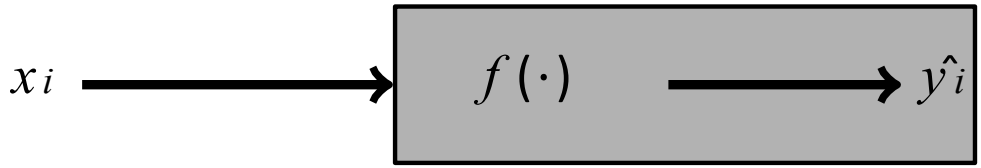
Prediction models are sometimes interpreted through a **causal lens**: the predictor is the **cause**, the label its **effect**. However this is **not correct**.

Our ability to build predictors is due to **association** between attributes, rather than **causation**. Two attributes in a dataset appear associated:

- If one causes the other (directly or indirectly).
- When both have a common cause.
- Due to the way we collect samples (sampling).

Take-home message: In machine learning we don't build causal models!

Mathematical notation



Population:

- x is the **predictor** attribute
- y is the **label** attribute

Dataset:

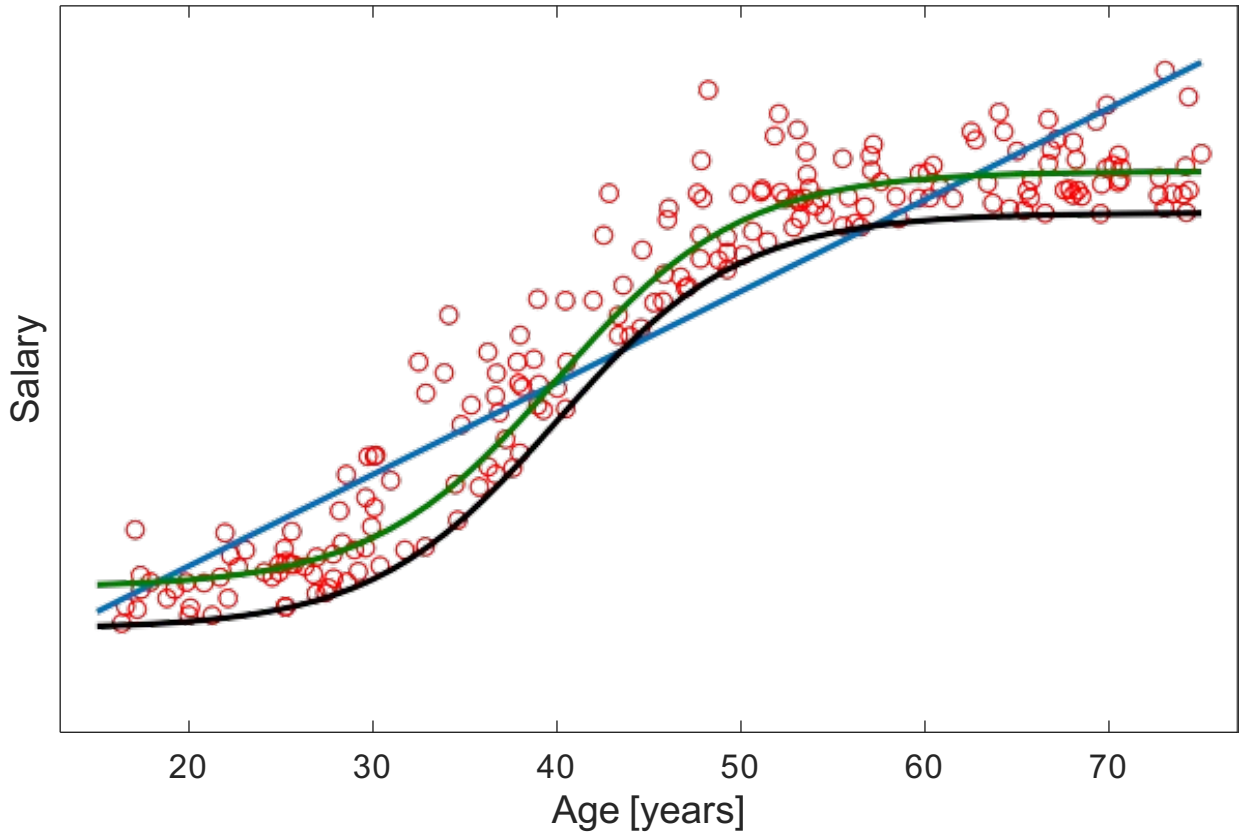
- N is the number of samples, i identifies each sample
- x_i is the predictor of sample i
- y_i is the actual label of sample i
- (x_i, y_i) is sample i, $\{(x_i, y_i) : 1 \leq i \leq N\}$ is the entire dataset

Model:

- $f(\cdot)$ denotes the model
- $\hat{y}_i = f(x_i)$ is the **predicted label** for sample i
- $y_i - \hat{y}_i$ is the **prediction error** for sample i

Candidate solutions

Which line is the *best* mapping of age to salary?



What is a good model?

In order for us to find the **best** model we need a notion of **model quality**.

The **squared error** $e_i^2 = (y_i - \hat{y}_i)^2$ is a common quantity used in regression to encapsulate the notion of single prediction quality.

Two quality metrics based on the squared error are the **sum of squared errors** (SSE) and the **mean squared error** (MSE), which can be computed using a dataset as:

$$E_{SSE} = e_1^2 + e_2^2 + \cdots + e_N^2 = \sum_{i=1}^{N} e_i^2$$

$$E_{MSE} = \frac{1}{N} \sum_{i=1}^{N} e_i^2$$

MSE: Example

A zero-error model?

Given a dataset, is it possible to find a model such that $\hat{y}_i = y_i$ for every instance i in the dataset, i.e. a model whose **error is zero**, $E_{MSE} = 0$?

(a) **Never**, there will always be a non-zero error

(b) It is **never guaranteed**, but might be possible for some datasets

(c) **Always**, there will always be a model complex enough that achieves this

The nature of the error

When considering a regression problem we need to be aware that:

- The chosen **predictors might not include all the factors** that determine the label.
- The chosen **model might not be able to represent** the true relationship between response and predictor (the pattern).
- **Random mechanisms** (noise) might be present.

Mathematically, we represent this discrepancy as

$$\begin{aligned} y &= \hat{y} + e \\ &= f(x) + e \end{aligned}$$

There will always be some discrepancy (error e) between the true label y and our model prediction $f(x)$. **Embrace the error!**

Regression as an optimisation problem

Given a dataset $\{(x_i, y_i) : 1 \leq i \leq N\}$, every candidate model f has its own E_{MSE}. Our goal is to find the **model with the lowest** E_{MSE}:

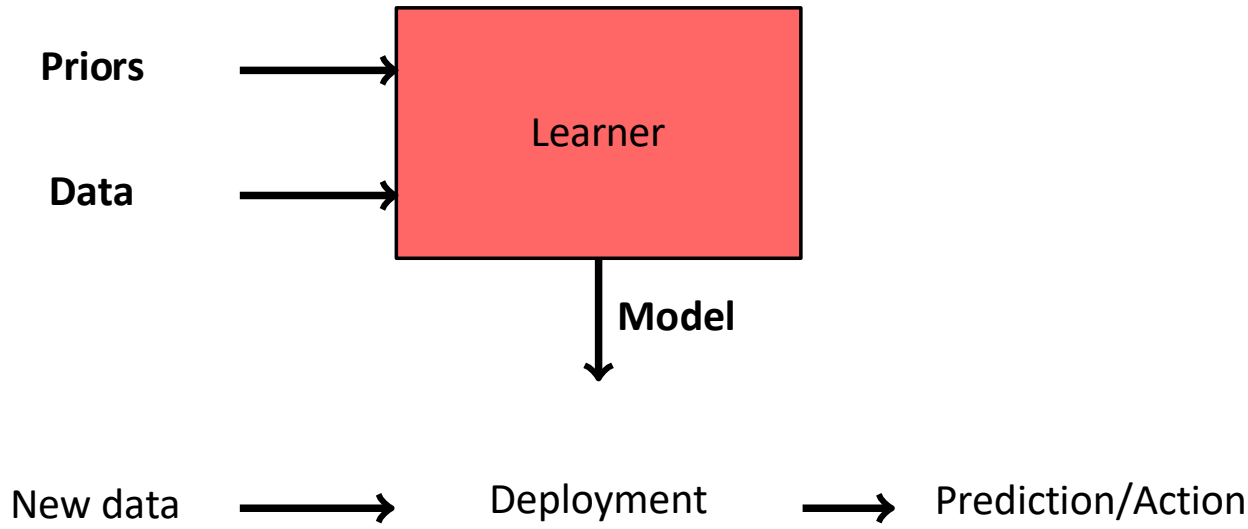
$$f_{best}(x) = \arg\min_f \frac{1}{N} \sum_{i=1}^{N} (y_i - f(x_i))^2$$

The question is, how do we find such model? Finding such a model is an **optimisation problem**.

Note that we are defining regression as finding the model that minimises E_{MSE} *on the dataset*, without considering what happens *once deployed*.

Agenda

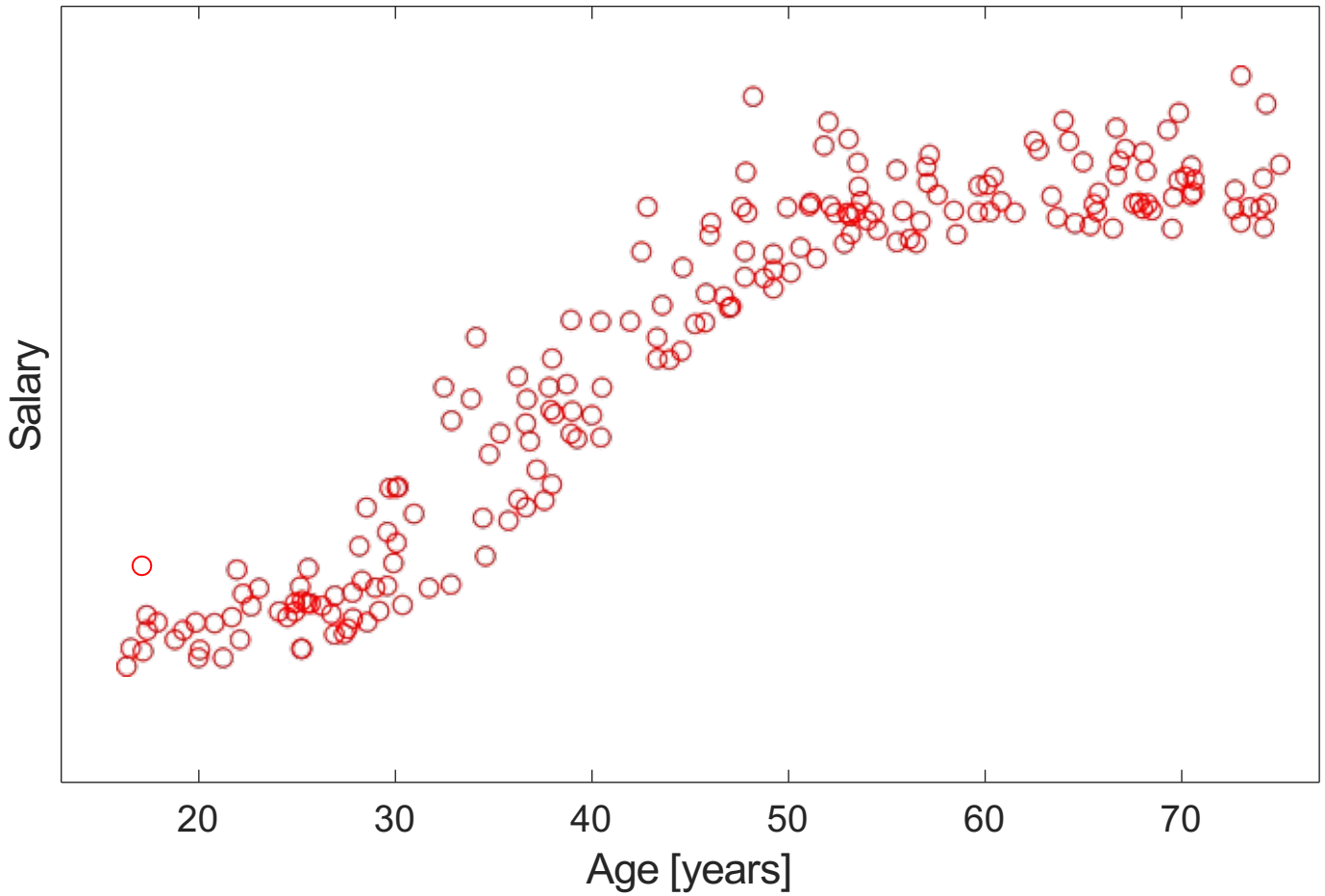
Our regression learner



- **Priors**: Type of model (linear, polynomial, etc). **Data**:
- Labelled samples (predictors and true label). **Model**:
- Predicts a label based on the predictors.

Simple regression

Simple regression considers **one predictor** x and one label y.



Simple linear regression

In simple **linear** regression, **models** are defined by the mathematical expression

$$f(x) = w_0 + wx_1$$

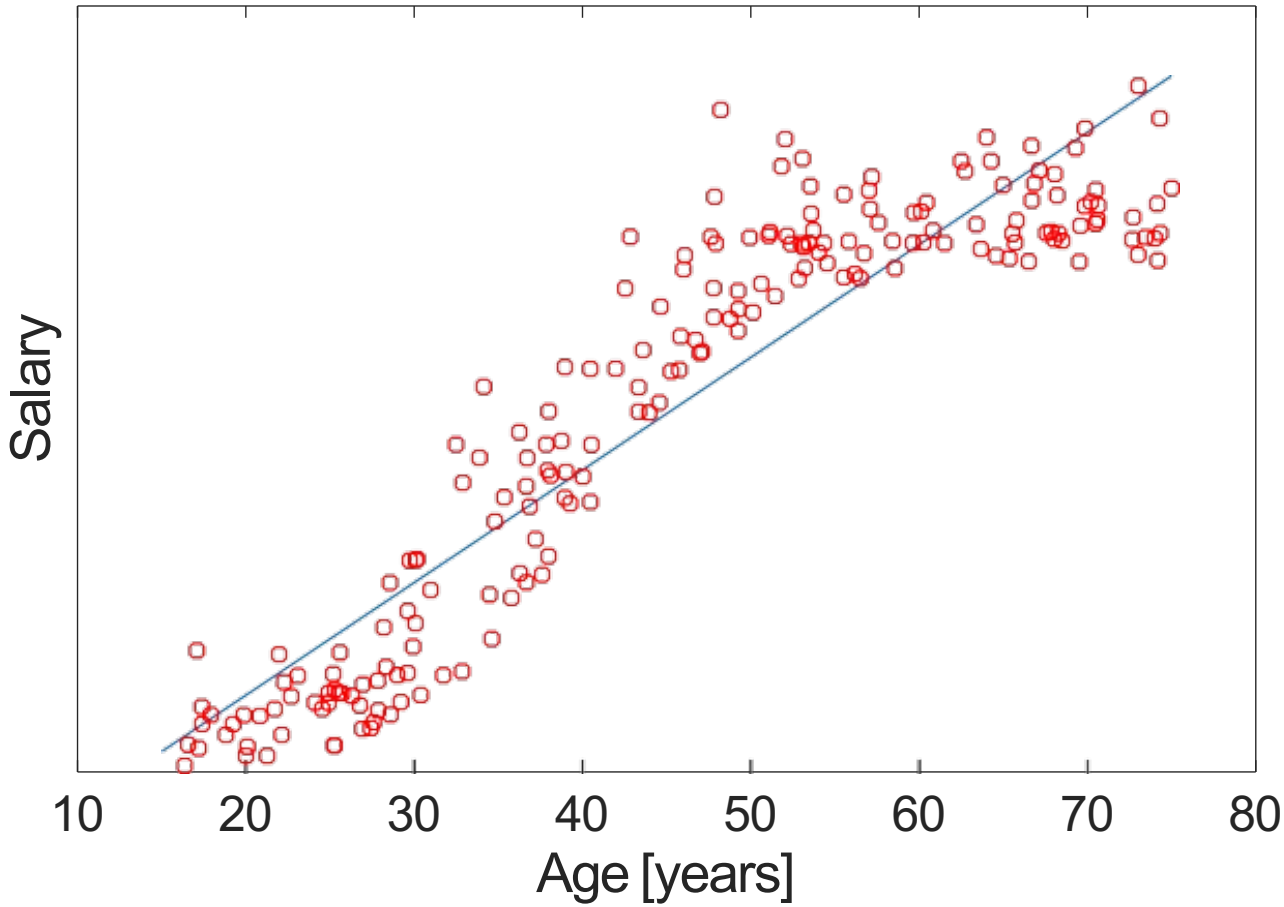
Hence, the predicted label $\hat{y}_i$ can be expressed as

$$\hat{y}_i = f(x_i) = w_0 + w_1 x_i$$

A linear model has therefore **two parameters** w_0 (intercept) and w_1 (gradient), which need to be **tuned** to achieve the highest quality.

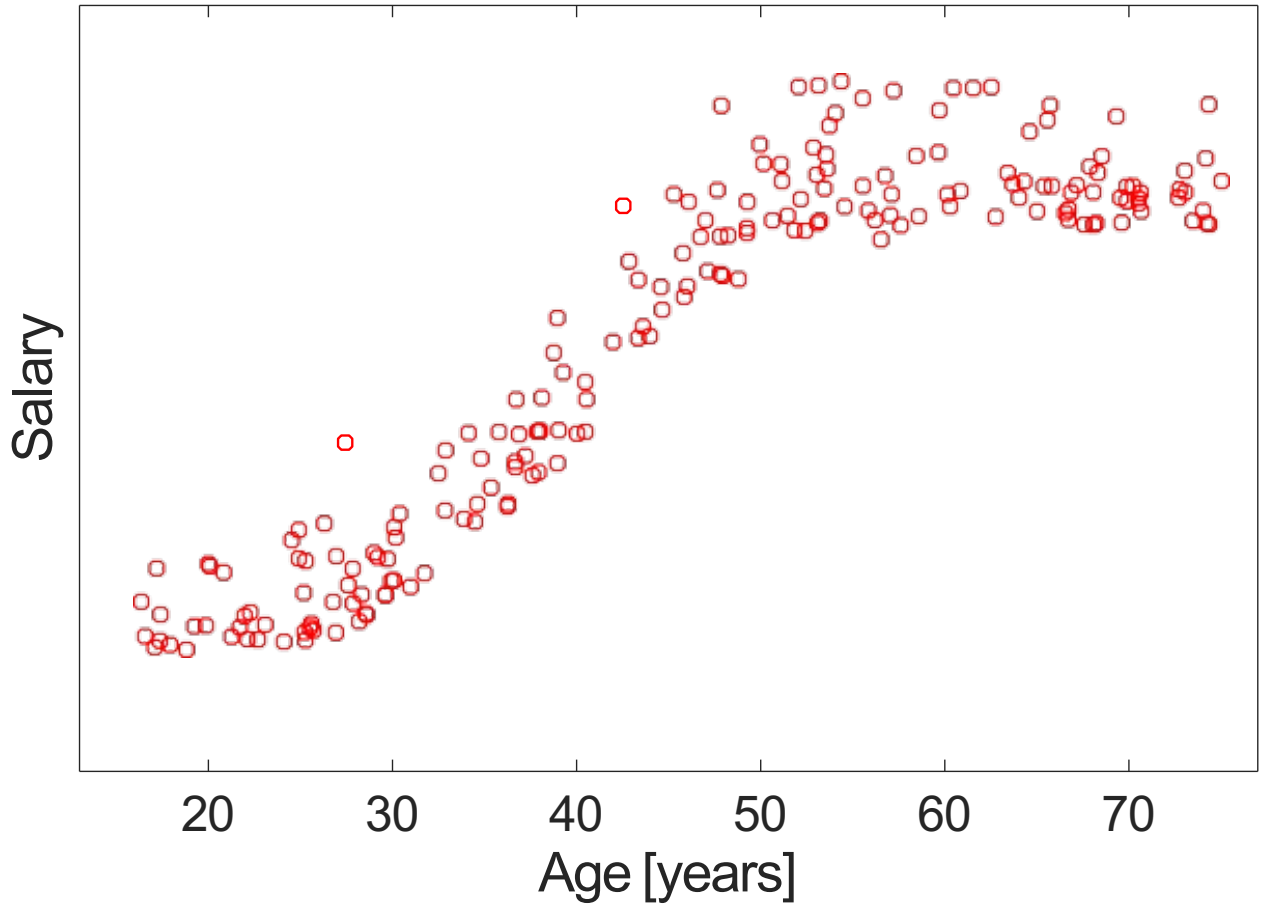
In machine learning, we use a **dataset** to tune the parameters. We say that we **train the model** or **fit the model** to the **training dataset**.

Linear solution: Example



Beyond linearity

Sketch the model that you would choose for the Salary Vs Age dataset and try to find a suitable mathematical expression.



Simple polynomial regression

The general form of a polynomial regression model is:

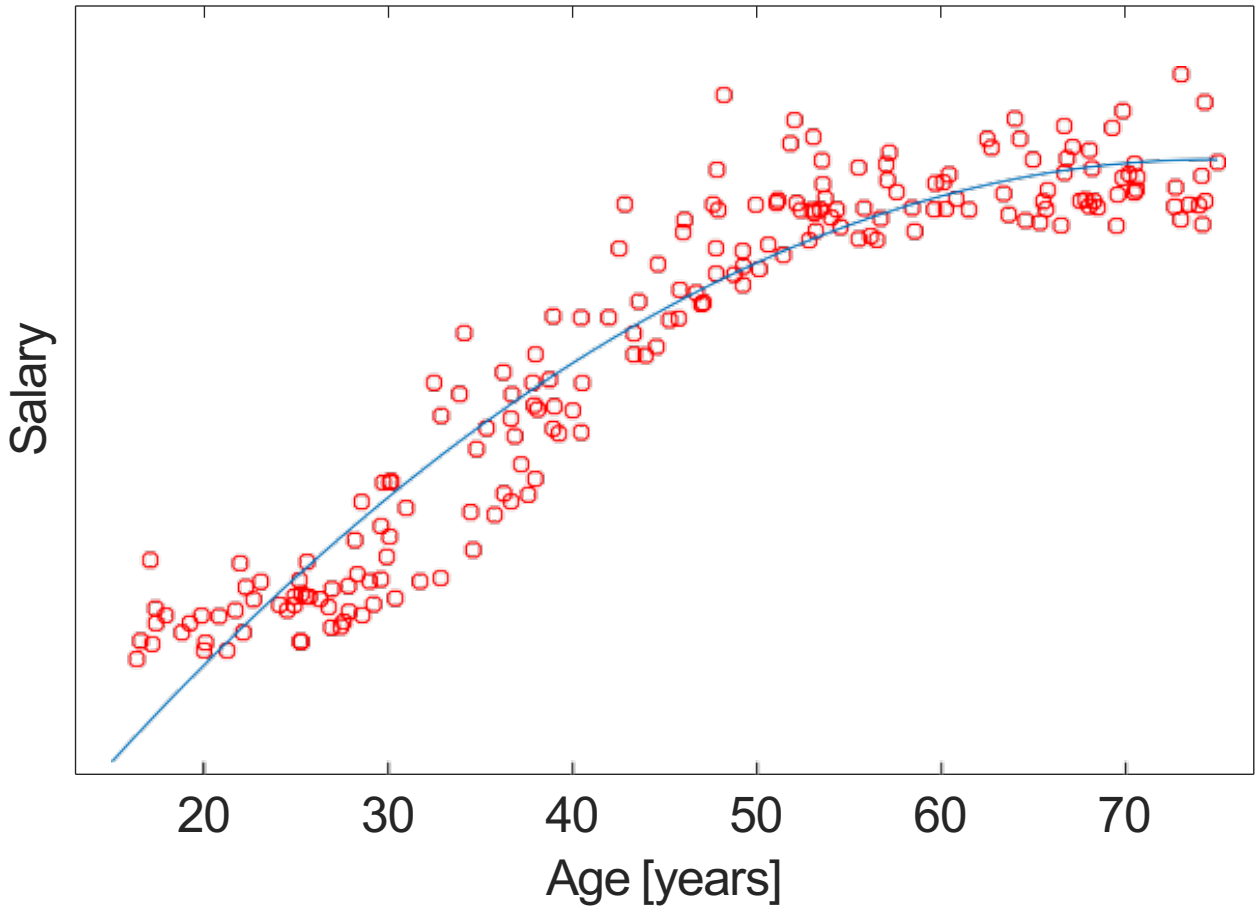
$$f(x_i) = w_0 + w_1 x_i + w_2 x_i^2 + \cdots + w_D x_i^D$$

where D is the degree of the polynomial.

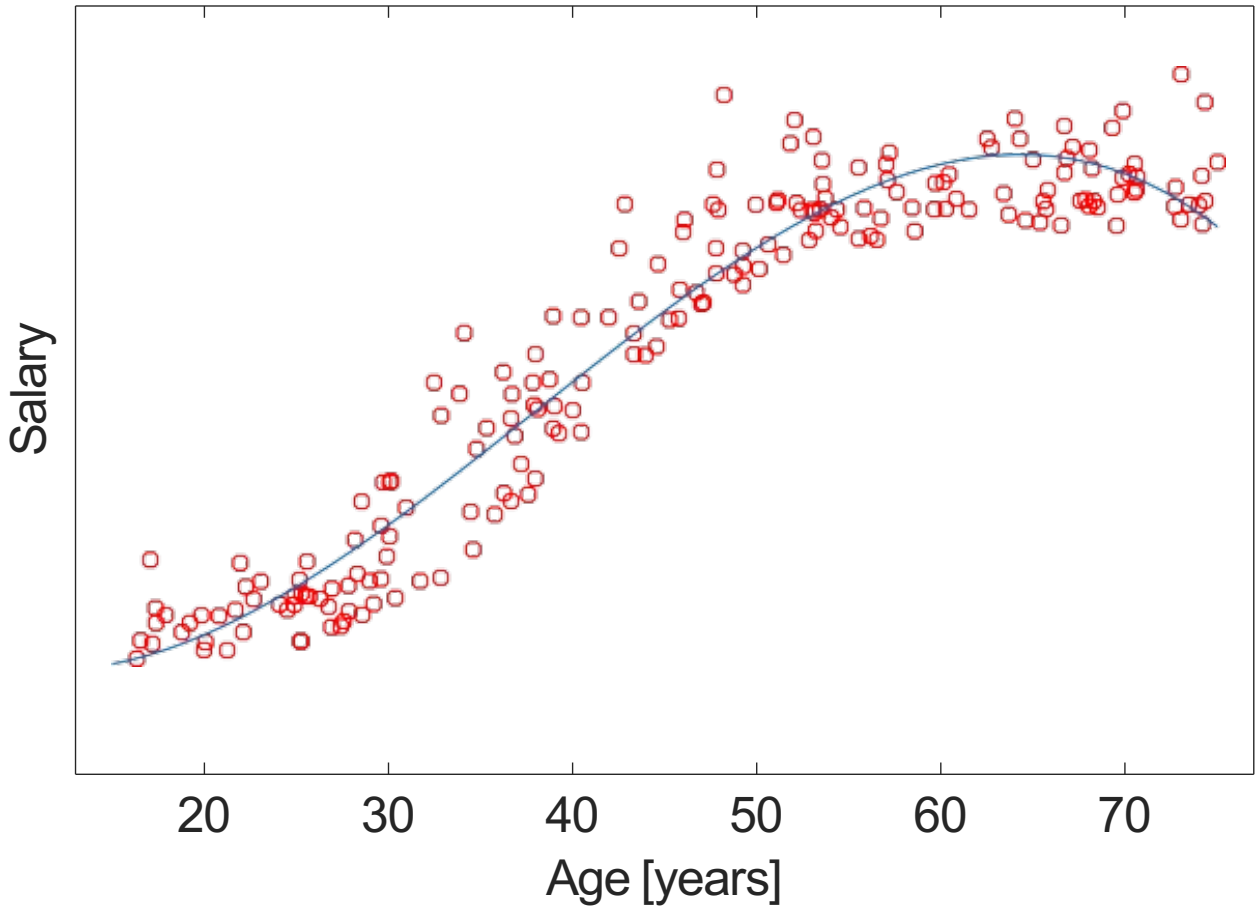
Polynomial regression defines a **family of families** of models. For each value of D, we have a different family: $D = 1$ corresponds to the linear family, $D = 2$ to the quadratic, $D = 3$ to the cubic, and so on.

We call D a **hyperparamenter**. What it means is that setting its value results in a different family, with a different collection of parameters.

Quadratic solution



Cubic solution



5-power solution