Linear Regression

A very teeny tiny primer

Example Dataset for Linear Regression

If you had studied longer, would your overall scores get any better

- One way of answering this question is by having data on how long you studied for and what scores you got.
- We can then try to see if there is a pattern in that data, and if in that pattern, when you add to the hours, it also ends up adding to the scores percentage
- For instance, say you have an hour-score dataset, which contains entries such as 1.5h and 87.5% score.
  - It could also contain 1.61h, 2.32h and 78%, 97% scores.
- The kind of data type that can have any intermediate, value (or any level of 'granularity') is known as continuous data.
- Based on the modality (form) of your data - to figure out what score you'd get based on your study time - , you'll perform regression or classification.

Regression vs Classification

- Regression is performed on continuous data, while classification is performed on discrete data.
- Regression can be anything from predicting someone's age, the house of a price, or value of any variable.
- Classification includes predicting what class something belongs to (such as whether a tumor is benign or, malignant).

Descriptive Info

	Hours	Scores,
count	25.000000	25.000000,
mean	5.012000	51.480000,
std	2.525094	25.286887,
min	1.100000	17.000000,
25%	2.700000	30.000000,
50%	4.800000	47.000000,
75%	7.400000	75.000000,
max	9.200000	95.000000

Regression and Classification

- For both regression and classification - we'll use data to predict labels, (umbrella-term for the target variables).
  - Labels can be anything from "B" (class) for classification tasks,
  - to 123 (number) for regression tasks.
- Because we're also supplying the labels - these are supervised learning algorithms (we have a dataset telling us what the score was according to number of hours studied)

Linear Regression

- Because of linearity, we will use **y = ax + b.**
  - y represents the score percentage,
  - x represents the hours studied,
  - b is where the line starts at the Y-axis (called Y-axis intercept),
  - a defines whether the line is going to move up/down in the graph
    - after all, it is called the **slope** for a reason)

How does it work?

- By adjusting the **slope** and **intercept** of the line, we can move it in any direction.
  - By figuring out the slope and intercept values, we can adjust a line to fit our data!
- That's the heart of linear regression
  - It really only figures out the values of the slope, and intercept.
  - It uses the values of x and y that we already have and varies the values of a and b.
  - By doing that, it fits multiple lines to the data points and returns the line that is closer to all the data points, or the best fitting line.
  - By modelling that linear relationship, our regression algorithm, is also called a model.
- In this process, when we try to determine, or predict the percentage (score), based on, the hours of study, it means that our **y** variable depends on the values of our x variable.