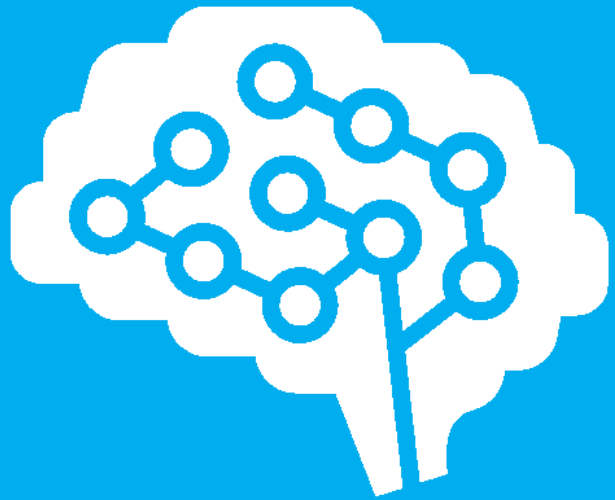
Warwick Business School

Data Science & Generative AI

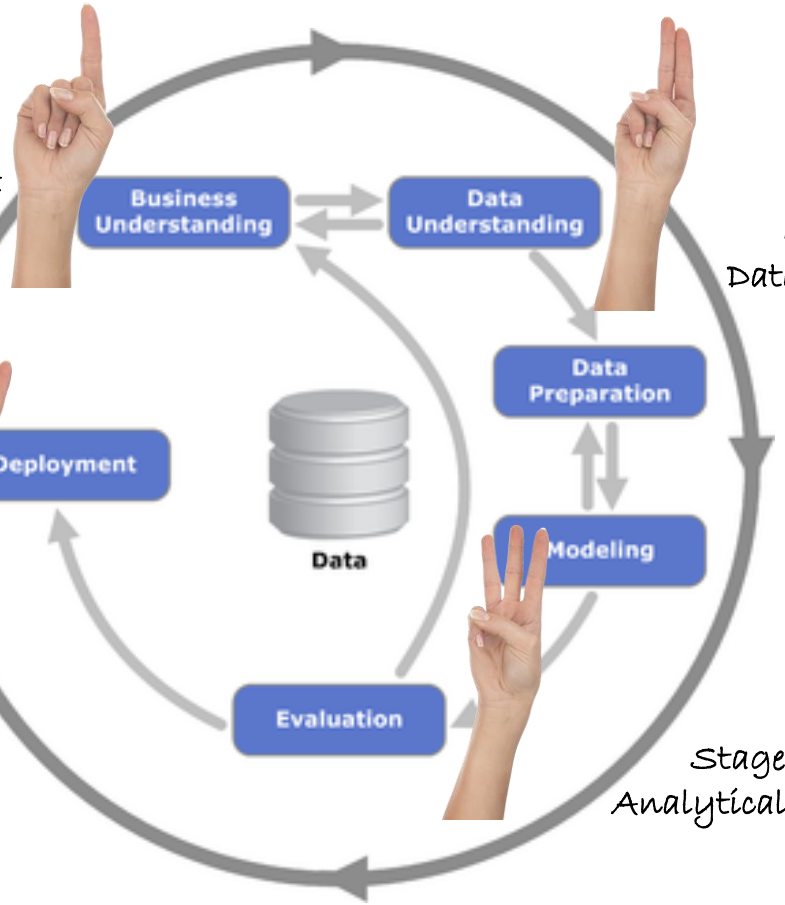
Dr Michael Mortenson
Associate Professor (Reader)
michael.mortenson@wbs.ac.uk



Session 3: Machine Learning Methodology

1.1 Stage Three

Stage one:
Business context

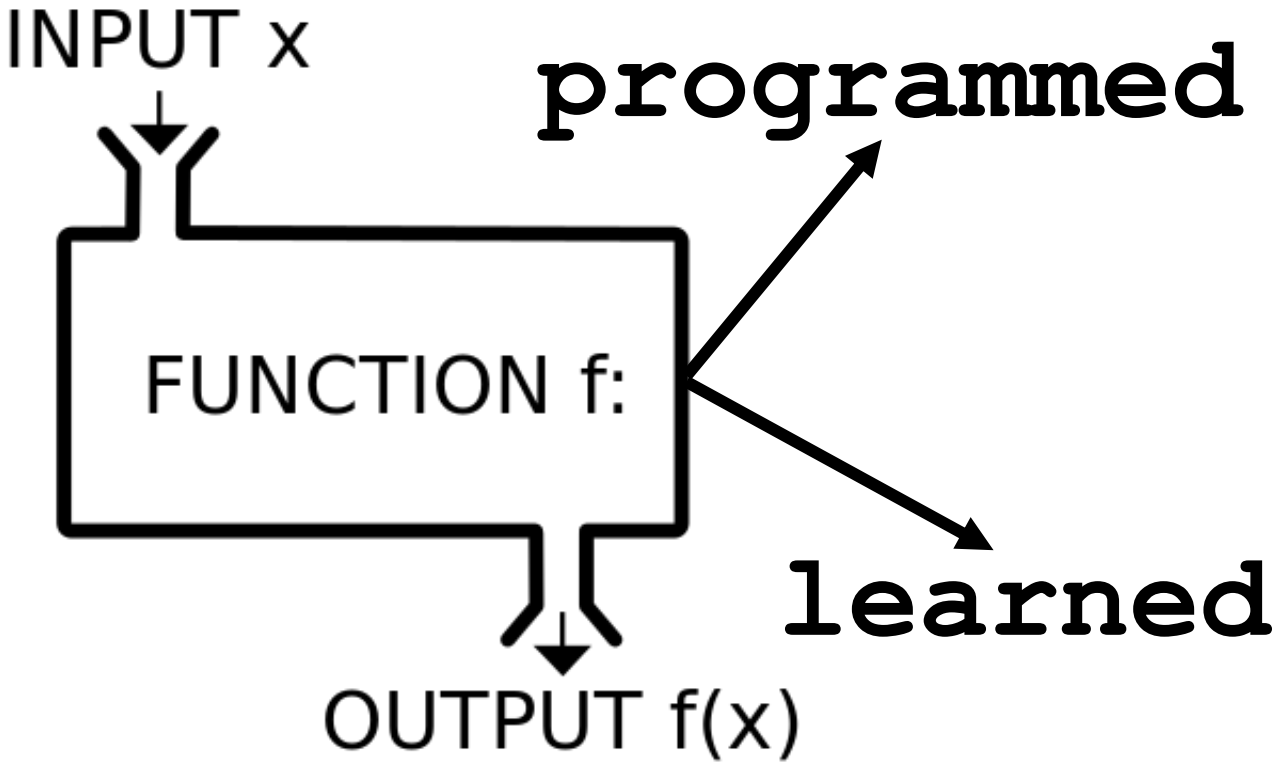


Stage two:
Data Engineering

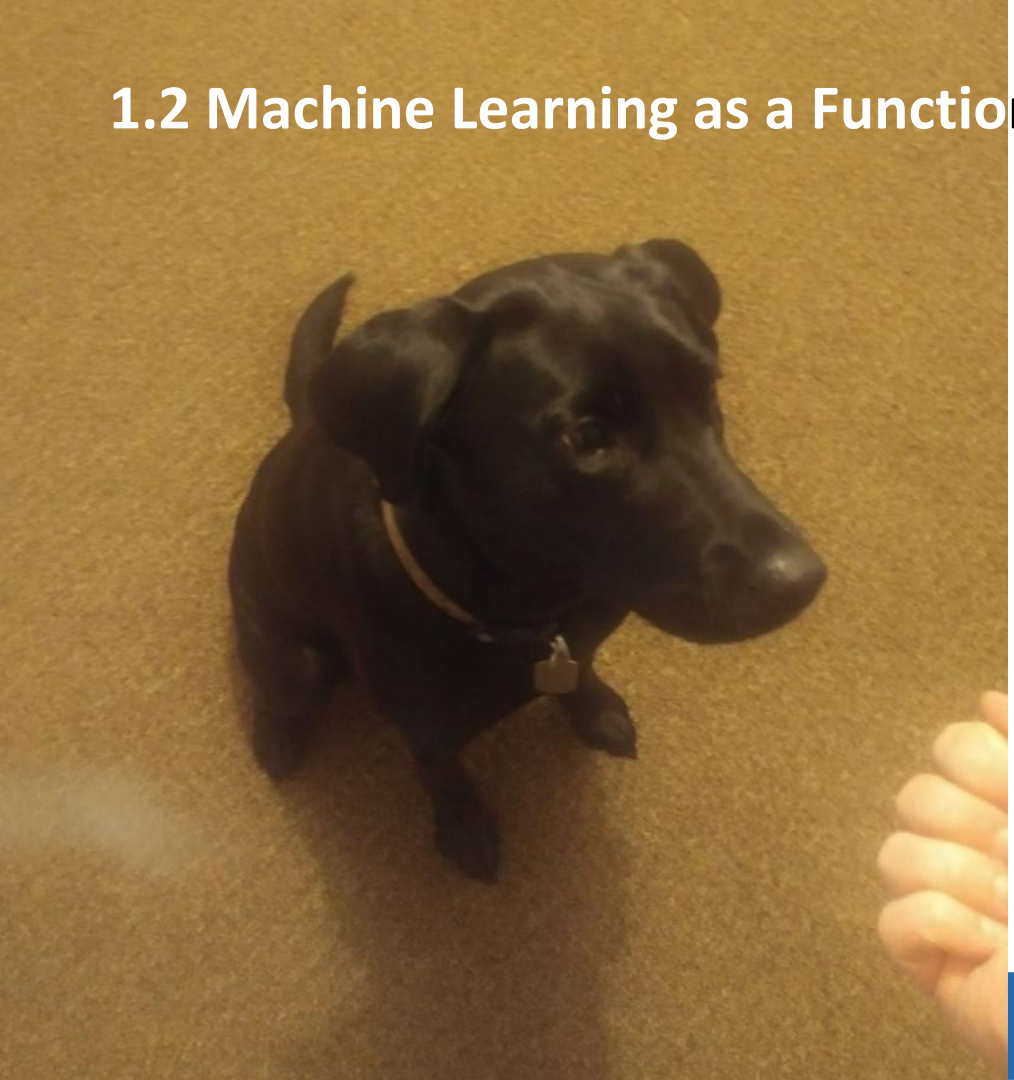
Stage four:
Implementation

Stage three:
Analytical Modelling

1.2 Machine Learning as a Function



1.2 Machine Learning as a Function



“SIT” “PAW”

“LIE DOWN”

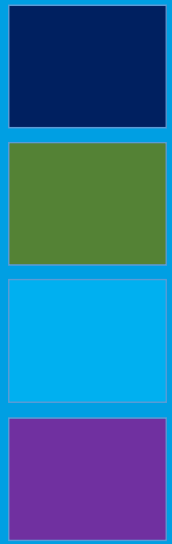
Session Aims

Introduction

Linear Regression

Bias-variance Trade-off & Regularisation

Logistic Regression



2.1 Linear Regression

- One of the most widely used algorithms/models across nearly all scientific fields;
- Models the relationship between a target variable (Y) and some predictor variable(s) ($\mathbb{X}$). We can describe Y as the dependent variable and $\mathbb{X}$ as the independent variable(s) in that changes in Y "depend" on changes in $\mathbb{X}$.

2.1 Linear Regression

- The formal equation is:
  $Y = \alpha + \beta_1 x_1 + \ \ldots + \ \beta_n x_n + \varepsilon$
- Where:
  - Y is the dependent value predicted by the model (the predicted value is denoted by $\hat{Y}$);
  - α is the intercept – the point where the regression line crosses the Y–axis;
  - x_i is the ith independent variable out of n variables in $\mathbb{X}$ (which could be only one variable);
  - β_i is the co-efficient/impact on the slope associated with x_i;
  - ε is the error of the prediction.

2.1 Linear Regression

Marketing	Sales
£0	£100,000
£100,000	£459,691
£200,000	£667,273
£300,000	£1,010,614
£400,000	£1,320,492
£500,000	£1,613,548
£600,000	£1,919,967
£700,000	£2,202,183
£800,000	£2,424,434
£900,000	£2,733,996
£1,000,000	£3,001,957

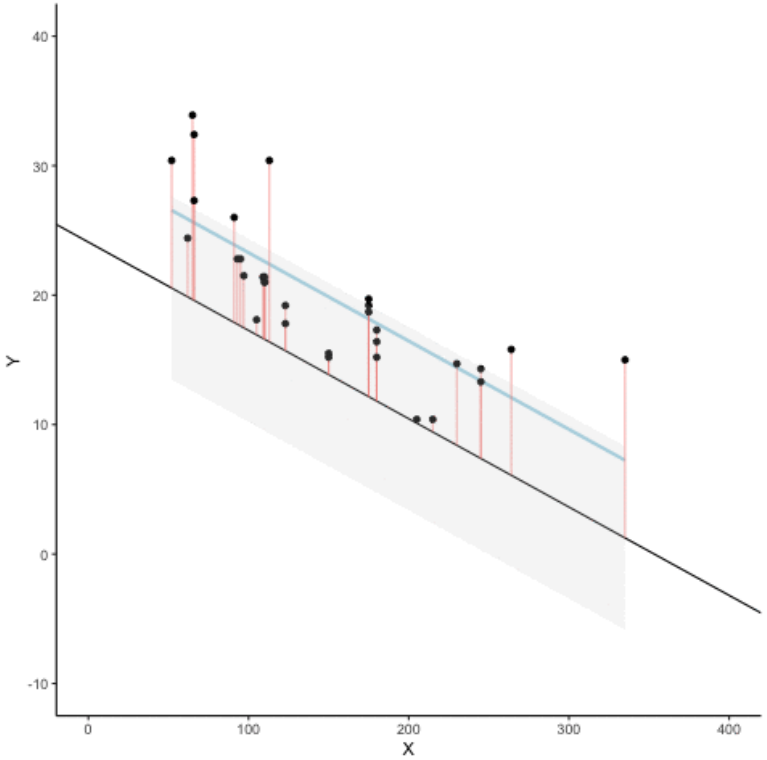
$$Y = \alpha + \beta x + \varepsilon$$

2.1 Linear Regression

Marketing	Sales
£0	£100,000
£100,000	£459,691
£200,000	£667,273
£300,000	£1,010,614
£400,000	£1,320,492
£500,000	£1,613,548
£600,000	£1,919,967
£700,000	£2,202,183
£800,000	£2,424,434
£900,000	£2,733,996
£1,000,000	£3,001,957

$$Y = 100{,}000 + 3x + \varepsilon$$

2.2 Linear Regression (Ordinary Least Squares)



$$minimise \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

Image Credit: https://www.crumplab.com/statistics/13-Gifs.html

2.3 Statistics: Definition

Statistics is the science concerned with developing and studying methods for collecting, analyzing, interpreting and presenting empirical data.

UCI Department of Statistics (2024). *What is Statistics?* https://www.stat.uci.edu/what-is-statistics/

2.4 Machine Learning: Definition

"Machine learning (ML) is the process of using mathematical models of data to help a computer learn without direct instruction"

Microsoft (2023)

https://azure.microsoft.com/en-us/resources/cloud-computing-dictionary/what-is-machine-learning-platform/

2.5 Statistics or Machine Learning?

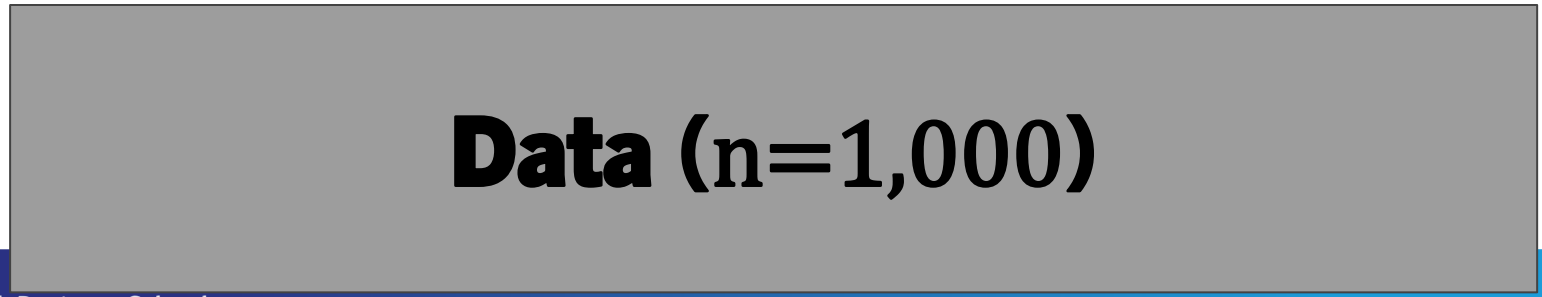
- Is linear regression a statistical approach or a machine learning approach????

3.1 What is Learning?

Image Credit:
Gair Rhydd

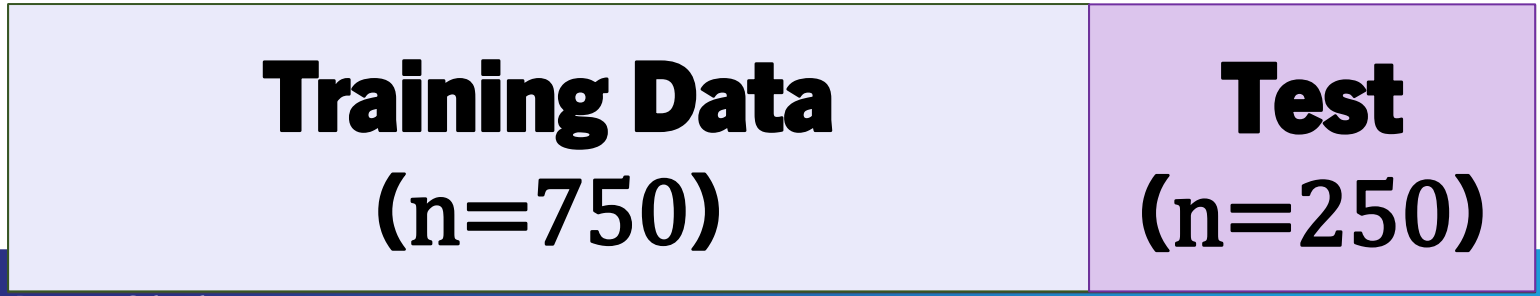
3.2 Generalisability and ML

- To assess the (generalisable) learning of a model we need it to predict data it has not seen.
- I.e. what we really care about (usually) is it's ability to predict future data.
- As a proxy for this we can reserve some data to use exclusively for testing.



3.2 Generalisability and ML

- To assess the (generalisable) learning of a model we need it to predict data it has not seen.
- I.e. what we really care about (usually) is it's ability to predict future data.
- As a proxy for this we can reserve some data to use exclusively for testing.

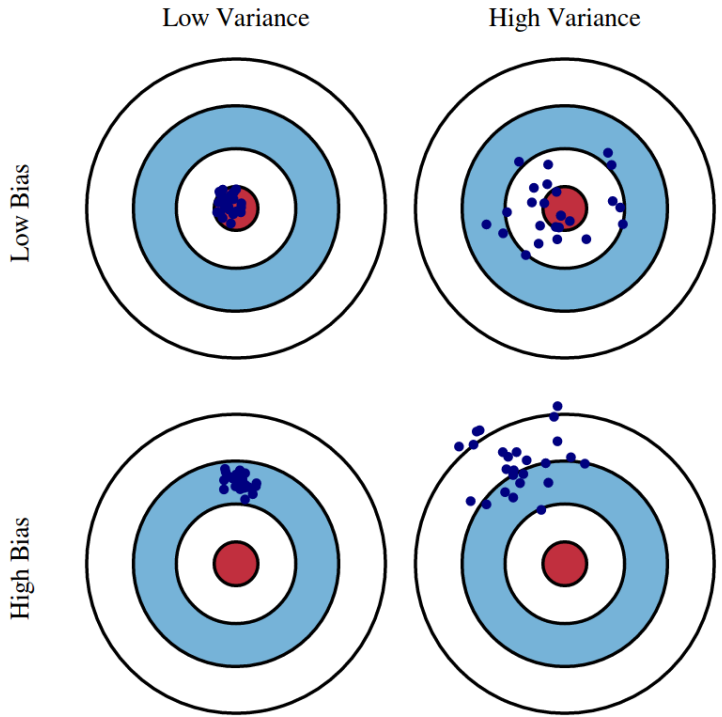


3.3 Bias-Variance Trade-Off

- **Underfitting**: the model fails to capture the underlying pattern in the data (typically the model is too simple).
- **Overfitting**: the model may captures the underlying pattern but also random variations (noise) in the training data not present in test/future data.
- **Ideal fitting**: the model captures the underlying pattern but ignores the random variations (noise).

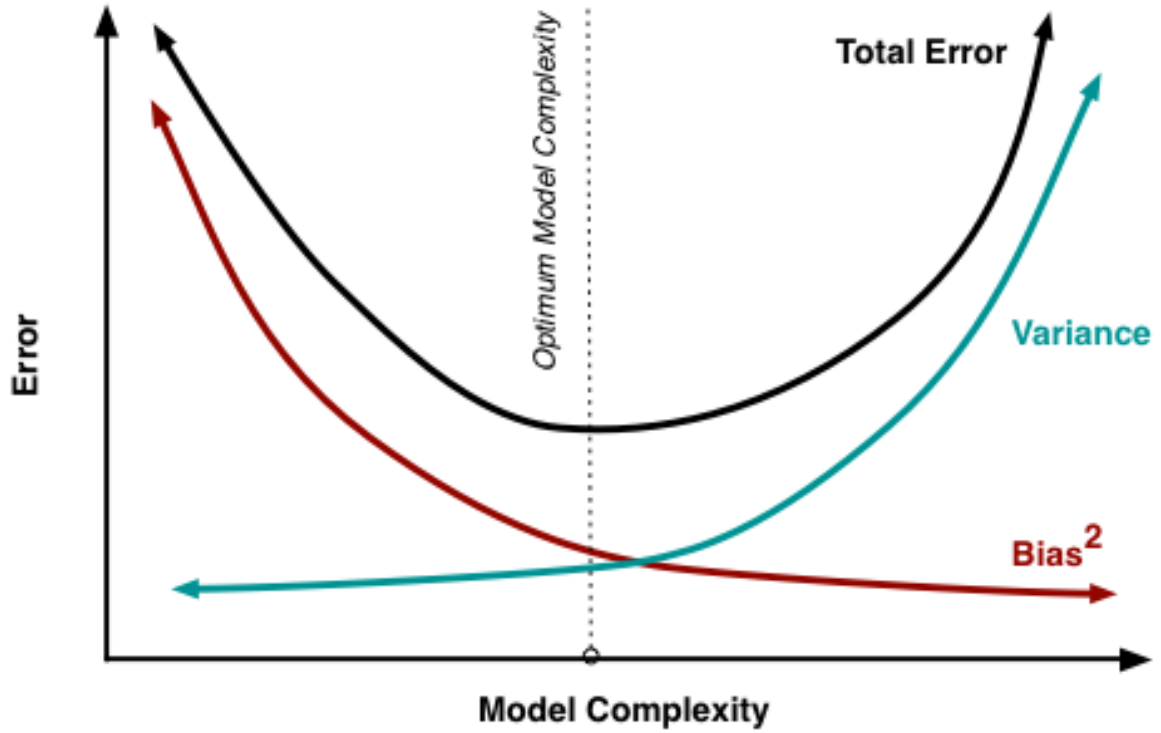
Image Credit: https://commons.wikimedia.org/wiki/Category:The_Three_Bears#/media/File:The_three_bears_pg_14.jpg

3.3 Bias-Variance Trade-Off



https://www.cs.cornell.edu/courses/cs4780/2015fa/web/lecturenotes/lecturenote12.html

3.3 Bias-Variance Trade-Off



https://www.cs.cornell.edu/courses/cs4780/2015fa/web/lecturenotes/lecturenote12.html

3.4 Regularisation

Image Credit: https://medium.com/summit-plus/decision-making-is-killing-organizations-b2bde26d5aa7

3.4 Regularisation

- Vanilla linear regression evaluates only the line that provides the “best fit” – the agreed pattern according to our ‘experts’;
- However, this can model both the underlying pattern in the data, but also the noise (random variations);
- We can modify the algorithm by applying a penalty that forces the algorithm to minimise parameters that may emphasise noise in our data (*regularisation*).

3.4 Regularisation

Facilitator

Policeman

Image Credit: https://medium.com/summit-plus/decision-making-is-killing-organizations-b2bde26d5aa7

3.5 LASSO (L1 Regression)

- LASSO (Least Absolute Shrinkage and Selection Operation) is an example of a modified linear regression that incorporates *regularisation* techniques;

- The approach adds a penalty equivalent to the sum of absolute value of the coefficients (β's):

$$OLS\ objective + \alpha \cdot \sum absolute\ value\ of\ coefficients$$

Where α controls how much we listen to the line of best fit (facilitator) and how much we penalise the coefficients (policeman). A big value of α means all coefficients are zero, a value close to 0 means its normal regression.

3.6 Statistics vs ML - Summary

- Statistical modelling looks for models to represent the data and make inferences about *causality*;
- ML modelling looks for models that can learn from data to perform tasks. Typically that task is predicting unseen data (future data) and models are assessed on their performance on test data;
- We particularly care about avoiding *overfitting* in ML … building models that learn *too* well from the data we have seen, and can't predict data we have not seen.

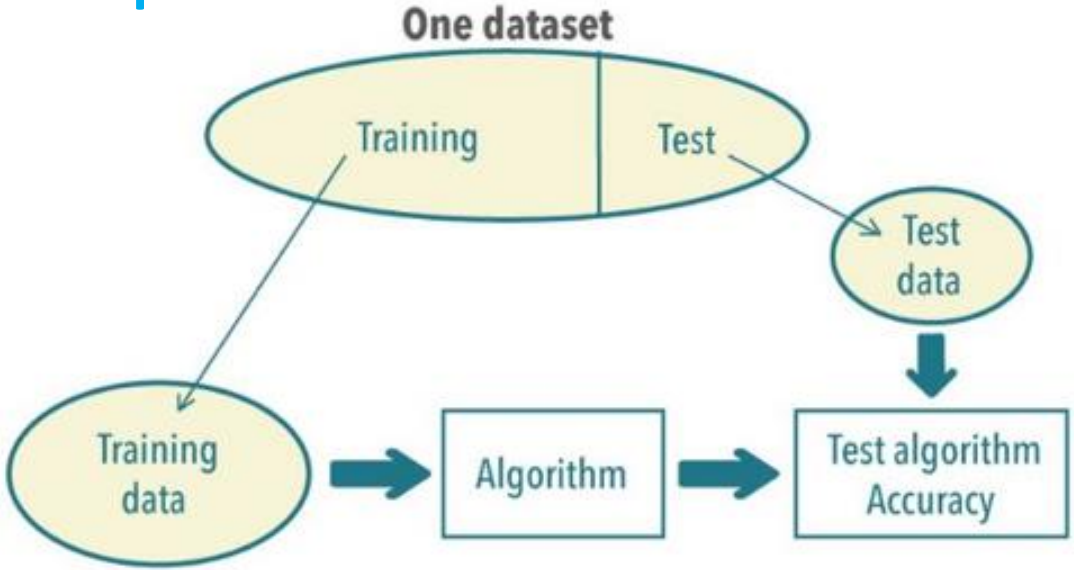
3.6 Statistics vs ML - Summary

- Conceptually we can consider the differences between the linear regression in each case as follows:
  - The statistical approach focuses on building a theoretical model of the data and validates it via hypothesis testing and measurement of model error. We typically optimise the model via a loss/cost function that minimises error;
  - The ML approach focuses on building a model that can predict future data and validates it via performance when predicting unseen data. Typically we will optimise the model via a loss/cost function that minimises error and limits overfitting.

4.1 Labels and Supervision

- Determining the type of machine learning approach we use depends on two factors. The first is the type of data we are predicting (Y):
  - Y is a continuous variable (i.e. a number) – e.g. profit; cost; time; marks earned in in an assessment; etc.
  - Y is a discrete variable (i.e. a category) – e.g. “good” or “bad”; “buy” or “not buy”; “fail”, “pass”, “merit” or “distinction”; etc.

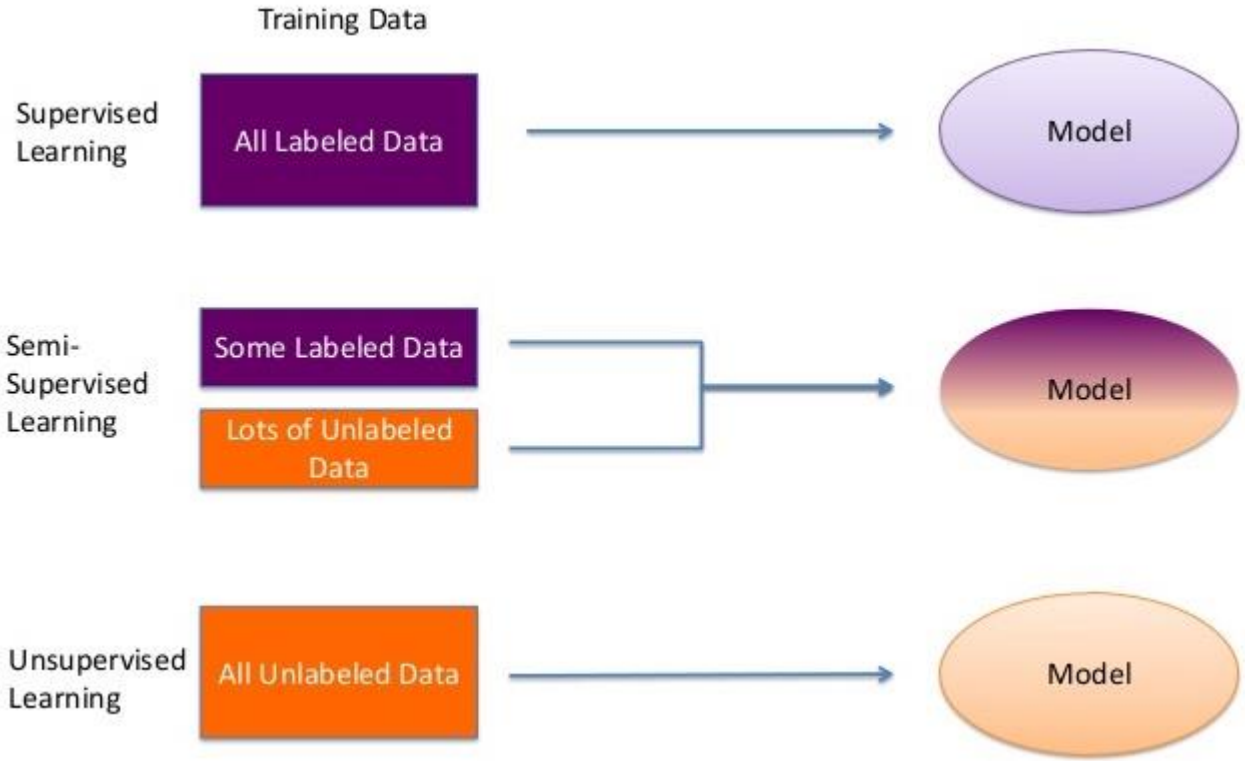
4.2 Labels and Supervision



Supervision refers to having training data that includes the value (label) to be predicted (Y). Learning is performed by minimising the cost/loss when predicting Y

We measure prediction performance by comparing predicted values with actual values in the test data (which was *unseen* during the model's training process)

4.2 Labels and Supervision



4.3 Types of Machine Learning

Regression

- Y is a vector of numerical values (typically continuous data);
- $\mathbb{X}$ is a vector/matrix of features;
- Training is performed on data which has Y values available.

Classification

- Y is a set of discrete values (labels/categories);
- $\mathbb{X}$ is a vector/matrix of features;
- Training is performed on data which has Y values available.

Dimension Reduction

- Y is a vector of continuous values;
- $\mathbb{X}$ is a vector/matrix of features;
- Training is performed on data with **no** Y values available.

Clustering

- Y is a set of discrete values (labels/categories);
- $\mathbb{X}$ is a vector/matrix of features;
- Training is performed on data with **no** Y values available.

4.3 Types of Machine Learning

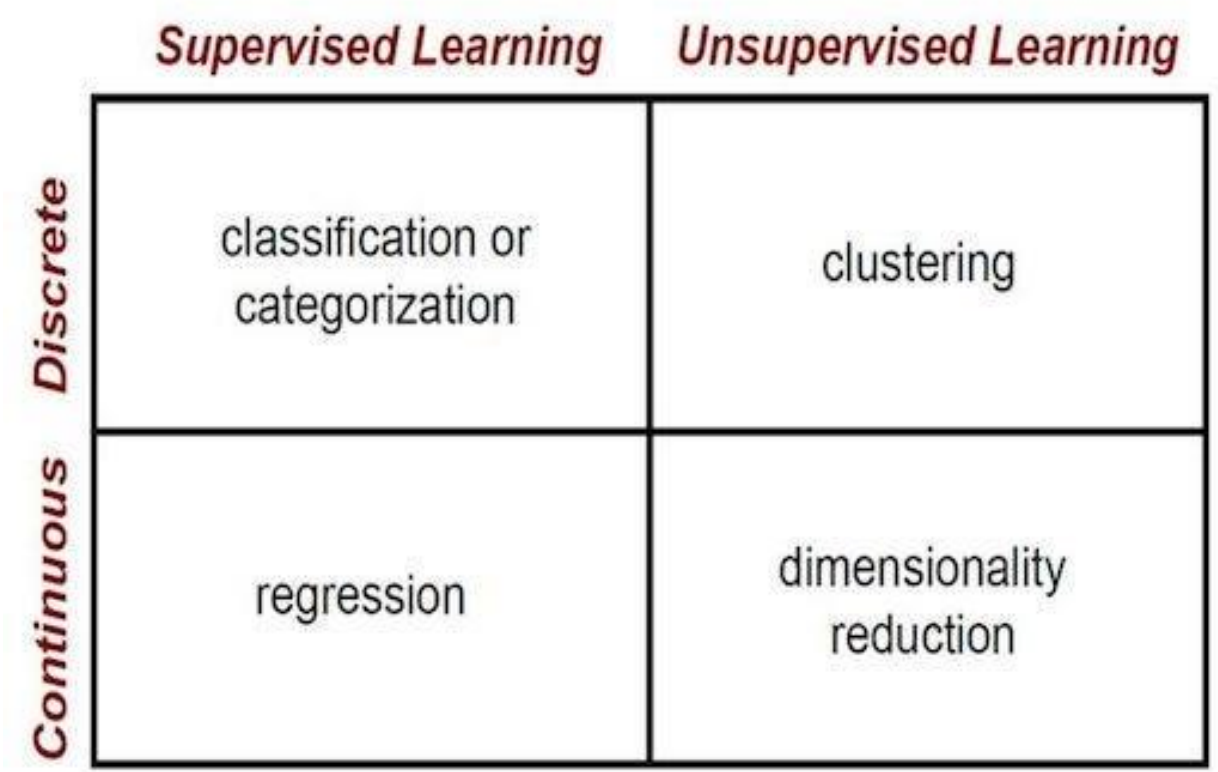


Image Credit: https://towardsdatascience.com/supervised-vs-unsupervised-learning-14f68e32ea8d

4.4 Logistic Regression

Logistic regression is (mostly) used where the Y value is a binary value – i.e. one of two options:

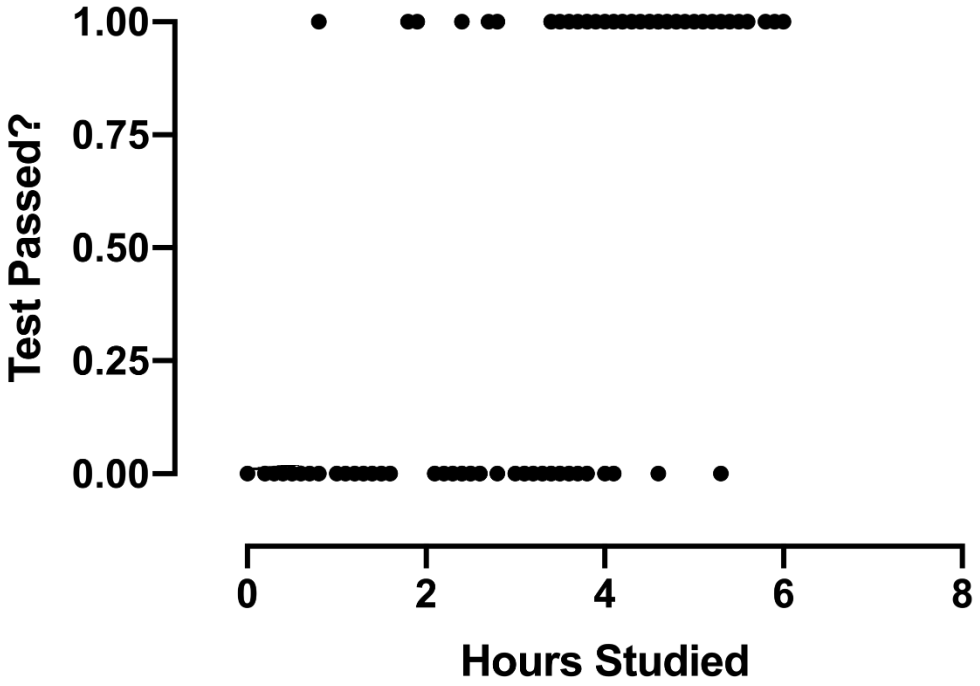
- Success or failure?
- Alive or dead?
- Purchase or not purchase?
- Pass or fail?

The logistic regression equation is written as:

$$\ln\left(\frac{P}{1-P}\right) = \alpha + \beta_1 x_1 + \ \ldots + \beta_n x_n + \varepsilon$$

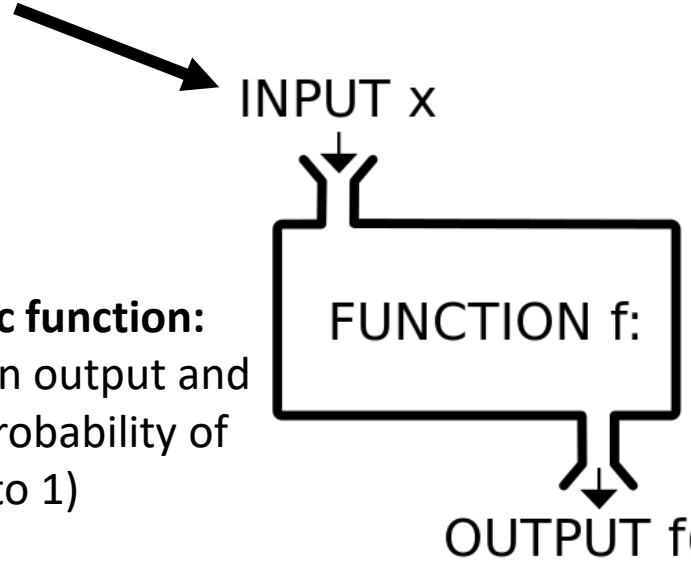
4.5 Logistic Regression by Eye

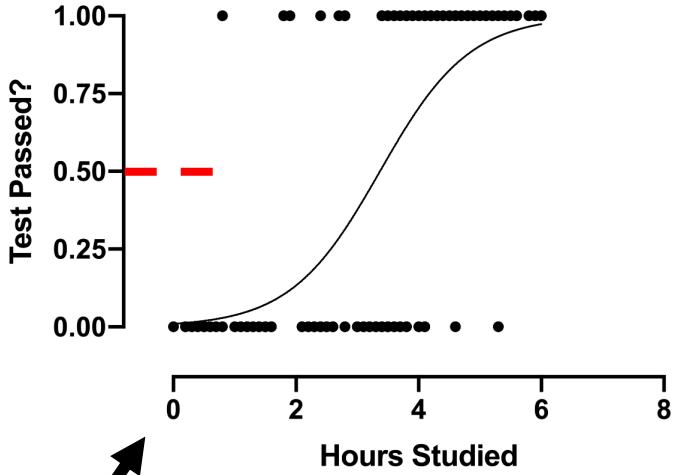
1. Plot the data onto a graph based on outcome (Y);
2. Fit a curved, S-shape line through the data;
3. Find the hours studied for new data and determine its probability based on the location on the Y-axis;
4. E.g. if the hours studied is 6, there is ~95% probability that the data will pass ☺



4.6 Logistic Regression (limited math version)

$$Y = \alpha + \beta_1 x_1 + \ ... + \ \beta_n x_n$$





Sigmoid / logistic function:
Takes the regression output and turns it into the probability of class 1 (0 to 1)

5.1 Summary

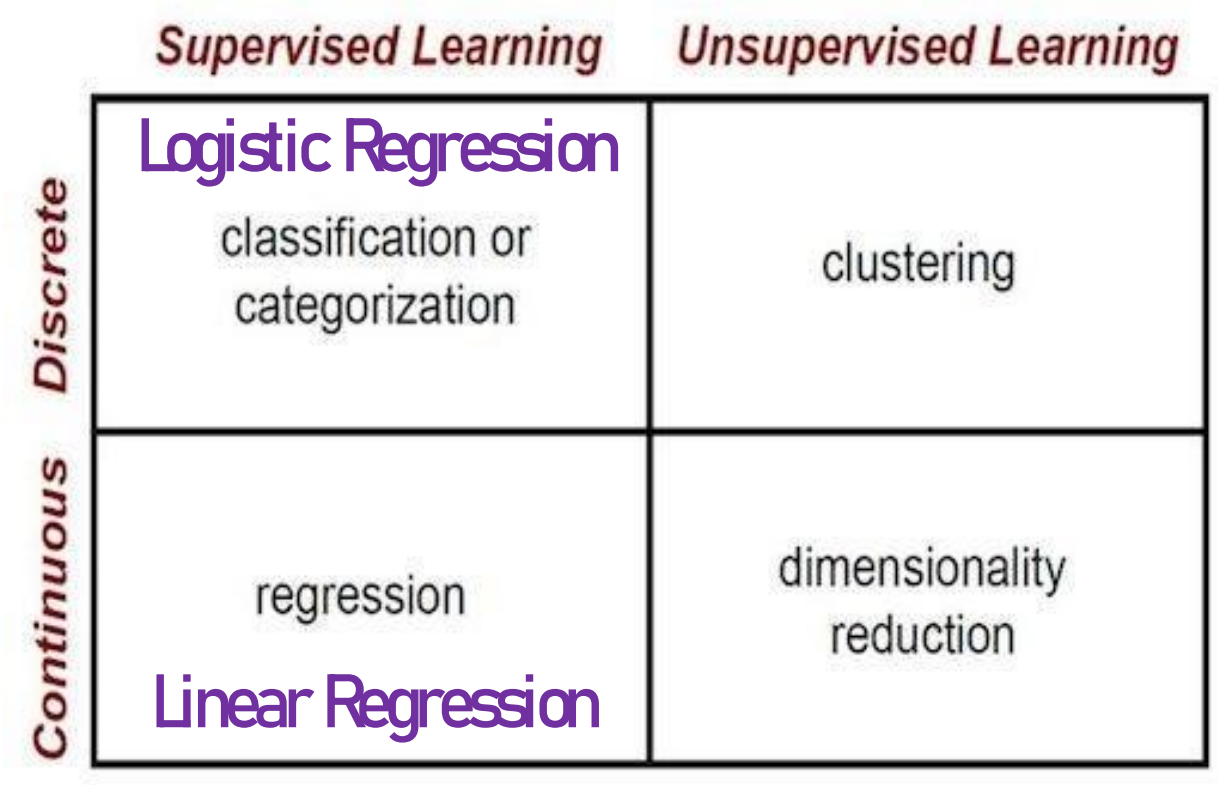


Image Credit: https://towardsdatascience.com/supervised-vs-unsupervised-learning-14f68e32ea8d

5.1 Summary