The basic structure of a data science pipeline

Imperial College London

What do data scientists build anyway?

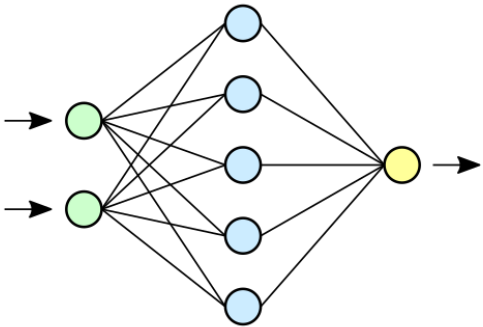


Figure 12: A neural network
https://commons.wikimedia.org/

Figure 13: R2D2, by Lucasfilm StarWars.com Encyclopedia

Imperial College London

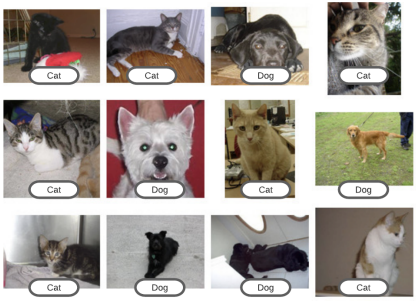
Supervised learning

Is this a cat or a dog?

predict(x) = "cat"

Imperial College London

Supervised learning



model. train(X, y)

Is this a cat or a dog?

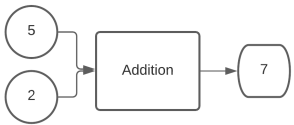
model. predict(x) = "cat"

Imperial College London

Code that writes itself

Algorithm

- A list of prespecified steps that complete a task (recipe).
- A computer program that for any given input produces an output with desired properties.

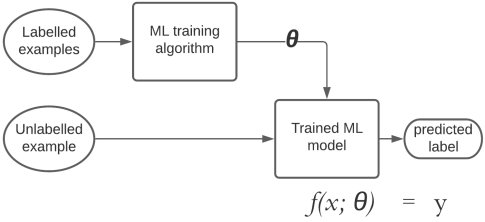


Imperial College London

Code that writes itself

Machine Learning model

- A computer program that gets trained using labelled examples and can then predict the labels of new examples.
- A computer program that programs parts of itself (represented by θ).



Imperial College London

Types of supervised machine learning tasks

Classification

- Predict one label from a finite set (e.g., "cat" vs "dog").
- In probabilistic classification estimate the probability of each label instead (e.g., risk of a heart attack in the next 6 months).

Regression

Predict a continuous quantity based on other characteristics (e.g., predict the yield of a crop based on the type of fertiliser, soil, and weather).

There are also unsupervised learning tasks, like identifying anomalous examples in a dataset, or clustering together examples that "look similar" within making use of labels.

Imperial College London

A data science pipeline

Imperial College London

A data science pipeline

Data collection

Data may reproduce historic biases or not represent the entire population equally well due to selection bias.

Imperial College London

A data science pipeline

Data preprocessing

Seemingly innocuous operations like filling in missing values can exacerbate biases in data.

Imperial College London

A data science pipeline

Problem formulation

Deciding how to turn a real-world problem into a supervised learning problem of the form $f(X) = y$ is an informal context-specific process that introduces risk of bias and harm that cannot easily be rectified later on.

Imperial College London

A data science pipeline

Feature engineering

Extracting features that are likely useful from unstructured data makes assumptions about the process that might be biased. Modern techniques rely on automation and further machine learning layers to make this task less manual, but this increases the dependence on bias-free data.

Imperial College London

A data science pipeline

Training

Training algorithms are programs that optimise some function like accuracy, but might be optimising the wrong. For example, an algorithm that is optimised to predict mortality risk from a virus will not have the incentive to do very well for patients from very rare diseases.

Imperial College London

A data science pipeline

Deployment

Even if the model is ethically designed, how it is actually deployed in practice might matter significantly. For example, if smartphone apps that predict heart failure become popular with doctors, the quality of healthcare of patients who cannot afford them might drop.

Imperial College London

A data science pipeline

Monitoring

Training data are almost never exactly representative of the real-world setting in which the model operates – this is known as "dataset shift". Monitoring the model post-deployment and triggering model reviews when there is indication that the performance is dropping is one step in the right direction.

Imperial College London

One example: health costs as a proxy for need

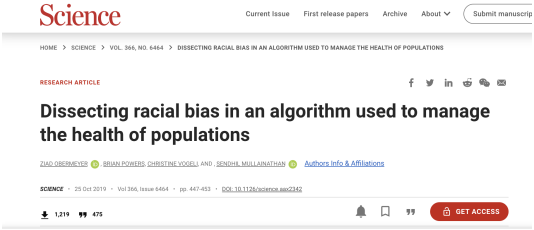


Figure 14: "Bias occurs because the algorithm uses health costs as a proxy for health needs. Less money is spent on Black patients who have the same level of need, and the algorithm thus falsely concludes that Black patients are healthier than equally sick White patients." https://www.science.org/doi/abs/10.1126/science.aax2342

Imperial College London

Types of bias

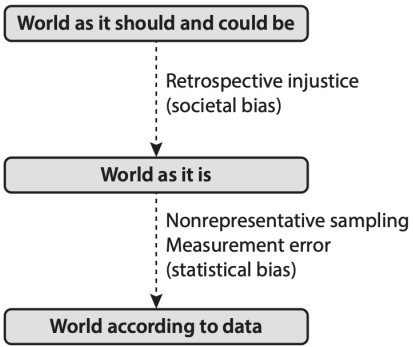


Figure 15: A distinction made between societal and statistical bias in Mitchell et al., 2021

Imperial College London

Summary

- Machine learning involves two algorithms: a training algo that takes as input a historical dataset; and a model which is "configured" using the output of the training algorithm and then used to predict the label of a new example.
- A data scientist produces a multi-step pipeline, not just a model.
- Pipelines have several stages, including data collection, feature engineering, training and monitoring. Each stage has its own risks from an ethical perspective.

Next, we will review codes of conduct and regulation.