



Introduction to Machine Learning

John Armstrong and Chris Boyle



CMAC
FUTURE MANUFACTURING
RESEARCH HUB

The background is a dark blue gradient. In the top right corner, there is a glowing blue capsule that is open, with a bright light and a cluster of glowing blue molecular structures emerging from it. Faint, larger molecular structures are visible in the upper left corner.

What is Machine Learning?

What is Machine Learning?

- Machine learning (ML) is the process of using statistical techniques to give computers the ability to perform a specific task *without* being explicitly told how to do the task
 - this is done through by formulating a model based on the data rather than using a mathematical model that can be explicitly written down – often referred to as *data-driven modelling*
 - ML fits the model to the data, while statistics fits the data to the model!
 - e.g. linear regression – the intercept and gradient are not explicitly given by the user but the computer finds the optimal value of this based on the data

Data-driven Models

- Data-driven models are very sensitive to the quality AND quantity of data that we have access to – a model trained on not enough data will learn a solution to a problem but may not be the desired solution and a model trained on poor quality data is never going to model what we want
- The following example will demonstrate how sensitive a data-driven model is to data quality/quantity as we will be using you as the machine to learn the solution

Data-driven Models

| |
|---|
| 4 |
| A |

| |
|----|
| 13 |
| B |

| |
|---|
| 6 |
| A |

| |
|----|
| 11 |
| B |

| |
|---|
| 8 |
| A |

| |
|---|
| 7 |
| ? |

Data-driven Models

| |
|---|
| 4 |
| A |

| |
|----|
| 13 |
| B |

| |
|---|
| 6 |
| A |

| |
|----|
| 11 |
| B |

| |
|---|
| 8 |
| A |

| |
|---|
| 7 |
| ? |

- Odd implies B
- < 10 implies A

Data-driven Models

| | | | | | |
|---|----|---|----|---|---|
| 4 | 13 | 6 | 11 | 8 | 2 |
| A | B | A | B | A | B |
| 7 | | | | | |
| ? | | | | | |

Data-driven Models

| |
|---|
| 4 |
| A |

| |
|----|
| 13 |
| B |

| |
|---|
| 6 |
| A |

| |
|----|
| 11 |
| B |

| |
|---|
| 8 |
| A |

| |
|---|
| 2 |
| B |

| |
|---|
| 7 |
| ? |

Prime numbers!

ML Fundamentals



What is Machine Learning Used For?

DOI: [10.1039/D2DD00033D](https://doi.org/10.1039/D2DD00033D) (Paper) *Digital Discovery*, 2022, Advance Article

Data mining crystallization kinetics[†]

Diego A. Maldonado , Antony Vassileiou , Blair Johnston , Alastair J. Florence and Cameron J. Brown *


EPSRC Future Manufacturing Research Hub for Continuous Manufacturing and Advanced Crystallisation (CMAC), University of Strathclyde, Technology and Innovation Centre, 99 George Street, Glasgow G1 1RD, UK. E-mail: cameron.brown.100@strath.ac.uk



Received 13th April 2022, Accepted 25th July 2022

First published on 1st August 2022

Control of Batch and Continuous Crystallization Processes using Reinforcement Learning

Brahim Benyahia ^a , Paul Danny Anandan ^a, Chris Rielly ^a

Show more 



+ Add to Mendeley  Share  Cite

<https://doi.org/>

A micro-XRT image analysis and machine learning methodology for the characterisation of multi-particulate capsule formulations

Frederik J.S. Doerr ^{a, b}, Alastair J. Florence ^{a, b} 

Show more 


+ Add to Mendeley  Share  Cite



Using Machine Learning to Predict Residence Time Distributions in Coiled Flow Inverter (CFI) Reactors

Maria Cecilia Barrera, Aleksander Josifovic, John Robertson, Blair Johnston, Cameron Brown, Alastair Florence.

Future Continuous Manufacturing and Advanced Crystallisation Research Hub, University of Strathclyde, Glasgow, UK.

 maria.barrera@strath.ac.uk

Machine Learning Workflows to Predict Crystallisability, Glass Forming Ability, mechanical properties of Small Organic Compounds.



ARTICULAR
MEDICINES MADE SMARTER

Vijay K. Srirambhatla,^{*1} Blair Johnston^{1,2}, Alastair Florence^{1,2}

¹EPSRC ARTICULAR, University of Strathclyde, Technology Innovation Centre, 99 George Street, Glasgow, G1 1RD, UK

²EPSRC CMAC Future Manufacturing Research Hub, University of Strathclyde, Glasgow, G1 1RD, U.K.

E-mail: Vijay.Srirambhatla@strath.ac.uk



ARTICULAR
MEDICINES MADE SMARTER

A Unified AI Framework for Solubility Prediction Across Organic Solvents

Antony D. Vassileiou^a, Murray N. Robertson^b, Bruce G. Wareham^c, Mithushan Soundaranathan^c, Sara Ottoboni^b, Alastair J. Florence^b, Thoralf Hartwig^d, Blair F. Johnston^{a,b,e}

Prediction of powder flow of pharmaceutical materials using machine learning

Laura Pereira Diaz, Cameron J. Brown, Alastair J. Florence

laura.pereira-diaz@strath.ac.uk

EPSRC CMAC Future Manufacturing Research Hub; Strathclyde Institute of Pharmacy & Biomedical Sciences

Prediction of mefenamic acid crystal shape by random forest classification

Siya Nakapraves^{1,2}, Monika Warzecha^{1,2}, Chantal Mustoe^{1,2}, Alastair J. Florence²

siya.nakapraves@strath.ac.uk

¹ Strathclyde Institute of Pharmacy and Biomedical Science (SIPBS), University of Strathclyde, UK

² EPSRC Future Manufacturing Research Hub for Continuous Manufacturing and Advanced Crystallisation (CMAC), University of Strathclyde, Technology and Innovation Centre, UK

Different Types of Machine Learning

- **Supervised vs. Unsupervised:**

- Supervised learning methods are used when the output of the problem is known
- e.g. determining whether an image contains a cat or a dog: each image is marked as either cat or dog and the algorithm learns by trying to estimate the correct animal
- Supervised learning is applicable in classification and regression tasks
- In supervised learning tasks, the algorithm learns by comparing its predictions with the known output and changing its internal parameters to match its predictions to the correct answer
- The comparison is typically through the mean squared error (MSE) function

Different Types of Machine Learning

- **Supervised vs. Unsupervised:**

- Unsupervised learning allows the computer to uncover patterns and correlations within the data without a particular constraint on what it is looking for
- This is applicable for data clustering or dimensionality reduction
- In unsupervised learning, the algorithm learns by making correlations based on the data provided to it e.g. PCA will calculate the principal components based on the provided data, it is then up to the user to look at those principal components and their explained variance to decide how many to keep.

Different Types of Machine Learning

- **Classical vs. Deep:**

- Classical ML are algorithms which do not involve the use of deep neural networks (with deep ML being those which do).
- Classical algorithms are good for descriptor-based data
- Deep algorithms are good for really high dimensional data such as images and videos

Machine Learning Algorithms

| | Supervised | Unsupervised |
|-----------|--|--|
| Classical | Decision Trees, Support Vector Machines, k-nearest neighbours | k-means, Agglomerative Clustering, PCA |
| Deep | Deep neural networks, CNNs, ResNets | GANs, VAEs, INNs, Normalising Flows |

Preparing Data for ML

- 90%* of doing “data science” is data manipulation
- This was seen this morning when combining and reducing our dataset into the columns we were interested in
- This is because applying an ML algorithm is easy, the hard part is making sure your data poses the problem you want your algorithm to learn
- However, what we done this morning was only really half* of what is needed to get data ready for ML

* These are not exact calculations, isn't hyperbole fun?

Preparing Data for ML

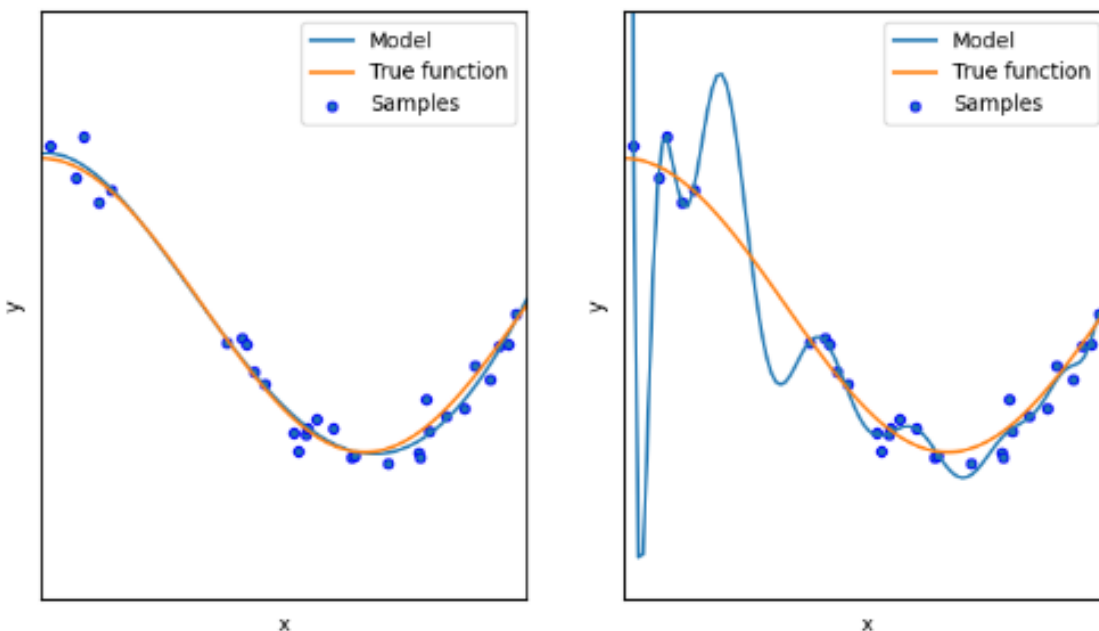
- **Train/Validation Split:**

- One of the key aims of training an ML model is its ability to *generalise* – that is, give robust predictions on data it has never seen before
- To test how well our model generalises we can separate a portion of the data we have at random and not include it in the training of the algorithm
- This then provides us with a small subset of data that the answer is known for but the algorithm has never seen before
- The data used for training the algorithm is known as the **training dataset** and the data used for testing generalisation is known as the **validation dataset**
- The validation can be anywhere from 10-20% of the data

Preparing Data for ML

- **Train/Validation Split:**

- Doing a train/validation split also helps avoid the model *overfitting*
- Overfitting occurs when the algorithm “memorises” the training data
- Essentially, the model predicts the training data very accurately but performs badly on the validation dataset



Source: https://scikit-learn.org/stable/auto_examples/model_selection/plot_underfitting_overfitting.html

Preparing Data for ML

- **Standardisation:**

- Most ML algorithms assume the input data given to it are drawn from a unit Gaussian distribution
- Given that for real data this is not the case, the data being passed to your ML algorithm of choice must be *standardised*
- This just means that each feature in the dataset has the mean of the feature subtracted and is divided by the standard deviation e.g.

$$\hat{x} = \frac{x - \mu_x}{\sigma_x}$$

Choosing an ML Algorithm

- You've narrowed down what type of machine learning you're working with (supervised vs. unsupervised) now how to do you decide which algorithm to use?
- How much time have you got?
 - If the answer is lots, then try a whole bunch and see which performs the best, then make sure you understand how the best performing model works
 - If the answers is not enough, try a couple that make sense to you and see which works best
- What kind of data do you have?
 - Tables of numbers, experimental data – start with classical methods and if they're not performing well, try deep learning
 - Images, videos – deep learning all the way

Algorithm Performance

- How do we know if an algorithm is learning?
 - Each algorithm will have its own way of monitoring performance based on the formulation of the error in the model e.g. MSE
 - If this error is decreasing for the training data in subsequent iterations we can assume the model is learning
 - If this error is decreasing for the validation data in subsequent iterations we can assume the model is generalising
 - If the error in the validation begins to increase or oscillate over multiple iterations then the model is overfitting
 - The training error may still decrease after the model has overfitted

Now Over to You! (yay?)

- The following exercises are outlined in the “machine_learning.ipynb” IPython notebook
- In this notebook you will further prepare data for ML purposes by splitting into training and validation and performing standardisation
- Two algorithms – random forest and neural network – will then be used to learn how to predict solubility
- Analysis of model performance will follow