MMLDT-CSET 2021

Short Course 2

Mechanistic Machine Learning for Engineering and Applied Science

(1) Introduction to Machine Learning

Instructors: Prof. J.S. Chen, Xiaolong He, Kristen Susuki (UC San Diego)

1 What is Machine Learning?

Simply put, it is the science of computer programming so that they can learn from data *without* explicit programming of rules to learn.

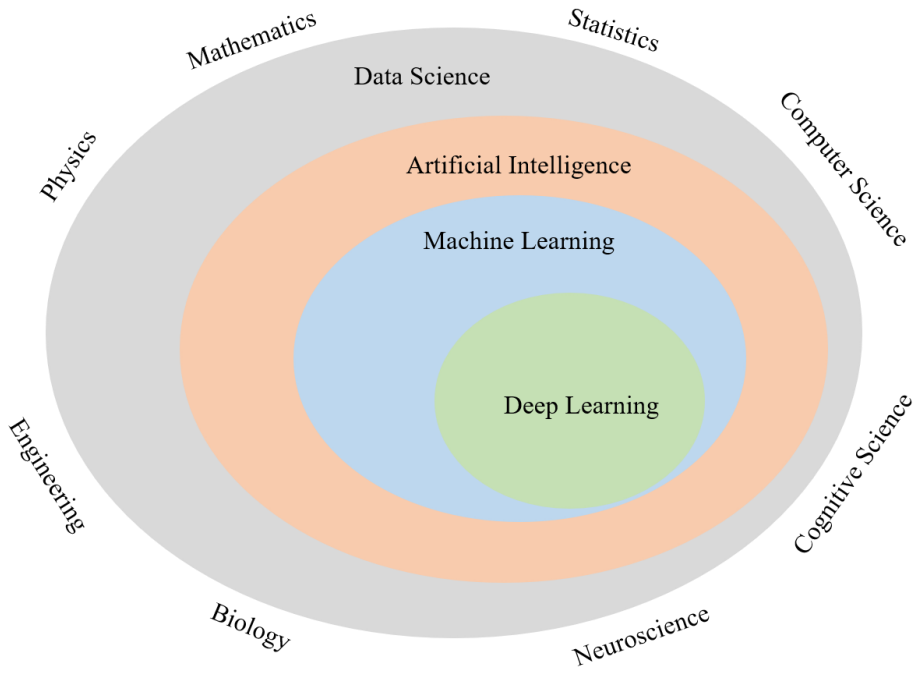

Figure 1: A multidisciplinary field.

2 How do Machines Learn?

Somewhat like humans, machines learn through experience E w.r.t some task T and their learning ability is measured through a performance measure P.

If its performance on T as measured by P improves with experience E and reaches a satisfactory level, we can say that the machine has learnt to perform task T through *training*. It is then that we can feed real world data to it and expect reliable outputs on general data.

In machine learning terminology, *E* usually means a data set containing data samples which the machine learning algorithm goes through to learn the task *T* and performance measure *P* is the accuracy of the algorithm's learning capability.

Table 1: Examples of Machine learning tasks, data sets and performance measures.

Task (*T*)	**Data set for Experience (*E*)**	**Performance Measure (P) × 100**
Recognise spam from regular email	User-flagged spam emails among regular emails in Inbox	$\frac{\text{Correctly classified emails}}{\text{Total emails}}$
Identify faulty composite panels	Images of faulty and healthy composite panels	$\frac{\text{Correctly classified samples}}{\text{Total samples}}$
Estimate pollution levels today	All previous pollution level, temperature, humidity etc.	$\frac{\text{Predicted value} - \text{True value}}{\text{True Value}}$

The workflow for a generic learning process is as shown in the **Error! Reference source not found.**

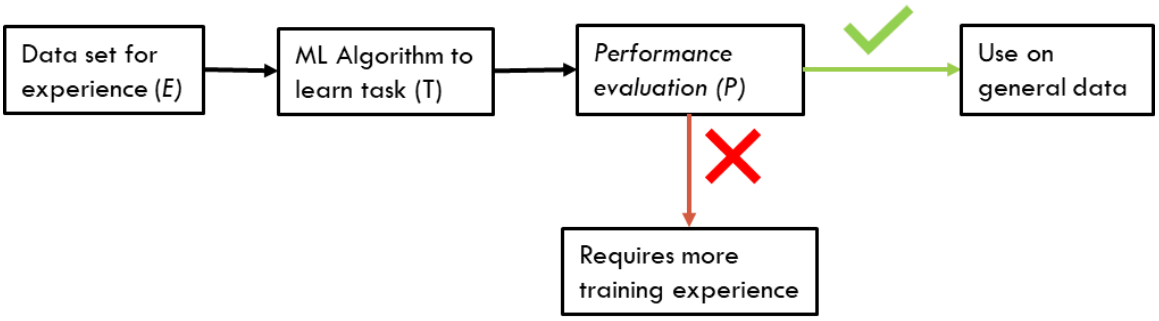

Figure 2: Workflow of a generic machine learning process.

3 Types of Machine Learning (ML)

There are a few categories that computer scientists use but broadly, the basic types we are going to study fall under two categories.

3.1 Supervised Learning

Each input data has an associated target output, which is called "labelled" data. The label of the data is used to measure the accuracy of prediction w.r.t. target output, which provides "supervision" to the learning process. Supervised learning tasks often aim to predict future outcome (prediction) or infer the relationships between input and output (inference). There are two main types of supervised learning:

3.1.1 Classification

Classification problems are to classify input data into target output categories or classes. Input could be qualitative or categorical for e.g. choice of cereal. The output variables can also be qualitative or categorical, e.g., person's gender. In Figure 3, the algorithm tries to classify if a steel joint sample shown in the image is a cracked or an un-cracked sample.

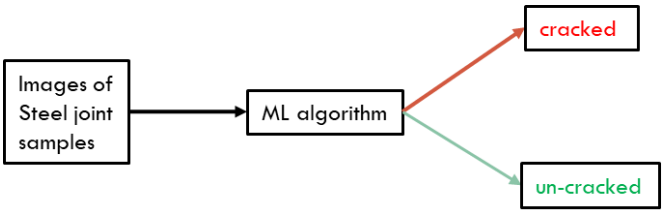

Figure 3: Example of a classification problem to identify cracked and un-cracked samples.

3.1.2 Regression

Regression problems are to predict numeric output quantity from given input features. The output variable is quantitative, e.g., person's age. In Figure 4, the algorithm tries to predict the failure strength of sample of glass when it is given some parameters of the experiment performed on the sample (magnitude of loading, rate of application of the load) and data about the composition of the material (% of sand, limestone and sodium carbonate).

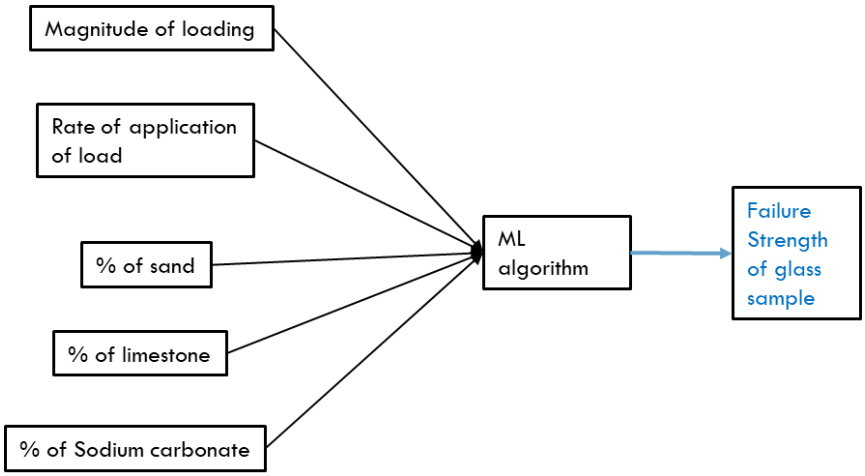

Figure 4: Example of regression problem to obtain strength of a glass sample.

3.2 Unsupervised Learning

Data for unsupervised learning does not has associated target output, which is called "unlabelled" data. Thus, there is no "supervision" or feedback during the learning process. It is usually applied to find hidden structure in data. Following shows two examples.

3.2.1 Clustering

Model learns to cluster data samples according to a user-defined number of clusters. Let's take the data set of the Iris flower for example. The goal of clustering is to segregate the data samples of the length and width of the Sepal of the Iris flower. This is done by agglomerating the data set into *k* groups by finding out *k* mean centres within the data set and segregating samples based on their distance from those *k* mean centres. This is shown in Figure 5 where the data set is clustered into 3 clusters shown by the red, blue, and green coloured dots. This type of clustering is called *k*-means clustering and there are many other types of clustering techniques available.

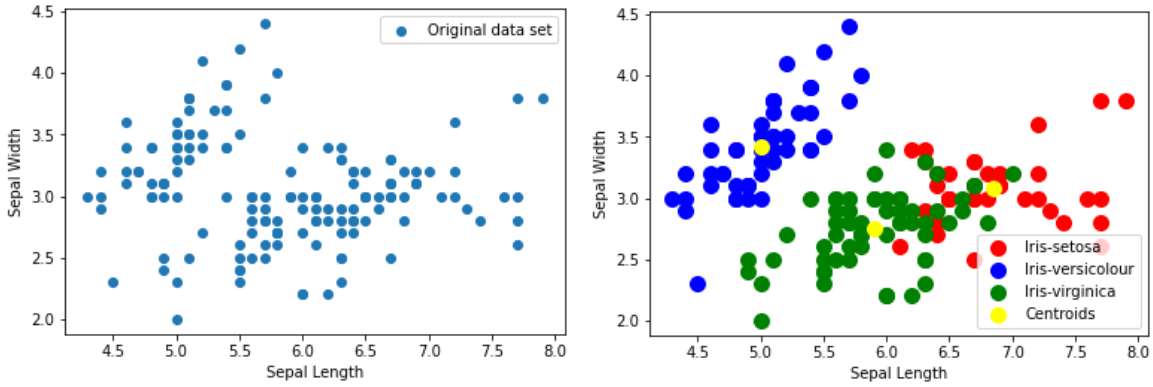

Figure 5: Clustering of flower data based on length and width of Sepal.

3.2.2 Feature Extraction/Dimensionality Reduction

Model extracts features by combining other features to more relevant features for better understanding of the data. In Figure 6, the algorithm tries to extract a new low-dimensional representative feature for wear and tear of heat resistant tiles on a space shuttle based on data regarding the missions the space shuttle has undergone, duration, what range of temperature it experienced, etc.

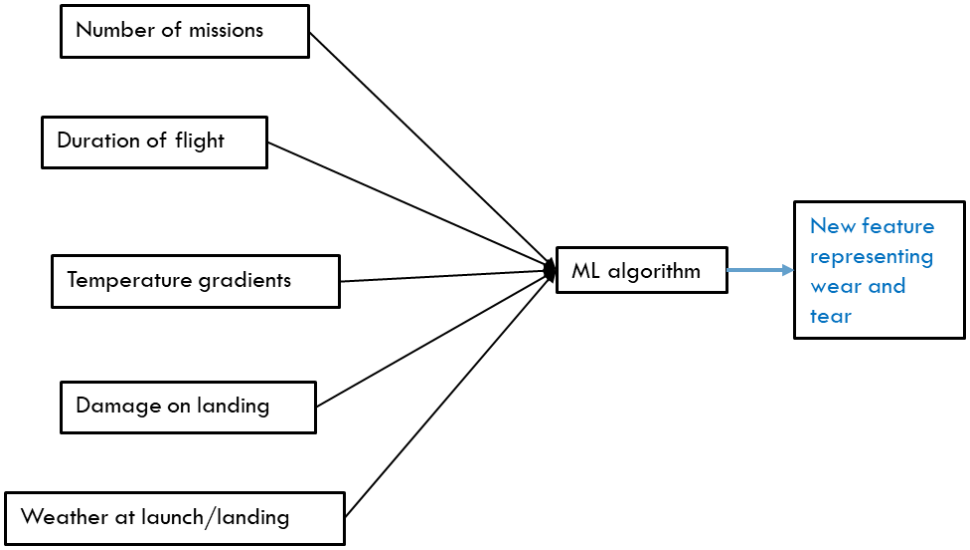

Figure 6: Feature extraction to represent wear and tear of heat resistant tiles on space shuttles.

3.3 Quick Overview of ML Algorithms

Most of statistical tools are ML algorithms. Following lists some of the popular ML algorithms for Supervised or Unsupervised learning. Some algorithms, e.g., Support Vector Machines and Artificial Neural Networks, can be used in both regression and classification problems. The K-means clustering algorithm is classified as an unsupervised learning technique. It can also be viewed as a technique for classification as it clusters data into different distinct classes/groups. For this course, we will focus on Supervised learning. The algorithms coloured in red will be covered in the following lectures.

- Supervised Learning:
  - Regression:
    - Least-Square Regression, Ridge Regression, LASSO, Sparse Regression
    - Support Vector Machines, Artificial Neural Networks
  - Classification:
    - Bayes Classifier, K-Nearest Neighbours, Logistic Regression, Linear Discriminant Analysis, Decision Tree, Random Forest
    - Support Vector Machines, Artificial Neural Networks
- Unsupervised Learning:
  - Clustering:
    - K-Means clustering, Mixture Models, Hidden Markov Model
  - Dimensionality Reduction / Manifold Learning:
    - Principal Component Analysis (PCA), Kernel-PCA, Auto-encoders,

4. Types of Data Sets used in Supervised Learning

4.1 Training Data

Training data contains the information and knowledge about desired tasks and is used to train ML algorithms. It is important to ensure a good quality of training data and that they can well represent the information and knowledge we expect the ML algorithms to learn since they are the only sources from which ML algorithms can gain experience and knowledge about desired tasks.

4.2 Validation Data

ML algorithms contains a lot of parameters (hyperparameters) responsible for its learning capabilities. To choose between models with varying hyperparameters, the algorithms need to be evaluated on data it has not seen yet. This data set is called *validation dataset*. It is usually obtained by keeping a part of the training data set as validation data before the start of the training procedure.

4.3 Test Data

Once final trained model has been chosen, model performance is evaluated on a *test data* set. This gives us the model testing (generalization) accuracy, i.e., if testing accuracy is 97%, when we feed it general real-world data, the model is expected to be accurate 97% of the time.

5 Main Challenges of ML

Below is a brief summary about main challenges of ML. More details will be discussed throughout the future lectures.

5.1 Size of training data

In general, the more complex the problem, the more samples are needed to train the ML algorithm to do its task. However, it is sometimes difficult or impossible to obtain a large amount of data due to many factors, e.g., time and cost.

Table 2: Examples of tasks and data set size used in ML algorithms.

Task	Number of samples in data set
Optical Character Recognition (OCR)	10^4
Clustering users for advertisement data	10^9

5.2 Non-representative data set

For good performance on real-world data (generalization), training data set must be representative of the new cases. For example, if model learns to classify input into 3 target categories but the real-world data sample belongs to none of those categories, a good answer should not be expected from the model.

5.3 Poor quality of training data

Real-world data often comes with noise and errors. Data may be too noisy, may contain many outliers or may have missing values/labels. Hence, data set should be cleaned up (pre-processing) by removing or taking some representative values of the suspect data before being fed to the algorithm for training.

5.4 Overfitting training data

Overfitting occurs when model performs well on training data set but does not give a good performance on the testing data set. Simply put, it means that the model has learned “too much”, picking up the patterns in between the individual samples rather than learning the general trend. This can be seen in Figure 7, where the model was trained to classify red and blue points in the

training data set. To overcome, one should try:
i. Simplifying the model and reducing the parameters/features.
ii. Increasing training data set size
iii. Reducing noise in data sets
iv. Resampling methods (will be covered later in this course)

Finding an optimal model with sufficient complexity or learning capabilities is challenging.

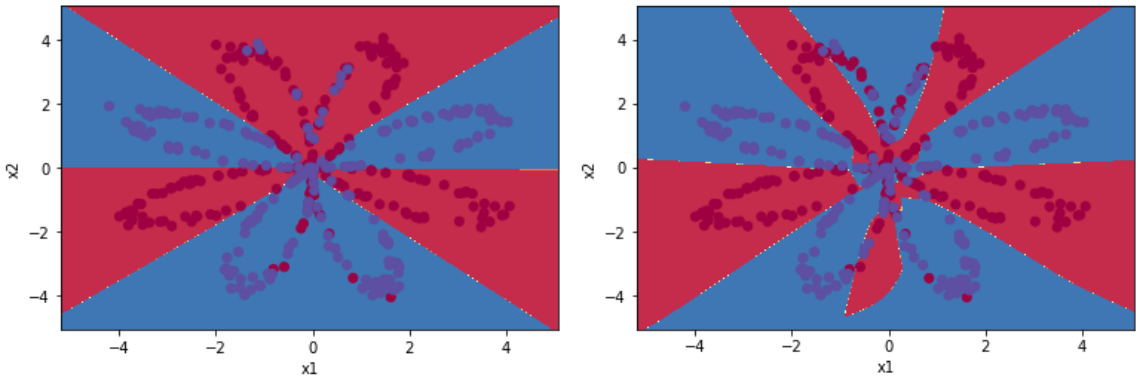

Figure 7: Example of a good general fit on the left and overfitting on the right in a classification task.

5.5 Underfitting training data

The model is too simple to learn the patterns in the data, as shown in Figure 8. To increase accuracy, one could increase the complexity of the model. But the question is how complex the model should be in order to have desirable accuracy. Various model configurations can be tried and techniques such as resampling help in choosing the right models.

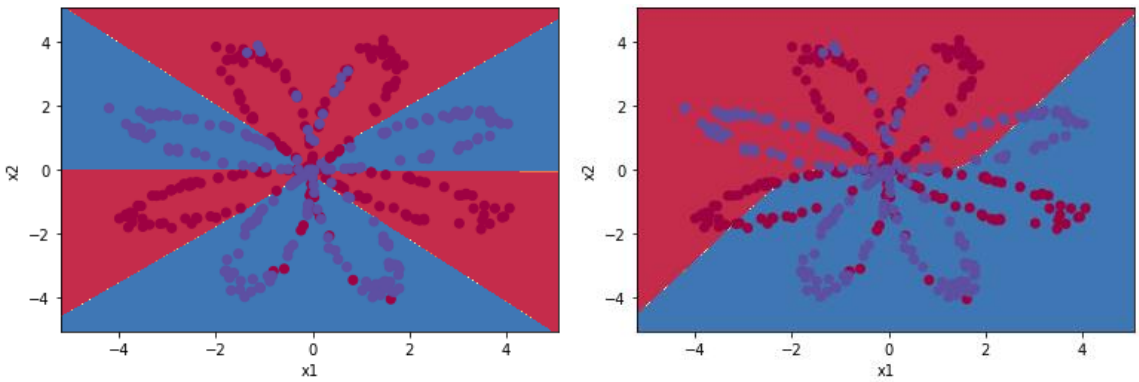

Figure 8: Example of a good general fit on the left and underfitting on the right for a classification task.

References:

1. Géron, A. (2019). *Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems*. O'Reilly Media.
2. Zhang, A., Lipton, Z. C., Li, M., & Smola, A. J. (2020). *Dive into Deep Learning*. https://d2l.ai.