

4. Machine Learning (ML) as a solution tool for DH and ULT DH

Machine learning is a fast-developing tool that allows you to handle large and complex systems with many parameters and data, with different desires for efficiency. It is a branch of artificial intelligence, and defined by Computer Scientist and machine learning pioneer Tom M. Mitchell in relation to his book Machine Learning (1997) [2] as:

“Machine learning is the study of computer algorithms that allow computer programs to automatically improve through experience.”

4.1. Introduction to Machine Learning (ML)

In recent years, Machine Learning (ML) has become one of the most used techniques when modelling relationships between different parameters. It allows, among others, researchers, data scientists and engineers to produce reliable, repeatable decisions and results and to uncover hidden insights through learning from historical relationships and trends in data that the more traditional analytic methods would not be able to grasp. The basis is pattern recognition and, e.g., imitation of neural networks in the human brain, and is strongly entangled with the concept of artificial intelligence (AI). Although ML has existed for decades, one of the major reasons behind its present success is the strong development in computational capacity of modern computers that enables very fast calculation and thereby promotes the possibility to better “train” ML-algorithms efficiently without the need of special and unique computer facilities at hand.

Numerous examples of the widespread application of ML exist today. They count exotic areas such as automatic voice- and face recognition, self-driving cars and artificial players in computer games. Daily tasks, such as reading the addresses on the letters at the post sorting office, ensuring it arrives at the right recipients, as well as the spam filter in your inbox, are based on ML.

Popularly phrased, what distinguishes ML techniques from other statistical methods, is that the algorithms are not specifically programmed; instead the purpose is to let the algorithms learn and improve from experience in order to obtain a well performing algorithm. In “Supervised learning”-techniques the algorithm uses labeled data to train models that are used for a multitude of different problems, the most common ones being classification and regression modelling. In other words, based on a set of training data it learns the general patterns and relations and uses this experience to make a prediction.

The predictions of a ML algorithm will not be fully accurate, and the precision must be evaluated on a test set of data. Along this line, ML falls in the category of “Soft computing” that differs from conventional (“Hard”) computing in that,



unlike hard computing, it is tolerant of imprecision, uncertainty, partial truth, and approximation. For example, an average of 95–99% accuracy typically achieved in the case of a so called “artificial neural network” (ANN) when compared with numerical results [3]. In this respect, ML and soft computing resembles the human mind.

Similarly, ML-algorithms cannot give a result which is outside its training area. If an algorithm is trained to distinguish between two classes of mails, regular and spam, it will not be able to distinguish between mails concerning botany or ornithology. This means that any ML-algorithm will only provide a sensible output in the field in which it has been trained.

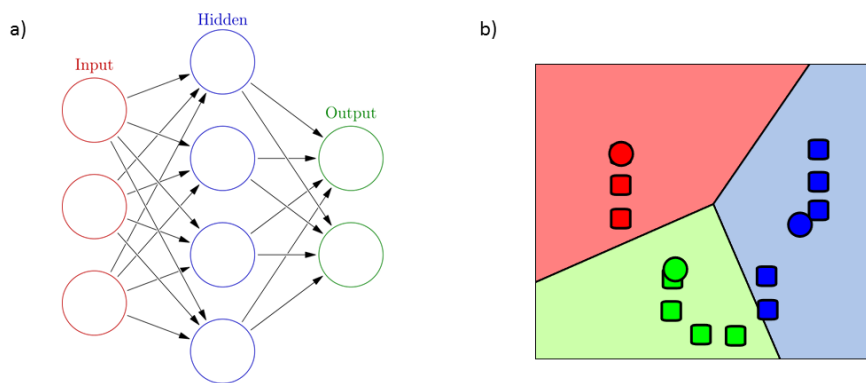


Figure 1: Illustration of two common ML techniques: a) An artificial neural network (ANN) and b) K-means clustering.

Sources: Reproduced without changes from (a) Glosser.ca and (b) Weston.pace, both licensed under the [Creative Commons Attribution-Share Alike 3.0 Unported license](https://creativecommons.org/licenses/by-sa/3.0/).

4.2. ML techniques and tools

The concept of ML covers a wide variety of specific techniques with the common denominator that they all rely on large amounts of data in order to generate a precise model.

Two examples are shown in Figure 1, illustrating the idea behind an artificial neural network (ANN) (a) in which input data is processed through hidden layers to produce an output in the end – a structure inspired by the working mechanisms of the neurons in the human brain. To the right (b) another approach through classification is shown, visualizing how a ML algorithm can recognize patterns and groups of data points to extract information from, at first sight, non- or less structured information.

Many of these techniques can be effectively used for the purpose of optimization. Neural networks and related predictive techniques are highly valuable, for instance, in the financial and economic sector to help predicting prices and work out prognoses etc. without a basis of solid and known model equations. At the other hand, classification methods can be effectively exploited to draw out anomalies, i.e. detection of abnormal behavior, in large amounts of information. As an example, this can be used to identify problems and start maintenance before a vital component fails in a windmill.

4.2.1. Specific techniques

Below, a handful of selected techniques relevant to the work performed in this report, or used by the references cited, are briefly described:

- **Linear regression models (LRMs)** are used to fit a predictive model to an observed data set of values from response and explanatory variables. LRMs include both unique variable regression with one input and output variable respectively, and multiple variable regression that includes more than one input variables. LRMs are not by definition machine learning techniques, however they might be applied together with for example decision trees (DT, see below).
- **Artificial Neural Networks (ANNs)** are a supervised information processing algorithm inspired by the way biological nervous systems, such as the brain, process information. The key element is the novel structure of the information processing system, comprising a number of so-called “Hidden layers”. It is composed of many highly interconnected processing elements (neurones) working together to solve specific problems. ANNs, like the human brain, learn by examples. An ANN is configured for a specific application, such as pattern recognition or data classification, through a learning process.



- **Decision Trees (DTs)** is a type of supervised learning algorithm that can be used in both regression and classification problems. The technique is based on a number of questions that can partition a data set and split the data. Many divisions of the data over such decision nodes will form a “tree” that may eventually (if the correct questions are made) constitute a predictive model.
- **Cluster analysis** or **clustering** is an unsupervised method that groups a set of objects in such a way that objects in the same group (cluster) are more similar to each other as judged by specified criteria than to those in other groups. It is a common technique for statistical data analysis and used in many fields. Various algorithms can be applied, for example by **k-means clustering** which is an iterative process that converges towards an final division of the input data, however with a build-in randomness due to the (random) choice of initial groups that forms the starting point for the iterative process.
- **Support-vector machines (SVMs)** are supervised learning techniques that can be used for both classification and regression analysis. It is a discriminative classifier in the way that it finds an optimal hyperplane which categorizes new data examples from training on labelled training data. The algorithm can be tuned in different ways, for example by using different “Kernels” that describe both linear and non-linear data separation functions.



4.2.2. Software tools and programming languages

Various software tools or programming languages are available for application of ML. The choice of tool depends on many factors such as preferred programming language, license- or freeware, degree of manual programming vs. ready-to-use packages as well as the options for distribution of worked out solutions and programs (e.g. export of executable files).

The most relevant ones in the present context are briefly presented below:

- **Matlab®** (see <https://se.mathworks.com/products/matlab.html>): A numerical computing environment and programming language developed by MathWorks. It allows matrix manipulations, plotting of functions and data, implementation of algorithms, creation of user interfaces, and interfacing with other programs. Matlab® is not a freeware, but is very widely used at universities and in relation to research and engineering.
- **R** (see <https://www.r-project.org/>): A language and environment for statistical computing and graphics. It provides a wide variety of statistical tools (linear and nonlinear modelling, classical statistical tests, time-series analysis, classification, clustering etc.). One of R's strengths is the ease of use and its availability as Free Software.
- **Python** (see <https://www.python.org/>): According to its homepage Python is an interpreted, object-oriented, high-level programming language with dynamic semantics. Its high-level built-in data structures combined with dynamic typing and dynamic binding make it very attractive for Rapid Application Development, as well as for use as a scripting or glue language to connect existing components together. Python's simple, easy-to-learn syntax emphasizes readability and therefore reduces the cost of program maintenance.

According to some references, Python is the most used language for application of ML techniques used commercially, whereas R is the second most used. Matlab® alone has, however, more than 3 million licensed users worldwide.

In this work Matlab® has, because of its research character, been used for the ML activities, while R is the basis for the research carried out in section 7.2.2 concerning Linear Regression Models (LRMs).

