BeeBryte
Energy Intelligence

10 min

Machine Learning … a piece of cake !

Lucas BOYER, Maxime LAHARRAGUE, Pierre-Jean LARPIN

April 16, 2018

Machine Learning … a piece of cake!

Introduction to key concepts

When we talk about machine learning and artificial intelligence, we often hear very different answers from one person to the next: a revolutionary technology? A robotization of daily tasks? In reality, behind these rather generic buzzwords are hiding mathematics sometimes more than 200 years old. In 1829, Gauss and Legendre formulated the “least squares method”, which is the basis of certain prediction algorithms widely used today. But why so many years have passed before the arrival of Machine Learning in our lives? This can partly be explained by the mind-blowing amounts of data and computational power required for the full exploitation of these algorithms, brought about by the computer advances over the late twentieth century.

First in a series of articles on the subject, this publication aims to introduce the main concepts of Machine Learning by illustrating them with meaningful examples.

WHAT IS MACHINE LEARNING?

Machine Learning is a particular field from the wide range of solutions grouped under the generic name of artificial intelligence, as well as expert systems and many optimization techniques. It is usually subdivided into three subdomains: Supervised Learning, Unsupervised Learning, and Reinforcement Learning.

In this article, we will focus on the study of Supervised Learning, which consists in designing a mathematical autonomous model (regressor, classifier, etc) capable of identifying after some training, relations that are sometimes very complex (correlation, causality...) between a large amount of contextual data (the "features") and some observable data (the "targets"). Once this learning is completed (and can be deepened regularly as the data gets richer) the model is able to generalize, that is to say: to anticipate, reproduce or extrapolate the observable values or categories in new contexts.

But like the “haute cuisine”, Machine Learning is not limited to some simple scientific recipe. It also supposes a good dose of "art"; behind his computer keyboard as a stove, the seasoned data-scientist is a “chef” who composes, tests and adjusts its recipe. The data are his ingredients, the algorithms his utensils. He chooses the best combinations, develops his recipe, uses his knowledge and experience to make the most of what he has in the kitchen.

All top chefs will tell you that a successful dish requires first good ingredients, good utensils, but also and especially the "magic" forged on the experience to link the preparations, subtle cooking and balanced mixtures. Similarly, a relevant and effective Machine Learning program requires data in sufficient quantity and quality level, appropriate IT techniques and resources, and all the expertise of a data-scientist to ensure the selection and sequencing of processing steps and the most appropriate methods to extract all the very essence (information) contained in the learning datasets at his disposal, and optimize the performance of the desired functionality (prediction, detection, classification, etc.).

Lucas BOYER, Maxime LAHARRAGUE, Pierre-Jean LARPIN

April 16, 2018

Machine Learning … a piece of cake!

1st STEP: the choice and preparation of ingredients

It is initially a question of "shopping around" and selecting datasets with the greatest care. In particular, it is necessary to proceed in stages and to reject data conveying (a priori) no information, those characterized by a total absence of causal link with the target, or those carrying redundant information.

If, for example, we are trying to predict the consumption of a building, it is probably irrelevant to take into account the average income of building employees, similarly it is perfectly useless to take into account both temperature readings in Celsius, and the same records expressed in Fahrenheit.

The next step is to ensure the proper formatting of clean and homogeneous datasets. This step is crucial because in "real life", data obtained from field measurements often have disparate qualities (missing values, outliers, different acquisition rates, etc.). This involves the processing of each parameter individually - data consistency, choice of a suitable format and unit, possible standardization (scaling of parameters between them). It is also necessary to distinguish the quantitative parameters (e.g. for temperature, 20 ° C is definitely higher than 15 ° C) of the qualitative parameters (e.g. day of the week: a Friday is not “higher” than a Tuesday) which must be treated differently.

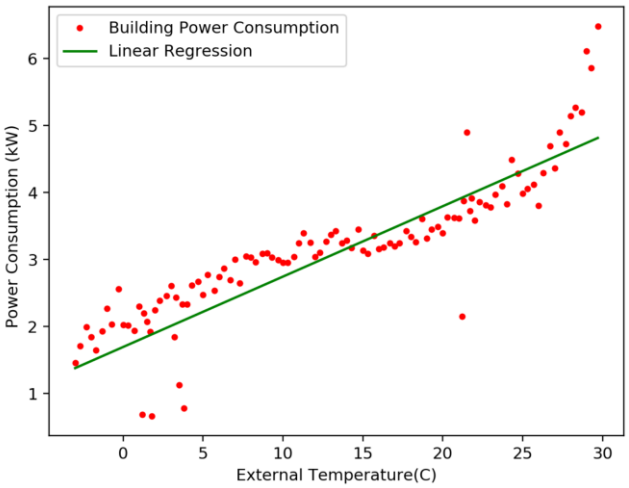
2nd STEP: the choice of utensils

The second step is to choose the most appropriate regression model (or estimator), and its setting parameters a priori (well, like spices, a dish will not express its full potential if not properly seasoned…). These choices must be made according to the type of target: quantitative or qualitative, and among the many existing models; we propose to quickly review the three most classic ones: linear regression, decision trees and neural networks.

LINEAR REGRESSION

In its simplest version, this linear model comes down to finding a linear combination of features to approach "the best" estimate of the target. In spite of its simplicity, this technique is remarkably effective and even generalizable, by using various mathematical "hacks" (typically the "kernel trick"), that we will not develop here, and which make it possible to transform highly non-linear systems into linear ones.

An example of simple linear regression is shown below. By identifying the straight line that best fits the points, we have a good approximation of the underlying relationship between the "feature" reported on the x-axis and the target reported on the y-axis.

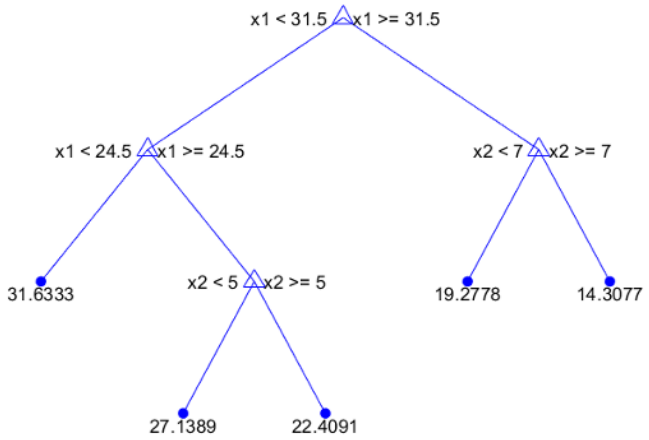


REGRESSION AND CLASSIFICATION TREE

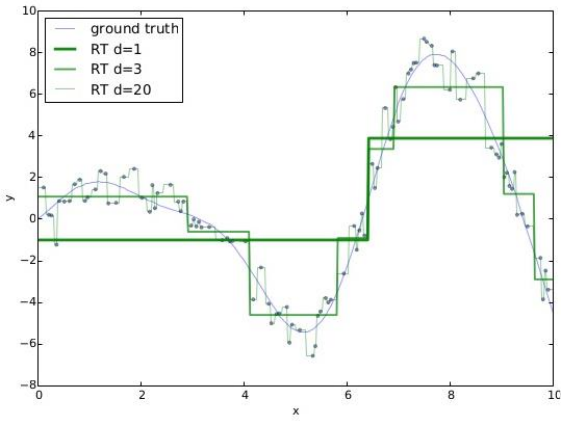
A regression tree (or classification) aims to successively partition the data sample according to a series of binary choices with respect to a threshold (e.g. $x < 3$ or $x \geq 3$). The result of this division is organized according to a graph looking like an inverted tree. An example of a parameter that needs to be chosen wisely is the depth of the tree, which determines the number of final "leaves" and thus the accuracy of the regression or classification.

Lucas BOYER, Maxime LAHARRAGUE, Pierre-Jean LARPIN

April 16, 2018



The training phase consists in identifying at each branch the most discriminating threshold in terms of information contained in the subsample resulting from the previous binary choice. In the inference phase, the value of the "feature" is injected into the tree, and drips along it according to the binary "switch" of each branch. The "leaf" reached at the end of its course is returned as the best estimate target value. The graph below shows the result of a regression with the evolution of the precision according to the depth of the tree.



In practice, regression trees are never used on their own today, but multiplied in the form of "forests", introducing randomness in their elaborations. The estimation of the target value is done by statistical exploitation of the results within the forest and is thus smoothed by “cluttering effect”, and turns out to be way better in generalization.

MULTILAYER PERCEPTRON, BETTER KNOWN AS THE “NEUTRAL NETWORK”

A neural network is a model inspired by our biological neurons. Its strength lies in its great versatility: theoretically, a well-sized neural network is able to reproduce any continuous function (universal approximation theorem).

We will not describe here this utensil, which will soon be the subject of a dedicated article.

3rd STEP: THE COOK’S KNOW-HOW

The “data-chef” must now choose the parameters of his model in order to optimize the accuracy and the relevance of his predictions, that is to say: to minimize their distance from the expected targets, without excess of complexity.

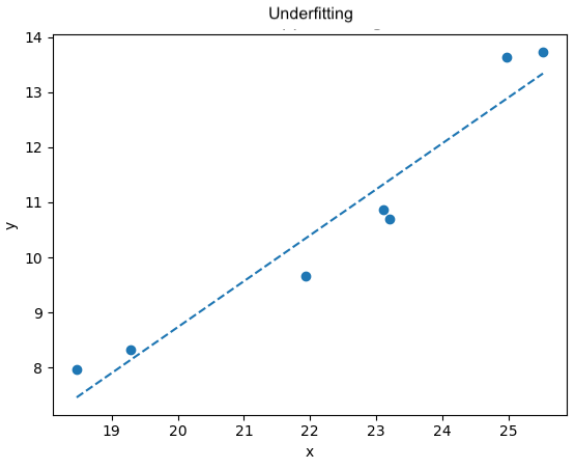
All the subtlety of the training phase lies in the model fitting to strike the right balance between (1) exploiting most of the information in the training dataset which is specific to the underlying model we are trying to approximate (2) and rejecting "parasite" information (noise, other influencing factors, etc.) that is not directly related to the model but contingent on the training sample itself.

Simply put: one needs to separate the general from the specific, precisely to obtain at the same time the best precision and the best possible generalization of the resulting approximate model. We will illustrate this point on a very simplified example:

UNDERFITTING

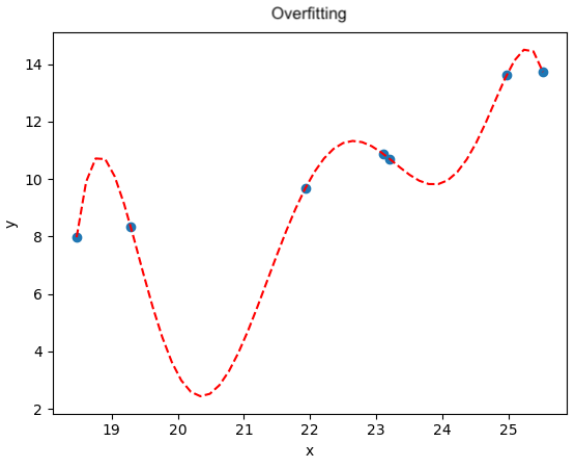
When the information contained in the learning sample is not sufficiently exploited, typically by using an overly simplistic model, the resulting estimator has good generalization property but unsatisfactory accuracy. This is referred to as "underfitting". The model has high bias and low variance.

Machine Learning … a piece of cake!



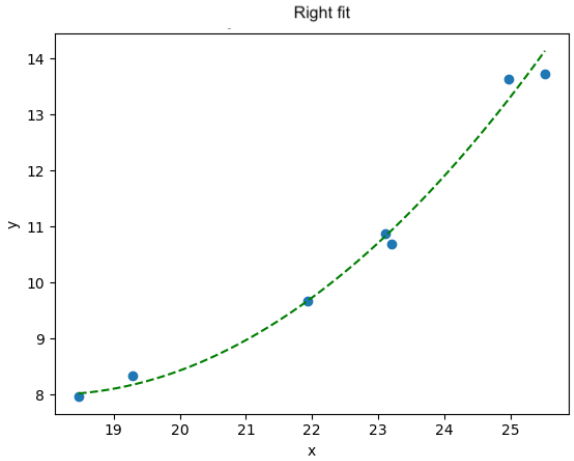
OVERFITTING

When we do not sufficiently discard the sample-specific information, the model fits the dataset well, even too well, but performs poorly in terms of prediction quality, and returns inconsistencies when applying to features away from the area of the training sample. We are talking about “overfitting”. The model has a low bias but high variance since it seeks to be accurate for a (too) large amount of data.



OPTIMAL CASE

Ideally, in data science as in cooking, everything is a matter of balance and a good training must lead to a well-balanced model, compromising between precision and generalization, between bias and variance



Through this article and its few culinary digressions, we hope that we clarified some basic concepts about Machine Learning and have aroused your curiosity for this vast domain! In the next article, we will focus on these fascinating tools that are the neural networks, which gave birth to deep learning and at the origin of the tremendous acceleration of machine learning performance in recent years.

Bee hungry, Bee foolish!

Lucas BOYER, Maxime LAHARRAGUE, Pierre-Jean LARPIN April 16, 2018