

The Role Of Error In Machine Learning And The Law: Challenges And Perspectives

Laurenço Ribeiro Grossi Araújo

*Ph.D. Candidate at the Graduate Program in Electrical
Engineering at UFMG*

MSc of Electrical Engineering at UFMG

Yuri Alexandre dos Santos

*MSc Candidate at the Graduate Program in Computer Science at
UFMG*

LL.B. at UFMG

Abstract

Error is a key element in machine learning, as every modern widely used machine learning algorithm performs some task to minimize - and not to nullify - a loss function, which basically measures the difference between the output of the predictive model and the function it tries to approximate. In that sense, machine learning algorithms, as they are today, will always make mistakes, even with large well-constructed datasets. This limitation, by itself, poses a challenge for lawmakers and jurists, as it may influence regulatory efforts or the possible usage of automated systems in real-life applications that may impact the legal field.

Keywords

Machine learning. Statistical learning theory. Law.

Introduction

From the early days of the Perceptron, through Backpropagation and Statistical Learning Theory, the error has been a key concept in machine learning literature. According to Tom Mitchell (1997), “Machine Learning is the study of computer algorithms that improve automatically through experience.”. It is a subfield of a much broader discipline known as artificial intelligence. Three main learning problems, that encompass a large spectrum of applications, are pattern recognition, regression estimation and density estimation (VAPNIK, 1999).

In this work, we approach learning from an optimization perspective (MITCHELL, 1997): learning is viewed as optimizing an objective function, typically associated with an error measurement, also known as a loss function. The loss function is a measurement of the difference between the output of the model and the function which we try to approximate (VAPNIK, 1999). A learning algorithm such as a neural network typically learns in epochs using a technique called gradient descent (or one of its many variations): for each iteration of the algorithm, the parameters (or neurons) of the neural network are changed in a direction (the opposite direction of the gradient, hence the name) that minimizes the loss function (HAYKIN, 2007).

When it comes to addressing *machine learning* from a legal perspective, legal literature tends to treat it either from a regulatory perspective, an approach that is probably influenced by the relevance given to Data Protection laws and regulations in the last few years, or from a “potential use in legal contexts” perspective.

Those who address the issue from the regulatory viewpoint are usually interested in how it is possible to legally regulate the use of machine learning models, discussing topics such as the impacts machine learning models may have on individuals’ rights (RASO et al., 2018), the difficulties in explaining them (and possible solutions) and

the possibilities of holding developers or companies accountable for decisions made based on their outcome (ROBERTO, 2020; TEPEDINO and SILVA, 2019), among other possible topics.

On the other hand, works that address machine learning from a “potential use in legal contexts” point of view are mainly concerned with the possibilities of using machine learning techniques in Law-related issues, aiding or even substituting professionals such as lawyers, judges, prosecutors and public officials. In this sense, it is worth mentioning the growing field of jurimetrics, which has been increasingly addressed by Brazilian legal researchers (MAIA and BEZERRA, 2020). Sometimes, this second approach overlaps with the first one.

In either case, errors that are inherent to algorithms covered by the term *machine learning* seem to be an overlooked topic, at least in Brazil, which, we argue, may lead to excessive optimism or pessimism in legal literature. In that sense, this work intends to contribute to the understanding of how error plays a key role in machine learning models, which, we believe, should impact the way legal literature and legal regulations address the matter.

It is worth mentioning that it is somewhat common for legal texts to address errors in machine learning, but from a very different perspective. Many recent researchers, certainly influenced by the growing field of fairness in machine learning, have been investigating mistakes in the context of explainability or accountability of automated decision-making systems (WACHTER et al., 2018; MORSE, 2019; DOSHI-VELEZ and KORTZ, 2017).

Even though these works are indeed very important, our aim is not to discuss ways to reduce mistakes or to explain decisions that could affect individuals. Nor is it to draw a regulatory framework applicable to artificial intelligence systems. What we do intend is to address the following question: how will the Law deal with the fact

that machine learning models will *always* make some mistakes, as error is a key subject in machine learning theory?

2. Statistical Learning Theory

Learning can be viewed as solving a function approximation problem (VAPNIK, 1999). An appropriate approximate function must be found, that is capable of approximating the behavior of a given function of interest. However, for finite datasets, there are infinite possible solutions.

The Statistical Learning Theory was developed by Vapnik and Chervonenkis and it builds a mathematical framework that allows a rigorous treatment of one of the main questions that machine learning scientists have to answer: ‘what is a good machine learning model?’ (VAPNIK, 1999; SCHÖLKOPF et al., 2002).

Some of the key concepts introduced by the Statistical Learning Theory are the Risk, the Empirical Risk and the Structural Risk. A more in-depth discussion of these concepts is beyond the scope of this work and the reader is referred to the work of Vapnik (1999) for definitions and proofs. However, an understanding of the concept of Risk and Empirical Risk is necessary for the following sections of this paper.

First, consider a typical machine learning training pipeline. For a given task, data is collected and preprocessed; it is then fed into a model, which learns from it; and, finally, the trained model is employed on unseen data.

In this case, the Risk would be the expected value for the loss when the model is applied on unseen data and the Empirical Risk would correspond to the loss calculated only on the training dataset (SCHÖLKOPF et al., 2002).

If one knew how to precisely compute the Risk, the best possible model would be obtainable, and performance on unseen data

would be maximum. However, training is limited to a finite amount of examples, which only allow the optimization of the Empirical Risk, which can converge, under some specific circumstances, to the real Risk (VAPNIK, 1999).

In that sense, the goal of *machine learning* is to find a model in which the Empirical Risk (which we can calculate) and the real Risk (which we cannot know for sure and can only estimate) are close to each other and are both low.

It is important to observe that, even when minimizing the real Risk, there is no way of guaranteeing a model with zero error in real-world applications. This result can be easily explained when the error decomposition (GEMAN et al., 1992) is performed, which results in three terms:

Bias and Variance are values that depend on the model and can be zeroed by a hypothetical perfect model; the irreducible error, however, as its name suggests can not be reduced, as it comes from noise and imprecision inherent to measurements.

3. Discussion

As we have discussed so far, the *error* appears in machine learning not only as an undesired effect of dataset limitations or poorly implemented machine learning algorithms, but as a fundamental concept in the way they function. Most major modern machine learning techniques rely on some form of error minimization, which is not the same as error nullification.

Systems that have empirical error equal to or very close to zero in the training set (i.e. the slice of the dataset used for training a machine learning model) may suffer from a problem called *overfitting*, which, in very general lines, happens when a model performs very well on - and *only* on - the data it has already seen (DIETTERICH, 1995). That means that even though a model may make very few or no

mistakes during its development, if it suffers from overfitting, it will behave poorly when it comes across previously unseen data, which is the whole point of using machine learning in the first place. A good model should, thus, generalize well, which implies accepting some degree of error.

Needless to say, machine learning models may be better or worse depending on the way they are implemented and on the quality of the data used for training them (EUROPEAN UNION AGENCY FOR FUNDAMENTAL RIGHTS, 2019). However, regardless of how good one dataset or algorithm may be, the fact is current state-of-the-art machine learning techniques will still make mistakes.

Given that, there is simply no way to guarantee a machine learning model will make no mistakes, and that poses a serious issue especially considering their outcome may not be fully understandable or explainable.

Model interpretability is desired in all applications of machine learning, but even more so in applications that involve human beings. Models with higher interpretability, however, tend to lack the capacity to treat complex problems. They are usually models with high bias, that lead to high error values when applied (MOLNAR, 2020).

That is not to say that a simpler, more interpretable model such as a decision tree or a linear regression should never be applied. Those models are perfectly capable of giving state-of-the-art solutions when applied to the correct classes of problems.

There are, however, problems that cannot be adequately solved by simple machine learning algorithms. More complex methods are needed and there is a significant decrease in terms of interpretability. The most famous example of non-interpretable black boxes are deep neural networks. Those neural networks are capable of astonishing results in areas such as image and sound processing, as well as natural language processing. They are composed of millions and millions of

parameters and it is really hard to understand the reasons for a given output, or the contribution of a subset of their neurons.

Even though an increasing number of studies on fairness, explainability, interpretability and transparency in machine learning has been done in recent years and are still being conducted by researchers from different areas, with some very interesting proposals³³, the fact is there is simply no current way of guaranteeing a machine learning model will not make mistakes in data it has not yet seen.

Even if the error of a certain machine learning model is very low, it is still possible that it makes many more mistakes than previously expected. Say, for example, that a certain model, during its training, makes 99,5% of its predictions correctly, meaning that, during its testing, it only makes mistakes in 0,5% of the cases - which is quite good for many tasks. When applied in the real world, with data it has never seen before, it is not possible to say, for sure, that those percentages will stay unchanged. In practice, it is perfectly possible that that same algorithm starts making more mistakes, say, with 2% or 3% likelihood.

Furthermore, machine learning algorithms are trained to minimize a loss function in a specific finite dataset, meaning that machine learning models, as accurate or well programmed as they may be, are not capable of interpreting deviant or new behavior that may affect the issue for which they were created in the first place. As a consequence, even if the error presented by the model during its development is very low, it will have no use if the correlations between the variables in its dataset are not present in a given new situation that presents itself to the system.

Nonetheless, there is a broader (and perhaps more philosophical) problem when it comes to the way machine learning

³³ For a very good example in this matter, refer to WACHTER et al. (2018). Their text provides an interesting discussion on the possibility of using counterfactual explanations in the context of the European General Data Protection Regulation (GDPR).

algorithms work. When it comes to models that somehow use information relating to human beings, training is always done with data about specific persons, i.e., to a group of people whose data is present in the dataset. However, such models may be used to make individual decisions that impact persons that are not part of that specific group. In that sense, it is possible to ask whether (and in which cases) certain patterns and correlations identified by an algorithm in a dataset restricted to a group of people should be used to make individual decisions for others.

As a matter of fact, machine learning techniques have many other limitations (MALIK, 2020) that are not addressed by this article and that should also be taken into account by anyone who intends to use, study or regulate them.

4. Conclusion

Machine learning techniques are not fail-proof and are still very limited for several applications, including the legal field. We are yet very far away from the development of a system that could, for example, substitute lawyers and judges, especially - but not only - in tasks that have particularities making them essentially different from previous ones.

The acknowledgment that machine learning algorithms have limitations and will inherently make mistakes is important to lower some expectations with the field and may pose some interesting questions to be addressed by jurists, practitioners and law-makers in their respective fields.

For instance, instead of immediately asking whether or not it is possible to use or to improve a machine learning technique for some specific task (with or without legal implications), it may be interesting to ask if, how and to what extent we *should* use them, considering all their limitations.

Furthermore, when addressing situations in which individuals are subject to decisions made by automated systems, the possibility of an error should always be taken into consideration, as there is no current way of eliminating it.

We do not intend to address the issue presented here entirely, but only to initiate a discussion that undoubtedly needs further investigation in future research.

References

DIETTERICH, Tom. *Overfitting and Undercomputing in Machine Learning*. Computing Surveys, v. 27, p. 326-327, 1995.

DOSHI-VELEZ, Finale; KORTZ, Mason A. *Accountability of AI Under the Law: The Role of Explanation*. Berkman Klein Center Working Group on Explanation and the Law. Berkman Klein Center for Internet & Society working paper, 2017.

EUROPEAN UNION AGENCY FOR FUNDAMENTAL RIGHTS. *Data Quality and Artificial Intelligence: Mitigating Bias and Error to Protect Fundamental Rights*. Publications Office of the European Union, 2019.

GEMAN, Stuart; BIENENSTOCK, Elie; DOURSAT, René. Neural networks and the bias/variance dilemma. *Neural computation*, v. 4, n. 1, p. 1-58, 1992.

HAYKIN, Simon. *Neural networks: a comprehensive foundation*. Prentice-Hall, Inc., 2007.

LIPTON, Zachary C. The mythos of model interpretability. *Queue*, v. 16, n. 3, p. 31-57, 2018.

MAIA, Marcos; BEZERRA, Cicero A. *Análise bibliométrica dos artigos científicos de jurimetria publicados no Brasil*. RDBCI: Digital Journal of Library and Information Science, v. 18, e020018. 2020.

MALIK, Momin M. *A Hierarchy of Limitations in Machine Learning*. arXiv preprint arXiv:2002.05193, 2020.

- MITCHELL, Tom. *Machine Learning*. McGraw Hill, 1997.
- MOLNAR, Christoph. *Interpretable Machine Learning*. 2020. Retrieved from <https://christophm.github.io/interpretable-ml-book/>.
- MORSE, Susan C. *When Robots Make Legal Mistakes*. Oklahoma Law Review, v. 72, n. 1, 2019.
- MULLAINATHAN, Sendhil; OBERMEYER, Ziad. *Does machine learning automate moral hazard and error?*. American Economic Review, v. 107, n. 5, p. 476-80, 2017.
- RASO, Filippo; HILLIGOSS, Hannah; KRISHNAMURTHY, Vivek; BAVITZ, Christopher; LEVIN, Kimberly. *Artificial Intelligence & Human Rights: Opportunities & Risks*. Berkman Klein Center for Internet & Society Research Publication, 2018.
- ROBERTO, Enrico. (2020). *Responsabilidade civil pelo uso de sistemas de inteligência artificial: em busca de um novo paradigma*. Internet & Sociedade, v. 1, n. 1, p. 121-143, fev. 2020.
- SCHÖLKOPF, Bernhard et al. *Learning with kernels: support vector machines, regularization, optimization, and beyond*. MIT press, 2002.
- TEPEDINO, Gustavo; SILVA, Rodrigo da Guia. *Desafios da inteligência artificial em matéria de responsabilidade civil*. Revista Brasileira de Direito Civil – RBDCivil, Belo Horizonte, v. 21, p. 61-86, jul./set. 2019.
- VAPNIK, Vladimir N. *An overview of statistical learning theory*. IEEE transactions on neural networks, v. 10, n. 5, p. 988-999, 1999.
- WACHTER, Sandra; MITTELSTADT, Brent; RUSSELL, Chris. *Counterfactual explanations without opening the black box: Automated decisions and the GDPR*. Harvard Journal of Law and Technology, v. 31, p. 841-887, abr. 2018.