aepd agencia española protección datos

EDPS

10

10 MISUNDERSTANDINGS ABOUT MACHINE LEARNING

The EU has identified artificial intelligence (AI) as one of the most relevant technologies of the 21st century and highlighted [1] its importance on the strategy for EU's digital transformation. Having a wide range of applications, AI can contribute in areas as disparate as helping in the treatment of chronic diseases, fighting climate change or anticipating cybersecurity threats.

"Artificial intelligence", however, is an umbrella term for technologies that aim at mimicking human reasoning capabilities, which can have very different applications and limitations. Frequently, technology vendors promote their systems claiming that they use AI, without specifying which type of AI.

Machine learning (ML) is a specific branch of AI, applied to the resolution of specific and limited problems - such as classification or prediction tasks. Unlike some other types of AI that try to distill human experience (e.g., expert systems [2]), the behaviour of machine learning systems is not defined by a predetermined set of instructions.

ML models are trained using datasets. During their training, ML systems adapt autonomously to the patterns found among the variables in the given dataset, creating correlations. Once trained, these systems will use the patterns learned to produce their output.
Unlike other types of AI systems [3] , the performance [4] of ML models depends greatly on the accuracy and representativeness of training data.

The aim of this document is to dispel common misconceptions surrounding ML systems, while underlining the importance of implementing these technologies in accordance with EU values, data protection principles and full respect of individuals.

---

1 Commission Communication, Artificial Intelligence for Europe, **https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=COM%3A2018%3A237%3AFIN**

2 Experts systems are computer programs designed to solve complex problems on specific areas. They rely on a knowledge base, which defines the rules for decision-making, and on an inference engine, which applies the rules.

3 In machine learning, the algorithm learns rules as it establishes correlations between inputs and outputs. In symbolic reasoning, the rules are created through human intervention. First humans must learn the rules by which two phenomena relate, and then hard-code those relationships into the symbolic reasoning system. Therefore, the accuracy of the symbolic AI system relies on the quality of the human-defined relationships, rather than the quality of the input dataset(s).

4 Simply put, the performance of an ML system is how "good" its predictions really are. Albeit being a simple concept, the complexity has to do with identifying what is considered "good". Several "performance metrics" exist, and they evaluate ML models differently: Accuracy is the fraction of predictions that a model got right; Precision is the ratio between the number of correct results and the number of all returned results; Recall is the ratio between the number of correct results and the number of results that should have been returned. Depending on the context of application, some performance metrics could be more relevant than others.

1 MISUNDERSTANDING

Correlation implies causality.

Fact: Causality requires more than finding correlations.

"Causality" is the relationship that exists between cause and effect. "Correlation" is the relationship that exists between two factors that occur or evolve with some synchronization.

ML systems are very efficient in finding correlations but lack the analytical ability to go beyond that and establish a causal relation[5].

For example, if given a dataset composed of IQ test scores and the related height of the individual (but not age), a ML model could erroneously predict that tall people are smarter than shorter people by establishing a correlation between increasing height and increasing IQ scores.

However, such a phenomenon could be rooted in the fact that children commonly score lower than adults could in IQ tests.

It is possible to train a ML system to infer diseases using a dataset with symptom-disease correlations. However, that same system might not be adequate to explain what is causing the inferred disease.

These examples illustrate that human supervision is necessary to ensure that ML systems identify the relevant variables (the causes) for a prediction or classification.

2 MISUNDERSTANDING

When developing machine learning systems, the greater the variety of data, the better.

Fact: ML training datasets must meet accuracy and representativeness thresholds.

The growing development of ML systems has led to a greater demand for sharing personal and non-personal data, because ML developers do not have enough data to improve the performance of their systems.

Typically, the training of ML systems requires large amounts of data, depending on the complexity of the task to be solved. However, adding more training data to a machine learning model development process will not always improve the system performance.

In fact, it could create new problems or worsen existing ones. For instance, adding more light-skinned male images to facial recognition training datasets will not help correct any existing gender, or ethnic, biases of the systems[6].

The GDPR requires the processing of personal data to be proportionate to its purpose. From a data protection perspective, it is not a proportionate practice to increase substantially the amount of personal data in the training dataset to have only a slight improvement in the performance of the systems.

More data will not necessarily improve the performance of ML models. On the contrary, more data could bring more bias.

5 **https://www.wired.com/story/ai-pioneer-algorithms-understand-why**

6 Find an analysis of gender and ethnic bias in: Gender Shades project **http://gendershades.org**

3 MISUNDERSTANDING

ML needs completely error-free training datasets.

Fact: Well-performing ML systems require training datasets above a certain quality threshold.

The performance of ML depends, among other factors, on the quality of its training, validation and test datasets. Therefore, training datasets should be able to describe an actual case in a comprehensive and accurate enough way.

However, statistical science suggests that despite the presence of individual errors in input data, it is possible to calculate accurately the average result when processing large amounts of data[7]. ML models are tolerant to occasional inaccuracies on individual records[8] because they rely in the overall quality of large datasets used to train them.

Some ML models are trained using synthetic data, i.e. artificially generated training datasets, which mimic real data. Even if no real data precisely matches the synthetic data, ML models trained on synthetic data can produce good performances[9].

Differential privacy is a technique that introduces noise into the training datasets to preserve data subjects' privacy. Despite the inaccuracies produced by differential privacy, ML models are able to achieve good performances[10].

7 "When you have plenty of data the law of big numbers tends to make sure the data is evenly distributed." (p. 43) Practical Machine Learning with H20 by Darren Cool tried to k, O'Really Media Inc. "Third, data anomalies were eliminated in the data cleansing process, due to the so-called law of big numbers." **https://link.springer.com/article/10.1186/s40537-019-0216-1**

8 In fact, due to the large volume of input variables (features) in some ML models, it is often necessary to use techniques that introduce noise into the input data such as Principal Component Analysis (PCA), a technique to aggregate features together. Of course, the noise introduced into the input data should be below the acceptable performance value of the application.

9 To examples of machine learning models trained on synthetic data: Amazon Alexa **https://www.amazon.science/blog/tools-for-generating-synthetic-data-helped-bootstrap-alexas-new-language-releases**
Google Waymo **https://blog.waymo.com/2019/08/learning-to-drive-beyond-pure-imitation_26.html**

10 'Recent research has shown, counterintuitively, that differential privacy can improve generalization in machine learning algorithms - in other words, differential privacy can make the algorithm work better!' **https://www.nist.gov/blogs/cybersecurity-insights/how-deploy-machine-learning-differential-privacy**

4 MISUNDERSTANDING

The development of ML systems requires large repositories of data or the sharing of datasets from different sources.

Fact: Federated learning allows the development of machine learning systems without sharing training data sets.

The pooling of both data and ML system into a cloud computing infrastructure controlled by the ML developer is a common solution for working around performance constraints. This is an architecture known as centralized learning. However, although centralized learning can mitigate performance constraints, there are still certain considerations that must be taken into account. One is that personal data requires both the data controller and the recipient of the data to comply with the GDPR principles of accountability, security, and purpose limitation, among others. Another is that larger repositories of personal data increase the interest of third parties to gain unauthorized access and exacerbates the impact of a personal data breach.

Distributed on-site learning and federated learning are alternative development architectures to centralized machine learning. In distributed on-site learning, each data controller server downloads a generic or pre-trained ML model from a remote server. Then each local server uses its own dataset to train and improve the performance of the generic model. After the remote server has distributed the initial model to the devices, no further communication is necessary. It involves the same techniques used in centralized learning but in the controller's servers. In federated learning, each data controller server trains a model with its own data and sends only its parameters[11] to a central server for aggregation. Data remains on-devices and knowledge is shared through an aggregated model with peers. No learning architecture fits all tasks. However, accumulating data on one, or many, servers is not always the best, nor the most efficient solution, and it could even become an obstacle for Small and Medium Enterprises' (SME) machine learning development[12].

11 In machine learning, parameters are the values that a learning algorithm can change independently as it learns. These values are optimized has the model learns, thus perfecting its reasoning.

12 Abdulrahman, Sawsan & Tout, Hanine & Ould-Slimane, Hakima & Mourad, Azzam & Talhi, Chamseddine & Guizani, Mohsen. (2020). A Survey on Federated Learning: The Journey From Centralized to Distributed On-Site Learning and Beyond. IEEE Internet of Things Journal. PP. 10.1109/JIOT.2020.3030072. **http://dx.doi.org/10.1109/JIOT.2020.3030072**

5

MISUNDERSTANDING

ML models automatically improve over time.

Fact: Once deployed, ML models performance may deteriorate and will not improve unless it receives further training.

During the training of a ML model, the algorithm is constantly tested. When the model is mature (i.e. it can correctly solve the problems for which it was designed), it is considered suitable to be deployed.

A model that is deployed and no longer trained will not "learn" further correlations from incoming data, no matter how much data it is given. This means that, unless ML models continue to be trained, they cannot be expected to evolve. This is a risk for the accuracy of the system, as its obsolescence towards reality can endanger its ability to make adjusted, and fair, judgements.

The predictive ability of ML models can deteriorate over time in two different ways: due to data drift (substantial changes in the input data) or due to concept drift (when our interpretation of the data changes while the general distribution of the data does not change).[13]

Since the context of the processing where the ML system works can evolve, it is necessary to monitor the system to detect any model deterioration and act on this decay (e.g., by further training the model with new data, while taking into account data protection requirements).

13 A Comprehensive Guide on How to Monitor Your Models in Production **https://neptune.ai/blog/how-to-monitor-your-models-in-production-guide**

6

MISUNDERSTANDING

Automatic decisions taken by ML algorithms cannot be explained.

Fact: A well-designed ML model can produce decisions understandable to all relevant stakeholders.

There are several approaches to provide explanations of AI-based decisions, and most of them can be applied to ML model decisions as well.

Some approaches clarify the model creation process, specifying which parameters and hyperparameters[14] were considered and how much influence each had in the resulting model. Others explain how the model interprets the characteristics of the incoming data,[15] allowing individuals to understand, and anticipate, how the system will behave in a particular situation. Some other approaches do not explain the overall behavior of the model, but instead focus on how a particular input influenced the achievement of a particular outcome[16].

Different degrees of explanation detail may be necessary, depending on the individuals and the context. The appropriate approach will be the one that can clearly describe to the audience the path taken to the decision-making since the training and creation of the model.

14 A hyperparameter is a parameter whose value is set before the machine learning process begins. In contrast, the values of other parameters are derived via training.

15 For example, the value of a patient's blood pressure is very relevant to detect a certain disease, while the age of the patient is not so relevant.

16 Arya, V. et al. "One Explanation Does Not Fit All: A Toolkit and Taxonomy of AI Explainability Techniques." ArXiv abs/1909.03012 'tr(2019): **https://arxiv.org/abs/1909.03012v2**

7 MISUNDERSTANDING

Transparency in ML violates intellectual property and is not understood by the user.

Fact: It is possible to provide meaningful transparency to AI users without harming intellectual property.

Individuals should receive sufficient information about how their personal data is handled, and AI systems should be no different. This type of transparency does not necessarily involve the disclosure of detailed technical information that, in most cases, would not be meaningful for the users.

In the same way that a medicine leaflet provides information about uses, misuses and side-effects, abstracting the user from the detailed chemical descriptions, a ML system should offer their users with meaningful information that makes them aware of the logic applied, as well as the importance and expected consequences of the processing.

When processing personal data using ML, data controllers should properly inform data subjects about the possible impacts in their daily lives.

Examples of meaningful information are certifications, limitations of the system, system's performance metrics, the personal data used for input and generated as output, impact of certain input data on the output, communications to third parties, and risks to rights and freedoms.

8 MISUNDERSTANDING

ML systems are less subject to human biases.

Fact: ML systems are subjects to different types of biases and some of these come from human biases.

ML models can be free of human bias or favoritism toward an individual or a group based on their inherent or acquired characteristics. However, ML systems are selected, designed, tuned, and trained with data, which, in most cases, was selected by humans. ML systems could be subject to more than twenty types of bias stemming from their data processing [17].

Some of the biases affecting ML systems replicate human biases (e.g. a model trained with historical CEO profiles will be biased towards male candidates). Other potential ML biases depend on human decisions such as how the training data is sampled or the output presented. Sometimes machine learning systems are used in contexts that are not the same for which the models were designed.

In short, the objective is for ML systems to build on the experience and insight provided by their creators.

However, the systems do not inherit the humanity required to handle exceptional situations: they do not have a global vision of the problem and have limited ability to adapt to contextual changes and to be flexible in the face of unforeseen circumstances.

17 Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. "A Survey on Bias and Fairness in Machine Learning." (2019) **https://arxiv.org/abs/1908.09635v2**

9 MISUNDERSTANDING

ML can accurately predict the future.

Fact: ML system predictions are only accurate when future events reproduce past trends.

ML takes into account data found in the datasets and uses it to draw projections of possible future outcomes.

Therefore, ML systems do not make guesses about the future, but rather forecasts, which are rooted on past events, and provided to the systems during training.

Some ML learning models could evolve to adapt to new data, such as profiling models in marketing or online media. However, they are unable to adapt to a completely new scenario or rapidly changing events. To adapt their predictions to such changes, most models will need large quantities of new data.

10 MISUNDERSTANDING

Individuals are able to anticipate the possible outcomes that ML systems can make of their data.

Fact: The ability for ML to find non-evident correlations in data can end up with the discovery of new data, unknown to the data subject.

ML systems are excellent at finding correlations in data and are able to identify patterns in personal data that have not been explicitly sought and are unknown even to the individuals concerned (e.g. a predisposition for a disease). This potential raises several concerns from a data protection point of view.

On the one hand, data subjects can be affected by decisions based on information they do not know and had no way to anticipate and/or react.
On the other hand, data subjects might receive ML-triggered information about them in places or situations where there might be an increased impact on their lives due to the specific context. For instance, by receiving mail discount coupons from a commercial store, based on their shopping habits, which could reveal a compulsory habit for lottery games.

When ML systems process personal data to create inferences beyond the stated purpose of the processing, for instance when doing some kind of profiling (predictions or classifications) of individuals, the controller still needs fulfil all data protection principles, including lawfulness, transparency (Art. 5(1)(a) of the GDPR) and purpose limitation (Art. 5(1)(b) of the GDPR).

Any type of further processing of personal data requires a legal basis and a clear purpose.