Explaining what learned models predict: in which cases can we trust machine learning models and when is caution required?

Marco Negrini
marco1.negrini@mail.polimi.it

Nowadays, different types of learning models are adopted. They differ from one another on the learning problem, the expected prediction, by the way of searching and building the data's underlying pattern, and the training method adopted if adopted at all.

One of these groups is composed for example of supervised learning models, that try to estimate the unknown model that maps known input data to labeled outputs (either a numerical value or class membership). Defining "supervised" such types of algorithms means that they will be checked and corrected during the training process. Moreover, they are task-oriented, because they are mostly adopted in specific contexts. Two tasks that can be performed by supervised learning systems are classification and regression. The main difference between classification and regression is that the latter, instead of giving labels as output through discrete values, produces real and continuous values.

Another group is composed of unsupervised learning models: unlike supervised ones, they work on their own trying to find structure and similarity between unlabeled input data. For this reason, it is possible to say that unsupervised learning is data-driven and the outcomes are controlled by the data and the way it is formatted. There are no correct answers or labeled outputs and no one is training the system. The goal is to find groups with similar aspects within the input data or to determine how the data is distributed in the space.

Dimensionality reduction is an unsupervised task essential to reduce the number of features, or the dimensions, of a dataset. This is needed because having a high-dimensional dataset translates into longer training times, more powerful machines with larger storage capacity, and a set of problems caused by the data weight. This is also known as "curse of dimensionality". When the dataset dimensionality increases, the volume of the space increases so fast that the available data becomes sparse, and it is not possible to form meaningful clusters.

Some types of machine learning models are black-boxes, which means that the steps the model went through to give a conclusion are not known. Outputs are taken blindly and assumed to be accurate, even if it is not always the case. Others are suitable for human interpretation, such as decision trees, linear regressors, or Bayesian classifiers.

Bayesian inference applied to machine learning models helps understanding model uncertainties. As a matter of fact, when analyzing a stochastic model the resulting output will be affected by uncertainties due to the complexity of the model, by the scarcity of data related to failures, and by the narrowness of adopted methods. When analyzing such models, aleatory and epistemic uncertainties can be identified. The former are related to phenomena whose occurrences are intrinsically stochastic, so they can be expressed with a probabilistic approach. On the other side, epistemic uncertainties are related to a lack of knowledge of model parameters. Bayesian classifiers exploit aleatory and epistemic uncertainties to give predictions on the described phenomena and a level of uncertainty on the given predictions.[KG17]

This kind of model improves accuracy in applications such as computer vision, where also black-box models are adopted. Indeed, black-box models are applied nowadays in several different fields, such as in web advertising, web search, image, or speech recognition. But in high-stakes situations model's accuracy and interpretability are fundamental, and understanding why a specific result was achieved is required. Sectors affected are for example finance, criminal justice, or health care. Why a doctor should trust a positive cancer diagnosis given by a black-box model? Which factors led the algorithm to that conclusion?

Local surrogate models are used to explain individual predictions of black-box models. LIME (Local Interpretable Model-Agnostic Explanations), introduced in 2016, is an implementation of local surrogate

model mainly used with tabular data, text, and images. "Local" reflects the fidelity of the classifier around the instance being predicted, while "Model-Agnostic" means that Lime can explain the model without diving into it. The idea behind LIME is perturbing the instance of interest in the dataset and see how black-box predictions change. But before it can be safely applied, some problems need to be solved. First, defining a meaningful neighborhood around a point is difficult: for each application, it is recommended to see for different settings of neighborhood if explanations make sense. Secondly, sometimes explanations of two close points vary greatly. This is called instability of explanations and means that a critical approach should be adopted towards this model.[RSG16]

Neural networks are highly vulnerable to this kind of instability, and adversarial examples can be problematic in safety-critical systems. For example, in image classification even epsilon-range perturbations on input dataset in state-of-the-art neural networks can produce different output labels, so these drawbacks limit their adoptions.

AI^2 is an analyzer for deep neural networks that proves safety properties, such as robustness of the model to adversarial examples. Tested in image classification, it creates an abstract element that captures all perturbed instances of the image to classify. By doing that, it avoids state space explosion, obtained as a consequence of testing neural networks extensively on all input perturbations. In the paper it is given as example the classification of the digit 8 on an 84 pixels representation: assuming a 64-bit floating-point number used to express pixel density, if brightening the value of pixels with intensity above 1-0.085 there are more than 10^{1154} possible perturbed images, unsustainable to verify completely. After creating the abstract element, defined abstract layers compute the effects of layers on this element. In the end, the abstract output is an over-aproximation of all possible concrete outputs, and this enables AI^2 to prove safety properties directly on the result and to certify the effectiveness of state-of-the-art defenses for neural networks.[GMDC$^+$18]

A problem concerning trust and fairness still to be solved is due to the fact that most machine learning systems carry out inductive reasoning: extracting information, create, correct and apply patterns based on previous experiences. They generalize and try to derive models from specific examples found in the training dataset. But sometimes these datasets are filled not only by factual and objective events but also they are impregnated with human inclinations and biases.

"Internet-trained models have internet-scale biases" researchers at OpenAI, an AI research and development company, noted in the paper introducing GPT-3, a new language model. Other than presenting and describing the new model, the paper discusses also broader societal impacts, including fairness, bias, and representation. The output quality of the text generated by GPT-3 is very high and OpenAI found that humans find it difficult to distinguish articles written by humans versus articles written by GPT-3. This carries both benefits and risks: GPT-3 is particularly good at generating extremist texts and could be adopted to create fake news and spread misinformation.[BMR$^+$20]

Building trust towards learned models is essential to take full advantage of this powerful technology. Crucial steps towards this goal are making them more accurate and interpretable, and other than identifying it is important starting to prevent biases in these models. Since they are all created and implemented by humans for humans, when giving an algorithm decision-making capabilities "It's far too easy to blame 'the algorithm' when something goes wrong, but ultimately humans bear the responsibility for these decisions, and they must understand where the potential pitfall lie" (David Cox, IBM Director at MIT-IBM Watson AI Lab).[1]

References

[1] https://www.morningbrew.com/emerging-tech/stories/2021/01/22/nine-experts-single-biggest-obstacle-facing-ai-algorithms-next-five-years.

[BMR$^+$20] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCan-

dlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners, 2020. OpenAI.

[GMDC+18] Timon Gehr, Matthew Mirman, Dana Drachsler-Cohen, Petar Tsankov, Swarat Chaudhuri, and Martin Vechev. Ai2 : Safety and robustness certification of neural networks with abstract interpretation, 2018. Department of Computer Science ETH Zurich, Switzerland.

[KG17] Alex Kendall and Yarin Gal. What uncertainties do we need in bayesian deep learning for computer vision?, 2017. University of Cambridge.

[RSG16] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. "why should i trust you?" explaining the predictions of any classifier, 2016. University of Washington.