The Role of Textualisation and Argumentation in Understanding the Machine Learning Process

Kacper Sokol and **Peter Flach**
Department of Computer Science, University of Bristol
{K.Sokol, Peter.Flach}@bristol.ac.uk

Abstract

Understanding data, models and predictions is important for machine learning applications. Due to the limitations of our spatial perception and intuition, analysing high-dimensional data is inherently difficult. Furthermore, black-box models achieving high predictive accuracy are widely used, yet the logic behind their predictions is often opaque. Use of textualisation – a natural language narrative of selected phenomena – can tackle these shortcomings. When extended with argumentation theory we could envisage machine learning models and predictions arguing persuasively for their choices.

1 Introduction

In recent years machine learning has witnessed a big technological leap and proliferation in everyday life. Predictive models initially flourished in the Internet fuelling shopping recommendations and internet search. Nowadays they are becoming a vital part of decision support systems used in legal matters, politics, finance, credit scoring and job appointments, among others. This widespread adaptation of machine learning algorithms and their influence on our every-day life is often criticised for unfairness[1]. In the wake of algorithms taking supposedly "optimal" decisions for human matters the European Union has introduced the "right to explanation"; it entitles involved parties to receive an explanation of the algorithm's decision[2]. Furthermore, protected features such as gender and race cannot be used in predictive models to prevent discrimination.

In the digital age data are easy to collect; machine learning models are simple to use, learn and deploy with packages such as `scikit-learn` and `weka`. Nevertheless, they are not often understood and inspected in detail before deployment as the main objective is to maximise the predictive accuracy, which rarely includes social costs. Given large amounts and high dimensionality of data, learnt models and the nature of their predictions can easily become incomprehensible. Improved understanding could lead to selecting the correct features and model for the task, hence guarantee fair predictions in deployment.

[1] "Weapons of Math Destruction" by Cathy O'Neil

[2] General Data Protection Regulation (EU 2016/679)

2 Research Problem

To address these issues machine learning experts have developed techniques to inspect data, models and predictions. Understanding data is the most difficult part of the process; researchers use correlation maps to discover dependencies and interactions between features, yet these are limited to pairwise correlation coefficient. Approaches such as t-SNE and PCA allow to project high-dimensional data into 2 or 3 dimensions that can be visually inspected [Maaten and Hinton, 2008]. Understanding models and predictions is a difficult task as well. Often, researchers use glass-box models when transparency is crucial. Decision trees and rule learners, for example, can be read as conjunctions of logical conditions. Predictions of linear regression can be described with corresponding feature weights (which, however, makes assumptions about the commeasurability of features). Black-box models such as deep neural networks, on the other hand, are very hard to interpret. In order to understand their predictions we need to resort to post-hoc methods. These usually build a simple representation of the decision criterion (e.g. a linear model) in the neighbourhood of the instance of interest.

In general, these techniques can be divided into two groups.

Model-dependent techniques are developed with a specific machine learning task in mind, e.g. a linear model and its feature weights or conjunction of logical conditions in a rule. However, they suffer from scalability issues as each model family requires its own approach.

Model-agnostic techniques are usually post-hoc and can be applied to any task, e.g. a local linear model. They are versatile but use glass-box models as backbone, hence they inherit their limitations.

Finally, all of these approaches to understand the data, models and predictions share one commonality: they use *visualisation* in the core of their descriptive power. This is powerful but also has limitations as our intuition in high-dimensional spaces is flawed – we often call this phenomenon the curse of dimensionality. Moreover, they *describe* but do not *explain*: they provide statistics and characteristics that quantify models' behaviour but not their reasoning. While

in some cases the model and the features are simple (e.g. a small and shallow decision tree) and their description is more or less equivalent to an explanation, such models are rarely the norm in today's data-rich world, where multidimensional data often renders glass-box models incomprehensible.

Therefore, we need explanations supported by *arguments* that lead to understanding. Explaining a model means providing a high-level insight into its decision system, e.g. self-driving cars stop before a zebra crossing if a moving object is detected nearby. Explaining a prediction means presenting a thought process supported with arguments, e.g. a self-driving car crossed the junction without stopping because the light was green and the pedestrian crossing was empty. Last but not least, explaining data means to understand their patterns as used by machine learning algorithms to make inferences, e.g. this picture shows a pedestrian crossing because there are orthogonal white stripes on the road.

3 Approach

The simplest and most common approach to characterise a machine learning component – data in particular – is *(statistical) summarisation*. These are usually numerical tables and vectors, which can be difficult to digest for non-experts. Their role is to describe the properties of data by providing information from the system to the user. A more advanced analytic tool is *visualisation*; a graphical representation of data in a form of plots and figures is more insightful. It is also descriptive but sometimes the communication can be bidirectional as in interactive plots. Visualisations are often supported by a small narrative in the form of a caption, which increases their informativeness. To overcome the curse of dimensionality we can use *textualisation* – narratives accompanied by statistics and figures. Natural language can express concepts of considerable complexity and dimensionality. Moreover, it is proven more insightful and effective than presenting raw, numerical and visual data [Portet *et al.*, 2009]. Finally, we suggest to make use of *formal argumentation* – structural, logical narratives accounting for every disputable statement. It provides explanation leading to understanding rather than informative description; a long overdue approach.

Using natural language to describe machine learning data, models and predictions is uncommon. Narrative was used to present data analysis in the Automatic Statistician project[3]. [Farrell *et al.*, 2015] generated reposts of control systems and used narrative to present anomalies in operating system logs. [Portet *et al.*, 2009] developed a system that synthesises medical reports from neonatal intensive care unit data to support medical decisions. Application of argumentation theory in machine learning is even less common despite its capability of argumenting classification choices, hence providing insight into the model reasoning [Možina *et al.*, 2007].

4 Contributions and Directions

Expressiveness and versatility of *textualisation* and logical reasoning behind *argumentation* suggest its wider adaptation in machine learning can be beneficial. We aim to improve the currently available data-to-text frameworks. Such systems mostly use (conditional) templating, hence they require manual engineering and are limited to a particular application domain. We aim to develop a flexible narrative generation platform that accepts a variety of data types. We will showcase the capabilities of our platform by automatically composing narrative accounts of experimental results for use in the *results* section of machine learning papers. For example, such a system would provide a textual narrative comparing performance of a novel algorithm against state-of-the-art solutions only based on accuracy results and meta-data.

Generating a summary of arbitrary data requires a unified feature and meta-data representation. We will introduce a feature annotation approach useful in explaining the data, models and predictions. For example, knowing that two features are length measurements expressed in the same unit would vouch for a model using their mean, while averaging timestamp and temperature is counter-intuitive.

Allowing the data-to-text framework to analyse feature interactions and dependencies could help avoid using combinations of features that are equivalent to a protected feature. If these interactions are complex, describing them with natural language is preferable to numerical coefficients, graphs and figures.

Finally, argumentation theory has not yet been applied in its full power to explain machine learning approaches. We will integrate it with other components of our data-to-text system to produce model-agnostic explanations that yield better understanding of our field.

5 Summary

Narrative is a promising direction for better understanding machine learning data, models and prediction. It is capable of describing concepts of considerable complexity and when accompanied with argumentation theory it can explain them. When combined with state-of-the-art statistical and machine learning models it can vastly improve our understanding of the algorithms that we use, as well as the predictions they produce and the data on which they are built.

References

[Farrell *et al.*, 2015] Rachel Farrell, Gordon Pace, and Michael Rosner. A framework for the generation of computer system diagnostics in natural language using finite state methods. In *Proceedings of the 15th European Workshop on Natural Language Generation, 52–56*, 2015.

[Maaten and Hinton, 2008] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9(Nov):2579–2605, 2008.

[Možina *et al.*, 2007] Martin Možina, Jure Žabkar, and Ivan Bratko. Argument based machine learning. *Artificial Intelligence*, 171(10-15):922–937, 2007.

[Portet *et al.*, 2009] François Portet, Ehud Reiter, Albert Gatt, Jim Hunter, Somayajulu Sripada, Yvonne Freer, and Cindy Sykes. Automatic generation of textual summaries from neonatal intensive care data. *Artificial Intelligence*, 173(7):789–816, 2009.

[3] https://www.automaticstatistician.com