Contextialized and personalized explainations of machine learning algorithms

Christin Seifert, University of Passau, Germany

Since the 1990' prominent inventions of the nearest neighbour algorithm and back-propagation for neural networks machine learning became ubiquitous, though often invisible, in everyday life. Machine-learning based applications range from speech generation to self-driving cars.

Some algorithms are fairly easy to understand, e.g. Decision Trees [15] and nearest neighbor methods [1], but more powerful models, such as neural networks [16] or support vector machines [4], are much more complex and thus much harder to comprehend. While comprehensibility of models is desired for both, experts and lay persons [18, 14], it is also one of the key challenges in data mining [9]. For general public users, comprehension would allow to built trust in the model [18] and therefore lead to higher usage and turnover rates. Prominent (yet simple) examples are keyword-in-context methods in information retrieval [7] and explanations in recommender systems [13]. With enhanced model comprehension, expert users could devise better models and "allow a user to discuss and explain the logic behind the model with colleagues, customers, and other users" [18]. Further, lack of model comprehensibility is one of the reasons why less powerful models are preferred in some application areas [9]. Explanation of simple machine learning models has been addressed by several authors for different model classes, e.g. for Naive Bayes classification [10], Decision Trees [19], simple Neural Networks [11] or based on basic common classifier properties, such as a-posteriori distributions [5].

The problem of model comprehension became even more prominent with the current resurgence of neural networks as deep neural networks (DNNs) in 2006. DNNs have been shown to outperform most state-of-the approaches in almost all application areas and already exhibit near-human performance in some domains [6]. DNNs have an even higher model complexity due to their large number of layers and their inherently distributed representation of knowledge. To address this problem, visualization techniques have already been applied to make the inner workings of DNNS accessible, however, only for expert users and in the domain of computer vision [17]. Thus, in order to exploit the full potential of machine learning applications that become more and more ubiquitous in everyday lives, these applications should be able to explain their reasoning and actions to human users.

Proposed Approach

To conceptualize the problem, consider the following example from the domain of medical diagnosis. When assessing tissue anomalies on x-ray images it is not important to detail the data normalisation and transformation step, but rather indicate the location and reason for detected anomalies. It also might be important to detail the limitations of the data acquisition procedure and potential causes for noise and artefacts. Further, the suitability of the explanation depends on the target group, domain experts (i.e., doctors) need different explanations than lay persons (i.e., patients). Therefore, in order to design suitable explanations [12] the following key questions need to be considered:

1. *What is there to explain?* (algorithm white-box view)
   It is necessary to first have a white-box view on the complete data processing pipeline, the data, the transformations, the features and the algorithms.
2. *What makes sense to explain?* (application scenario)
   The amount of exposure of white-box knowledge depends on the application, as shown in the medical diagnoses example above.
3. *What can be understood?* (user model)
   This is directly influenced by prior knowledge, investable time, implicit questions to the process and knowledge acquisition preferences of users.

4. *How to explain?* (explanation strategy)
   Depending on the application scenario, the user model and the white-box view the concrete explanation strategy can be devised. This step is theoretically grounded on the theory of human understanding.

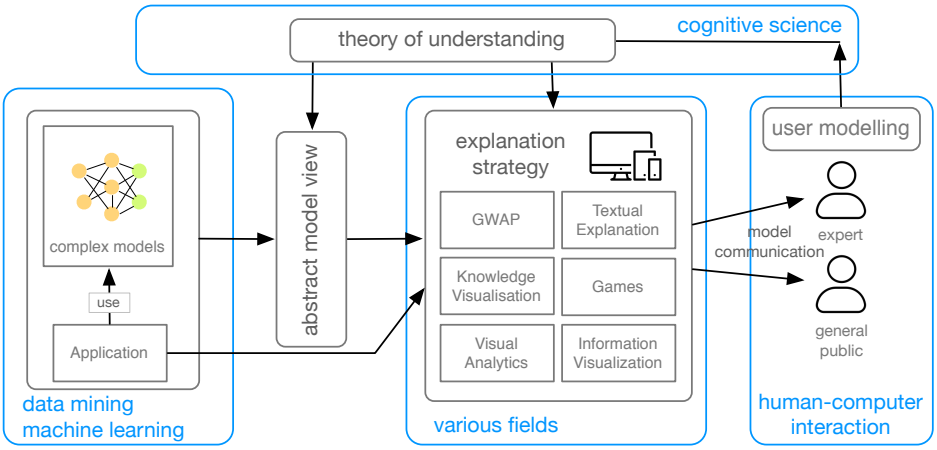

Figure 1: Conceptual overview of Explainable Machine Learning indicating relevant research fields

From these key question the necessary conceptual components can be derived: i) algorithm white-box view, ii) application scenario, iii) user models, iv) explanation strategy, v) theory of human understanding. Figure 1 illustrates the interplay between these conceptual components. Machine learning models are used within an application. A common theory of understanding and explanation guides the abstraction of the model, and the selection of the explanation strategy. User models are derived from different user groups informing the abstraction and selection process. The explanation strategy also depends on the application and on the available end user devises (i.e., their display and interaction capabilities) and encompass for instance games with a purpose (GWAP) [2], Knowledge Visualizations [3] or full Visual Analytics Applications [8].

References

[1] David W. Aha, Dennis Kibler, and Marc K. Albert. "Instance-Based Learning Algorithms". In: *Mach. Learn.* 6.1 (1991), pp. 37–66.

[2] Luis von Ahn. "Games with a Purpose". In: *Computer* 39.6 (2006), pp. 92–94.

[3] Stefan Bertschi et al. "What is Knowledge Visualization? Perspectives on an Emerging Discipline". In: *Proceedings of the 2011 15th International Conference on Information Visualisation*. IV '11. Washington, DC, USA: IEEE Computer Society, 2011, pp. 329–336.

[4] Bernhard E. Boser, Isabelle M. Guyon, and Vladimir N. Vapnik. "A Training Algorithm for Optimal Margin Classifiers". In: *Proceedings of the Fifth Annual Workshop on Computational Learning Theory*. COLT '92. Pittsburgh, Pennsylvania, USA: ACM, 1992, pp. 144–152.

[5] Henrico Dolfing. "A Visual Analytics Framework for Feature and Classifier Engineering". MA thesis. University of Konstanz, 2007.

[6] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. http://www.deeplearningbook.org. MIT Press, 2016.

[7] Marti A. Hearst. *Search User Interfaces*. 1st. New York, NY, USA: Cambridge University Press, 2009.

[8] Daniel A. Keim et al. "Visual Analytics: Combining Automated Discovery with Interactive Visualizations". In: *Discovery Science (DS 2008)*. LNAI. 2008, pp. 2–14.

[9] Ron Kohavi. *Data Mining and Visualization*. Invited talk at the National Academy of Engineering US Frontiers of Engineers (NAE). Sept. 2000.

[10] Todd Kulesza et al. "Principles of Explanatory Debugging to Personalize Interactive Machine Learning". In: *Proceedings of the 20th International Conference on Intelligent User Interfaces*. IUI '15. Atlanta, Georgia, USA: ACM, 2015, pp. 126–137.

[11] Jean-Baptiste Lamy and Rosy Tsopra. "Translating visually the reasoning of a perceptron: the weighted rainbow boxes technique and an application in antibiotherapy". In: *Proc. of International Conference Information Visualisation*. 2017.

[12] Tania Lombrozo. "The structure and function of explanations". In: *Trends in Cognitive Sciences* 10.10 (Oct. 2006), pp. 464–470.

[13] David Mcsherry. "Explanation in Recommender Systems". In: *Artificial Intelligence Review* 24.2 (2005), pp. 179–197.

[14] Brett Poulin et al. "Visual Explanation of Evidence in Additive Classifiers". In: *Proc. Conference on Innovative Applications of Artificial Intelligence (IAAI06)*. Bostan, MA, July 2006, pp. 1822–1829.

[15] J Ross Quinlan et al. *Discovering rules by induction from large collections of examples*. Expert systems in the micro electronic age. Edinburgh University Press, 1979.

[16] David E. Rumelhart, Geoffrey E. Hinton, and Ronald J. Williams. "Neurocomputing: Foundations of Research". In: ed. by James A. Anderson and Edward Rosenfeld. Cambridge, MA, USA: MIT Press, 1988. Chap. Learning Representations by Back-propagating Errors, pp. 696–699.

[17] Christin Seifert et al. "Visualizations of Deep Neural Networks in Computer Vision: A Survey". In: *Transparent Data Mining for Big and Small Data*. Ed. by Tania Cerquitelli, Daniele Quercia, and Frank Pasquale. Cham: Springer, 2017, pp. 123–144.

[18] Kurt Thearling et al. "Information Visualization in Data Mining and Knowledge Discovery". In: ed. by Usama Fayyad, Georges Grinstein, and Andreas Wierse. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 2001. Chap. Visualizing data mining models, pp. 205–222.

[19] Malcolm Ware et al. "Interactive machine learning: letting users build classifiers". In: *International Journal of Human-Computer Studies* 55.3 (2001), pp. 281–292.