TUTORIALS, Session 1: Interpretable Machine Learning

Wojciech Samek, Fraunhofer Institute for Telecommunications, Berlin, Germany & Klaus-R. Müller, TU Berlin, Germany

Abstract: Complex nonlinear models such as deep neural networks (DNNs) have become an important tool for image classification, speech recognition, natural language processing, and various other applications. At the same time, these powerful algorithms are conceived as black box methods, because it is difficult to intuitively and quantitatively understand how and why they arrive at a particular response. This is a major drawback for applications which require human verification, e.g., medical diagnosis. Also in the sciences the interpretability aspect is crucial, because there the goal is not necessarily to predict as accurately as possible, but to better understand the underlying physical or biological processes. From an engineer's perspective interpretability is also a valuable feature, because it helps to identify the strengths and weaknesses of a model.

In this tutorial we will review recent advances on interpretable machine learning, starting from simple linear models which are often applied in neuroscience and ending with state-of-the-art multi-layer neural networks used in computer vision. We will discuss different approaches to interpret DNNs, e.g., work dedicated to visualize particular neurons and work focusing on explaining a given prediction by sensitivity analysis or Taylor decomposition. Finally, we will address the question of how to objectively measure the quality of explanations and how to use this information for improving a given model.