Explaining the predictions of deep neural networks

Gregoire Montavon

Machine learning models such as deep neural networks are able to transform large amounts of data into complex predictive models. In practice, it is desirable not only to achieve high accuracy, but also to explain why the learned model predicts in a certain way, for example, which input features the model uses to support its prediction.

In this talk, approaches such as sensitivity analysis and sum-decomposition of the prediction score will be presented, with a focus on methods based on Taylor decomposition. In particular, the deep Taylor decomposition method, which can be used to explain deep neural networks, will be presented.

Because ground-truth explanations are usually not available, validating an explanation method can be difficult. Emphasis will therefore be placed on how to characterize the explanation problem, and how the proposed method naturally arises from this characterization.