Statistical relevance explanation models and modern methods of interpretable machine learning.

In this work we want to demonstrate how latest machine learning interpretation tools could be based on classical statistical relevance explanation model[1]. SR model is based upon two main ideas – homogenous partition of phenomena properties and statistical inference from properties to phenomena outcome. The latter is an inseparable part of any machine learning algorithm and, in order to adapt SR explanation model, we just need to find relevant partition of phenomena properties.

We argue that lack of explanatory power in neural networks models is caused by irrelevance of the information for explanation (which means that features were not representing homogeneous partition of phenomena) which the model uses.

Let’s take image recognition algorithms, for example. In most versions of the computer vision algorithms, every pixel of the image is used for the recognition task. Pixels of the image are homogeneously partitioned – all images with similar resolution are different in every pixel. However, for objects in images some of the pixels are totally irrelevant – thus explaining image in terms of all pixels could not be a plausible explanation for this particular object recognition.

In this logic, machine learning interpretation tools (LIME[2], SHAP[3]) provide explanations via setting up relevance score for every feature in an already trained model. Setting up threshold and visualizing of only the relevant part for the object imitates homogeneous partition and thus - SR-based explanation. While these tools could be effective for singular explanations, thresholds generated by this tool are proven to be unstable[4]. In order to overcome this instability, there are attempts to generate relevant feature partition during learning. For example, SENN[5] architecture uses additional sub-net with autoencoder for this task.

SR model demonstrated its capability of uniting different interpretability algorithms. Adopting its ideas could help to develop better explanations for machine learning models and, more importantly, to foresee where such explanations would fail thanks to the thorough discussion of this model.

1. Salmon, W., 1971a, ‘Statistical Explanation’, in *Statistical Explanation and Statistical Relevance*, W. Salmon, (ed.), 29–87, Pittsburgh: University of Pittsburgh Press.
2. Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). Model-agnostic interpretability of machine learning. arXiv preprint arXiv:1606.05386.
3. Lundberg, S. M., & Lee, S. I. (2017). A unified approach to interpreting model predictions. In Advances in Neural Information Processing Systems (pp. 4765-4774).
4. Alvarez-Melis, D., & Jaakkola, T. S. (2018). On the robustness of interpretability methods. arXiv preprint arXiv:1806.08049.
5. Alvarez-Melis, D., & Jaakkola, T. S. (2018, December). Towards robust interpretability with self-explaining neural networks. In Proceedings of the 32nd International Conference on Neural Information Processing Systems (pp. 7786-7795). Curran Associates Inc.