Why Explainable AI Should Move from Influence to Contextual Importance and Utility

Kary Främling

Department of Computing Science, Umeå University

Contextual Importance and Utility (CIU) is a method originally developed by Kary Främling in his PhD thesis [1]. CIU's objective was to enable similar explanation concepts for all possible decision support models, ranging from linear models such as the weighted sum, to rule-based systems, decision trees, fuzzy systems, neural networks and any machine learning-based models.

CIU arithmetic is defined using utility functions, which are a cornerstone in Decision Theory and notably in Multi-Attribute Utility Theory. Both Contextual Importance (CI) and Contextual Utility (CU) values are in the range $[0, 1]$ and have absolute (i.e. not relative) interpretations. $CI = 0$ signifies that the studied input feature (or set of features i) cannot change the output value (or its utility, in practice), no matter how much the value of the input feature changes. $CI = 1$ again signifies that the output value can change totally within the possible value range. $CU = 0$ signifies that the current value of the input feature(s) is the worst possible for the output value, whereas $CU = 1$ signifies that it's the best possible value.

Most current outcome explanation methods belong to the family of additive feature attribution methods, which in practice produce what we call an *influence* value ϕ. Contrary to CI and CU, ϕ is a relative value that is more difficult to define and interpret. Furthermore, it can be shown that influence only is incapable of producing any explanation even in simple settings. The presentation shows why CIU should be the method of choice for producing model-agnostic explanations of black-box outcomes. CIU is also defined for so-called intermediate concepts, which makes it possible to structure explanations in a hierarchical way that influence-based methods might not be able to provide.

References

[1] Främling, K.: Modélisation et apprentissage des préférences par réseaux de neurones pour l'aide à la décision multicritère. PhD thesis, INSA de Lyon (1996)