Leila Amgoud - Explaining classifiers under constraints

Abductive explanations (AXp's) are widely used for understanding decisions of classifiers. Existing definitions are suitable when features are independent. However, we show that ignoring constraints when they exist between features may lead to an explosion in the number of redundant or superfluous AXp's. We propose three new types of explanations that take into account constraints and that can be generated from the whole feature space or from a sample (such as a dataset). For each type, we analyse the complexity of finding an explanation and investigate its formal properties. The final result is a catalogue of different forms of AXp's with different complexities and different formal guarantees.

ANITI DAYS

ANITI