Interpretação de Algoritmos de Aprendizado de Máquina

Algoritmos de aprendizado de máquinas e inteligência artificial têm a reputação de serem caixas pretas. De alguma forma, cunhando a ideia de que precisamos renunciar à interpretação do modelo em nome dos ganhos de precisão e acurácia. Tais ideias podem ter ocasionado alguma resistência na aplicação dos algoritmos de aprendizado de máquina em análises econômicas. Embora as *features* de um algoritmo de aprendizado de máquina não sejam tão diretamente interpretáveis quanto os betas de uma regressão linear, ainda assim existem ferramentas que tornam a interpretação possível. Neste trabalho pretendemos revisar os métodos de interpretação de *features importance*, demonstrando que esses podem tornar os algoritmos de aprendizado de máquina tão analisáveis analiticamente quanto os modelos de estatística clássica.

Introdução