Resumo

Interpretabilidade de sistemas de aprendizado de máquina é um característica almejada dada a crescente criticidades das aplicações de tais sistemas, desde diagnóstico médico à controle automatizado de instalações críticas e monitoramento de sistemas, é importante que especialistas de domínios sejam capazes de entender quais fatores levaram um determinado sistema a tomar uma determinada descisão ou chegar em um determinado valor.

Métodos baseados em modelos-proxy podem ser usados para fornecer explicações baseadas em importância de atributos para um dada instância de uma tarefa de aprendizado de máquina, modelos-proxy são modelos mais simples, treinados sobre um conjunto de dados modificado $Y^* : (x_1, \ldots, x_n) \rightarrow y^*$ onde y^* é a *label* atribuída à instância $(x_1, \ldots, x_n)$ pelo modelo sendo analisado, então, passamos a interpretar o modelo-proxy, que passa a ser uma aproximação da fronteira de descisão do modelo original.

Proposta

Baseado no trabalho de Ribeito *et al.*[1] este projeto se propõe a estudar como uma abordagem de modelo-proxy fundamentada no conceito de Informação Mútua pode ser utilizada para interpretar modelos de aprendizado de máquina. Para isso, diferentes modelos serão treinados em tarefas de classificação, em diferentes conjuntos de dados e explicações serão geradas para instâncias aleatórias, essas explicações serã comparadas com explicações do geradas por LIME [1] com o objetivo de comparar as diferenças decorrentes do uso de um modelo que não assume lineraridade nas vizinhanças da instâncias.

Referências

[1] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. "why should I trust you?": Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, August 13-17, 2016*, pages 1135–1144, 2016.