Interpretación de Modelos de Machine Learning para Texto: Un caso de Estudio con Documentos Legales

Jorge Poco
FGV, Brasil
jorge.poco@fgv.br

Machine Learning ha revolucionado la tecnología del lenguaje en los últimos años, y constituye el estado del arte en dominios que van desde la traducción automática y la respuesta a preguntas hasta el reconocimiento del habla y la generación de música. A pesar de su adopción generalizada, los modelos de aprendizaje automático siguen siendo en su mayoría cajas negras. Con esa potencia y popularidad, surgen nuevas responsabilidades y preguntas: ¿cómo garantizamos la fiabilidad, evitamos sesgos indeseables y proporcionamos información sobre cómo un sistema llega a un resultado concreto? ¿Cómo aprovechar los conocimientos especializados y las opiniones de los usuarios para mejorar aún más los modelos? En todas estas cuestiones, la "interpretabilidad" de los modelos de aprendizaje profundo es clave. En esta charla vamos a definir el problema de interpretación de modelos de Machine Learning para textos. Además, describiremos los fundamentos matemáticos para este tipo de técnicas, en específico la técnica Local Interpretable Model-agnostic Explanations (LIME). Finalmente mostraré un caso de estudio usando documentos de precedentes jurídicos del Supremo Tribunal Federal de Brasil.