Descripción del Problema

Kevin Manzano Rodríguez
Roger Fuentes Rodríguez

1. Introducción

En este trabajo, exploramos un modelo de aprendizaje automático (ML) que recibe grafos como entrada, denominado caja negra M. Específicamente, nos enfocamos en el problema de búsqueda de contrafactuales, que consiste en encontrar una perturbación G' a partir de un grafo inicial G tal que las salidas del modelo para ambas entradas sean diferentes.

2. Definición del Problema

Nuestro objetivo es identificar contrafactuales "buenos", es decir, aquellos que son cercanos a G. Formalmente, dado un modelo M y un grafo G, buscamos devolver G' lo más cercano posible a G cumpliendo con la condición de que las salidas del modelo para G y G' sean distintas.

3. Complejidad del Problema

Este problema es intrínsecamente complicado, similar a los problemas NP-completos en términos de decisión. En el contexto de optimización, se clasifica como NP-duro.

4. Aplicaciones Prácticas

Las aplicaciones son diversas; por ejemplo, en redes sociales donde los algoritmos de recomendación pueden generar sugerencias inesperadas. La búsqueda de contrafactuales permite entender cómo pequeñas modificaciones pueden alterar significativamente las recomendaciones ofrecidas. Además, el estudio de contrafactuales proporciona información sobre la robustez y confiabilidad de las predicciones del modelo, permitiendo evaluar hasta qué punto se puede modificar un objeto sin afectar su resultado final.