SEMINARIO

Mario Villaizán Vallelado

Universidad de Valladolid

MTabGen: Modelo de difusión para la imputación y generación de datos tabulares. Un enfoque basado en una atención condicionada y el poder de los Transformers

Abstract: La imputación de datos y la generación de datos sintéticos son tareas fundamentales en diversos dominios, como la salud y las finanzas, donde los datos incompletos o faltantes pueden comprometer la precisión de los análisis y la toma de decisiones. En este contexto, los modelos de difusión han surgido como potentes modelos generativos capaces de capturar distribuciones complejas de datos en modalidades como imágenes, audio y series temporales. Recientemente, estos modelos han sido adaptados para abordar los desafíos específicos de los datos tabulares.

MTabGen presentará un modelo de difusión diseñado específicamente para datos tabulares, el cual introduce tres innovaciones clave: (1) un mecanismo de atención condicionada, (2) una red transformer encoder-decoder para la tarea de eliminación de ruido, y (3) un sistema de enmascarado dinámico. El mecanismo de atención condicionada mejora la capacidad del modelo para capturar las relaciones entre los datos de condición y los datos sintéticos generados, mientras que las capas transformer permiten modelar las interacciones dentro de los datos condicionados (en el encoder) y los datos sintéticos (en el decoder). El enmascarado dinámico, por su parte, permite que el modelo aborde eficientemente tanto la imputación de datos faltantes como la generación de datos sintéticos dentro de un marco unificado.

El rendimiento de MTabGen ha sido comparado con el rendimiento de otras técnicas de vanguardia, como los Autoencoders Variacionales (VAEs), las redes generativas adversarias (GANs) y otros modelos de difusión, utilizando conjuntos de datos de referencia. La evaluación se centrará en tres criterios fundamentales: (1) la eficiencia en tareas de aprendizaje automático, (2) la similitud estadística de los datos generados respecto a los datos reales, y (3) la mitigación de riesgos de privacidad.

Este enfoque innovador permite abordar de manera robusta los problemas de imputación y generación de datos tabulares, ofreciendo una solución integrada y adaptable a múltiples aplicaciones en contextos críticos como la salud y las finanzas.

Seminario del Departamento de Estadística e Investigación Operativa
8 de Noviembre de 2024 (11:00)
Organiza: Departamento de Estadística e Investigación Operativa