Actividad N° 3

Docente: Fred Torres Cruz
Estudiante: Waldir Yeison Velásquez Quispe

TÉCNICAS DE ESCALADO EN CONJUNTOS DE DATOS CON VARIABLES HETEROGÉNEAS

1. Introducción

La elección de una técnica de escalado adecuada en conjuntos de datos con variables heterogéneas es crucial para el éxito de los modelos de aprendizaje automático y el análisis de datos. La heterogeneidad de los datos se refiere a la presencia de variables con diferentes rangos, unidades de medida o distribuciones, lo que puede afectar negativamente el rendimiento de muchos algoritmos.

2. Factores a Considerar

Al abordar la heterogeneidad de los datos, es fundamental considerar varios factores.

Primero, la naturaleza de las variables es primordial. Las variables pueden ser numéricas, categóricas, ordinales, o incluso datos no estructurados como texto o imágenes. Cada tipo de variable puede requerir un enfoque de escalado diferente. Por ejemplo, las variables numéricas a menudo se benefician de la normalización o estandarización, mientras que las variables categóricas pueden necesitar codificación one-hot o técnicas de incrustación [1].

Segundo, el algoritmo de aprendizaje automático a utilizar influye directamente en la elección de la técnica de escalado. Algunos algoritmos, como las máquinas de vectores de soporte (SVM) o las redes neuronales, son sensibles a la escala de las características y requieren que los datos estén escalados para converger eficientemente y evitar que las características con rangos más grandes dominen el proceso de aprendizaje [2]. Otros algoritmos, como los árboles de decisión, son menos sensibles a la escala.

Tercero, la distribución de los datos es un factor importante. Si los datos siguen una distribución normal, la estandarización (escalado Z-score) puede ser apropiada. Sin embargo, si los datos tienen una distribución sesgada o contienen valores atípicos, técnicas como el escalado robusto (RobustScaler) o la transformación de cuantiles (QuantileTransformer) pueden ser más adecuadas, ya que son menos sensibles a los valores extremos [3].

Cuarto, la preservación de la información es crucial. Algunas técnicas de escalado pueden alterar la distribución original de los datos o la relación entre las variables. Es importante seleccionar una técnica que preserve las propiedades esenciales de los datos relevantes para el problema en cuestión.

3. Conclusión

En resumen, la selección de una técnica de escalado para conjuntos de datos con variables heterogéneas es una decisión multifacética que depende de la naturaleza de las variables, el algoritmo de aprendizaje automático, la distribución de los datos, la necesidad de preservar la información y la interpretabilidad del modelo. Una consideración cuidadosa de estos factores es esencial para optimizar el rendimiento del modelo y garantizar resultados robustos y significativos.

4. Referencias

[1] G. Kumar, S. Basri, A. A. Imam, S. A. Khowaja, L. F. Capretz, "Data harmonization for heterogeneous datasets: a systematic literature review,.ªpplied Sciences, vol. 11, no. 17, p. 8275, 2021.

[2] E. G. Radhika, G. S. Sadasivam, .ª review on prediction based autoscaling techniques for heterogeneous applications in cloud environment,"Materials Today: Proceedings, vol. 46, pp. 6745-6750, 2021.

[3] S. Borowicc, S. Alves-Souza, "Heterogeneous Data Integration: A Literature Scope Review,ïn 26th International Conference on Enterprise Information Systems (ICEIS 2024), 2024.