Tarea 7 Aprendizaje no supervisado

Universidad Autonoma de Nuevo León
Alanis Mares, Victor Hugo

18 de Julio, 2024

1 Introducción

Se pretende analizar un conjunto de series de tiempo con diferentes indicadores macroeconomicos para implementar un sistema de clasificacion no supervisada.

2 Descripción de los datos

Los datos a trabajar cuentan con 16 variables de las cuales 11 son referentes a la serie de tiempo multivariada y 5 son categoricas. Para propositos de este estudio de clusterizacion se utilizaran las 11 variables numericas, las cuales son:

- Year: Año de observacion
- Under_five_deaths: Cantidad de muertes en infantes menores de 5 años por cada 1000 habitantes
- Adult_mortality: Cantidad de muertes en adultos por cada 1000 habitantes
- Alcohol_consumption: Litros anualizados de consumo de alcohol por capita
- Hepatitis_B: Cantidad porcentual de niños de 1 año inmunizados con la vacuna para la Hepatitis B
- BMI: Indice de grasa corporal promedio
- Polio: Cantidad porcentual de niños de 1 año inmunizados con la vacuna para la Polio
- Diptheria: Cantidad porcentual de niños de 1 año inmunizados con la vacuna para tetanos
- Incidents_HIV: Casos de VIH por cada 1000 habitantes
- GDP_per_capita: Producto interno bruto per capita

- Schooling: Cantidad de años promedio que un adulto (25+) ha pasado estudiando
- Life_expectancy: expectativa de vida al nacimiento

2.1 Origen de los datos

Los datos fueron obtenidos de un registro público de conjuntos de datos, los cuales provienen de estudios de múltiples hospitales, a continuación se encuentra el enlace al conjunto de datos .

2.2 Preprocesamiento

Basado en un analisis de caracteristicas se decidio omitir ciertas variables del conjunto, para más informacion al respecto referirse a este repositorio

3 Metodologia

3.1 Aprendizaje no supervisado

Para este analisis se usara un algoritmo de gas neuronal creciente (o GNS por sus siglas en ingles) partiendo de la libreria neupy para la aplicacion del algoritmo, numpy para algebra matricial necesaria y pandas para manejo de datos.

3.1.1 Descripción del algoritmo GNS

Citado por primera vez en 1995 el algoritmo GNS parte de vectores con la misma dimensionalidad y busca generar un grafo el cual explique la distribucion de los puntos al mismo tiempo que minimiza el error.

3.1.2 Fundamento matematico

Partiendo de 2 puntos aleatorios conectados entre si se sigue el siguiente algoritmo en cada iteracion con cada vector ingresado por el conjunto datos:

- 1: Se calcula la distancia entre el vector ingresado y los dos nodos más cercanos
- 2: El error del nodo más cercano es sumado al error del vector (el cual inicializa en 0)
- 3: El nodo más cercano y todo nodo conectado a el es movido en funcion de los errores acumulados de cada nodo, en direccion del vector ingresado.
- 4: La variable "edad" de cada arista ya existente se aumenta en 1.
- 5: En caso de que el nodo más cercano y el segundo nodo más cercano esten conectados entre si su arista es igualada a 0 , en caso contrario se crea una arista entre ambos nodos.

- 6: Se comprueba la edad de cada arista, en caso de que sea mayor a una variable previamente decidida se elimina el arista.
- 7: En caso de que la iteracion actual sea multiplo de una variable previamente decida se agrega un nodo al azar al sistema conectando el nodo con el error más grande y su vecino más cercano, despues se elimina el arista que conecta a ambos nodos.
- 8: Se decrementa el error acumulado de todos los nodos en el sistema por un factor constante
- 9: Se repite el proceso hasta que el algoritmo haya llegado a las iteraciones previamente definidas

3.2 Aprendizaje supervisado

3.2.1 Metodología

Para este análisis se tomará como variable objetivo la expectativa de vida para cada país, se utilizara un regresor de potenciación de gradiente apoyándonos de la librería `skforecast`. La razón principal de elección de la librería previamente mencionada es su método `ForecasterAutoregMultiSeries` el cual implementa un regresor de potenciación de gradiente multivariado, este método recibe dos argumentos:

- Un conjunto de series de tiempo referentes a nuestra variable objetivo.
- Un conjunto de series de tiempo referentes a variables exógenas a la variable objetivo las cuales se presuponen afectan a nuestra variable objetivo.

Debido a la baja cantidad de observaciones por serie de tiempo se utilizara un regresor que solo toma en cuenta las ultimas 5 observaciones de la serie de tiempo.

References

[1] Bernd Fritzke (1995) A Growing Neural Gas Network Learns Topologies.

[2] Halil Ertan (2022) Multivariate Time Series Clustering Using Growing Neural Gas and Spectral Clustering

[3] Joaquín Amat Rodrigo, Javier Escobar Ortiz (2022) Modelos de forecasting globales: modelado de múltiples series temporales con machine learning