Instituto Politécnico Nacional

Escuela Superior de Cómputo

Práctica 4

Proyección de datos con método de Sammon, Autocodificadores y análisis espectral de datos periódicos.

Alumno: López Fabián Jesús Manuel
Profesor: Flores Estrada Ituriel Enrique
Materia: Análitica y visualización de datos
Grupo: 5AV1

13 de junio de 2024

1. Transformación de datos.

Convertir las dimensiones categóricas binarias en dimensiones numéricas.

Se realizo una función que identifica la cantidad de valores únicos en una dimensión y si la cantidad es igual a dos entonces se considera una dimensión categórica binaria. Entonces, bajo estas reglase se identificaron las siguientes dimensiones binarias:

- gender
- private
- freepoor
- freerepat
- nchronic
- lchronic

De estas seis dimensiones binarias la variable 'gender' tiene los valores de 'female' y 'male'. En cambio, las otras dimensiones tienen valores de 'yes' y 'no'.

	rownames	visits	gender	age	income	illness	reduced	health	private	freepoor	freerepat	nchronic	lchronic
0	1	1	female	0.19	0.55	1	4	1	yes	no	no	no	no
1	2	1	female	0.19	0.45	1	2	1	yes	no	no	no	no
2	3	1	male	0.19	0.90	3	0	0	no	no	no	no	no
3	4	1	male	0.19	0.15	1	0	0	no	no	no	no	no
4	5	1	male	0.19	0.45	2	5	1	no	no	no	yes	no

Figura 1: Conjunto de datos previo a la transformación de datos.

	rownames	visits	gender	age	income	illness	reduced	health	private	freepoor	freerepat	nchronic	lchronic
0	1	1	0	0.19	0.55	1	4	1	1	0	0	0	0
1	2	1	0	0.19	0.45	1	2	1	1	0	0	0	0
2	3	1	1	0.19	0.90	3	0	0	0	0	0	0	0
3	4	1	1	0.19	0.15	1	0	0	0	0	0	0	0
4	5	1	1	0.19	0.45	2	5	1	0	0	0	1	0

Figura 2: Conjunto de datos después a la transformación de datos.

2. Creación de nuevo espacio dimensional a través de método de Sammon, visualización de nuevo espacio dimensional, medición de tiempo necesario para realizar la proyección, y despliegue del resultado de la función de estrés.

El mapeo de Sammon es una técnica de reducción de dimensionalidad utilizada para visualizar datos de alta dimensión en un espacio de menor dimensión, generalmente en dos o tres dimensiones.

El principio fundamental del mapeo de Sammon es mantener las distancias relativas entre los puntos en el espacio original tanto como sea posible en el espacio reducido. En otras palabras, intenta preservar las relaciones de proximidad entre los puntos de datos.

2.1. Utilizando en el método los datos originales.

Antes de aplicar el mapeo de Sammon es necesario que en el conjunto de datos no existan elementos duplicados ya que al ser datos exactamente identidicos la distancia entre ellos seria cero. Adicionalmente

se elimino la dimensión 'rownames' ya que al ser un identificador único para cada observación no aporta información al análisis.

El tiempo necesario para realizar la proyección fue de 8 minutos y 7 segundos, adicionalmente se obtuvo un estrés de 0.015

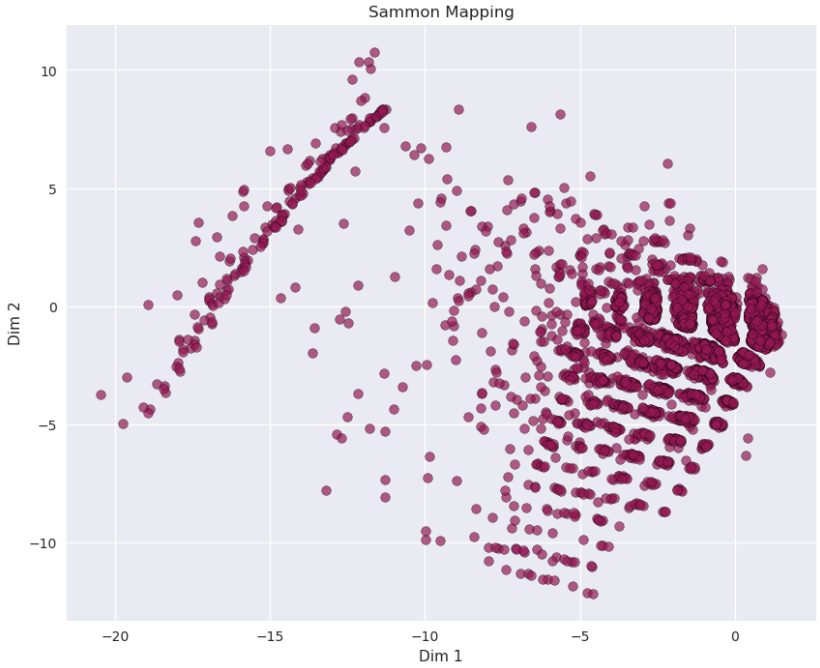

Figura 3: Mapeo de Sammon utilizando el conjunto de datos originales.

2.2. Utilizando en el método los datos transformados a través de PCA.

Para que el mapeo de Sammon utilice los datos transformados a través de PCA debemos de indicar 'pca' en el parametro 'init' de la función sammon().

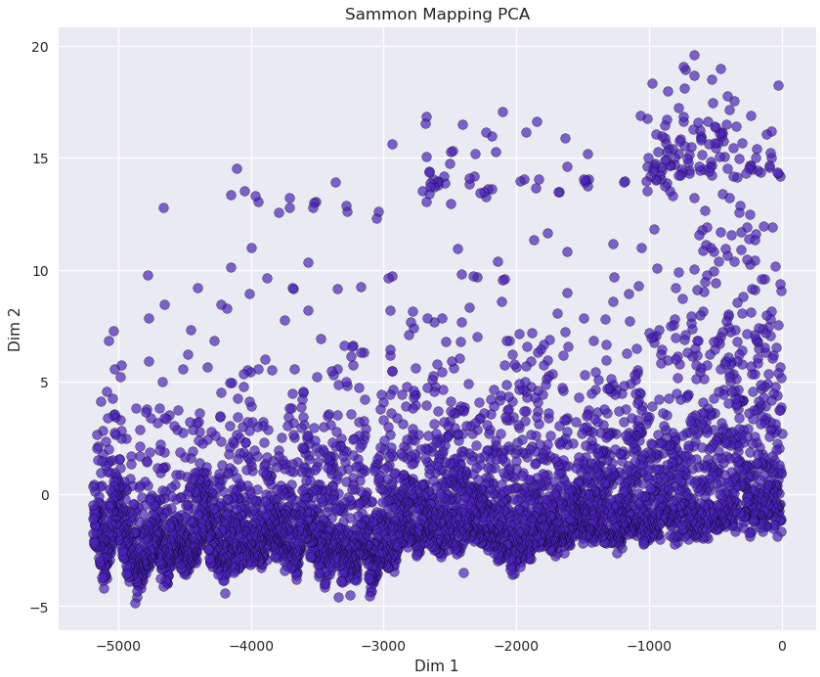

Figura 4: Mapeo de Sammon utilizando el conjunto de datos al aplicarle PCA.

El tiempo necesario para realizar la proyección fue de 7 minutos y 8 segundos, adicionalmente se obtuvo un estrés de $1{,}01 \times 10^{-7}$

3. Conclusiones.

Cuando se utiliza PCA para inicializar el mapeo de Sammon, se obtiene una estimación inicial basada en componentes principales. PCA es una técnica lineal de reducción de dimensionalidad que intenta preservar la mayor varianza posible en las primeras componentes. Al proporcionar estas coordenadas iniciales, el mapeo de Sammon tiene una buena base desde la cual comenzar, lo que puede mejorar la convergencia y reducir el número de iteraciones necesarias. La visualización inicial con PCA suele estar cerca del resultado final, ya que PCA ya intenta preservar la estructura de los datos. Por lo tanto, el algoritmo de Sammon puede realizar ajustes más finos desde esta base bien formada, resultando en una visualización que refleja mejor la estructura de los datos con menor distorsión.

En contraste, cuando el mapeo de Sammon se inicializa sin PCA, las coordenadas iniciales son iguales al conjunto de datos inicia. En este caso, la convergencia puede ser más lenta, ya que el algoritmo de Sammon tiene que encontrar una estructura óptima desde una base menos informada. La visualización inicial puede diferir significativamente del resultado final, y el algoritmo puede necesitar más iteraciones para alcanzar una representación que preserve adecuadamente las distancias originales. Aunque finalmente puede converger a una solución similar, el proceso puede ser menos eficiente y requerir más ajustes iterativos para reducir el error de Sammon (stress).

Al comparar las coordenadas reducidas obtenidas con y sin PCA, observamos diferencias importantes. La inicialización con PCA generalmente conduce a una mayor precisión desde el comienzo, ya que las coordenadas iniciales están más alineadas con la estructura de los datos. Además, el uso de PCA puede mejorar la eficiencia del algoritmo de Sammon, reduciendo el número de iteraciones y el tiempo de cómputo necesario para converger. Aunque ambas técnicas eventualmente pueden producir resultados similares, la calidad inicial y el ajuste fino tienden a ser mejores cuando se utiliza PCA, dado que PCA ya proporciona una buena base que preserva gran parte de la varianza de los datos originales.