Aplicación de Modelos PCA y SOM

Quiñones Mayorga Rodrigo

INTRODUCCIÓN

En el análisis de datos, es común encontrarse con conjuntos de datos que tienen un gran número de características, lo que puede dificultar la visualización y comprensión de los patrones subyacentes. Dos técnicas populares para abordar este problema son el Análisis de Componentes Principales (PCA) y los Mapas Autoorganizados (SOM). Ambos métodos permiten reducir la complejidad de los datos, aunque lo hacen de formas distintas y con propósitos diferentes.

La primera es PCA, una técnica que reduce la dimensionalidad que transforma variables en componentes. Estos componentes están ordenados de tal forma que el primero explica la mayor cantidad de varianza en los datos, seguido del segundo, y así sucesivamente. Al reducir la dimensionalidad, PCA facilita la identificación de patrones en los datos al tiempo que preserva la mayor parte de la información posible.

Por otro lado, los **Mapas Autoorganizados (SOM)** son un tipo de red neuronal no supervisada que organiza los datos en una representación topológica. Este modelo agrupa observaciones similares en un mapa visual, lo que facilita la interpretación de las relaciones entre las distintas muestras de datos.

Lo primero a realizas fue buscar dos conjuntos de datos, en este caso el primero habla de los datos socio-economicos de una población con respecto al crimen, mientras que el segundo nos habla del ingreso anual de adultos que recibirán mas de 50K al año.

PCA

Para aplicar el modelo PCA, primero fue necesario preparar los datos adecuadamente. Se seleccionaron las variables numéricas relevantes del conjunto de datos "adult.data":

- age (edad),
- education-num (número de años de educación),
- capital-gain (ganancia de capital),
- capital-loss (pérdida de capital),
- hours-per-week (horas trabajadas por semana).

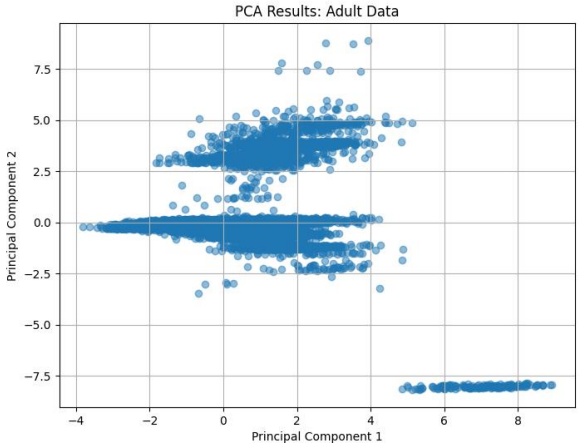
Dado que PCA es sensible a la escala de las variables, fue fundamental estandarizar los datos. La estandarización garantiza que todas las variables tengan una media de 0 y una desviación estándar de 1, evitando que alguna variable con mayor escala domine los componentes principales.

Para lograr esto, se utilizó el método StandardScaler de la biblioteca scikit-learn, que transforma los datos para que todas las variables numéricas tengan la misma influencia en el modelo PCA.

Una vez que los datos fueron estandarizados, se aplicó el modelo PCA con el objetivo de reducir la dimensionalidad a **dos componentes principales**.

Tras aplicar el modelo, se obtuvieron dos componentes principales (PC1 y PC2), que son combinaciones lineales de las variables originales. Estos componentes capturan la mayor parte de la varianza presente en los datos, reduciendo las cinco dimensiones originales a solo dos.

Para visualizar los resultados, se generó un gráfico de dispersión (scatter plot) que muestra cada observación proyectada en los ejes definidos por los dos primeros componentes principales. En el gráfico, cada punto representa una observación del conjunto de datos transformada en este nuevo espacio de características.



El gráfico muestra posibles agrupamientos o patrones en los datos, lo que indica que las variables seleccionadas tienen una influencia fuerte en la estructura de los datos.

El grupo en la esquina inferior derecha podría representar un subconjunto de la población que es muy diferente del resto, probablemente en términos de variables como ingreso, edad, o nivel de educación.

Las bandas horizontales sugieren que algunas variables están correlacionadas, lo que ha sido capturado por PC1 y PC2.

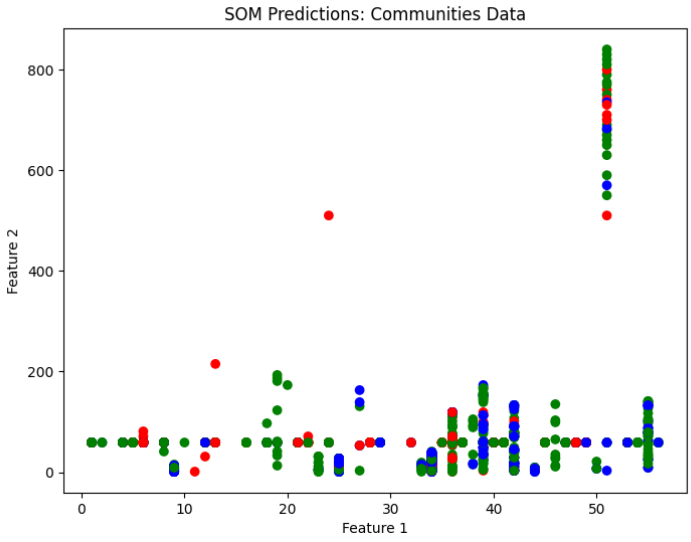
SOM

El primer paso consistió en cargar el archivo de datos "communities.data" y reemplazar los valores faltantes (representados por ?) por NaN, lo que permite identificarlos y procesarlos correctamente. El conjunto de datos contenía valores no numéricos y valores faltantes que impedían un análisis directo. Utilicé un imputador para reemplazar estos valores faltantes con la media de las columnas correspondientes. Este proceso garantiza que no se eliminen demasiadas observaciones y permite que SOM pueda trabajar con un conjunto de datos completo.

Antes de aplicar el modelo SOM, estandaricé los datos. Esto significa que todas las características numéricas se transformaron para tener una media de 0 y una desviación estándar de 1, utilizando el StandardScaler de sklearn. La estandarización es un paso importante porque SOM es sensible a las escalas de las características, y sin estandarización, las características con mayores rangos podrían dominar el entrenamiento.

El Mapa Autoorganizado (SOM) es un tipo de red neuronal no supervisada que organiza datos en un mapa topológico, agrupando puntos de datos similares entre sí. Para este análisis, utilicé un SOM de 3x1, lo que significa que intentamos agrupar las observaciones en tres categorías o clusters. El número de dimensiones se ajustó automáticamente al número de características disponibles en los datos.

Después de entrenar el SOM, obtuve las predicciones, que asignan cada observación del conjunto de datos a uno de los tres grupos definidos en la cuadrícula. Visualicé estas predicciones utilizando un gráfico de dispersión (scatter plot), donde las dos primeras características numéricas del conjunto de datos están representadas en los ejes X e Y. Los diferentes colores (rojo, verde, azul) representan los grupos asignados por el SOM.



Podemos observar como la mayoría de los puntos se agrupan en la parte inferior mientras un grupo es el único que se concentra en la superior sin embargo en el mismo eje x.

En la variable feature 1 tenemos el estado, mientras que en la 2 es el condado.