COLOMBIA POTENCIA DE LA VIDA

TIC

UNIVERSIDAD SERGIO ARBOLEDA

Análisis de Datos
Nivel Integrador

TALENTO TECH

UNIVERSIDAD SERGIO ARBOLEDA

PREPARACIÓN SEMANA 9

TÉCNICAS DE ANÁLISIS EXPLORATORIO AVANZADAS

Dentro de las técnicas de análisis exploratorio avanzadas tenemos:

- PCA: Análisis de componentes principales
- Análisis de clúster
- Análisis de series temporales
- Análisis de corresponsales múltiples (MCA)

¿QUÉ ES PCA?

El Análisis de Componentes Principales (PCA) es una técnica estadística utilizada para reducir la dimensionalidad de un conjunto de datos, conservando al mismo tiempo la mayor cantidad posible de su variabilidad. El objetivo principal de PCA es transformar las variables originales en un nuevo conjunto de variables (llamadas componentes principales) que sean ortogonales entre sí y que capturen la mayor variabilidad posible.

Aplicaciones en el Análisis de Datos:

1. **Reducción de Dimensionalidad:** PCA es utilizado para reducir la cantidad de características en un conjunto de datos, manteniendo la información esencial.
2. **Visualización de Datos:** Permite visualizar datos de alta dimensión en gráficos de dos o tres dimensiones, facilitando la interpretación.
3. **Eliminación de Correlaciones:** Ayuda a eliminar la multicolinealidad al transformar variables correlacionadas en componentes no correlacionados.
4. **Simplificación de Modelos:** En el análisis predictivo, PCA puede ser usado para simplificar modelos al trabajar con un conjunto de características reducido.

¿QUÉ ES PCA?

En el contexto de PCA, los componentes principales representan nuevas variables que son combinaciones lineales de las variables originales. Estas nuevas variables están diseñadas de manera que capturen la máxima varianza presente en el conjunto de datos. Los componentes principales están ordenados de manera descendente en función de la cantidad de varianza que explican.

1. Primer Componente Principal (PC1):

- **Definición:** Es la combinación lineal de variables originales que maximiza la varianza.
- **Importancia:** PC1 explica la mayor parte de la variabilidad presente en los datos.

2. Segundo Componente Principal (PC2):

- **Definición:** Es la siguiente combinación lineal que es ortogonal (no correlacionada) con PC1 y maximiza la varianza restante.
- **Importancia:** PC2 captura la varianza no explicada por PC1.

¿QUÉ ES PCA?

Relación con la Varianza de los Datos:

- **Varianza Total:** La suma de las varianzas explicadas por cada componente principal es igual a la varianza total de los datos originales.
- **Porcentaje de Varianza Explicada:** Cada componente principal tiene asociado un porcentaje de varianza explicada, que indica cuánto contribuye ese componente a la variabilidad total. Al sumar estos porcentajes, se puede determinar cuánta varianza se conserva al considerar los primeros k componentes principales.
- **Selección de Componentes Principales:** En la práctica, a menudo se selecciona un número suficiente de componentes principales para conservar un alto porcentaje (por ejemplo, el 95%) de la varianza total, lo que permite una reducción significativa de la dimensionalidad sin perder información crítica.

¿QUÉ ES PCA?

- **PROCESO DE PCA**

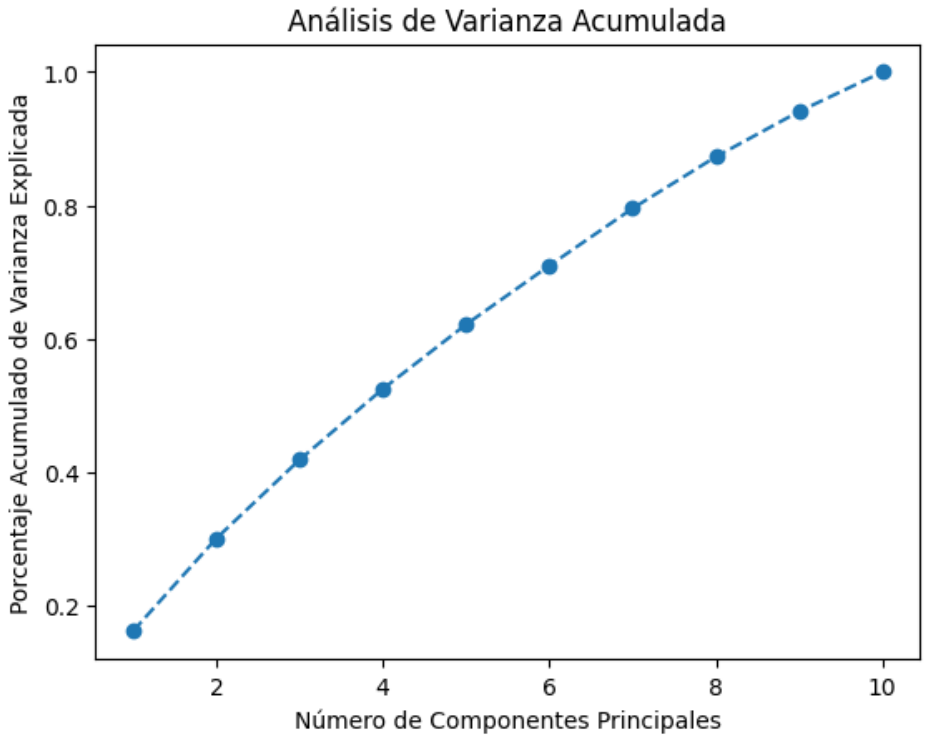
1. **Estandarización de Datos:** Es crucial estandarizar los datos antes de aplicar PCA para asegurar que todas las variables tengan la misma escala.
2. **Cálculo de la Matriz de Covarianza:** Se calcula la matriz de covarianza para entender cómo las variables originales se relacionan entre sí.
3. **Obtención de Valores y Vectores Propios:** Se encuentran los valores y vectores propios de la matriz de covarianza, que representan la dirección y magnitud de la máxima variabilidad.
4. **Selección de Componentes Principales:** Se seleccionan los componentes principales basados en la proporción de varianza explicada que aportan.

¿QUÉ ES PCA?

Ejemplo: Este ejemplo muestra cómo calcular y visualizar el porcentaje acumulado de varianza explicada por los componentes principales, ayudando a decidir cuántos componentes seleccionar para conservar una cantidad significativa de información

```python
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
from sklearn.decomposition import PCA

# Crear un conjunto de datos simulado con 10 variables
np.random.seed(42)
data = np.random.rand(100, 10)

# Convertir a un DataFrame de Pandas para mayor comodidad
df = pd.DataFrame(data, columns=[f'Variable_{i}' for i in range(1, 11)])

# Análisis de Componentes Principales (PCA)
pca = PCA()
pca.fit(df)

# Varianza explicada por cada componente principal
varianza_explicada = pca.explained_variance_ratio_

# Porcentaje acumulado de varianza explicada
porcentaje_acumulado_varianza = np.cumsum(varianza_explicada)

# Gráfico de la varianza explicada acumulada
plt.plot(range(1, len(porcentaje_acumulado_varianza)+1), porcentaje_acumulado_varianza, marker='o', linestyle='--')
plt.xlabel('Número de Componentes Principales')
plt.ylabel('Porcentaje Acumulado de Varianza Explicada')
plt.title('Análisis de Varianza Acumulada')
plt.show()
```

¿QUÉ ES PCA?



Recuerda que el código de todos los ejemplos que mostremos acá lo puedes encontrar en: https://acortar.link/1IvmmS

SELECCIÓN DE COMPONENTES PRINCIPALES

En la selección de componentes principales es crucial determinar el número óptimo de componentes a retener para lograr un equilibrio entre la reducción de dimensionalidad y la retención de información esencial. Aquí hay algunos criterios y técnicas comunes:

1. **Porcentaje de Varianza Explicada:**

- **Criterio:** Seleccionar un número suficiente de componentes para explicar un alto porcentaje de la varianza total.
- **Implementación:** Visualizar el porcentaje acumulado de varianza explicada y seleccionar un umbral, por ejemplo, el 95% o 99%.

2. **Codo en el Gráfico de Varianza Explicada:**

- **Criterio:** Buscar el "codo" en el gráfico de varianza explicada acumulada.

Implementación: Graficar la varianza explicada acumulada y seleccionar el punto donde la ganancia adicional disminuye significativamente (codo).

SELECCIÓN DE COMPONENTES PRINCIPALES

3. **Autovalores y Autovalores Proporcionales:**

- **Criterio:** Observar los autovalores (eigenvalues) y los autovalores proporcionales.
- **Implementación:** Los autovalores indican la cantidad de varianza explicada por cada componente. Seleccionar componentes con autovalores significativos.

4. **Criterios de Retención de Información:**

- **Criterio:** Establecer un umbral para la retención de información.
- **Implementación:** Utilizar criterios como el porcentaje acumulado de varianza explicada o el error de reconstrucción.

IMPLEMENTACIÓN EN PYTHON

En este ejemplo se muestra una visión completa de como realizar y entender un análisis de componentes principales(PCA) en Python, el cual realiza:

1. Se crea un conjunto de datos simulado con 20 variables y 1000 observaciones.
2. Se realiza un análisis de componentes principales (PCA) sobre el conjunto de datos.
3. Se visualiza la varianza explicada acumulada para determinar el número óptimo de componentes.
4. Se determina el número de componentes necesarios para explicar al menos el 95% de la varianza.
5. Se transforma el conjunto de datos original al nuevo espacio de características reducido usando el número óptimo de componentes.
6. Se muestran las primeras filas del conjunto de datos transformado.

IMPLEMENTACIÓN EN PYTHON

VEAMOS EL PASO A PASO, 1:

```
# Crear un conjunto de datos simulado con 20 variables y 1000 observaciones
np.random.seed(42)
data = np.random.rand(1000, 20)

# Convertir a un DataFrame de Pandas para mayor comodidad
df = pd.DataFrame(data, columns=[f'Variable_{i}' for i in range(1, 21)])
```

```
Primeras filas del DataFrame:
   Variable_1  Variable_2  Variable_3  Variable_4  Variable_5  Variable_6  \
0    0.374540    0.950714    0.731994    0.598658    0.156019    0.155995
1    0.611853    0.139494    0.292145    0.366362    0.456070    0.785176
2    0.122038    0.495177    0.034389    0.909320    0.258780    0.662522
3    0.388677    0.271349    0.828738    0.356753    0.280935    0.542696
4    0.863103    0.623298    0.330898    0.063558    0.310982    0.325183

   Variable_7  Variable_8  Variable_9  Variable_10  Variable_11  Variable_12  \
0    0.058084    0.866176    0.601115     0.708073     0.020584     0.969910
1    0.199674    0.514234    0.592415     0.046450     0.607545     0.170524
2    0.311711    0.520068    0.546710     0.184854     0.969585     0.775133
3    0.140924    0.802197    0.074551     0.986887     0.772245     0.198716
4    0.729606    0.637557    0.887213     0.472215     0.119594     0.713245

   Variable_13  Variable_14  Variable_15  Variable_16  Variable_17  \
0     0.832443     0.212339     0.181825     0.183405     0.304242
1     0.065052     0.948886     0.965632     0.808397     0.304614
2     0.939499     0.894827     0.597900     0.921874     0.088493
3     0.005522     0.815461     0.706857     0.729007     0.771270
4     0.760785     0.561277     0.770967     0.493796     0.522733

   Variable_18  Variable_19  Variable_20
0     0.524756     0.431945     0.291229
1     0.097672     0.684233     0.440152
2     0.195983     0.045227     0.325330
3     0.074045     0.358466     0.115869
4     0.427541     0.025419     0.107891
```

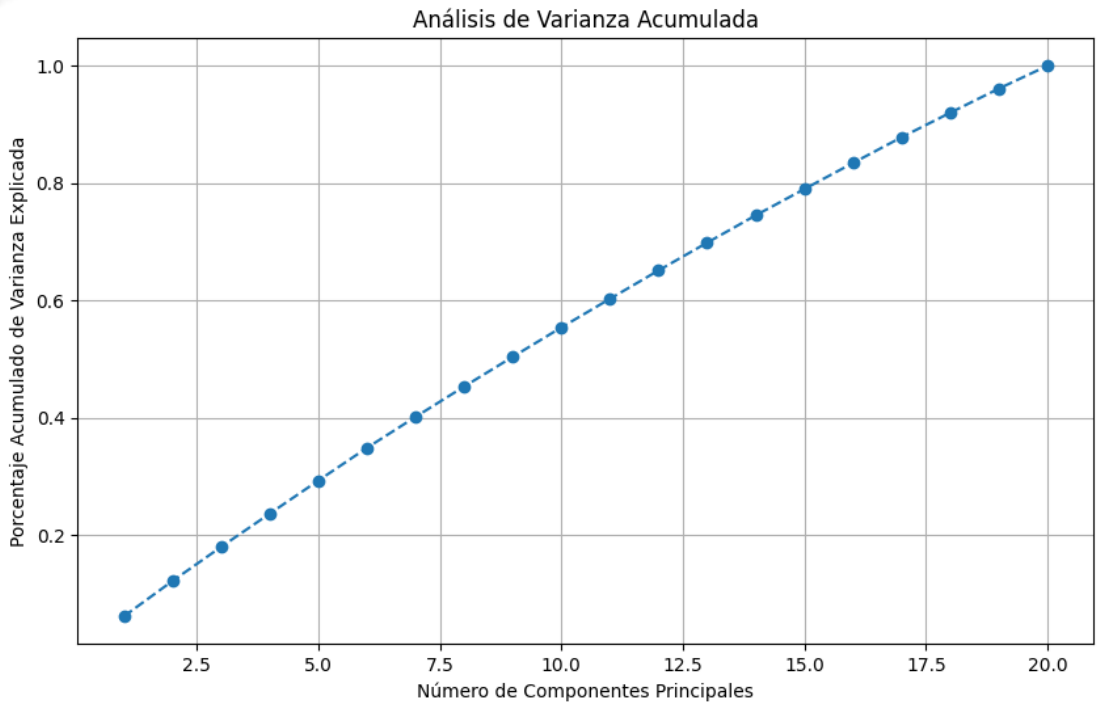
IMPLEMENTACIÓN EN PYTHON

VEAMOS EL PASO A PASO, 2:

```
# Análisis de Componentes Principales (PCA)
pca = PCA()
pca.fit(df)
```

IMPLEMENTACIÓN EN PYTHON

VEAMOS EL PASO A PASO, 3:



IMPLEMENTACIÓN EN PYTHON

VEAMOS EL PASO A PASO, 4:

```
Número de componentes para explicar al menos el 95% de la varianza: 19
```

IMPLEMENTACIÓN EN PYTHON

VEAMOS EL PASO A PASO, 5-6:

```
Primeras filas del conjunto de datos transformado:
   Componente_1  Componente_2  Componente_3  Componente_4  Componente_5  \
0     -0.177021      0.196946      0.380536      0.600370      0.004683
1     -0.098799     -0.051236     -0.113252     -0.249617     -0.003372
2      0.495869      0.307295      0.334187      0.191079      0.386715
3     -0.545123      0.048805     -0.100849     -0.419703     -0.329299
4      0.052469      0.693380      0.156894      0.103554      0.084592

   Componente_6  Componente_7  Componente_8  Componente_9  Componente_10  \
0      0.436867      0.121947      0.000379      0.179966      -0.183514
1     -0.558524      0.184328     -0.089373      0.127126       0.073608
2      0.196728     -0.168530     -0.039448     -0.067065       0.009247
3     -0.149264      0.772666     -0.165841      0.132679       0.180556
4      0.393504      0.180040     -0.270969     -0.102544       0.245617

   Componente_11  Componente_12  Componente_13  Componente_14  Componente_15  \
0       0.644760      -0.188979      -0.039150       0.467873       0.250104
1      -0.780358       0.316044       0.120547      -0.292930      -0.044968
2      -0.487295       0.900056      -0.207312      -0.102993       0.194279
3      -0.240551       0.422799      -0.464322       0.025663      -0.063199
4       0.097898      -0.194609       0.022140      -0.053978      -0.310793

   Componente_16  Componente_17  Componente_18  Componente_19  Componente_20
0      -0.007587       0.335770      -0.228148      -0.297542       0.063501
1      -0.246376       0.416154      -0.204946       0.128982      -0.233995
2      -0.292764       0.078175      -0.312757      -0.157516       0.081915
3      -0.057665       0.110335      -0.031923      -0.182094       0.364166
4       0.011090       0.255747      -0.030667       0.082299      -0.539410
```

¿Preguntas?