Cuestionario Dos

Isabel Ruiz, Aondra Vega, Alejandro Zamora, Nadia Flores, Abel Cano

May 27, 2024

Ejercicio 1. Explica en qué consiste el análisis por correspondencias. Realiza una tabla comparativa con respecto al análisis por componentes principales.

Característica	**Análisis de Correspondencias (CA)**	**Análisis de Componentes Principales (PCA)**
Tipo de Datos de Entrada	Datos categóricos (tablas de contingencia)	Datos continuos (matriz de características)
Objetivo Principal	Identificar y visualizar asociaciones entre categorías	Reducir dimensionalidad y conservar variabilidad
Salida	Coordenadas de filas y columnas en espacio reducido	Nuevas variables (componentes principales)
Representación Gráfica	Mapa de correspondencias	Gráfica de componentes principales
Distancias	Similitud entre categorías de filas y columnas	Similitud entre observaciones
Aplicaciones	Estudios de mercado, encuestas, sociología, biología	Compresión de datos, reducción de ruido, visualización de datos, preprocesamiento para ML
Técnica Subyacente	Descomposición de valores singulares (SVD) en tablas de contingencia	Descomposición de valores singulares (SVD) en matrices de datos continuos

Ejercicio 2. Realiza un análisis de correspondencias múltiple con los datos de iris. ¿Cuáles son tus conclusiones? Colab

Ejercicio 3. Realiza un análisis del análisis factorial y su significado. Puedes tomar como referencia: https://online.stat.psu.edu/stat505/lesson/12 Colab

Ejercicio 4. Del libro de Rencher, realizar los siguientes ejercicios:

13.1, 13.2, 13.3, 13.5

Respuesta Ejercicio 13.1:

$$\begin{aligned}\text{var}(y_i) &= \text{var}(y_i - \mu_i) = \text{var}(\lambda_{i1}f_1 + \lambda_{i2}f_2 + \cdots + \lambda_{im}f_m + \varepsilon_i) \\ &= \sum_{j=1}^{m} \lambda_{ij}^2 \text{var}(f_j) + \text{var}(\varepsilon_i) + \sum_{j\neq k} \lambda_{ij}\lambda_{ik}\text{cov}(f_j, f_k) \\ &\quad + \sum_{j=1}^{m} \lambda_{ij}\text{cov}(f_j, \varepsilon_i) \\ &= \sum_{j=1}^{m} \lambda_{ij}^2 + \psi_i.\end{aligned}$$

La última igualdad se sigue de las suposiciones $\text{var}(f_j) = 1$, $\text{var}(\varepsilon_i) = \psi_i$, $\text{cov}(f_j, f_k) = 0$, y $\text{cov}(f_j, \varepsilon_i) = 0$.

Respuesta Ejercicio 13.2:

$$\begin{aligned}\text{cov}(\mathbf{y}, \mathbf{f}) &= \text{cov}(\mathbf{\Lambda f} + \varepsilon, \mathbf{f}) && \text{[sección (13.3)]} \\ &= \text{cov}(\mathbf{\Lambda f}, \mathbf{f}) && \text{[sección (13.10)]} \\ &= \mathbb{E}[(\mathbf{\Lambda f} - \mathbb{E}(\mathbf{\Lambda f}))(\mathbf{f} - \mathbb{E}(\mathbf{f}))'] && \text{[sección (3.31)]} \\ &= \mathbb{E}[(\mathbf{\Lambda f} - \mathbf{\Lambda}\mathbb{E}(\mathbf{f}))(\mathbf{f} - \mathbb{E}(\mathbf{f}))'] \\ &= \mathbf{\Lambda}\mathbb{E}[(\mathbf{f} - \mathbb{E}(\mathbf{f}))(\mathbf{f} - \mathbb{E}(\mathbf{f}))'] \\ &= \mathbf{\Lambda}\text{cov}(\mathbf{f}) = \mathbf{\Lambda} && \text{[sección (13.7)]}\end{aligned}$$

Respuesta Ejercicio 13.3:

$$\begin{aligned}E(f^*) = E(T'f) = T'E(f) = T'0 = 0, \\ \text{cov}(f^*) = \text{cov}(T'f) = T'\text{cov}(f)T = T'IT = I\end{aligned}$$

Respuesta Ejercicio 13.5:

$$\sum_{i=1}^{p}\sum_{j=1}^{m} \hat{\lambda}_{ij}^2 = \sum_{i=1}^{p}\left[\sum_{j=1}^{m} \hat{\lambda}_{ij}^2\right] = \sum_{i=1}^{p} \hat{h}_i^2 \quad \text{[sección (13.28)]}$$

Al intercambiar el orden de suma, tenemos

$$\sum_{i=1}^{p}\sum_{j=1}^{m}\hat{\lambda}_{ij}^2 = \sum_{j=1}^{m}\sum_{i=1}^{p}\hat{\lambda}_{ij}^2 = \sum_{j=1}^{m}\theta_j \quad \text{[sección (13.29)]}.$$

Ejercicio 5. Considera el siguiente conjunto de puntos $(0,0)$, $(0,1)$, $(-1,2)$, $(2,0)$, $(3,0)$, $(4,-1)$.

1. Calcula la matriz de disimilaridades.

	$(0,0)$	$(0,1)$	$(-1,2)$	$(2,0)$	$(3,0)$	$(4,-1)$
$(0,0)$	0	1	$\sqrt{5}$	2	3	$\sqrt{17}$
$(0,1)$	1	0	$\sqrt{2}$	$\sqrt{5}$	$\sqrt{10}$	$\sqrt{20}$
$(-1,2)$	$\sqrt{5}$	$\sqrt{2}$	0	$\sqrt{13}$	$\sqrt{20}$	$\sqrt{34}$
$(2,0)$	2	$\sqrt{5}$	$\sqrt{13}$	0	1	$\sqrt{5}$
$(3,0)$	3	$\sqrt{10}$	$\sqrt{20}$	1	0	$\sqrt{2}$
$(4,-1)$	$\sqrt{17}$	$\sqrt{20}$	$\sqrt{34}$	$\sqrt{5}$	$\sqrt{2}$	0

Realiza un K-means. Usa $(0,0)$ y $(4,-1)$ como centroides iniciales ($K = 2$). ¿A qué clúster pertenece el punto $(1,1)$?

Iteración 1:

Centroides actuales:

$$\begin{bmatrix} 0 & 0 \\ 4 & -1 \end{bmatrix}$$

Clusters asignados:

$$\begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 1 \\ 1 \end{bmatrix}$$

Nuevos centroides:

$$\begin{bmatrix} 0.25 & 0.75 \\ 3.5 & -0.5 \end{bmatrix}$$

Iteración 2:

Centroides actuales:

$$\begin{bmatrix} 0.25 & 0.75 \\ 3.5 & -0.5 \end{bmatrix}$$

Clusters asignados:

$$\begin{bmatrix} 0 \\ 0 \\ 0 \\ 1 \\ 1 \\ 1 \end{bmatrix}$$

Nuevos centroides:

$$\begin{bmatrix} -0.333 & 1 \\ 3 & -0.333 \end{bmatrix}$$

Iteración 3:
Centroides actuales:

$$\begin{bmatrix} -0.333 & 1 \\ 3 & -0.333 \end{bmatrix}$$

Clusters asignados:

$$\begin{bmatrix} 0 \\ 0 \\ 0 \\ 1 \\ 1 \\ 1 \end{bmatrix}$$

Nuevos centroides:

$$\begin{bmatrix} -0.333 & 1 \\ 3 & -0.333 \end{bmatrix}$$

Alcanzada la convergencia.
Resultados finales:
Clusters asignados:

$$\begin{bmatrix} 0 \\ 0 \\ 0 \\ 1 \\ 1 \\ 1 \end{bmatrix}$$

Centroides finales:

$$\begin{bmatrix} -0.333 & 1 \\ 3 & -0.333 \end{bmatrix}$$

El punto $(1, 1)$ **pertenece al cluster 1.**

Ejercicio 6. Para un conjunto de puntos $(x_i)_{i=1}^n$ en $\mathbb{R}^m$, demuestra que la media muestral $\hat{\mu}$ es la solución al problema de optimización

$$\hat{\mu} = \arg\min_{\mu \in \mathbb{R}^m} \sum_{i=1}^{n} d_2(x_i, \mu)^2. \tag{1}$$

Sea x_i un conjunto de puntos en $\mathbb{R}^m$, y $\hat{\mu}$ la media muestral definida como:

$$\hat{\mu} = \frac{1}{n} \sum_{i=1}^{n} x_i$$

Demostraremos que $\hat{\mu}$ es la solución al siguiente problema de optimización:

$$\hat{\mu} = \arg\min_{\mu \in \mathbb{R}^m} \sum_{i=1}^{n} d^2(x_i, \mu)^2$$

donde $d(x_i, \mu)$ es la distancia euclidiana entre x_i y μ.
Expandiendo la función de pérdida, obtenemos:

$$\begin{aligned}\|x_i - \mu\|^2 &= (x_i - \mu)^T (x_i - \mu) \\ &= x_i^T x_i - 2x_i^T \mu + \mu^T \mu\end{aligned}$$

Derivando e igualando a cero para minimizar, tenemos:

$$\frac{\partial}{\partial \mu} \sum_{i=1}^{n} \|x_i - \mu\|^2 = \sum_{i=1}^{n} (-2x_i) + 2n\mu = 0$$

Despejando μ, obtenemos $\mu = \frac{1}{n} \sum_{i=1}^{n} x_i = \hat{\mu}$.
Por lo tanto, $\hat{\mu}$ es la solución al problema de optimización.

Ejercicio 7. Supongamos que se tienen n observaciones, cada una con p características. Determina cuáles de los siguientes enunciados son verdaderos con respecto al análisis de clústers:

1. Podemos agrupar las n observaciones sobre la base de las p características para identificar subgrupos entre las observaciones.
2. Podemos agrupar las p características sobre la base de las n- observaciones para descubrir subgrupos entre las características.
3. El análisis por clústers es parte del aprendizaje supervisado y es parte del análisis exploratorio de datos.

Justifica tus respuestas.

Respuestas.

1. **Verdadero.** Este es el proceso típico de análisis de clúster, en el cual se agrupan las observaciones en base a sus características para identificar patrones o subgrupos dentro de los datos
2. **Falso.** El análisis de clúster se centra en agrupar las observaciones, no las características, pero podrá ser en el PCA para reducir la dimensionalidad de los datos y encontrar patrones entre las características.
3. **Falso.** El análisis de clúster es parte del análisis exploratorio de datos porque se utiliza para comprender la estructura de los datos y descubrir patrones subyacentes sin la necesidad de etiquetas previas, pero no forma parte del aprendizaje supervisado, ya que en el aprendizaje supervisado se utilizan etiquetas para entrenar modelos predictivos.

Ejercicio 8. Realiza un algoritmo de K-means (K = 2) dadas las asignaciones siguientes:

Observación	X_1	X_2	**Clúster inicial**
1	1	3	2
2	0	4	1
3	6	2	2
4	5	2	2
5	1	6	1

Determina las asignaciones finales de los agrupamientos.

Iteración 0: Asignaciones iniciales de agrupamientos:

$$[2, 1, 2, 2, 1]$$

Centroides iniciales:

$$\begin{bmatrix} 0.93403344 & 0.58421908 \\ 0.5 & 5 \end{bmatrix}$$

Iteración 1: Asignaciones de agrupamientos:

$$[1, 1, 0, 0, 1]$$

Centroides:

$$\begin{bmatrix} 5.5 & 2 \\ 0.66666667 & 4.33333333 \end{bmatrix}$$

Iteración 2: Asignaciones de agrupamientos:

$$[1, 1, 0, 0, 1]$$

Centroides:

$$\begin{bmatrix} 5.5 & 2 \\ 0.66666667 & 4.33333333 \end{bmatrix}$$

Asignaciones finales de agrupamientos:

$$[1, 1, 0, 0, 1]$$

Centroides finales:

$$\begin{bmatrix} 5.5 & 2 \\ 0.66666667 & 4.33333333 \end{bmatrix}$$

Ejercicio 9. Realiza este ejercicio a mano y en Python. Considera los siguientes puntos:

(2,10), (2,5), (8,4), (5,8), (7,5), (6,4), (1,2), (4,9)

1. Usando el algoritmo de k-means, agrupa puntos en 3-clústers.
2. Realiza un análisis de clúster usando single-link, complete-link y average-link para agrupar los puntos dados.

1. **Inicialización**: Seleccionar k centros iniciales. Aquí $k = 3$.
2. **Asignación**: Asignar cada punto al centro más cercano.
3. **Actualización**: Calcular los nuevos centros como el promedio de los puntos asignados a cada clúster.
4. **Repetición**: Repetir los pasos de asignación y actualización hasta que las asignaciones no cambien.

Para simplificar, empezaremos con los siguientes centros iniciales (esto puede variar):

- Centro 1: (2, 10)
- Centro 2: (5, 8)
- Centro 3: (1, 2)

En Python, el código sería el siguiente:

```
import numpy as np
from sklearn.cluster import KMeans
import matplotlib.pyplot as plt

# Puntos dados
points = np.array([
    [2, 10], [2, 5], [8, 4], [5, 8],
    [7, 5], [6, 4], [1, 2], [4, 9]
])

# Aplicar k-means con 3 clusteres
kmeans = KMeans(n_clusters=3, random_state=0).fit(points)
labels = kmeans.labels_
centers = kmeans.cluster_centers_

print("Centros de clusteres:")
print(centers)
print("Etiquetas de cluster:")
print(labels)

# Visualizacion
plt.scatter(points[:, 0], points[:, 1], c=labels, cmap='viridis')
plt.scatter(centers[:, 0], centers[:, 1], s=300, c='red')
plt.show()
```

Realiza un análisis de clúster usando *single-link*, *complete-link* y *average-link* para agrupar los puntos dados. En Python, el código sería el siguiente:

```
import numpy as np
import matplotlib.pyplot as plt
from scipy.cluster.hierarchy import dendrogram, linkage

# Puntos dados
points = np.array([
    [2, 10], [2, 5], [8, 4], [5, 8],
    [7, 5], [6, 4], [1, 2], [4, 9]
])

# Single-link
```

```
Z_single = linkage(points, 'single')
plt.figure(figsize=(10, 7))
plt.title("Dendrograma Single-link")
dendrogram(Z_single)
plt.show()

# Complete-link
Z_complete = linkage(points, 'complete')
plt.figure(figsize=(10, 7))
plt.title("Dendrograma Complete-link")
dendrogram(Z_complete)
plt.show()

# Average-link
Z_average = linkage(points, 'average')
plt.figure(figsize=(10, 7))
plt.title("Dendrograma Average-link")
dendrogram(Z_average)
plt.show()
```

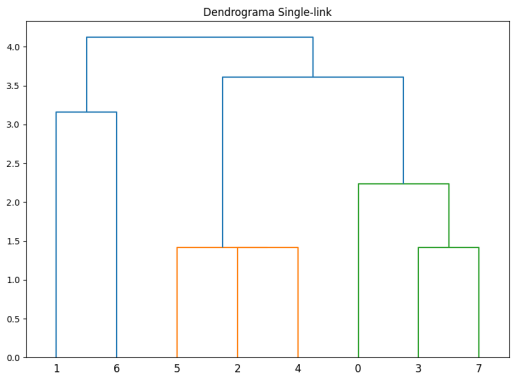
Resultados de los dendrogramas:

Ejercicio 10. Construye una base de datos de 20 tickers (que incluyan 10 empresas con tickers terminación .MX) en un umbral de tiempo de 6 años. Considera las siguientes estadísticas:

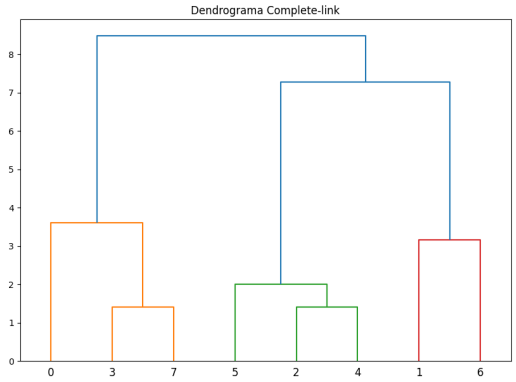
1. Movimientos finales. Se obtienen a través de los precios diarios tomando el precio más alto menos el precio más bajo en ese día.

2. Rendimientos. Se obtiene como el precio a la fecha corriente entre el precio del día anterior menos 1.

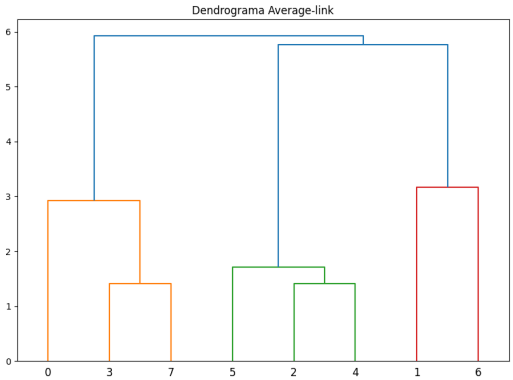
Realiza un análisis de clúster con respecto a tales estadísticas. ¿Cuáles son tus conclusiones? También, realiza una simulación Monte-Carlo para obtener la frontera eficiente de Markowitz. ¿Cuál es el índice de Sharpe? ¿Qué pesos son los óptimos en tu portafolio?



(a) Dendrograma Single-link



(b) Dendrograma Complete-link



(c) Dendrograma Average-link

Figure 1: Dendrogramas de diferentes métodos de enlace