Cuestionario final

Angel Maya, Andrea Cañas, Dante Rodriguez, Adrián García, Ian Benitez

May 2024

1 Introduction

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

Análisis por Correspondencias (AC)

El Análisis por Correspondencias (AC) es una técnica estadística multivariante que se utiliza para analizar y visualizar las relaciones entre dos variables categóricas. Se basa en una tabla de contingencia que contiene las frecuencias de co-ocurrencia de las categorías de ambas variables. El objetivo principal del AC es transformar esta tabla en un mapa donde las filas y las columnas de la tabla se representan como puntos, y la proximidad de los puntos refleja la asociación entre las categorías.

Pasos principales en el AC:

1. **Construcción de la tabla de contingencia:** Crear una tabla de frecuencias que cruce las categorías de las dos variables.
2. **Cálculo de las frecuencias relativas:** Transformar las frecuencias absolutas en relativas.
3. **Cálculo de los perfiles de filas y columnas:** Determinar los perfiles que son las frecuencias relativas divididas por los totales de filas y columnas.
4. **Cálculo de las distancias chi-cuadrado:** Calcular la distancia chi-cuadrado entre los perfiles.
5. **Descomposición en valores singulares:** Realizar una descomposición en valores singulares (SVD) para obtener las coordenadas de los puntos en el espacio de menor dimensión.

Análisis por Componentes Principales (ACP)

El Análisis por Componentes Principales (ACP) es una técnica de reducción de dimensionalidad que transforma un conjunto de variables posiblemente correlacionadas en un conjunto de valores no correlacionados llamados componentes principales. El objetivo del ACP es capturar la mayor cantidad de variabilidad en los datos con el menor número de componentes.

Pasos principales en el ACP:

1. **Estandarización de los datos:** Normalizar las variables para que tengan media cero y varianza uno.
2. **Cálculo de la matriz de covarianza:** Construir la matriz de covarianza de las variables estandarizadas.
3. **Cálculo de los vectores y valores propios:** Determinar los vectores y valores propios de la matriz de covarianza.
4. **Selección de los componentes principales:** Seleccionar los componentes que capturan la mayor cantidad de varianza.
5. **Transformación de los datos:** Proyectar los datos originales en el espacio de los componentes principales seleccionados.

Tabla Comparativa

Característica	**Análisis por Correspondencias (AC)**	**Análisis por C**
Tipo de datos	Categóricos	
Objetivo	Analizar relaciones entre categorías	Reducir la di
Matriz inicial	Tabla de contingencia	
Distancia utilizada	Chi-cuadrado	
Visualización	Mapa de correspondencias	Gráfica
Aplicaciones comunes	Marketing, estudios de mercado, ciencias sociales	Análisis fin
Reducción de dimensionalidad	Implícita (proyección en un espacio de menor dimensión)	Explícita (sel

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

Análisis de Correspondencias Múltiple (ACM)

Para realizar un Análisis de Correspondencias Múltiple (ACM) con los datos de Iris, primero necesitamos transformar los datos continuos en categorías, ya que el ACM se utiliza principalmente para variables categóricas. El conjunto de datos de Iris contiene las siguientes variables: *sepal length*, *sepal width*, *petal length*, *petal width* y *species*.

Pasos para realizar el ACM:

1. **Cargar los datos de Iris:** Utilizaremos un conjunto de datos conocido como el conjunto de datos de Iris que está disponible en muchas bibliotecas de Python, como *sklearn*.

2. **Transformar los datos continuos en categorías:** Necesitamos convertir las variables continuas en categorías para poder aplicar ACM.

3. **Realizar el ACM:** Utilizaremos la biblioteca *prince* para realizar el análisis de correspondencias múltiples.

4. **Interpretar los resultados:** Generar gráficos y tablas para interpretar los resultados del análisis.

Conclusiones

- **Separación de las especies:** En el gráfico de correspondencias múltiples, se observa una clara separación entre las tres especies de Iris (*setosa*, *versicolor* y *virginica*), lo que sugiere que las características medidas (longitud y anchura de sépalo y pétalo) son efectivas para diferenciar entre las especies.

- **Variables importantes:** Las variables relacionadas con el pétalo (*petal length* y *petal width*) parecen tener una mayor contribución en la diferenciación de las especies, ya que las categorías correspondientes están bien separadas en el espacio bidimensional.

- **Asociaciones entre categorías:** Las categorías *corto* y *estrecho* para las variables de longitud y anchura de sépalo y pétalo están más asociadas con la especie *setosa*. Las categorías *largo* y *ancho* están más asociadas con las especies *versicolor* y *virginica*.

Ejercicio 3. Solución en Python **Ejercicio 4.** Solución en Python
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.

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

Parte a: Calcular la matriz de disimilaridades

La distancia euclidiana entre dos puntos (x_1, y_1) y (x_2, y_2) se calcula como:

$$d = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2}$$

Calculamos las distancias entre todos los pares de puntos:

Puntos	*Distancia*
$(0,0),(0,1)$	$\sqrt{(0-0)^2+(1-0)^2}=1$
$(0,0),(-1,2)$	$\sqrt{(-1-0)^2+(2-0)^2}=\sqrt{1+4}=\sqrt{5}\approx 2.24$
$(0,0),(2,0)$	$\sqrt{(2-0)^2+(0-0)^2}=2$
$(0,0),(3,0)$	$\sqrt{(3-0)^2+(0-0)^2}=3$
$(0,0),(4,-1)$	$\sqrt{(4-0)^2+(-1-0)^2}=\sqrt{16+1}=\sqrt{17}\approx 4.12$
$(0,1),(-1,2)$	$\sqrt{(-1-0)^2+(2-1)^2}=\sqrt{1+1}=\sqrt{2}\approx 1.41$
$(0,1),(2,0)$	$\sqrt{(2-0)^2+(0-1)^2}=\sqrt{4+1}=\sqrt{5}\approx 2.24$
$(0,1),(3,0)$	$\sqrt{(3-0)^2+(0-1)^2}=\sqrt{9+1}=\sqrt{10}\approx 3.16$
$(0,1),(4,-1)$	$\sqrt{(4-0)^2+(-1-1)^2}=\sqrt{16+4}=\sqrt{20}\approx 4.47$
$(-1,2),(2,0)$	$\sqrt{(2-(-1))^2+(0-2)^2}=\sqrt{9+4}=\sqrt{13}\approx 3.61$
$(-1,2),(3,0)$	$\sqrt{(3-(-1))^2+(0-2)^2}=\sqrt{16+4}=\sqrt{20}\approx 4.47$
$(-1,2),(4,-1)$	$\sqrt{(4-(-1))^2+(-1-2)^2}=\sqrt{25+9}=\sqrt{34}\approx 5.83$
$(2,0),(3,0)$	$\sqrt{(3-2)^2+(0-0)^2}=1$
$(2,0),(4,-1)$	$\sqrt{(4-2)^2+(-1-0)^2}=\sqrt{4+1}=\sqrt{5}\approx 2.24$
$(3,0),(4,-1)$	$\sqrt{(4-3)^2+(-1-0)^2}=\sqrt{1+1}=\sqrt{2}\approx 1.41$

La matriz de disimilaridades es entonces:

	$(0,0)$	$(0,1)$	$(-1,2)$	$(2,0)$	$(3,0)$	$(4,-1)$
$(0,0)$	0	1	2.24	2	3	4.12
$(0,1)$	1	0	1.41	2.24	3.16	4.47
$(-1,2)$	2.24	1.41	0	3.61	4.47	5.83
$(2,0)$	2	2.24	3.61	0	1	2.24
$(3,0)$	3	3.16	4.47	1	0	1.41
$(4,-1)$	4.12	4.47	5.83	2.24	1.41	0

Parte b: Realizar un K-means con $K = 2$

Usamos los puntos $(0,0)$ y $(4,-1)$ como centroides iniciales.

Paso 1: Centroides iniciales

- $C1 = (0,0)$
- $C2 = (4,-1)$

Paso 2: Asignar cada punto al centroide más cercano

Calculamos la distancia de cada punto a los centroides iniciales y los asignamos al centroide más cercano.

$Punto$	$DistanciaaC1$	$DistanciaaC2$	$CentroideAsignado$
$(0,0)$	0	4.12	$C1$
$(0,1)$	1	4.47	$C1$
$(-1,2)$	2.24	5.83	$C1$
$(2,0)$	2	2.24	$C1$
$(3,0)$	3	1.41	$C2$
$(4,-1)$	4.12	0	$C2$

Paso 3: Recalcular los centroides

Nuevo $C1$**:** Promedio de $[(0,0),(0,1),(-1,2),(2,0)]$

$$C1_x = \frac{0+0-1+2}{4} = 0.25, \quad C1_y = \frac{0+1+2+0}{4} = 0.75$$

Nuevo $C1 = (0.25, 0.75)$

Nuevo $C2$**:** Promedio de $[(3,0),(4,-1)]$

$$C2_x = \frac{3+4}{2} = 3.5, \quad C2_y = \frac{0-1}{2} = -0.5$$

Nuevo $C2 = (3.5, -0.5)$

Paso 4: Reasignar puntos según los nuevos centroides

Calculamos la distancia del punto $(1,1)$ a los nuevos centroides:

$$DistanciaaC1 = \sqrt{(1-0.25)^2 + (1-0.75)^2} = \sqrt{0.5625 + 0.0625} = \sqrt{0.625} \approx 0.79$$

$$DistanciaaC2 = \sqrt{(1-3.5)^2 + (1-(-0.5))^2} = \sqrt{6.25 + 2.25} = \sqrt{8.5} \approx 2.92$$

El punto $(1,1)$ pertenece al Clúster $C1$, ya que la distancia a $C1$ es menor que la distancia a $C2$.

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

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

Donde $d_2(x_i, \mu)$ es la distancia euclidiana entre los puntos x_i y μ.

Solución

Paso 1: Definición del problema

La distancia euclidiana al cuadrado entre x_i y μ es:

$$d_2(x_i, \mu)^2 = \|x_i - \mu\|^2$$

Por lo tanto, el problema de optimización se puede reescribir como:

$$\hat{\mu} = \arg\min_{\mu \in R^m} \sum_{i=1}^{n} \|x_i - \mu\|^2$$

Paso 2: Expansión de la función objetivo

Expandimos el término $\|x_i - \mu\|^2$:

$$\|x_i - \mu\|^2 = (x_i - \mu)^T (x_i - \mu)$$

Esto se expande a:

$$(x_i - \mu)^T (x_i - \mu) = x_i^T x_i - 2x_i^T \mu + \mu^T \mu$$

Por lo tanto, la función objetivo se convierte en:

$$f(\mu) = \sum_{i=1}^{n} (x_i^T x_i - 2x_i^T \mu + \mu^T \mu)$$

Paso 3: Simplificación de la función objetivo

Podemos separar los términos que dependen de μ y los que no:

$$f(\mu) = \sum_{i=1}^{n} x_i^T x_i - 2\sum_{i=1}^{n} x_i^T \mu + \sum_{i=1}^{n} \mu^T \mu$$

Los términos independientes de μ se agrupan en una constante C:

$$C = \sum_{i=1}^{n} x_i^T x_i$$

Por lo tanto, la función objetivo es:

$$f(\mu) = C - 2\sum_{i=1}^{n} x_i^T \mu + n\mu^T \mu$$

Paso 4: Derivada y condiciones de optimalidad

Para encontrar el mínimo, derivamos $f(\mu)$ con respecto a μ y lo igualamos a cero:

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

Simplificando, tenemos:

$$-\sum_{i=1}^{n} x_i + n\mu = 0$$

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

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

Conclusión

Hemos demostrado que la solución al problema de optimización es la media muestral $\hat{\mu}$:

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

Por lo tanto, la media muestral es la que minimiza la suma de las distancias euclidianas cuadradas de los puntos al centroide, lo que concluye la demostración.

1. **Podemos agrupar las n observaciones sobre la base de las p características para identificar subgrupos entre las observaciones.**

   Verdadero.

   Justificación: El análisis de clústers se utiliza principalmente para agrupar las observaciones (en este caso, los puntos de datos) basándose en sus características. Este enfoque es uno de los métodos más comunes en análisis de datos multivariados y se emplea para descubrir patrones y subgrupos naturales dentro de los datos. Algoritmos como K-means, jerárquicos y DBSCAN se utilizan para este propósito, dividiendo las n observaciones en grupos o clústers donde las observaciones dentro del mismo clúster son más similares entre sí en términos de las p características medidas.

2. **Podemos agrupar las p características sobre la base de las n observaciones para descubrir subgrupos entre las características.**

   Verdadero.

   Justificación: Aunque es menos común que agrupar las observaciones, también es posible agrupar las características. Este tipo de análisis puede ser útil para identificar características que tienden a variar juntas o que tienen patrones similares a través de las observaciones. Este enfoque se llama análisis de clúster de variables. Técnicas como el análisis de conglomerados de variables pueden identificar grupos de características que están altamente correlacionadas entre sí. Esta información puede ser valiosa para reducir la dimensionalidad del problema, interpretar relaciones entre variables, o seleccionar subconjuntos de características relevantes.

3. **El análisis por clústers es parte del aprendizaje supervisado y es parte del análisis exploratorio de datos.**

   Parcialmente verdadero.

   Justificación: El análisis por clústers es parte del aprendizaje no supervisado, no del aprendizaje supervisado. En el aprendizaje supervisado, los algoritmos se entrenan usando datos etiquetados, es decir, con un conocimiento previo de la variable objetivo. En contraste, el análisis por clústers se realiza sin etiquetas de clase, intentando descubrir la estructura inherente de los datos.

   Sin embargo, es cierto que el análisis por clústers es una parte fundamental del análisis exploratorio de datos (EDA, por sus siglas en inglés). En el EDA, los analistas buscan patrones, anomalías y relaciones interesantes en los datos antes de aplicar modelos predictivos. El análisis de clústers ayuda a identificar grupos naturales en los datos y puede revelar información importante que influye en el desarrollo de modelos posteriores o en la toma de decisiones.

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.**

   Verdadero.

   Justificación: El análisis de clústers se utiliza principalmente para agrupar las observaciones (en este caso, los puntos de datos) basándose en sus características. Este enfoque es uno de los métodos más comunes en análisis de datos multivariados y se emplea para descubrir patrones y subgrupos naturales dentro de los datos. Algoritmos como K-means, jerárquicos y DBSCAN se utilizan para este propósito, dividiendo las n observaciones

en grupos o clústers donde las observaciones dentro del mismo clúster son más similares entre sí en términos de las p características medidas.

2. **Podemos agrupar las p características sobre la base de las n observaciones para descubrir subgrupos entre las características.**

   Verdadero.

   Justificación: Aunque es menos común que agrupar las observaciones, también es posible agrupar las características. Este tipo de análisis puede ser útil para identificar características que tienden a variar juntas o que tienen patrones similares a través de las observaciones. Este enfoque se llama análisis de clúster de variables. Técnicas como el análisis de conglomerados de variables pueden identificar grupos de características que están altamente correlacionadas entre sí. Esta información puede ser valiosa para reducir la dimensionalidad del problema, interpretar relaciones entre variables, o seleccionar subconjuntos de características relevantes.

3. **El análisis por clústers es parte del aprendizaje supervisado y es parte del análisis exploratorio de datos.**

   Parcialmente verdadero.

   Justificación: El análisis por clústers es parte del aprendizaje no supervisado, no del aprendizaje supervisado. En el aprendizaje supervisado, los algoritmos se entrenan usando datos etiquetados, es decir, con un conocimiento previo de la variable objetivo. En contraste, el análisis por clústers se realiza sin etiquetas de clase, intentando descubrir la estructura inherente de los datos.

   Sin embargo, es cierto que el análisis por clústers es una parte fundamental del análisis exploratorio de datos (EDA, por sus siglas en inglés). En el EDA, los analistas buscan patrones, anomalías y relaciones interesantes en los datos antes de aplicar modelos predictivos. El análisis de clústers ayuda a identificar grupos naturales en los datos y puede revelar información importante que influye en el desarrollo de modelos posteriores o en la toma de decisiones.

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.

1. **Calcular los centroides iniciales**

   Clúster 1: Observaciones: 2, 5

$$Centroide1 = \left(\frac{0+1}{2}, \frac{4+6}{2}\right) = \left(\frac{1}{2}, 5\right) = (0.5, 5)$$

   Clúster 2: Observaciones: 1, 3, 4

$$Centroide2 = \left(\frac{1+6+5}{3}, \frac{3+2+2}{3}\right) = \left(\frac{12}{3}, \frac{7}{3}\right) = (4, 2.33)$$

2. **Reasignar observaciones a los centroides más cercanos**

   Distancias al Centroide 1 $(0.5, 5)$ **y Centroide 2** $(4, 2.33)$

$Observación$ $DistanciaaCentroide2$	X_1 $ClústerAsignado$	X_2	
1	1	3	$\sqrt{(1-0.5)^2 +$
$\sqrt{(1-4)^2 + (3-2.33)^2} = \sqrt{9 + 0.4489} = \sqrt{9.4489} \approx 3.07$	1		
2	0	4	$\sqrt{(0-0.5)^2 +$
$\sqrt{(0-4)^2 + (4-2.33)^2} = \sqrt{16 + 2.6689} = \sqrt{18.6689} \approx 4.32$	1		
3	6	2	$\sqrt{(6-0.5)^2 + ($
$\sqrt{(6-4)^2 + (2-2.33)^2} = \sqrt{4 + 0.1089} = \sqrt{4.1089} \approx 2.03$	2		
4	5	2	$\sqrt{(5-0.5)^2 + ($
$\sqrt{(5-4)^2 + (2-2.33)^2} = \sqrt{1 + 0.1089} = \sqrt{1.1089} \approx 1.05$	2		
5	1	6	$\sqrt{(1-0.5)^2 +$
$\sqrt{(1-4)^2 + (6-2.33)^2} = \sqrt{9 + 13.1889} = \sqrt{22.1889} \approx 4.71$	1		

3. **Recalcular los centroides con las nuevas asignaciones**

   Nuevas asignaciones:

   - Clúster 1: Observaciones 1, 2, 5
   - Clúster 2: Observaciones 3, 4

   Nuevos centroides:

   Clúster 1: Observaciones: 1, 2, 5

$$Centroide1 = \left(\frac{1+0+1}{3}, \frac{3+4+6}{3}\right) = \left(\frac{2}{3}, \frac{13}{3}\right) = (0.67, 4.33)$$

   Clúster 2: Observaciones: 3, 4

$$Centroide2 = \left(\frac{6+5}{2}, \frac{2+2}{2}\right) = \left(\frac{11}{2}, \frac{4}{2}\right) = (5.5, 2)$$

4. **Reasignar observaciones a los nuevos centroides**

   Distancias al nuevo Centroide 1 $(0.67, 4.33)$ **y Centroide 2** $(5.5, 2)$

Observación	X_1	X_2	D
*DistanciaaCentroide*2	*ClústerAsignado*		
1	1	3	$\sqrt{(1-0.67)^2+(3-$
$\sqrt{(1-5.5)^2+(3-2)^2} = \sqrt{20.25+1} = \sqrt{21.25} \approx 4.61$	1		
2	0	4	$\sqrt{(0-0.67)^2+(4-$
$\sqrt{(0-5.5)^2+(4-2)^2} = \sqrt{30.25+4} = \sqrt{34.25} \approx 5.85$	1		
3	6	2	$\sqrt{(6-0.67)^2+(2-$
$\sqrt{(6-5.5)^2+(2-2)^2} = \sqrt{0.25+0} = \sqrt{0.25} = 0.5$	2		
4	5	2	$\sqrt{(5-0.67)^2+(2-$
$\sqrt{(5-5.5)^2+(2-2)^2} = \sqrt{0.25+0} = \sqrt{0.25} = 0.5$	2		
5	1	6	$\sqrt{(1-0.67)^2+(6-$
$\sqrt{(1-5.5)^2+(6-2)^2} = \sqrt{20.25+16} = \sqrt{36.25} \approx 6.02$	1		

5. **Asignaciones finales:**

   - Clúster 1: Observaciones 1, 2, 5
   - Clúster 2: Observaciones 3, 4

Observación	X_1	X_2	Clúster final
1	1	3	1
2	0	4	1
3	6	2	2
4	5	2	2
5	1	6	1

Ejercicio 9.2. Realiza un análisis de clúster usando *single-link*, *complete-link* y *average-link* para agrupar los puntos dados.

Puntos a analizar

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

Enlace Simple (Single-Link)

El método de enlace simple agrupa los puntos de manera que la distancia mínima entre los puntos de los clústeres es la menor posible.

Enlace Completo (Complete-Link)

El método de enlace completo agrupa los puntos de manera que la distancia máxima entre los puntos de los clústeres es la menor posible.

Enlace Promedio (Average-Link)

El método de enlace promedio agrupa los puntos basándose en la distancia promedio entre todos los pares de puntos en los clústeres.

Procedimiento

(a) Calcular la matriz de distancias.

(b) Aplicar cada uno de los métodos de enlace (single-link, complete-link, average-link).

(c) Generar el dendrograma correspondiente.

Paso 1: Calcular la matriz de distancias

Calculamos la distancia euclidiana entre cada par de puntos:

$$d((x_1, y_1), (x_2, y_2)) = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2}$$

Paso 2: Aplicar métodos de enlace y generar dendrogramas

Usaremos Python para aplicar los métodos de enlace y generar los dendrogramas correspondientes.

Interpretación de los resultados

(a) **Enlace Simple (Single-Link):** Agrupa los puntos de manera que la distancia mínima entre los clústeres sea la menor posible. Es útil para detectar cadenas de puntos conectados, pero puede ser sensible al ruido y a los valores atípicos.

(b) **Enlace Completo (Complete-Link):** Agrupa los puntos de manera que la distancia máxima entre los clústeres sea la menor posible. Tiende a crear clústeres compactos y es menos sensible al ruido que el enlace simple.

(c) **Enlace Promedio (Average-Link):** Agrupa los puntos basándose en la distancia promedio entre todos los pares de puntos en los clústeres. Es un compromiso entre los métodos de enlace simple y completo, proporcionando clústeres relativamente compactos y menos sensibles al ruido.

Estos métodos permiten identificar la estructura de agrupación en los datos, y el dendrograma generado por cada método ofrece una visualización clara de cómo se agrupan los puntos en diferentes niveles de distancia.

Ejercicio 10 Solución en Python