# Sistemas de Inteligencia Artificial
# **Análisis de Componentes Principales**

2020

# Organización

# Introducción

**Tenemos**

Un conjunto de datos, con variables o atributos (en las columnas) y observaciones (en las filas).

**Muchas variables**

Conjunto de datos multivariado.

# Ejemplo: Competencia de Natación en el río

Tiempos en cada uno de cuatro tramos:

Cuadro: Tiempos de Nadadores

| nadador | tr1 | tr2 | tr3 | tr4 |
|---|---|---|---|---|
| 1 | 10 | 10 | 13 | 12 |
| 2 | 12 | 12 | 14 | 15 |
| 3 | 11 | 10 | 14 | 13 |
| 4 | 9 | 9 | 11 | 11 |
| 5 | 8 | 8 | 9 | 8 |
| 6 | 8 | 9 | 10 | 9 |
| 7 | 10 | 10 | 8 | 9 |

# Ejemplo: Países de Europa

| Country | Area | GDP | Inflation |
|---|---|---|---|
| Austria | 83871 | 41600 | 3.5 |
| Belgium | 30528 | 37800 | 3.5 |
| Bulgaria | 110879 | 13800 | 4.2 |
| Croatia | 56594 | 18000 | 2.3 |

# Comenzamos con un ejemplo

Deseamos conocer cuáles son los factores relacionados con el **riesgo de enfermedad coronaria**.
Del conocimiento previo sabemos que el riesgo está relacionado con los siguientes factores:

- presión arterial
- edad
- obesidad
- tiempo desde que se ha diagnosticado hipertensión arterial
- el pulso
- stress

# Ejemplo

Para una investigación se seleccionan al azar 20 pacientes hipertensos sobre los que se midieron las siguientes variables:

- $X_1$: Presión arterial media (mm Hg)
- $X_2$: Edad (años)
- $X_3$: Peso (Kg).
- $X_4$: Superficie corporal ($m^2$)
- $X_5$: Duración de la Hipertensión (años)
- $X_6$: Pulso (pulsaciones/minuto)
- $X_7$: Medida del stress.

# Ejemplo

**La Pregunta:**

¿es posible definir un índice que cuantifique la situación de riesgo cardíaco de un paciente con hipertensión arterial?

# Ejemplo

Cuadro: Riesgo Cardíaco

| caso | presión | edad | peso | superf | hipert | pulso | stress |
|---|---|---|---|---|---|---|---|
| 1 | 105 | 47 | 85.4 | 1.75 | 5.1 | 63 | 33 |
| 2 | 115 | 49 | 94.2 | 2.1 | 3.8 | 70 | 14 |
| 3 | 116 | 49 | 95.3 | 1.98 | 8.2 | 72 | 10 |
| 4 | 117 | 50 | 94.7 | 2.01 | 5.8 | 73 | 99 |
| 5 | 112 | 51 | 89.4 | 1.89 | 7 | 72 | 95 |
| 6 | 121 | 48 | 99.5 | 2.25 | 9.3 | 71 | 10 |
| 7 | 121 | 49 | 99.8 | 2.25 | 2.5 | 69 | 42 |

# Medidas descriptivas

- Media Muestral: Dado un conjunto de datos de una variable $\mathbf{X} = \{\mathbf{x_1}, , \mathbf{x_2}, \ldots, \mathbf{x_n}\}$

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

# Medidas descriptivas

- Varianza Muestral:

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

es una medida de dispersión.

# Medidas descriptivas

- Desviación estándar muestral: Es otra medida de dispersión

$$s_{dv} = \sqrt{s^2}$$

# Medidas descriptivas

- Covarianza muestral: es una medida de asociación lineal entre los datos de dos variables. Dado el conjunto de datos:

$$\left[\begin{array}{c|c|c|c} \mathbf{X_1} & \mathbf{X_2} & \dots & \mathbf{X_n} \\ \hline x_{11} & x_{12} & \dots & x_{1n} \\ x_{21} & x_{22} & \dots & x_{2n} \\ \vdots & & & \\ x_{m1} & x_{m2} & \dots & x_{mn} \end{array}\right]$$

$$s_{ik} = \frac{1}{n}\sum_{j=1}^{m}(x_{ji} - \bar{x_i}) * (x_{jk} - \bar{x_k})$$

# Matriz de Covarianzas

## Las covarianzas forman una matriz simétrica definida positiva

$$S = \begin{bmatrix} s_{11} & s_{12} & \dots & s_{1n} \\ s_{21} & s_{22} & \dots & s_{2n} \\ \vdots & & & \\ s_{m1} & s_{m2} & \dots & s_{mn} \end{bmatrix}$$

# Interpretación de la Covarianza muestral

## Interpretación:

$s_{ik} > 0$ indica una asociación lineal positiva entre los datos de las variables.
$sik < 0$ indica una asociación lineal negativa entre los datos de las variables.
$s_{ik} = 0$ indica que no hay una asociación lineal entre los datos de las variables.

## Nota

La varianza muestral es la covarianza muestral entre los datos de la i-ésima variable con ella misma, algunas veces se denota como $s_{ii}$

# Variables estandarizadas

$$\tilde{\mathbf{x}}_1 = \frac{\mathbf{x}_1 - \bar{\mathbf{x}}_1}{\mathbf{s}_1}$$

A cada variable le resto su propia media y divido por su propia desviación estándar

# Correlación muestral: Calcular las covarianzas con las variables estandarizadas

Es otra medida de asociación lineal. Para los datos de la i-ésima y k-ésima variable se define como:

$$r_{ik} = \frac{s_{ik}}{\sqrt{s_{ii}}\sqrt{s_{kk}}}$$

La correlación está acotada entre -1 y 1.

# Organización

# Análisis de Componentes Principales

**De qué se trata**

Si las variables están muy correlacionadas, entonces poseen información redundante.

La idea de este método es eliminar la redundancia.

# Objetivo

**¿Cómo?**

Transformar el conjunto original de variables en otro conjunto de nuevas variables que sean combinaciones lineales de las anteriores pero que no estén correlacionadas entre sí, llamado **conjunto de componentes principales.**

# Componentes Principales

## Historia

La técnica de componentes principales es debida a Hotelling (1933), aunque sus orígenes se encuentran en los ajustes ortogonales por mínimos cuadrados introducidos por K. Pearson (1901).

# Componentes Principales

**Dadas $p$ variables originales**

Se buscan $q < p$ variables que sean combinaciones lineales de las $p$ originales, recogiendo la mayor parte de la información o variabilidad de los datos.

Si las variables originales no están correlacionadas, entonces no tiene sentido realizar un análisis de componentes principales.

# La característica que maximiza la variabilidad...

Es un buen factor para diferenciar objetos de un conjunto de datos.

**Por ejemplo**

Conjunto de datos con información de vehículos

- Característica 1: Cantidad de ruedas.
- Característica 2: Longitud del vehículo. (Mayor variabilidad)

Utilizando la segunda nos daríamos cuenta si el registro corresponde a un auto o un colectivo.

# Transformación PCA

Supongamos que se dispone de los valores de $p$-variables en $n$ elementos de una población dispuestos en una matriz $X$ de dimensiones $n \times p$, donde las columnas contienen las variables y las filas contienen los elementos.

Variables:

$$\{x_1, \ldots, x_p\}$$

# Transformación PCA

**¿Qué buscamos?**

PCA realiza una transformación del conjunto de datos $X$ que consiste de una traslación y una rotación, definidas de manera tal que la varianza del nuevo conjunto de variables sea máxima.

# Transformación PCA: la Primera componente

## Por ejemplo

$$y_1 = \sum_{j=1}^{p} a_{1j}(x_j - \bar{x}_j) = a_{11}(x_1 - \bar{x}_1) + ... + a_{1p}(x_p - \bar{x}_p)$$

con

$$\vec{a_1} = (a_{11}, a_{12}, ..., a_{1p}) \in \mathbb{R}^p$$

# Transformación PCA

**El conjunto de Componentes principales**

es una combinación lineal de las variables originales.

# Transformación PCA

Buscamos que $\vec{a_1}/\left\|\vec{a_1}\right\| = 1$ y que la $Var(y_1)$ resulte máxima en el conjunto de las combinaciones lineales posibles de las variables originales.

**Cargas**

Los coeficientes $a_{ji},\ i = 1, \ldots, p, j = 1, \ldots, p$ se denominan **cargas (o loadings)**.

¿Cómo hallar los valores de las cargas?
Son los autovectores de la matriz de covarianzas

# Este problema se resuelve

**Buscamos los $v1, \ldots, v_n$ y los $\lambda_1, \ldots, \lambda_n$**

$$det(S_X - \lambda_i I) = 0$$

y los $v_i$ tal que

$$S_X v_i = \lambda_i v_i$$

Entonces

$$y_1 = v_{11}x_1 + \cdots + v_{n1}x_n$$

$$y_2 = v_{12}x_1 + \cdots + v_{n2}x_n$$

y así con todos.
Además, el autovalor $\lambda_i$ es la varianza de la componente $i$.

# Transformación PCA

**Ordenando los autovalores de $S_X$**

de mayor a menor, $\lambda_1 \geq \lambda_2 \geq ... \geq \lambda_n$ podemos reducir la dimensionalidad tomando los autovectores correspondientes a los primeros $q$ autovalores, que son los que proveen mayor información (en términos de variabilidad).

# Transformación PCA

## Además

El porcentaje de variabilidad que poseen las primeras $q$ componentes es

$$Variabilidad = \frac{\sum_{i=1}^{q} \lambda_i}{\sum_{i=1}^{n} \lambda_i}$$

# Transformación PCA

**Entonces**

Para reducir la dimensionalidad de un data set considerando sus características más importantes, deberíamos requerir que la proyección cubra, por ejemplo, el 95 % de la variabilidad, o sea que

$$\textit{Variabilidad} \geq 0{,}95$$

# Algoritmo para calcular Componentes Principales

1. Paso 1:Tomar un conjunto de Datos $X$. Las variables deben estar en las columnas.
2. Paso 2: Restar la media de cada conjunto de variables. Calcular $X - \bar{X}$. Obtener un conjunto de datos con media cero.
3. Paso 3: Calcular la matriz de Covarianzas.
4. Paso 4: Calcular autovalores y autovectores de la mariz de covarianzas y ordenar los autovalores de mayor a menor.
5. Paso 5: Formar la matriz $E$ tomando los autovectores correspondientes a los mayores autovalores.
6. Paso 6: Calcular las nuevas variables $Y = (X - \bar{X})E$

# Organización

# Matriz Covarianzas vs. Matriz de Correlación

Si alguna de las variables, por ejemplo la primera, tiene valores mayores que las demás, la manera de aumentar la varianza es hacer tan grande como podamos la coordenada asociada a esta variable.

**Por ejemplo**

pasamos de medir en km. a medir en metros, el peso de esa variable en el análisis aumentará.

# Matriz Covarianzas vs. Matriz de Correlación

**Entonces**

Cuando las escalas de medida de las variables son muy distintas, la maximización de la varianza dependerá decisivamente de estas escalas de medida y las variables con valores más grandes tendrán más peso en el análisis.

# Matriz Covarianzas vs. Matriz de Correlación

**Si queremos evitar este problema**

Conviene utilizar las variables estandarizadas para calcular las componentes principales, de manera que las magnitudes de los valores numéricos de las variables originales sean similares.

# Matriz Covarianzas vs. Matriz de Correlación

## Lo cual es lo mismo que

Aplicar el análisis de componentes principales utilizando la matriz de correlaciones en lugar de la matriz de covarianzas.

## Cuando

Las variables tienen las mismas unidades, ambas alternativas son posibles.

# Ejemplo

Sea la matriz de covarianzas

$$S = \begin{pmatrix} 3 & 1 & 1 \\ 1 & 3 & 1 \\ 1 & 1 & 5 \end{pmatrix}$$

correspondiente a las variables aleatorias $(X_1, X_2, X3)$ con media cero.

1. Escribir las variables $(Y_1, Y_2, Y_3)$ de componentes principales y calcular qué proporción de la varianza total explica cada componente.

# Solución

Autovalores: $\lambda_1 = 6$ , $\lambda_2 = 3$ , $\lambda_3 = 2$
Autovectores:

$$E = \begin{pmatrix} V_1 & V_2 & V_3 \\ \hline -0{,}40 & -0{,}57 & 0{,}70 \\ -0{,}40 & -0{,}57 & -0{,}70 \\ -0{,}81 & 0{,}57 & 0 \end{pmatrix}$$

$$Y_1 = -0{,}40X_1 + (-0{,}40)X_2 + (-0{,}81)X_3$$
$$Y_2 = -0{,}57X_1 + (-0{,}57)X_2 + 0{,}57X_3$$
$$Y_3 = 0{,}70X_1 + (-0{,}70)X_2$$

# Solución

## Proporción de varianza de cada componente

Autovalores: $\lambda_1 = 6\ , \lambda_2 = 3\ , \lambda_3 = 2$

$$de\ Y_1 = \frac{6}{11}$$
$$de\ Y_2 = \frac{3}{11}$$
$$de\ Y_3 = \frac{2}{11}$$

# Interpretación de la Primera Componente

## Encuesta de Presupuestos Familiares

Los datos de una encuesta de presupuestos familiares en un país, presentan los gastos medios de las familias para las 51 provincias que lo componen, en el año 1980, utilizando seis variables:

# Interpretación de la Primera Componente

## Encuesta de Presupuestos familiares

- $X_1$ = alimentación
- $X_2$ = vestido y calzado
- $X_3$ = vivienda
- $X_4$ = mobiliario doméstico
- $X_5$ = salud
- $X_6$ = educación y cultura

# Interpretación de la Primera Componente

Calculamos el primer autovector y obtenemos:

**Encuesta de Presupuestos familiares**

La primera componente principal:
$Y_1 = 0{,}50x_1 + 0{,}22x_2 + 0{,}35x_3 + 0{,}33x_4 + 0{,}48x_5 + 0{,}49x_6$

$Y_1$ es una suma ponderada de todos los gastos, con mayor carga, de las variables $x_1$ (Educación y cultura), $x_5$ salud y $x_6$ educación. El menor peso lo tiene el gasto en $x_2$ vestido y calzado.

# Ejemplo teórico: Interpretación de la Primera Componente

- Si calculamos los valores de $Y_1$ para cada provincia y las ordenamos por esta nueva variable las provincias quedan ordenadas por su gastos.
- Explicación inmediata: muestra la capacidad de gasto de cada provincia.

# ¿Qué información aportan las cargas o *loadings*?

- Si la carga (coeficiente o *loading*) de una variable en la componente principal es positiva, significa que la variable y la componente tienen una correlación positiva.

# ¿Qué información aportan las cargas o *loadings*?

- Si por el contrario, la carga es negativa, este hecho indica que dicha variable se correlaciona en forma negativa con la primera componente.

# Interpretación:

La primera componente representa un índice (o una característica) por el cual se pueden ordenar los registros.

# Ejercicio Obligatorio: Utilizar una librería para calcular las componentes principales. Interpretar la primera componente.

El conjunto de datos europe.csv corresponde a características económicas, sociales y geográficas de 28 países de Europa. Las variables son:

- *Country*: Nombre del país.
- *Area*: área.
- *GDP*: producto bruto interno.
- *Inflation*: inflación anual.
- *Life.expect*: Espectativa de vida media en años.
- *Military*
- *Pop.growth*: tasa de crecimiento poblacional.
- *Unemployment*: tasa de desempleo.