Cuestionario 1 Estadística

Ian, Angel, Andrea, Adrian, Dante

May 2, 2024

1 Ejercicio 1

1. Diga si las siguientes enunciados son verdaderos o falsos. Argumente su respuesta

(1) VERDADERO: el PCA halla la máxima variabilidad y con eso forma la Cámara de Weyl

(2) VERDADERO: porque los eigenvectores asociados a los eigenvalores son ortogonales y los eigenvalores son la varianza

(3) VERDADERO: Hotteling-Fisher, trabajar los datos X reduciendo lo mas posible la dimensión. Aplicar una transformación ortogonal de tal manera que la pérdida de información sea controlable y pueda medirse.

2 Proceso del PCA

(1) Estandarización de los datos: restar la media y dividir por la desviación estándar de la muestra.

(2) Calcular la matriz de covarianza o correlación

(3) Obtener los eigenvalores y ordenarlos de mayor a menor (Cámara de Weyl)

(4) Obtener los eigenvectores

(5) Realizar descomposición espectral

(6) Obtener la varianza explicada de las k componentes

3 Ejercicio 3

Obs	X_1	X_2
1	-2	2
2	2	-2

$b_{11} = 0.7071$

$$b_{11} = \begin{bmatrix} 0.7071 \\ b_{12} \end{bmatrix} (b_{12}) < 0$$

$b_{11}^2 + b_{12}^2 = 1$

$$Xb_{11} = \begin{bmatrix} -2 & 2 \\ 2 & -2 \end{bmatrix} \begin{bmatrix} 0.7071 \\ -0.7071 \end{bmatrix}$$

b_{12}= -0.7071

3.1 Código de Python

Código de Python:

```
import matplotlib.pyplot as plt

import pandas as pd
from sklearn.preprocessing import StandardScaler
from sklearn import datasets # import base de datos
from sklearn.decomposition import PCA # descomposici n del PCA

df = pd.read_csv(
    filepath_or_buffer='https://archive.ics.uci.edu/ml/machine-learning-database
    header=None,
    sep=',). # conjuntos de datos

#Procesamiento de datos

df.columns=['sepal_len', 'sepal_wid', 'petal_len', 'petal_wid', 'class']
df.dropna(how="all", inplace=True)
df.to_csv("iris_download.csv", index = False)
df.head()

X = df.loc[:, 'sepal_len':'petal_wid'].values
y = df.loc[:, 'class'].values

pca = PCA()
X_r = pca.fit(X).transform(X)

target_names = df.iloc[:,4].unique()
target_names
```

```
def pca_scatter(pca1, pca2):
    plt.close
    plt.figure()
    colors = ['red', 'cyan', 'blue']
    lw = 2


    for color, target_name in zip(colors, target_names):
        plt.scatter(X_r[y == target_name, pca1], X_r[y == target_name, pca2], co
                    label=target_name)
    plt.legend(loc='best', shadow=False, scatterpoints=1)
    plt.title('PCA of IRIS Dataset: Components {} and {}'.format(pca1+1, pca2+1)
    plt.xlabel('Component {}'.format(pca1+1))
    plt.ylabel('Component {}'.format(pca2+1))
    plt.show()

pca_scatter(0, 1)

X_r = pca.fit(X).transform(X)
print('\nEigenvalores \n%s' %pca.explained_variance_)
print('Eigenvectores \n%s' %pca.components_)

def scree_plot():
    from matplotlib.pyplot import figure, show
    from matplotlib.ticker import MaxNLocator

    ax = figure().gca()
    ax.plot(pca.explained_variance_)
    ax.xaxis.set_major_locator(MaxNLocator(integer=True))
    plt.xlabel('Principal Component')
    plt.ylabel('Eigenvalue')
    plt.axhline(y=1, linewidth=1, color='r', alpha=0.5)
    plt.title('Scree Plot of PCA: Component Eigenvalues')
    show()

scree_plot()

def var_explicada():
    import numpy as np
    from matplotlib.pyplot import figure, show
    from matplotlib.ticker import MaxNLocator

    ax = figure().gca()
    ax.plot(np.cumsum(pca.explained_variance_ratio_))
    ax.xaxis.set_major_locator(MaxNLocator(integer=True))
    plt.xlabel('Nu mero de componentes')
```

```
    plt.ylabel('Varianza explicada acumulativa')
    plt.axvline(x=1, linewidth=1, color='r', alpha=0.5)
    plt.title('Varianza explicada del PCA por componente')
    show()

var_explicada()
```

3.2 Imagenes y Explicación del código

En el código y en las imágenes podemos ver que se realiza el PCA a la base de datos Iris la cual cuenta con datos correlacionados y lo que se busca es separar esos datos en las componentes principales y que con esto las variables ya no esten correlacionadas y asi poder obtener la varianza explicada de los datos.

4 Ejercicio 4

Explicación de imágenes

Se realiza un PCA a los datos de las figuras mostradas, el cual consiste en rotar los ejes para eliminar las correlaciones de los datos y el hecho de rotar los ejes hace que los puntos se vuelvan ortogonales.

5 Ejercicio 7

1. Diga si las siguientes enunciados son verdaderos o falsos. Argumente su respuesta

(1) Verdadero. Cada componente principal captura una porción única de la varianza en los datos. A medida que se agregan más componentes principales, se explora y explica más varianza en los datos. Por lo tanto, la proporción de varianza explicada por un componente adicional nunca disminuirá.

(2) Falso. Aunque es común que la proporción acumulativa de varianza explicada aumente a medida que se agregan más componentes principales, puede haber casos en los que agregar un componente adicional no mejore significativamente la capacidad del modelo para explicar la variabilidad en los datos. En tales casos, la proporción acumulativa podría dejar de crecer o incluso disminuir ligeramente.

(3) Falso. Incluir todas las posibles componentes principales no siempre conduce a un mejor entendimiento de los datos. Algunas de las componentes pueden contener ruido o información redundante, lo que no contribuye significativamente al entendimiento de la estructura subyacente de los datos. Además, utilizar todas las componentes puede llevar al sobreajuste y dificultar la interpretación del modelo.

(4) Verdadero. La gráfica scree es una representación gráfica de los valores propios (eigenvalues) de los componentes principales en orden descendente. Los valores propios representan la cantidad de varianza explicada por cada componente principal. En la gráfica scree, el punto donde la pendiente de la curva se aplana indica el número óptimo de componentes principales a retener, ya que representa un punto de inflexión donde la ganancia marginal en la varianza explicada es mínima.

6 Ejercicio 8

¿Cuáles de las aseveraciones pueden ser demostradas visulamente?

(1) x1 si esta más correlacionada con X2 qué con X3 ya que X1 y X2 están en la misma dirección y a una distancia cercana

(2) X3 si tiene la varianza más alta, pues es el vector de mayor longitud y en el biplot la longitud indica la varianza de la variable.

(3) Esta aseveración no se puede demostrar gráficamente.

7 Ejercicio 9

¿Cuáles de las aseveraciones son verdaderas?

(1) Esta afirmación es verdadera. Ambas partes de la ecuación representan la varianza total de los datos explicada por los dos componentes principales. Según la propiedad de los componentes principales, la suma de los cuadrados de los coeficientes de carga de cada componente principal debe ser igual a 1.

(2) Esta afirmación es verdadera. Dado que los componentes principales son ortogonales entre sí, los coeficientes de carga de componentes principales ortogonales (en este caso, b_{j1} y b_{j2}) deben sumar cero.

(3) Esta afirmación es verdadera. Como se mencionó anteriormente, la suma de los cuadrados de los coeficientes de carga de cada componente principal debe ser igual a 1, ya que representan la proporción de varianza explicada por cada componente principal.

8 Ejercicio 10

Determina cuáles de los siguientes (eigen)-vectores representa mejor la primer componente principal

(A) (1, 1, 1, 1): Este vector asigna el mismo peso a todas las variables. Puede ser apropiado si todas las variables tienen la misma importancia y variabilidad, pero en la mayoría de los casos esto no es cierto.

(B) (0.5, -0.5, 0.5, -0.5): Este vector asigna pesos positivos y negativos alternativamente a las variables. Esto podría ser útil si hay algunas variables que se correlacionan positivamente con la primera componente principal y otras que se correlacionan negativamente, pero generalmente esto no es común.

(C) (1, -1, 1, -1): Similar al caso anterior, este vector asigna pesos positivos y negativos alternativamente, lo que no es ideal para representar la primera componente principal.

(D) (0.7071, 0, (-0.7071), 0): Este vector asigna un peso positivo a las variables de edad y altura, y un peso negativo a las variables de peso e ingresos. Esto parece más prometedor ya que muestra una clara diferencia en la contribución de cada variable a la primera componente principal.

(E) (0.5, 0.5, 0.5, 0.5): Este vector asigna el mismo peso positivo a todas las variables, lo que podría no ser adecuado si algunas variables tienen una mayor variabilidad que otras.

Basándonos en esta evaluación, el vector que mejor representa la primera componente principal parece ser (D) (0.7071, 0, -0.7071, 0), ya que asigna pesos positivos a las variables que probablemente tengan una mayor variabilidad en los datos y pesos negativos a las variables con menor variabilidad

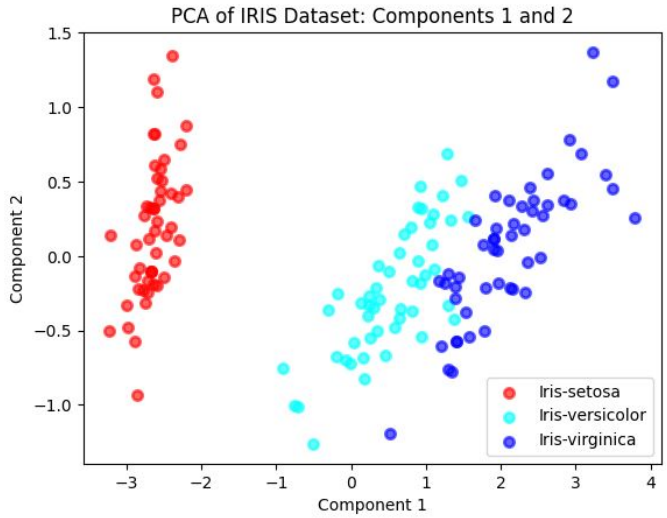


Figure 1: Componentes principales del PCA

```
Eigenvalores
[4.22484077 0.24224357 0.07852391 0.02368303]
Eigenvectores
[[ 0.36158968 -0.08226889  0.85657211  0.35884393]
 [ 0.65653988  0.72971237 -0.1757674  -0.07470647]
 [-0.58099728  0.59641809  0.07252408  0.54906091]
 [ 0.31725455 -0.32409435 -0.47971899  0.75112056]]
```

Figure 2: Eigenvalores y Eigenvectores obtenidos por el PCA

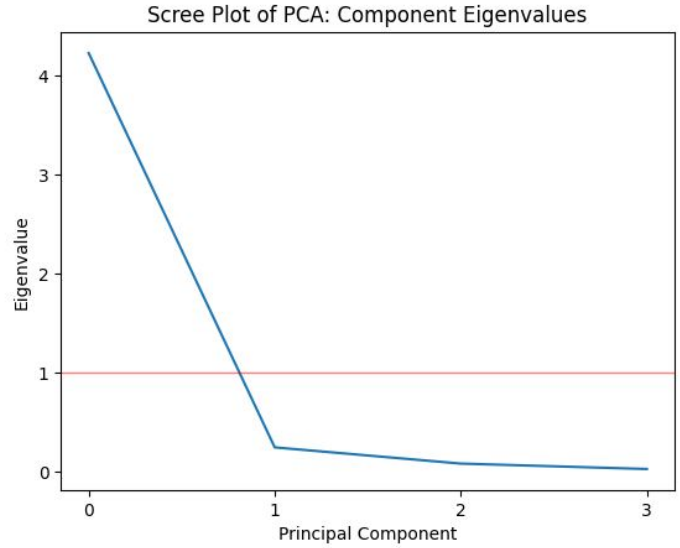
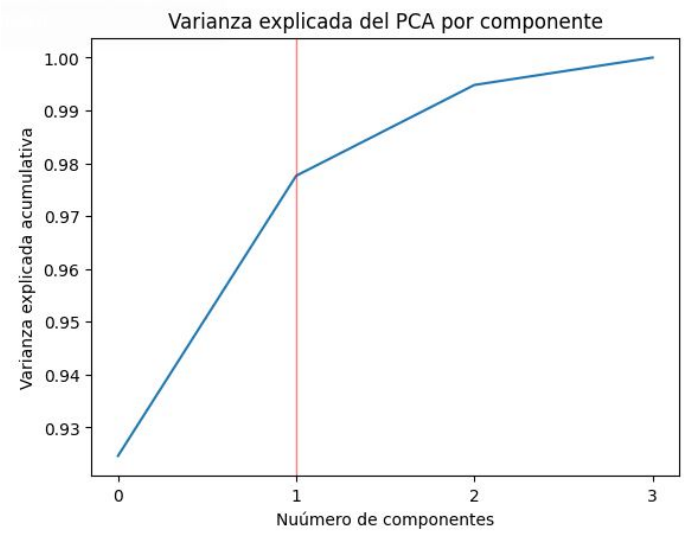


Figure 3: Scree Plot del PCA (Eigenvalores)


Varianza explicada del PCA por componente
Varianza explicada acumulativa
1.00
0.99
0.98
0.97
0.96
0.95
0.94
0.93
0
1
2
3
Nuúmero de componentes