Laboratorio de Datos

Primer Cuatrimestre 2024

Práctica N° 8: Componentes principales

Componentes Principales

1. Dada la siguiente tabla de datos correspondientes a la longitud y el ancho de las tortugas pintadas

Longitud	93	94	96	101	102	103	104	106
Ancho	76	78	80	84	85	82	83	83

   a) Normalizar las variables para que tengan media 0 y hacer el diagrama de dispersión. Estimar la presencia de correlación entre las variables a partir de este gráfico

   b) Calcular la matriz de covarianzas y hallar sus autovalores y autovectores.

   c) Hallar las componentes principales.

   d) Decidir si la información está mayormente representada en una de estas dos componentes.

   e) Indicar la proporción de la variabilidad explicada por cada una de ellas. ¿ A que conclusión puede llegar?

2. Sea $A = \begin{pmatrix} 3 & 1 & 1 \\ 1 & 3 & 1 \\ 1 & 1 & 5 \end{pmatrix}$ la matriz de covarianzas de una cierta muestra de datos cuya media es cero.

   a) Hallar los autovalores y autovectores de la matriz de covarianzas.

   b) Dar la expresión de las componentes principales z_1, z_2, z_3 e indicar la proporción de la variabilidad explicada por cada una de ellas.

   c) Hallar los *scores* de las primeras dos componentes principales correspondientes a la observación $x_1 = 2, x_2 = 2, x_3 = 1$ (es decir, los valores de z_1 y z_2 para dicha observación.

3. Implementar un programa que reciba como input un archivo de datos y un número p_acum y devuelva la mínima cantidad de componentes principales que deben considerarse para que el porcentaje de varianza acumulada sea mayor o igual que p_acum.

4. Considerando el archivo de datos `p8-chalets.csv` se pide:

   a) Graficar los diagramas de dispersión de las variables de a pares. Estimar la presencia de correlación entre las variables a partir de estos gráficos.

   b) Calcular la matriz de covarianzas.

   c) A partir de lo observado, resulta razonable pensar en un análisis de componentes principales para reducir la dimensión del problema?

d) Hallar la primera componente principal.

e) Indicar qué porcentaje de variabilidad total logra explicar esta componente.

5. Considerar el dataset `p8-iris.txt`, que representa información del largo y ancho del pétalo y del sépalo de diversas muestras de flores de la especie Iris, la cual se puede distinguir en varias subespecies. Aplicar el programa del ejercicio anterior para determinar la menor cantidad de componentes principales necesarias para alcanzar un 90 % de variabilidad. Graficar los datos transformados que se obtienen luego de reducir variables.

6. Con el objetivo de obtener índices útiles para la gestión hospitalaria basados en técnicas estadísticas multivariantes descriptivas se recogió información del Hospital de Algeciras correspondiente a los ingresos hospitalarios del período 2007-2008. Se estudiaron las siguientes variables habitualmente monitorizadas por el Servicio Andaluz de Sald del Sistema Nacional de Salud Español:

   NI: número de ingresos
   MO: tasa de mortalidad
   RE: tasa de egresos
   NE: número de consultas externas
   ICM: índice cardíaco máximo
   ES: número de estancias

   Las variables se midieron en un total de 22486 ingresos. En el archivo `p8-hospitales.csv` se aprecia la distribución de los valores obtenidos en las variables listadas por los servicios del hospital de Algeciras. Como hay variables con distintos órdenes de magnitud, en el archivo `p6-hospitales-escalado.csv` se escalaron las variables a un mismo rango. Utilizando el archivo escalado:

   a) Calcular las dos primeras componentes principales.

   b) ¿Qué porcentaje de variabilidad logra captar cada una de ellas?

   c) ¿Considera adecuado considerar dos componentes principales?

   d) Hallar la correlación entre las nuevas variables y las originales (por medio de R^2).

   e) ¿Se obtienen las mismas conclusiones si se utiliza el archivo sin escalar?

7. Se realiza un estudio sobre la calidad del agua en 4 ríos distintos del país. En el estudio se miden las concentraciones de 4 sustancias presentes en el agua, realizándose una medición por día durante los 365 días de un año. Los datos recolectados (generados artificialmente en este ejercicio) se encuentran en en el archivo `p8-calidad-agua.csv`. Cada columna representa una de las 4 variables medidas: x_1, x_2, x_3 y x_4.

   a) Realizar un gráfico de dispersión de las variables x_1 y x_2. ¿Cuántos clusters puede observar?

   b) Realizar un gráfico de dispersión de las variables x_3 y x_4. ¿Cuántos clusters puede observar?

   c) Realizar la descomposición en componentes principales de los datos y realizar un gráfico de dispersión de las dos primeras componentes principales z_1 y z_2. ¿Cuántos clusters puede observar?

d) Utilizando el método de clustering que considere apropiado, clasificar a los datos en 4 clusters utilizando solo las variables z_1 y z_2 y realizar nuevamente el gráfico de dispersión de z_1 y z_2 coloreando cada punto según el cluster al que pertenece.

Importante: al llamar al algoritmo de clustering utilice una matriz W que contenga como columnas solo las variables z_1 y z_2. En todos los puntos, puede utilizar sus propias funciones y código o las funciones de los paquetes de Python.