

PRINCIPAL COMPONENTS ANALYSIS

Laura del Pino Díaz

Concurso “Crea tu
programa para la
CASIO FX-CP400”

1 INTRODUCCIÓN

Principal Components Analysis (PCA) es un algoritmo aplicado ampliamente en el campo de la minería de datos, para la clusterización de datos, es decir para hacer un resumen de los mismos. Este resumen se caracteriza por plasmar los datos de entrada en unas nuevas dimensiones calculadas sobre las varianzas de las dimensiones originales.

La presente participación es una implementación de este algoritmo teniendo en cuenta las limitadas características sobre las que se ejecuta el programa, pero, a su vez, explotando su potencial mostrando las gráficas de los datos a la entrada del algoritmo y tras la ejecución del mismo. El fin último del algoritmo es introducir a los alumnos universitarios a la minería de datos de forma gráfica, permitiéndoles introducir una pequeña base de datos en forma de matriz y mostrarle los cambios.

2 Tabla de contenido

1	INTRODUCCIÓN	1
3	TUTORIAL DE USO	3
3.1	DATOS DE ENTRADA.....	3
3.2	GRÁFICOS DE SALIDA.....	4
3.3	INTERPRETACIÓN DE LA SALIDA.....	5
3.4	GUÍA PASO POR PASO	6
4	FUNCIONAMIENTO DEL PROGRAMA	7

3 TUTORIAL DE USO

El algoritmo de análisis de los componentes principales (*Principal Components Analysis o PCA*) presentado tiene como objetivo realizar una introducción gráfica a la minería de datos a estudiantes universitarios.

Todo algoritmo de minería de datos necesita de una base de datos como dato de entrada sobre la que realizar los cálculos. Se le mostrará al alumno los datos introducidos en un gráfico de nube de puntos bidimensional donde cada punto será representado con un cuadrado. Una vez finalizado el algoritmo se mostrarán sobre el mismo gráfico los datos obtenidos en las nuevas dimensiones.

3.1 DATOS DE ENTRADA

El algoritmo de análisis de los componentes principales (*Principal Components Analysis o PCA*) necesita de una base de datos. Esta base de datos se representa como una matriz que contiene una observación por cada fila y dos columnas¹ que representa las características observadas.

Se ha restringido el número total de elementos de la matriz a 60, en otras palabras 30 observaciones o 30 filas, puesto que en versiones tempranas de desarrollo se empleó una base de datos con un número de elementos superior y el sistema mostró mensajes de falta de espacio para cálculos que el usuario no debe afrontar. De forma que si se restringe a las 30 observaciones el único mensaje que recibe el usuario es el de error del propio programa pca advirtiéndole de las dimensiones excesivas que tiene la matriz de entrada, hecho fácilmente corregible por el usuario.

La matriz puede tener cualquier nombre que el usuario quiera. El programa le pedirá dicho nombre al usuario como se muestra en la Ilustración 1.

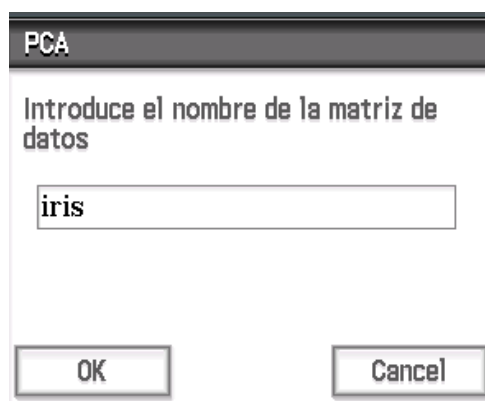


Ilustración 1: Ejemplo de introducción del nombre de la matriz.

¹ La restricción de las columnas está asociada a la representación gráfica que también es bidimensional.

3.2 GRÁFICOS DE SALIDA

El algoritmo mostrará al terminar un gráfico bidimensional en donde quedarán representados los datos introducidos a la entrada y los obtenidos por el algoritmo.

Los datos se representan de la siguiente forma:

- Con cuadrados: los datos de entrada.
- Con cruces: los datos de salida del algoritmo.

Esta información se le proporciona al usuario en la pantalla de salida tal y como se muestra en la Ilustración 2.

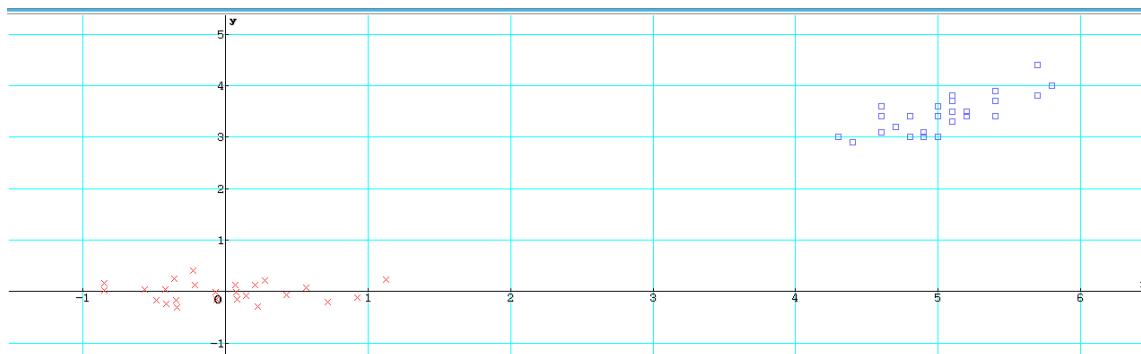


Ilustración 2: Ejemplo de representación obtenida.

3.3 INTERPRETACIÓN DE LA SALIDA

Los datos de salida representan los datos de entrada en unas nuevas dimensiones distintas de las originales.

Estas nuevas dimensiones están establecidas con respecto a las varianzas de los valores que toman los datos en cada una de las componentes. En la Ilustración 3 podemos ver representado con flechas las variaciones de los datos.

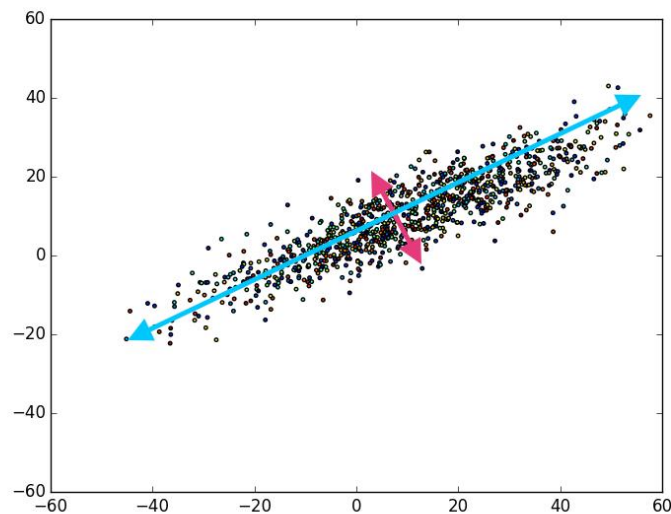


Ilustración 3: Representación de las nuevas dimensiones sobre el conjunto de datos

La intersección de ambas flechas recuerda a la intersección de los ejes X e Y comúnmente utilizados en matemáticas. Este hecho no es fortuito ya que con ellas se representarán las nuevas dimensiones del conjunto de datos.

PCA vuelve a calcular los datos sobre estas nuevas dimensiones, situando los nuevos ejes en la misma posición que los originales. De esta forma los datos a la salida quedan rotados y situados en torno al origen de coordenadas.

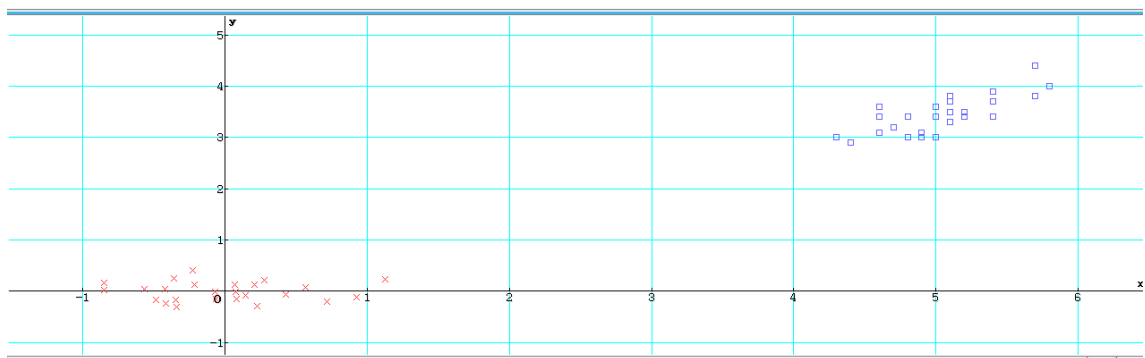


Ilustración 4: Ejemplo de la modificación de los datos, con los resultados del propio programa implementado

3.4 GUÍA PASO POR PASO

En la aplicación principal

Primero construya una matriz que tenga treinta o menos registros y dos columnas ya sea introduciéndola a mano o generándola con una función.

```
iris_gen N
[[[5.1,3.5],[4.9,3],[4.7,3.2],[4.6,3.1],[5.3,6],[5.4,3.9],[4.6,3.4],[5.3,4],[4.4,2.9],[4.9,3.1],[5.4,3.7],[4.8,3.4],
[4.8,3],[4.3,3],[5.8,4],[5.7,4.4],[5.4,3.9],[5.1,3.5],[5.7,3.8],[5.1,3.8],[5.4,3.4],[5.1,3.7],[4.6,3.6],[5.1,3.3],[
4.8,3.4],[5.3],[5.3,4],[5.2,3.5],[5.2,3.4]]iris
```

Ilustración 5: Programa generador de la matriz que contiene los primero 30 elementos de la base de datos de las flores de iris.

Seguidamente ejecute el programa `pca()`. Le saldrá un diálogo pidiendo el nombre de la matriz que acaba de crear.

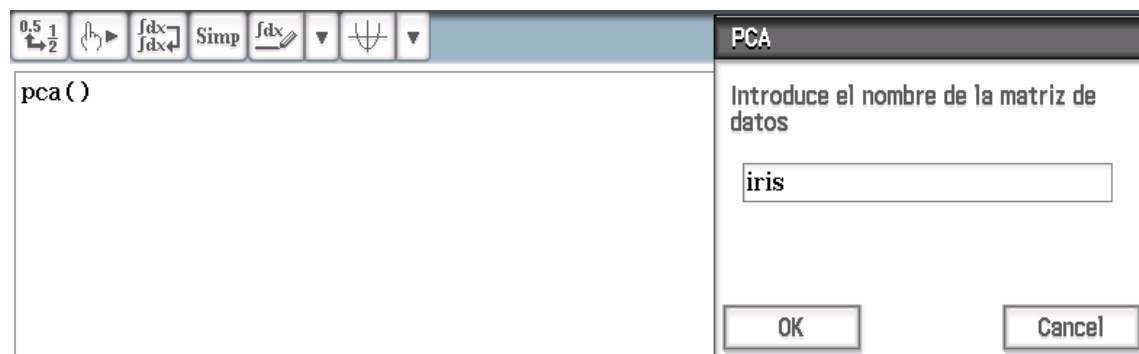


Ilustración 6: Introducción del nombre de la matriz

Espere a que el programa termine para ver las gráficas.

Los cuadrados son los datos originales. Las cruces son los datos obtenidos por el PCA.

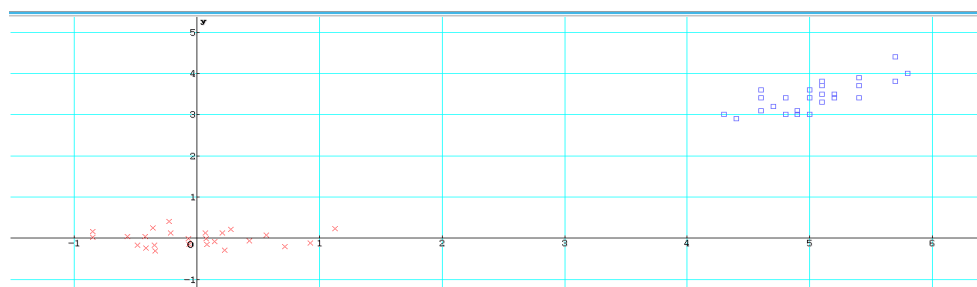


Ilustración 7: Resultados del programa

4 FUNCIONAMIENTO DEL PROGRAMA

La implementación del algoritmo PCA se lleva a cabo de la siguiente forma:

- Se calcula los centroides o medias de cada uno de las características o columnas de la matriz. $C_j = \frac{1}{m} \sum_{i=1}^m X_{ij}$
- Se normalizan los valores restándoles a los valores de cada columna su centroide correspondiente. $X_{ij} = X_{ij} - C_j$
- Se calcula las covarianzas multiplicando la matriz de datos normalizados por su traspuesta y dividiendo entre el número de observaciones. $Z = X^t X / m$
- Se calculan los autovalores y autovectores (V) de la matriz de covarianzas.
- Se ordenan los autovectores según el orden descendente de los autovalores.
- Se obtienen los resultados sobre los nuevos ejes. $Y = XV$

Antes de la ejecución anterior se realiza la elaboración del gráfico con los datos originales y al terminar la ejecución se elabora el gráfico con los resultados para después mostrarlos ambos.

La base de datos original, aquella que tiene el nombre que se ha introducido al comienzo queda intacta para el uso posterior que le quiera dar el usuario.

La matriz de autovectores utilizada para el cálculo de los nuevos valores y la matriz con los resultados se eliminan por el espacio de memoria que podrían ocupar y que dificultaría el funcionamiento correcto de las demás aplicaciones.