Lab 8. Reducción de dimensionalidad con PCA

Objetivos

En esta práctica se programará el algoritmo del Análisis de Componentes Principales (PCA) y se probará con el conjunto de datos Iris Flower dataset, que caracteriza tres tipos de flores (Iris setosa, Iris virginica e Iris versicolor) mediante cuatro variables: longitud y anchura de sépalos y de pétalos. El código que habrá que completar y entregar se encuentra en el script `main_PCA.m`. Este script y el archivo de datos (`datos_iris.mat`) se encuentran en el archivo comprimido Lab8_PCA.zip.

Como siempre, los datos están contenidos en la variable de Matlab `X` y sus clases se indican en la variable de Matlab `Y`. Los patrones estarán contenidos en las columnas de las matrices de datos (es decir, cada patrón está en una columna distinta de la variable `X`).

1. Reducción de dimensionalidad mediante Análisis de Componentes Principales (PCA)

El script de esta práctica `main_PCA.m` realiza la lectura de un fichero de datos que contiene el set de datos de las flores iris que es muy utilizado en comprobación de algoritmos de clasificación.

Este código tiene que realizar una reducción de la dimensionalidad por medio de un análisis de componentes principales (PCA) para eliminar variables que puedan introducir redundancia en los datos. Este código está estructurado en tres partes.

1 – Lectura del fichero de datos.

2 – Plot tridimensional de las variables en las que cada coordenada se corresponde con una variable, el tamaño de los puntos con la cuarta variable y el color nos indica la clase a la que pertenece cada muestra. En la Figura siguiente se ve dicha gráfica.

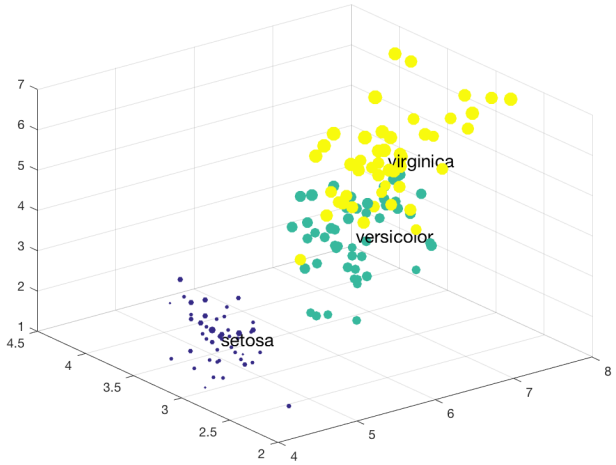

Figura 1. Plot tridimensional de los datos Iris

3 – Análisis de componentes principales de los datos para reducirlos a una dimensionalidad *k*.

4 – Plot de los resultados de la PCA sobre un mapa de dos dimensiones, siempre y cuando *k=2*.

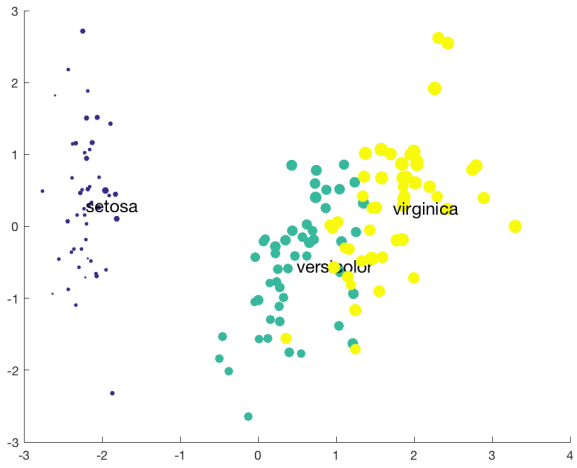

Figura 2. Plot en dos dimensiones de los datos Iris

En el código de la práctica falta por programar el algoritmo del PCA, por lo que habrá que completarlo con las instrucciones necesarias. En el código se indica con comentarios los pasos del PCA y dónde hay que escribir las instrucciones para llevarlos a cabo, de acuerdo al algoritmo visto en clase. Estos pasos, y algunas funciones útiles para llevarlos a cabo son:

1 – Normalización de los datos.

2 – Cálculo de la matriz de covarianza (**función útil:** `cov`).

3 – Obtención de los autovalores y autovectores de la matriz de covarianza (**función útil:** `eig` o `svd`).

4 – Ordenación de los autovectores en función del valor sus autovalores asociados de mayor a menor.

5 – Selección de los autovectores que tienen los k autovalores mayores, que serán los componentes principales. De esta manera se crea la matriz de transformación.

6 – Obtención de los nuevos datos. Para ello se multiplica la matriz de transformación por la matriz de datos.

El resultado tiene que ser similar al de la Figura 2, aunque puede ocurrir que los ejes estén invertidos, en función de cómo se obtengan los autovectores.