INFORME IRIS DATASET



Fuente: "Análisis de datos exploratorios en R, The University of Waikato, 2013"

Marta Divassón

Profesor: J.MLópez Zafra

El Iris dataset está compuesto por 150 observaciones de especies de flores conocidas como Iris Setosa, Iris Versicolor e Iris Virginica. De estas conoceremos 4 características esenciales, que son: la longitud y la anchura de pétalos y sépalos. De cada una se extraen 50 observaciones medidas con variables y atributos como:

- Tipo de flor (variable categórica)
- El largo y ancho del pétalo (varibale numérica)
- El largo y ancho del sépalo (variable numérica).

Este dataset se utiliza en R y se carga mediante el siguiente comando:

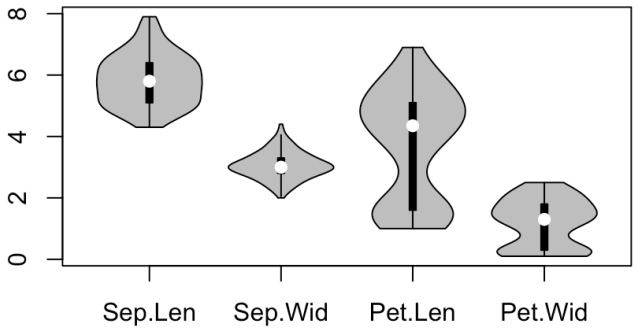
```
>library(datasets)
>str(iris)
```

A grandes rasgos, de los datos obtenidos al cargar las 5 filas del subconjunto que contienen las características: "Sepal leght, Sepal witdh, Petal Legth, Petal Width y Especies". Si hacemos una comparación entre los datos obtenidos hay una diferencia significativa en la longitud del pétalo de la setosa, siendo estos más pequeños que la versicolor y virginica.

Esto puede comprobarse ejecutándose un comando en el que se verifique qué longitudes de pétalo son inferiores a 2 o a 3, por ejemplo, dando el nombre de las especies que cumplan esto. En efecto, saldrá la setosa.

Bien, para conocer datos estadísticos relacionados con este dataset, ejecutamos el comando, >summary(iris), que nos da para cada característica, un mínimo, un máximo, la media y mediana y el primer y tercer cuartil.

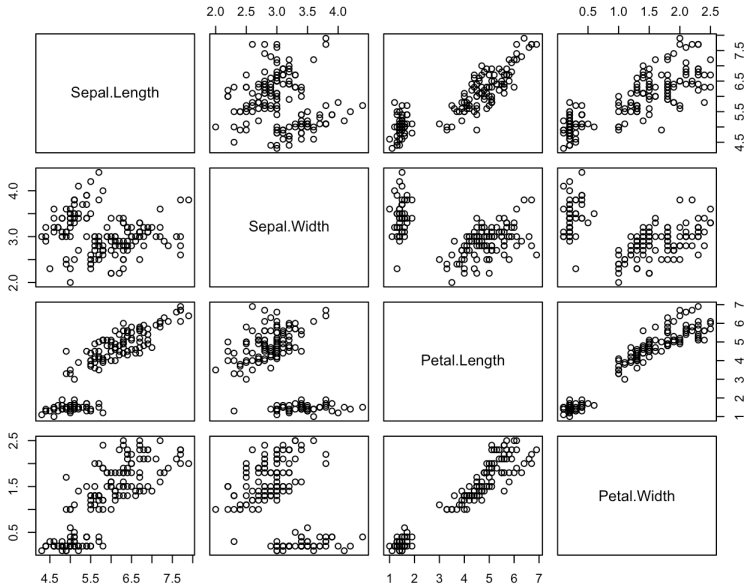
Para tener una idea más clara de los números, vamos a representarlos gráficamente según las 4 características. En el podemos observar la media de la longitud y anchura de los pétalos y sépalos entre los 3 diferentes tipos de flores.



Para conocer la relación entre variables, hay que utilizar medidas de dispersión para obtener el nivel de relación entre variables. Para ello, primero vamos a observar la correlación entre las 4 combinaciones de variables obteniendo la tabla reflejada en el Apéndice. Los resultados positivos indican correlación, el 1 significa que están perfectamente correlacionadas mientras que los negativos indican correlación inversa. En este caso las más relacionadas son el *largo del pétalo y sépalo* (exceptuando la relación entre variables que son iguales), y las inversamente relacionadas son el *largo del pétalo y el ancho del sépalo.* La correlación mide la fuerza como la relación lineal cuando comparamos.
En segundo lugar, la covarianza nos va a dar medidas no estandarizadas (es decir, que no estén entre 1 y -1) sino que pueden oscilar desde números infinitos. En el caso dela longitud del pétalo, la relación entre ambas da 3.1162779. Como el coeficiente de correlación siempre depende de la covarianza, por tanto, una covarianza positiva

indica una correlación positiva, en este caso mayor el la relación. En caso contrario, la anchura y longitud del sépalo tienen una covarianza negativa, -0.0424340, lo que indica que están negativamente correlacionadas.

El siguiente gráfico de dispersión incluye todas las variables y resultados, al que habría que darle color para tener resultados más específicos y fáciles de observar.



De lo aplicado, extrapolamos que Iris es una herramienta sencilla a la hora de empezar a utilizar R. Solo tiene una combinación de 150 variables y es una forma relativamente fácil de empezar visto con estos gráficos. La finalidad de este dataset es asignar y clasificar diferentes algoritmos a los 3 tipos de flores para guardarlos con los nuevos atributos en Iris para que otros usuarios dispongan de ello en formato CSV.

APÉNDICE:

```
> library(datasets)
> str(iris)
'data.frame':   150 obs. of  5 variables:
 $ Sepal.Length: num  5.1 4.9 4.7 4.6 5 5.4 4.6 5 4.4 4.9 ...
 $ Sepal.Width : num  3.5 3 3.2 3.1 3.6 3.9 3.4 3.4 2.9 3.1 ...
 $ Petal.Length: num  1.4 1.4 1.3 1.5 1.4 1.7 1.4 1.5 1.4 1.5 ...
 $ Petal.Width : num  0.2 0.2 0.2 0.2 0.2 0.4 0.3 0.2 0.2 0.1 ...
 $ Species     : Factor w/ 3 levels "setosa","versicolor",..: 1 1 1 1 1 1 1 1 1 1 ...
> View(iris)
>
> # Obtener 5 primeras filas de cada subconjunto
> subset(iris, Species == "setosa")[1:5,]
  Sepal.Length Sepal.Width Petal.Length Petal.Width Species
1          5.1         3.5          1.4         0.2  setosa
2          4.9         3.0          1.4         0.2  setosa
3          4.7         3.2          1.3         0.2  setosa
4          4.6         3.1          1.5         0.2  setosa
5          5.0         3.6          1.4         0.2  setosa
> subset(iris, Species == "versicolor")[1:5,]
   Sepal.Length Sepal.Width Petal.Length Petal.Width    Species
51          7.0         3.2          4.7         1.4 versicolor
52          6.4         3.2          4.5         1.5 versicolor
53          6.9         3.1          4.9         1.5 versicolor
54          5.5         2.3          4.0         1.3 versicolor
55          6.5         2.8          4.6         1.5 versicolor
> subset(iris, Species == "virginica")[1:5,]
    Sepal.Length Sepal.Width Petal.Length Petal.Width   Species
101          6.3         3.3          6.0         2.5 virginica
102          5.8         2.7          5.1         1.9 virginica
103          7.1         3.0          5.9         2.1 virginica
104          6.3         2.9          5.6         1.8 virginica
105          6.5         3.0          5.8         2.2 virginica
>
> # De la columna especies para todas las filas en las que "Petal.Legth" sea inferior a 2
o 3"
> subset(iris, Petal.Length < 2)[,"Species"]
 [1] setosa setosa setosa setosa setosa setosa setosa setosa setosa setosa setosa setosa
setosa
[14] setosa setosa setosa setosa setosa setosa setosa setosa setosa setosa setosa setosa
setosa
[27] setosa setosa setosa setosa setosa setosa setosa setosa setosa setosa setosa setosa
setosa
[40] setosa setosa setosa setosa setosa setosa setosa setosa setosa setosa setosa setosa
Levels: setosa versicolor virginica
>
> # Ver un gráfico especial
> library(vioplot)
> vioplot(iris$Sepal.Length,iris$Sepal.Width,iris$Petal.Length,iris$Petal.Width,
+         names=c("Sep.Len","Sep.Wid","Pet.Len","Pet.Wid"),
+         col="gray")
> print("Correlacion entre variables")
[1] "Correlacion entre variables"
> correlacion<- cor(iris[,1:4])
> round(correlacion,3)
             Sepal.Length Sepal.Width Petal.Length Petal.Width
```

```
Sepal.Length        1.000      -0.118       0.872       0.818
Sepal.Width        -0.118       1.000      -0.428      -0.366
Petal.Length        0.872      -0.428       1.000       0.963
Petal.Width         0.818      -0.366       0.963       1.000

> print("Covarianza entre variables")
[1] "Covarianza entre variables"
> covarianza<- cov(iris[,1:4])
> round(covarianza,3)
             Sepal.Length Sepal.Width Petal.Length Petal.Width
Sepal.Length        0.686      -0.042        1.274       0.516
Sepal.Width        -0.042       0.190       -0.330      -0.122
Petal.Length        1.274      -0.330        3.116       1.296
Petal.Width         0.516      -0.122        1.296       0.581

> # Scatterplots
> pairs(iris[,1:4])
```