K-Vecinos más cercanos iris (kNN)

Lino Oswaldo Sanchez

27/5/2022

Introducción

Con este análisis de K-Vecinos(vecino más próximo) es un método para clasificar observaciones "casos" basando en su parecido a las otras observaciones.Es un método de clasificación no paramétrico. Es decir, no requiere asumir ninguna distribución para variable aleatoria. La idea es buscar, para una nueva observación que se requiere clasificar, sus k vecinos más cercanos. Es decir, las observaciones más cercanas respecto a una medida de distancia.

Librerías necesarias

Convertimos la base en un data frame y la renombramos **Z**

```
Z<-as.data.frame(iris)
colnames(Z)
```

```
## [1] "Sepal.Length" "Sepal.Width"  "Petal.Length" "Petal.Width"  "Species"
```

Para este Anáslisis usaremos la matriz **iris** base precargada en R, esta contiene las variables de que variables que están dentro son datos sobre los tipos de flores tales cono largo del petalo, ancho del petalo, largo del sepalo, ancho del sepalo y especie.

Tenemos 5 variables de las cuales 4 son cuantitativas numéricas y una carácter.

Definición de variables

```
x<-Z[,1:4]
y<-Z[,5]
```

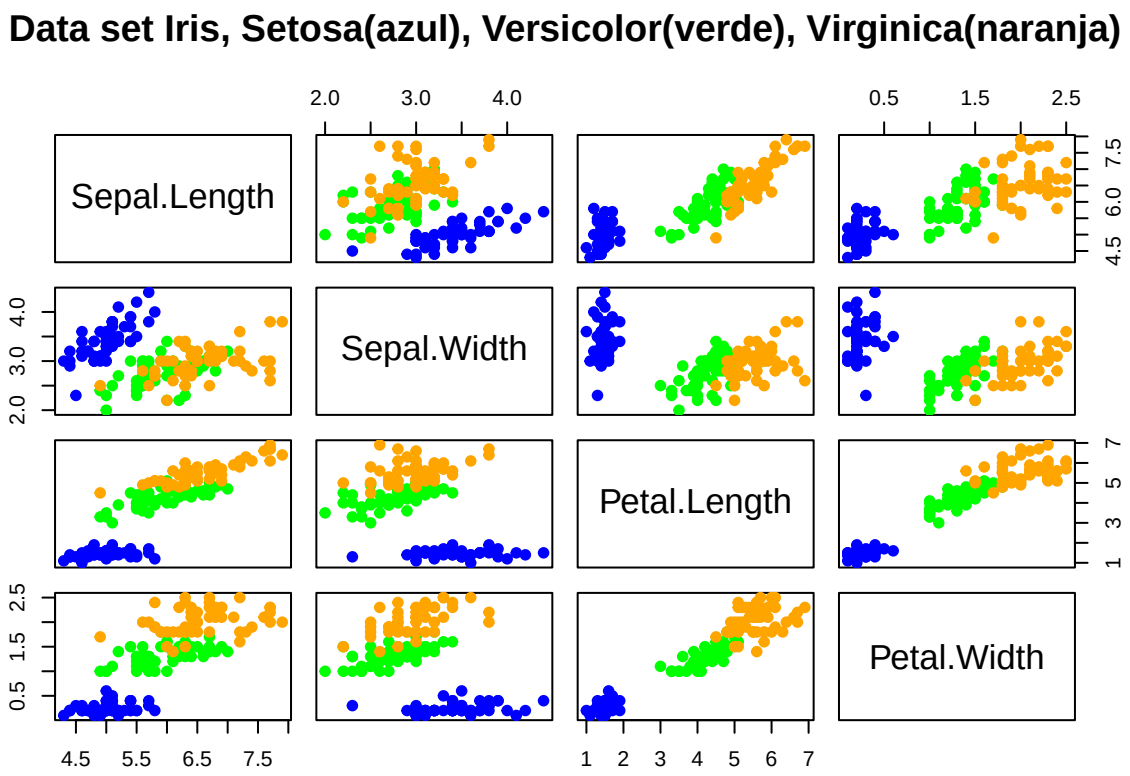
Definimos la matriz de datos que trabajaremos con las variables de la 1 a la 4 y la 5 como variable respuesta, que contiene las clasificaciones.

```
n<-nrow(x)
p<-ncol(x)
```

Se definen las variables y las observaciones

Gráfico scatter plot

```
# Creacion de un vector de colores
col.iris<-c("blue","green","orange")[y]
pairs(x, main="Data set Iris, Setosa(azul), Versicolor(verde), Virginica(naranja)", pch=19,col=col.iris
```



Con el scatter plot observamos que si hay correlación entre algunas de las variables, exceptuando algunas que se encuentran en la parte exterior del gráfico, donde la dispersión es nula parea alguna especies en especifico para setosa.

Algoritmo k-vecinos más próximos

Librería necesaria

Se fija una "semilla" *para obtener los mismos valores al replicar el ejercicio.*

```
set.seed(1000)
```

Creación de los ciclos

Inicialización de una lista vacía de tamaño 20

```
knn.class<-vector(mode="list",length=20)
knn.tables<-vector(mode="list", length=20)
```

Clasificaciones erróneas

```
knn.mis<-matrix(NA, nrow=20, ncol=1)
```

crecion de una funcion

En la que contiene las listas vacías que creamos anterior mente.

```
for(k in 1:20){
  knn.class[[k]]<-knn.cv(x,y,k=k)
  knn.tables[[k]]<-table(y,knn.class[[k]])
  # la suma de las clasificaciones menos las correctas
  knn.mis[k]<- n-sum(y==knn.class[[k]])
}
```

Número óptimo de k-vecinos

```
which(knn.mis==min(knn.mis))
```

```
## [1] 14 18 19
```

Se visualizan los resultados que nos arrojó el ciclo con el error más bajo.

```
knn.tables[[14]]
```

```
##
## y            setosa versicolor virginica
##   setosa         50          0         0
##   versicolor      0         48         2
##   virginica       0          1        49
```

```
knn.tables[[18]]
```

```
##
## y            setosa versicolor virginica
##   setosa         50          0         0
##   versicolor      0         48         2
##   virginica       0          1        49
```

```
knn.tables[[19]]
```

```
##
## y            setosa versicolor virginica
##   setosa         50          0         0
##   versicolor      0         48         2
##   virginica       0          1        49
```

En los tres casos el resultado de la clasificación es igual, clasifica bien **setosa** pero par el caso de **versicolor** solo 48 de estas están bien clasificadas como esa especie y dos de las misma están confundidas como "virginica",y para **virginica** 49 estan bien clasificadas y una esta confundida como "versicolor".

Esta interpretación es la misma para los tres resultados pues arrogan la misma clasificación.

Se señala el k mas eficiente:

En este caso es el 14, por el echo de ser más bajo lo elegimos

```
k.opt<-14
knn.cv.opt<-knn.class[[k.opt]]
```

Se visualiza la tabla de contingencia con las clasificaciones buenas y malas

```
knn.tables[[k.opt]]
```

```
##
## y            setosa versicolor virginica
##   setosa         50          0         0
##   versicolor      0         48         2
##   virginica       0          1        49
```

Vemos lo que anterior mente fu explicado.

Cantidad de observaciones mal clasificadas

```
knn.mis[k.opt]
```

```
## [1] 3
```

Como podemos apreciar solo 3 están mal clasificadas al tener este valor y tomando en cuenta que la muestra es de 150 observaciones(flores) las observaciones mal clasificadas son muy bajas así que podemos trabajar con la clasificación sin problema.

Error de clasificacion (MR)

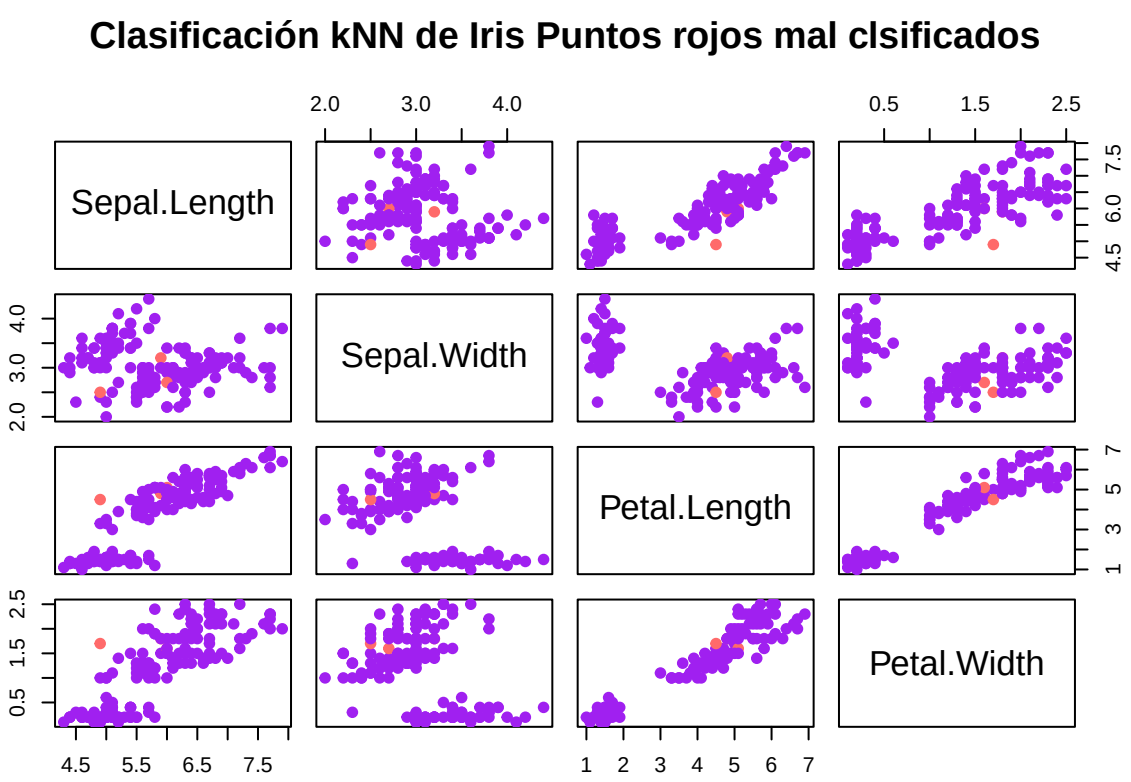
```
knn.mis[k.opt]/n
```

```
## [1] 0.02
```

El error resultante es muy baso así que confirmamos que la clasificación es buena y puede ser utilizada, ya queda como criterio individual del investigador si encuentra mas factible re-acomodar las observaciones mal clasificadas al grupo que se les fue asignado o simplemente descartar esas observaciones.

Gráfico identificando las clasificaciones correctas y erróneas.

```
# Grafico de clasificaciones
col.knn.iris<-c("indianred1","purple")[1*(y==knn.cv.opt)+1]
pairs(x, main="Clasificación kNN de Iris Puntos rojos mal clsificados",
      pch=19, col=col.knn.iris)
```



Es apreciable que las observaciones mal clasificadas son muy pocas casi imperceptibles por eso es que el investigador deberá tomar una decisión el estadístico ya ha echo su trabajo, y emitido sus recomnedaciones.