Tarea - KNN regression

David Cardoner & Arnau Mercader

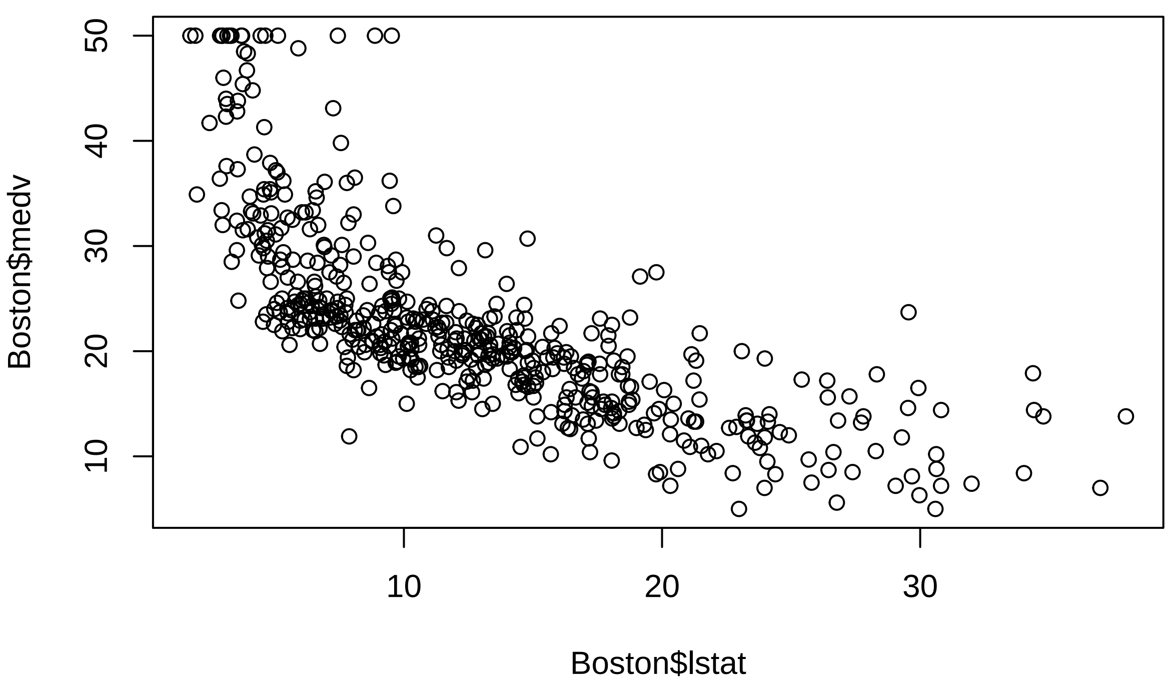
Objetivo

Aplicar el algoritmo *knn* en el dataset Boston, disponible mediante la librería **MASS** de R. Tenemos que explicar la variable **medv** en función de **lstat**. Veamos las primeras líneas de nuestro dataset y un gráfico bivariante del modelo a analizar.

```
library(MASS)
data(Boston)
head(Boston,3)
```

```
     crim zn indus chas   nox    rm  age    dis rad tax ptratio  black
1 0.00632 18  2.31    0 0.538 6.575 65.2 4.0900   1 296    15.3 396.90
2 0.02731  0  7.07    0 0.469 6.421 78.9 4.9671   2 242    17.8 396.90
3 0.02729  0  7.07    0 0.469 7.185 61.1 4.9671   2 242    17.8 392.83
  lstat medv
1  4.98 24.0
2  9.14 21.6
3  4.03 34.7
```

```
plot(Boston$lstat, Boston$medv)
```



Función KNN

```
knn_r <- function(xy,p,resp,k=3){
  d_st_xy <- as.matrix( dist(c(p,xy)) )[1,-1]
  d_st_xy_k <- sort(d_st_xy,partial=k)[k]
  N_st_k <- unname( which(d_st_xy <= d_st_xy_k) )
  return(pred=sum(resp[N_st_k]/k))
}
```

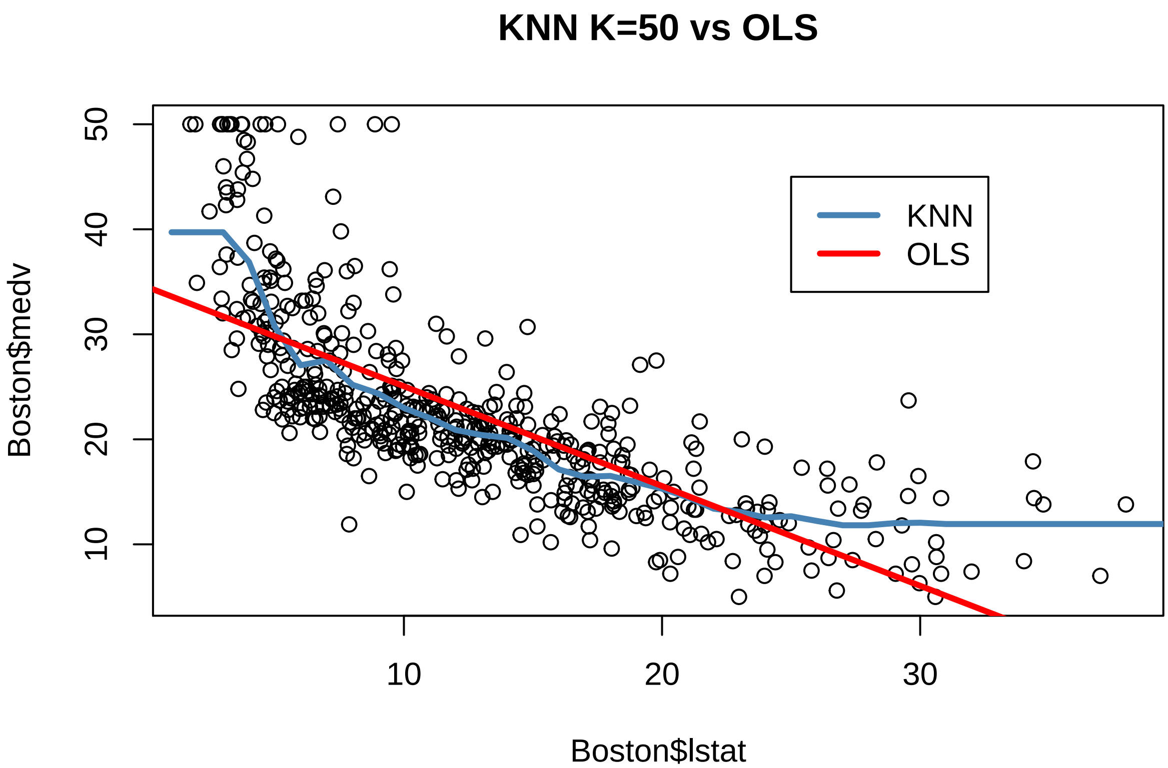
Generamos una secuencia de 1 a 40 y usamos la función **knn_r** descrita anteriormente con k=50.

```
t <- seq(1,40,1)
t_size <- length(t)
y_pred <- rep(0, t_size)
k <- 50
for (i in 1:t_size){
y_pred[i] <- knn_r(xy=Boston[, 13],resp=Boston[,14], p=t[i],k=k)
}
```

KNN K=50 vs. OLS

A continuación se muestra un gráfico con la regresión obtenida con KNN usando K=50, y el clásico modelo lineal. Vemos como el modelo lineal es mucho más rígido.

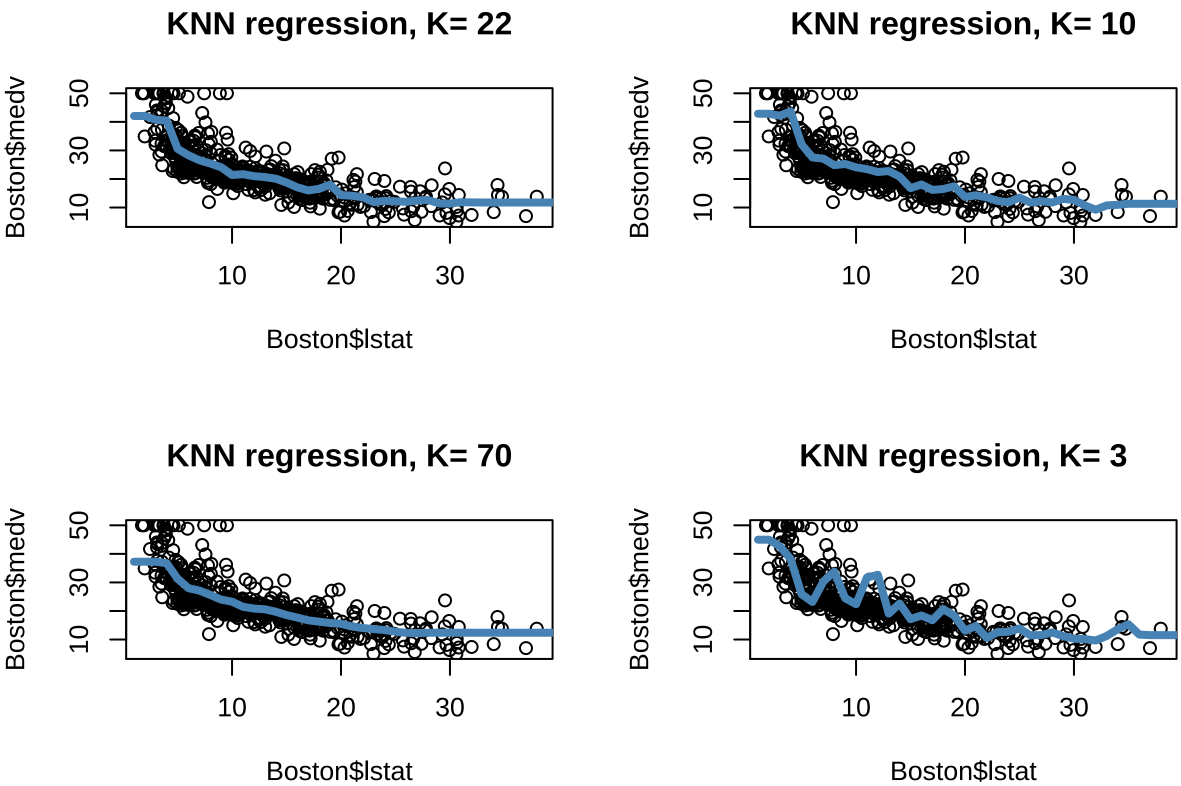
```
lm0 <- lm(medv ~ lstat ,data = Boston)

plot(Boston$lstat, Boston$medv,main = 'KNN K=50 vs OLS')
legend(x=25,45,c('KNN','OLS'),
       lwd=c(3,3),col=c('steelblue','red'))

points(t, y_pred, type="l", col='steelblue', lwd=3)
abline(lm0,col='red',lwd=3)
```



Distintos valores K

En la práctica, se suele usar k=$\sqrt{n}$, donde n son el nº de observaciones. En este caso k=22.4944438. Asi que podemos usar K=22 como una opción de K.

```
par(mfrow=c(2,2))
t <- seq(1,40,1)
t_size <- length(t)
y_pred <- rep(0, t_size)
k_opt <- c(22,10,70,3)
for (kk in 1:length(k_opt)) {
for (i in 1:t_size){
y_pred[i] <- knn_r(xy=Boston[, 13],resp=Boston[,14], p=t[i],k=k_opt[kk])
}
plot(Boston$lstat, Boston$medv,main = paste('KNN regression, K=',k_opt[kk]))
points(t, y_pred, type="l", col='steelblue', lwd=4)
}
```



Comentarios

Vemos como el valor K que se escoge, influye mucho en nuestro análisis. También se podrían estudiar diferentes métricas para calcular las distancias como también normalizar o transformar los datos. El método KNN es bastante flexible ya que al ser un método no paramétrico no asume ninguna distribución adyacente a los datos. Un incoveniente es que requiere tratar bastante las variable categóricas que se incorporen en el modelo, como también require tiempo estudiar el comportamiento del algoritmo en función de distintas métricas y valores K.