Sesión 3 - Introducción a los K Vecinos más cercanos

Técnica de clasificación y regresión:

1. Clasificación

Sea { ( X1 , Y1) , … , ( Xn , Yn) } una m.a donde Xi  ϵ RP y Y i ϵ{ C1, … ,CJ } =C

Ahora sea X m.a nueva observación cuya clase Y ϵ C

se desea predecir,

Primero se calculan las instancias de X a X1, … , Xn :

{ X, i = || X - Xi || , i=1, … , n }

Ahora se define E X , (1) = min i E X , i (La menor distancia)

E X , (2) = min i E X , i \ E X , (I) (La segunda menor distancia)

.

.

.

E X , (n) = max i E X , i (La mayor distancia)

Los K Vecinos más cercanos a X son los Xj que satisfacen || X –Xj || < E X , (K)

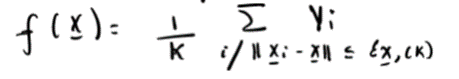
Referimos el clasificador de los K Vecinos más cercanos como

d: RP 🡪 C

d( X ) 🡪 arg max j Σ { yi = Cj }, donde || Xi – X || < E X , (K)

1. Regresión: Es este caso Yi ϵ R

La función de regresión es:



Si X0 es un valor particular, entonces f (X0) es una aproximación de E [ Y | X = X0]

¿Qué es E [ Y| X = X0] ?

Sea ( Y, X ) una tupla de r.a con fdp conjunta f Y,X ( Y, X).

Dado X = X0 , queremos aproximar Y por g ( X ), con g: RP 🡪 R.

El costo cuadrado de g es



En la expresión azul, X está fija y Y es la variable de integración. Cuando X está fija entonces

f (Y, X) = f Y|X = X (y) fx (X).

Recordemos que la fdp condicional de Y dado X = X0 es



es una función solo de y. Así





La formula (2) indica que para minimizar el costo cuadrático de g() se puede hacer la minimización condicionado sobre X y luego tomando la esperanza con respecto a X:



Para minimizar esta expresión con respecto a C, derivamos con respecto a C e igualamos la derivada a O:

Si se reemplaza (5) en (2) se obtiene la pérdida cuadrática de g:



Por otro lado el MSE o Pérdida Cuadrática del estimador (4) del parámetro δ es:



Esto quiere decir que el MSE de g() se puede ver como:

