Introducción a los K-vecinos más cercanos

Técnica de clasificación y regresión.

1) Clasificación: Sea $\{(\underline{X}_1, Y_1), \dots, (\underline{X}_n, Y_n)\}$ una m.a. donde $\underline{X}_i \in \mathbb{R}^p$ y $Y_i \in \{c_1, \dots, c_J\} = \mathcal{C}$

Ahora sea $\underline{X}$ una nueva observación cuya clase $Y \in \mathcal{C}$ se desea predecir.

Primero se calculan las distancias de $\underline{X}$ a $\underline{X}_1, \dots, \underline{X}_n$:

$$\varepsilon_{\underline{X},i} = \| \underline{X} - \underline{X}_i \| \,, \quad i = 1, \dots, n.$$

Ahora se define

$$\varepsilon_{\underline{X},(1)} = \min_i \varepsilon_{\underline{X},i} \quad \text{(la menor distancia)}$$

$$\varepsilon_{\underline{X},(2)} = \min_i \varepsilon_{\underline{X},i} \setminus \varepsilon_{\underline{X},(1)} \quad \text{(la segunda menor distancia)}$$

$$\vdots$$

$$\varepsilon_{\underline{X},(n)} = \max_i \varepsilon_{\underline{X},i} \quad \text{(la mayor distancia)}$$

Los K vecinos más cercanos a $\underline{X}$ son los $\underline{X}_j$ que satisfacen $\| \underline{X} - \underline{X}_j \| \leq \varepsilon_{\underline{X},(K)}$

Definimos el clasificador de los K vecinos más cercanos como

$$d : \mathbb{R}^p \longrightarrow \mathcal{C}$$
$$d(\underline{X}) \longrightarrow \arg\max_j \sum_{\| \underline{X}_i - \underline{X} \| \leq \varepsilon_{\underline{X},(K)}} \mathbb{1}_{\{Y_i = c_j\}} \,,$$ donde

2) Regresión: en este caso $Y_i \in \mathbb{R}$.

La función de regresión es

$$f(\underline{X}) = \frac{1}{K} \sum_{i / \| \underline{X}_i - \underline{X} \| \leq \varepsilon_{\underline{X},(K)}} Y_i$$

Si $\underline{X}_0$ es un valor particular, entonces $f(\underline{X}_0)$ es una aproximación de $E[Y \mid \underline{X} = \underline{X}_0]$

Tomado de Introduction To Statistical Learning 2nd ed. Statistical Decision Theory, pág. 18. Hastie et al

¿Qué es $E[Y \mid \underline{X} = \underline{x}_0]$?

Sea $(Y, \underline{X})$ una tupla de v.a. con fdp conjunta $f_{Y,\underline{X}}(y, \underline{x})$.

Dado $\underline{X} = \underline{x}_0$, queremos aproximar Y por $g(\underline{X})$, con $g: \mathbb{R}^p \to \mathbb{R}$.

El costo cuadrático de g es

$$E[(Y - g(\underline{X}))^2]$$
$$= \int_{\mathbb{R}^p} \int_{\mathbb{R}} (y - g(\underline{x}))^2 f_{Y,\underline{X}}(y, \underline{x})\, dy\, d\underline{x} \quad (1)$$

En la expresión azul, $\underline{x}$ está fija y y es la variable de integración. Cuando $\underline{x}$ está fija entonces

$$f_{Y,\underline{X}}(y, \underline{x}) = f_{Y|\underline{X}=x}(y)\, f_{\underline{X}}(\underline{x}).$$

Recordemos que la fdp condicional de Y dado $\underline{X} = x_0$ es

$$f_{\underline{Y}|\underline{X}=\underline{x}_0}(y) = \left.\frac{f_{Y,\underline{X}}(y, x)}{f_{\underline{X}}(\underline{x})}\right|_{\underline{x}=x_0}$$

$$= f_{Y,\underline{X}}(y, x_0) / f_{\underline{X}}(\underline{x}_0).$$

$f_{Y|\underline{X}=\underline{x}_0}(y)$ es una función solo de y. Así

$$E[(Y - g(X))^2] = \int_{\mathbb{R}^p} \int_{\mathbb{R}} (y - g(\underline{x}))^2 f_{Y|\underline{X}=x}(y)\, f_{\underline{X}}(\underline{x})\, dy\, d\underline{x}$$

$$= \int_{\mathbb{R}^p} \left[\int_{\mathbb{R}} (y - g(\underline{x}))^2 f_{Y|\underline{X}=x}(y)\, dy\right] f_{\underline{X}}(\underline{x})\, d\underline{x}$$

$$= \int_{\mathbb{R}^p} E_{Y|\underline{X}=\underline{x}}[(Y - g(\underline{x}))^2]\, f_{\underline{X}}(\underline{x})\, d\underline{x}$$

$$= E_{\underline{X}}\left[E_{Y|\underline{X}}\left[(Y - g(\underline{X}))^2 \right]\right] \quad (2)$$

La fórmula (2) indica que para minimizar el costo cuadrático de $g(\,)$ se puede hacer la minimización condicionando sobre X y luego tomando la esperanza con respecto a $\underline{X}$:

$$E_{Y|\underline{X}=x_0}\left[(Y - g(\underline{x}_0))^2 \right] = E_{Y|\underline{X}=x_0}\left[Y^2 - 2g(\underline{x}_0)Y + g^2(\underline{x}_0) \right]$$

$$= E_{Y|\underline{X}=x_0}\left[Y^2 - 2cY + c^2 \right], \quad \text{con} \quad c = g(\underline{x}_0).$$

$$= \quad - 2c E_{Y|X=x_0}[Y] + c^2. \quad (3)$$

Para minimizar esta expresión con respecto a c, derivamos con respecto a c e igualamos la derivada a 0:

$$\frac{d}{dc} E_{Y|\underline{X}=x_0}\left[(Y - g(\underline{x}_0))^2 \right] = -2E_{Y|X=x_0}[Y] + 2c \quad (4)$$

$$\frac{d}{dc} E_{Y|\underline{X}=x_0}\left[(Y - g(\underline{x}_0))^2 \right] = 0 \iff c = E_{Y|X=x_0}[Y].$$

Así $g(\underline{x}_0) = E_{Y|\underline{X}=\underline{x}_0}[Y] \quad (5).$

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

$$E_{\underline{X}}\left[E_{Y|\underline{X}}\left[(Y - E_{Y|X}[Y^2])^2 \right]\right] =$$

$$= E_{\underline{X}}\left[V[Y|\underline{X}] \right]$$

Por otro lado el MSE ó pérdida cuadrática del estimador Θ del parámetro θ es

$$MSE(\Theta) = E[(\Theta - \theta)^2]$$

$$= E[((\Theta - E[\Theta]) + (E[\Theta] - \theta))^2]$$

$$= E[(\Theta - E[\Theta])^2] + 2E[(\Theta - E[\Theta])(E[\Theta] - \theta)] + E[(E[\Theta] - \theta)^2]$$

$$= V[\Theta] + B^2[\Theta], \quad B[\Theta] = E[\Theta] - \theta$$

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

$$MSE(g) = V[g(\underline{X})] + B^2[g(\underline{X})],$$

donde $B[g(\underline{X})] = E[g(\underline{X})] - Y$