k-Vecinos más cercanos

Carlos Malanche

12 de abril de 2018

El méodo de *machine learning* de hoy es uno de los más intuitivos, sólo que lo trataré de presentar de manera formal.

1. Clasificación múltiple

Ahora nos enfrentamos a un problema más duro. Tenemos una serie de pares de datos $S = \{y_i, \underset{\rightarrow}{x}_i\}_{i=1}^N$, y sabemos que $y_i \in C = \{\text{Clase1}, \text{Clase2}, \cdots, \text{ClaseH}\}$, es decir, hay H clases a las que cada y_i puede pertenecer.

2. k-Vecinos más cercanos (*k-Nearest Neighbours, k-NN*)

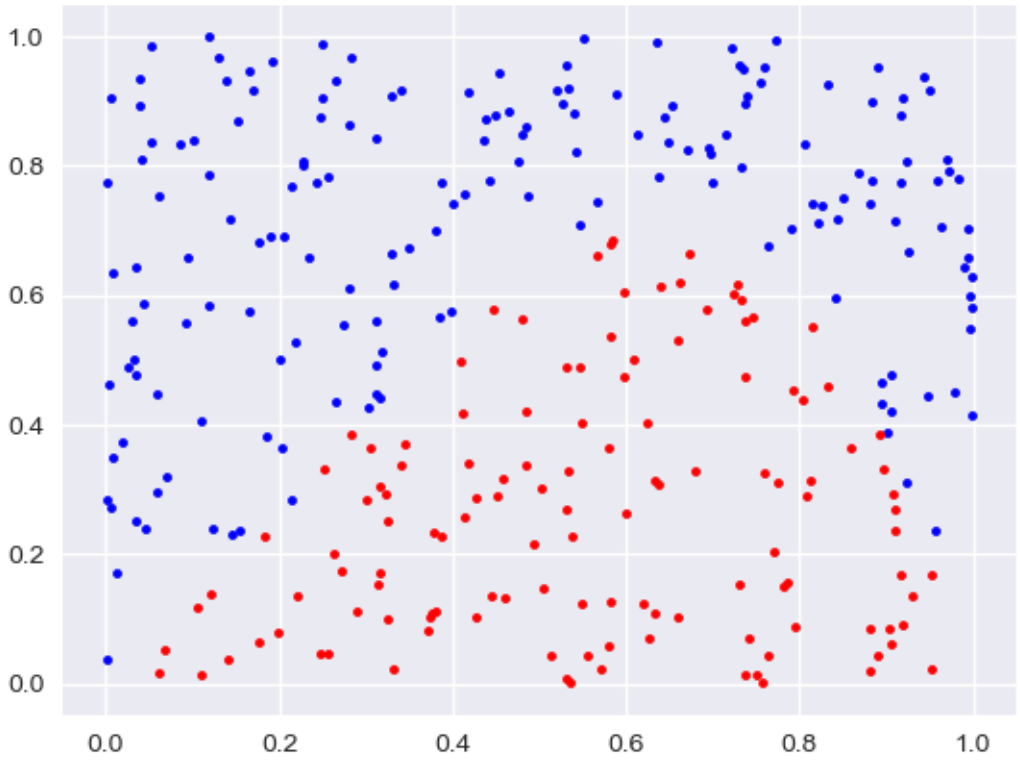
El único requisito para que este método funcione es que las variables $\underset{\rightarrow}{x}_i$ vivan en algún espacio U al que se le pueda asignar una métrica, de tal modo que la cantidad $D(\underset{\rightarrow}{x}_i, \underset{\rightarrow}{x}_j)$, la distancia entre dos elementos del espacio, esté bien definida para cualesquiera dos elementos $\underset{\rightarrow}{x}_i, \underset{\rightarrow}{x}_j \in U$. Definimos por $\dot{S}_{\underset{\rightarrow}{x}}$ la permutación de la serie S en donde se cumple que

$$D((S_{\underset{\rightarrow}{x}})_i, \underset{\rightarrow}{x}) \leq D((S_{\underset{\rightarrow}{x}})_{i+1}, \underset{\rightarrow}{x})), \text{ para } i = 1, \cdots, N-1$$

En donde $(\dot{S}_{\underset{\rightarrow}{x}})_i$ es el elemento número i de la lista. En el caso de que todas las componentes del vector $\underset{\rightarrow}{x}$ estén en los reales, $D(\underset{\rightarrow}{x}_i, \underset{\rightarrow}{x}_j) \coloneqq \|\underset{\rightarrow}{x}_i - \underset{\rightarrow}{x}_j\|$. Por último, denotamos por $\dot{S}^k_{\underset{\rightarrow}{x}}$ al conjunto que contiene los primeros k elementos de $\dot{S}_{\underset{\rightarrow}{x}}$. Un estimador de k-NN queda escrito entonces como:

$$\hat{f}(\underset{\rightarrow}{x}) = \frac{1}{k} \sum_{i:x_i \in S^k_{\underset{\rightarrow}{x}}} y_i \tag{1}$$

Es decir, nuestra estimación es el promedio de los k elementos más cercanos a $\underset{\rightarrow}{x}$, según la métrica establecida. Cuando se trata de clasificaciones, basta con tomar el elemento de voto mayoritario.


1.0
0.8
0.6
0.4
0.2
0.0
0.0
0.2
0.4
0.6
0.8
1.0