Clasificadores K-NN

10.1 Introducción

En este tema vamos a estudiar un paradigma clasificatorio conocido como K-NN (K-*Nearest Neighbour*). La idea básica sobre la que se fundamenta este paradigma es que un nuevo caso se va a clasificar en la clase más frecuente a la que pertenecen sus K vecinos más cercanos. El paradigma se fundamenta por tanto en una idea muy simple e intuitiva, lo que unido a su fácil implementación hace que sea un paradigma clasificatorio muy extendido.

Después de introducir el algoritmo K-NN básico y presentar algunas variantes del mismo, en este tema se estudian métodos para la selección de prototipos.

10.2 El algoritmo K-NN básico

La notación a utilizar (véase la Figura 1) en este tema es la siguiente:

		X_1	$\ldots$	X_j	$\ldots$	X_n	C
$(\mathbf{x}_1, c_1)$	1	x_{11}	$\ldots$	x_{1j}	$\ldots$	x_{1n}	c_1
	$\vdots$	$\vdots$		$\vdots$		$\vdots$	$\vdots$
$(\mathbf{x}_i, c_i)$	i	x_{i1}	$\ldots$	x_{ij}	$\ldots$	x_{in}	c_i
	$\vdots$	$\vdots$		$\vdots$		$\vdots$	$\vdots$
$(\mathbf{x}_N, c_N)$	N	x_{N1}	$\ldots$	x_{Nj}	$\ldots$	x_{Nn}	c_N
$\mathbf{x}$	$N+1$	$x_{N+1,1}$	$\ldots$	$x_{N+1,j}$	$\ldots$	$X_{N+1,n}$	?

Figura 1: Notación para el paradigma K-NN

- D indica un fichero de N casos, cada uno de los cuales está caracterizado por n variables predictoras, $X_1, \ldots, X_n$ y una variable a predecir, la clase C.
- Los N casos se denotan por

$$\begin{aligned} (\mathbf{x}_1, c_1), \ldots, (\mathbf{x}_N, c_N) \quad & \text{donde} \\ \mathbf{x}_i = (x_{i,1} \ldots x_{i,n}) \quad & \text{para todo } i = 1, \ldots, N \\ c_i \in \{c^1, \ldots, c^m\} \quad & \text{para todo } i = 1, \ldots, N \end{aligned}$$

  $c^1, \ldots, c^m$ denotan los m posibles valores de la variable clase C.
- El nuevo caso que se pretende clasificar se denota por $\mathbf{x} = (x_1, \ldots, x_n)$.

En la Figura 2 se presenta un pseudocódigo para el clasificador K-NN básico. Tal y como puede observarse en el mismo, se calculan las distancias de todos los casos ya clasificados al nuevo caso, $\mathbf{x}$, que se pretende clasificar. Una vez seleccionados los K casos ya clasificados, $D_{\mathbf{x}}^k$ más cercanos al nuevo caso, $\mathbf{x}$, a éste se le asignará la

COMIENZO
  Entrada: $D = \{(\mathbf{x}_1, c_1), \ldots, (\mathbf{x}_N, c_N)\}$
        $\mathbf{x} = (x_1, \ldots, x_n)$ nuevo caso a clasificar
  PARA todo objeto ya clasificado (x_i, c_i)
        calcular $d_i = d(\mathbf{x}_i, \mathbf{x})$
  Ordenar $d_i (i = 1, \ldots, N)$ en orden ascendente
  Quedarnos con los K casos $D_{\mathbf{x}}^K$ ya clasificados más cercanos a $\mathbf{x}$
  Asignar a $\mathbf{x}$ la clase más frecuente en $D_{\mathbf{x}}^K$
FIN

Figura 2: Pseudocódigo para el clasificador K-NN

clase (valor de la variable C) más frecuente de entre los K objetos, $D_{\mathbf{x}}^k$. La Figura 3 muestra de manera gráfica un ejemplo de lo anterior. Tal y como puede verse en

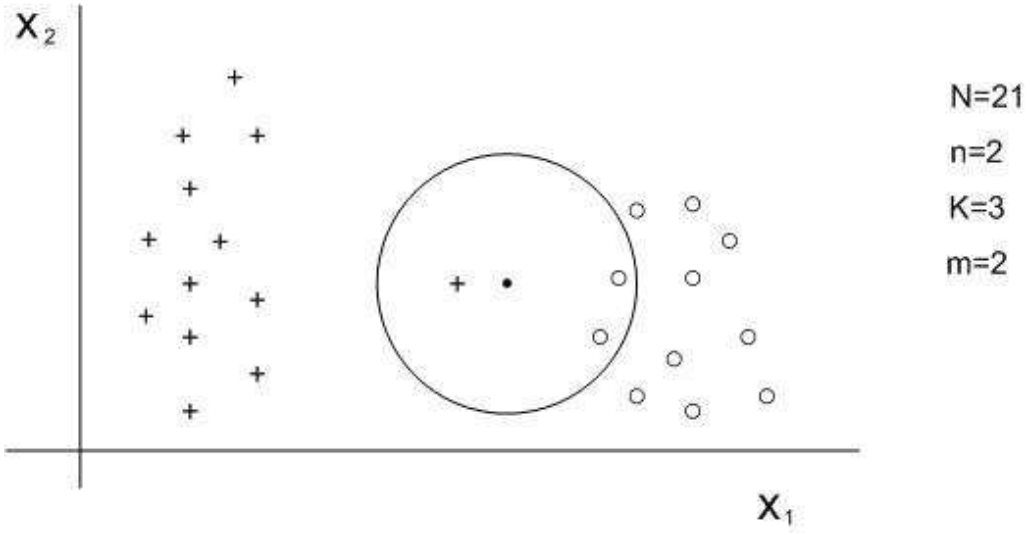

Figura 3: Ejemplo de aplicación del algoritmo K-NN básico

la Figura 3 tenemos 24 casos ya clasificados en dos posibles valores ($m = 2$). Las variables predictoras son X_1 y X_2, y se ha seleccionado $K = 3$. De los 3 casos ya clasificados que se encuentran más cercanos al nuevo caso a clasificar, $\mathbf{x}$ (representado por $\bullet$), dos de ellos pertenecen a la clase $\circ$, por tanto el clasificador 3-NN predice la clase $\circ$ para el nuevo caso. Nótese que el caso más cercano a $\mathbf{x}$ pertenece a la clase $+$. Es decir, que si hubiésemos utilizado un clasificador 1-NN, $\mathbf{x}$ se hubiese asignado a $+$.

Conviene aclarar que el paradigma K-NN es un tanto atípico si lo comparamos con el resto de paradigmas clasificatorios, ya que mientras que en el resto de paradigmas la clasificación de un nuevo caso se lleva a cabo a partir de dos tareas, como son la *inducción* del modelo clasificatorio y la posterior *deducción* (o aplicación) sobre el nuevo caso, en el paradigma K-NN al no existir modelo explícito, las dos tareas anteriores se encuentran colapsadas en lo que se acostumbra a denominar *transinducción*.

En caso de que se produzca un *empate* entre dos o más clases, conviene tener una *regla*

heurística para su ruptura. Ejemplos de reglas heurísticas para la ruptura de empates pueden ser: seleccionar la clase que contenta al vecino más próximo, seleccionar la clase con distancia media menor, etc.

Otra cuestión importante es la determinación del valor de K. Se constata empíricamente que el porcentaje de casos bien clasificados es no monótono con respecto de K (véase Figura 4), siendo una buena elección valores de K comprendidos entre 3 y 7.

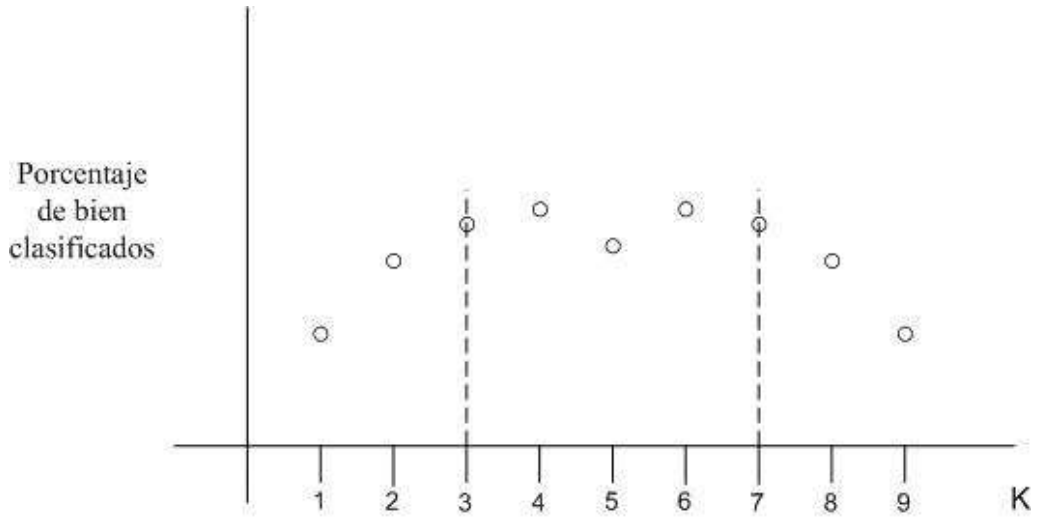

Figura 4: Ejemplo de la no monotocidad del porcentaje de bien clasificados en función de K

10.3 Variantes sobre el algoritmo básico

En este apartado vamos a introducir algunas variantes sobre el algoritmo básico.

10.3.1 K-NN con rechazo

La idea subyacente al K-NN con rechazo es que para poder clasificar un caso debo de tener ciertas garantías. Es por ello por lo que puede ocurrir que un caso quede sin clasificar, si no existen ciertas garantías de que la clase a asignar sea la correcta.

Dos ejemplos utilizados para llevar a cabo clasificaciones con garantías son los siguientes:

- el número de votos obtenidos por la clase deberá superar un *umbral prefijado*. Si suponemos que trabajamos con $K = 10$, y $m = 2$, dicho umbral puede establecerse en 6.
- establecimiento de algún tipo de *mayoría absoluta* para la clase a asignar. Así, si suponemos que $K = 20$, $m = 4$, podemos convenir en que la asignación del nuevo caso a una clase sólo se llevará a cabo en el caso de que la diferencia entre las frecuencias mayor y segunda mayor supere 3.

10.3.2 K-NN con distancia media

En el K-NN con distancia media la idea es asignar un nuevo caso a la clase cuya distancia media sea menor. Así que en el ejemplo de la Figura 5, a pesar de que 5 de los 7 casos más cercanos al mismo pertenecen a la clase $\circ$, el nuevo caso se clasifica como $+$, ya que la distancia media a los dos casos $+$ es menor que la distancia media a los cinco casos $\circ$.