Tema 5: Clasificadores K-NN

Abdelmalik Moujahid, Iñaki Inza y Pedro Larrañaga

Departamento de Ciencias de la Computación e Inteligencia Artificial

Universidad del País Vasco

http://www.sc.ehu.es/isg/

Introducción

- K-NN (K-*Nearest Neighbour*)
- Un nuevo caso se va a clasificar en la clase más frecuente a la que pertenecen sus K vecinos más cercanos
- Idea muy simple e intuitiva
- Fácil implementación
- No hay modelo explícito
- *Case Based Reasoning (CBR)*

Introducción

		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
$\mathbf{x}$	$N+1$	$x_{N+1,1}$	$\ldots$	$x_{N+1,j}$	$\ldots$	$x_{N+1,n}$	?

Notación para el paradigma K-NN

El algoritmo K-NN básico

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

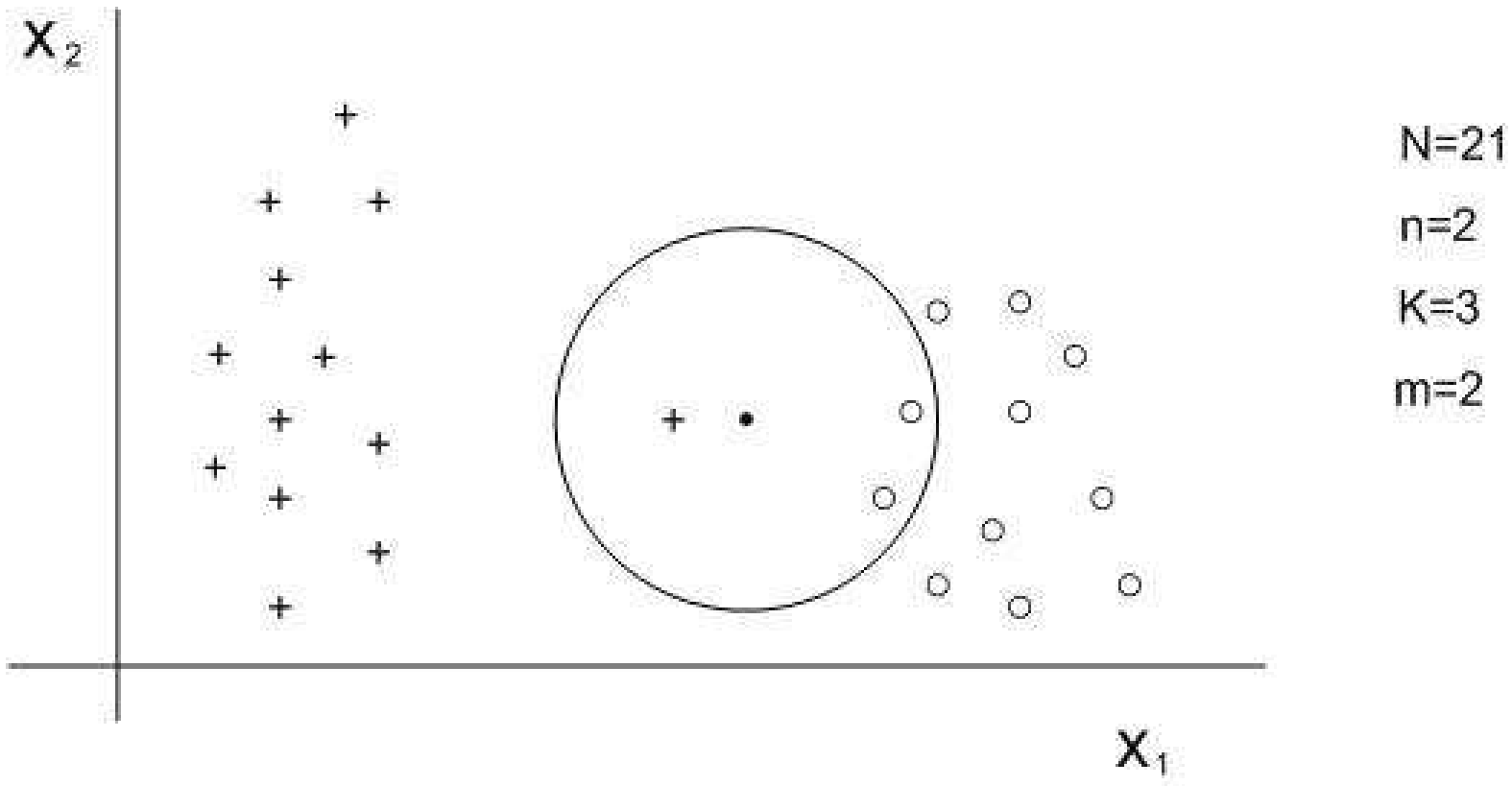
Quedarnos con los K casos $D_{\mathbf{x}}^K$ ya clasificados más cercanos a $\mathbf{x}$

Asignar a $\mathbf{x}$ la clase más frecuente en $D_{\mathbf{x}}^K$

FIN

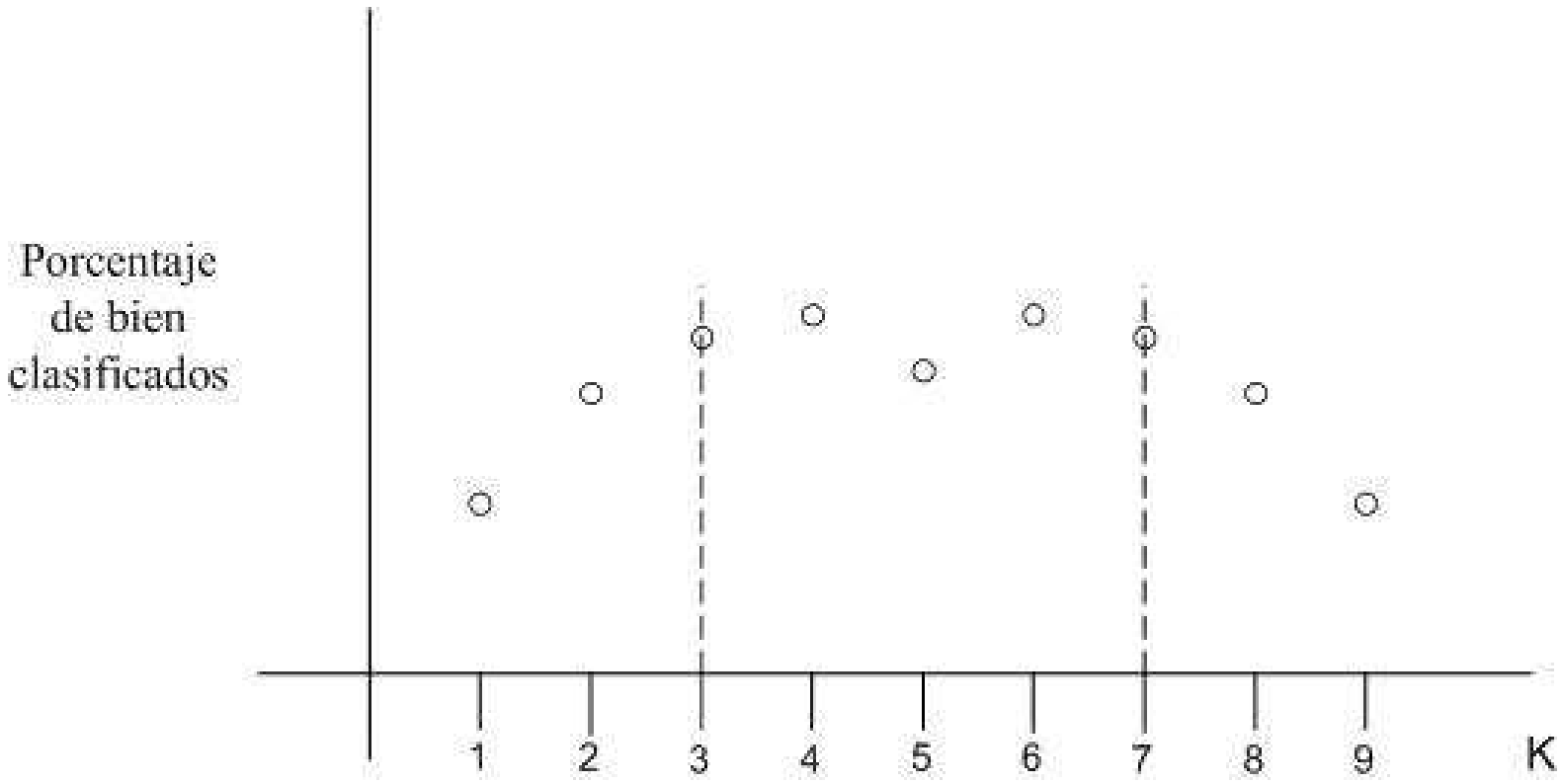
Pseudocódigo para el clasificador K-NN

El algoritmo K-NN básico



Ejemplo de aplicación del algoritmo K-NN básico

El algoritmo K-NN básico



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

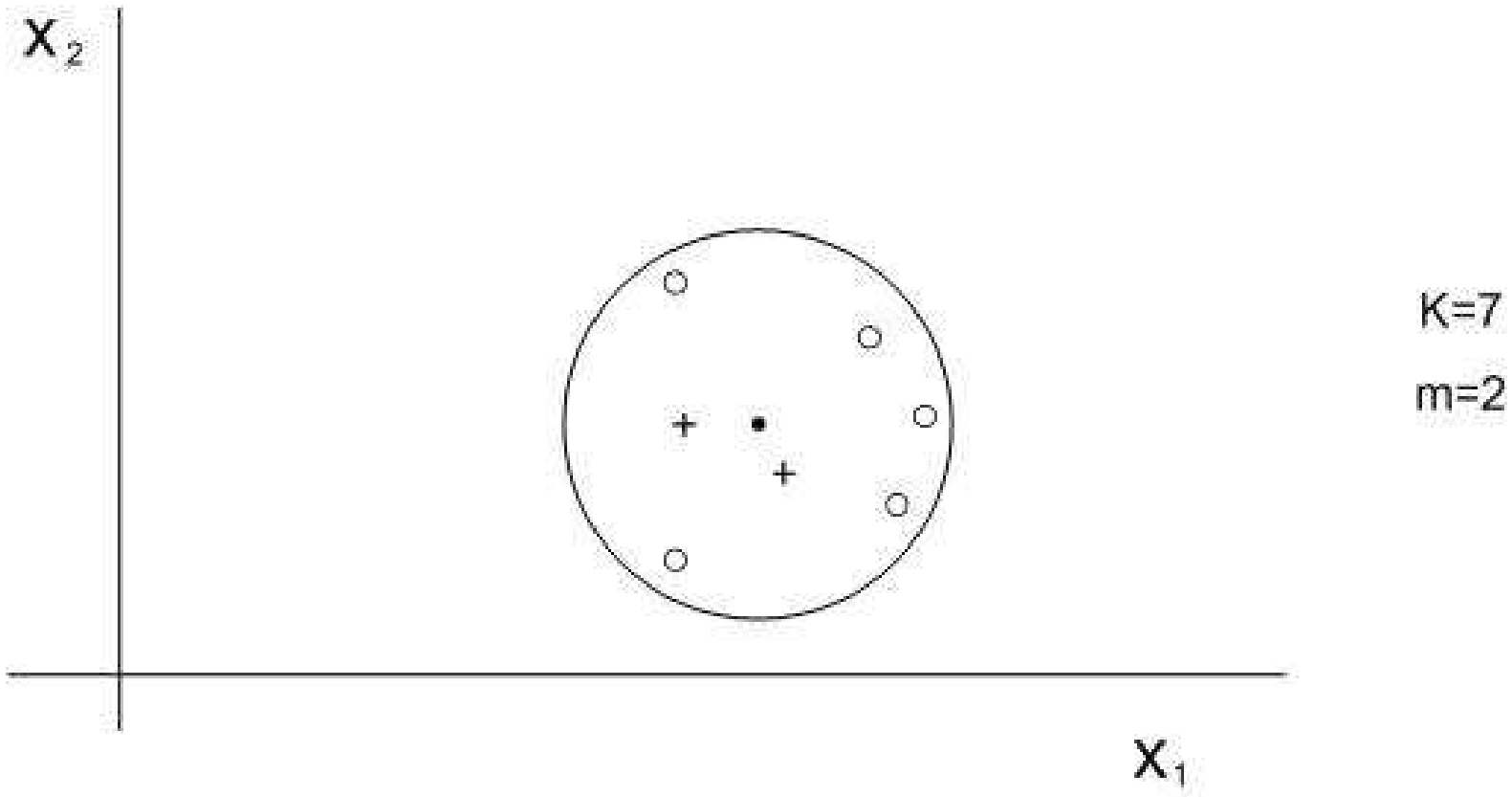
Variantes del algoritmo K-NN básico

- K-NN con rechazo
- K-NN con distancia media
- K-NN con distancia mínima
- K-NN con pesado de vecinos
- K-NN con pesado de variables

K-NN con rechazo

- Para clasificar un caso exigo ciertas garantías
- Si no las tengo puedo dejar el caso sin clasificar
- Umbral prefijado
- Mayoría absoluta

K-NN con distancia media

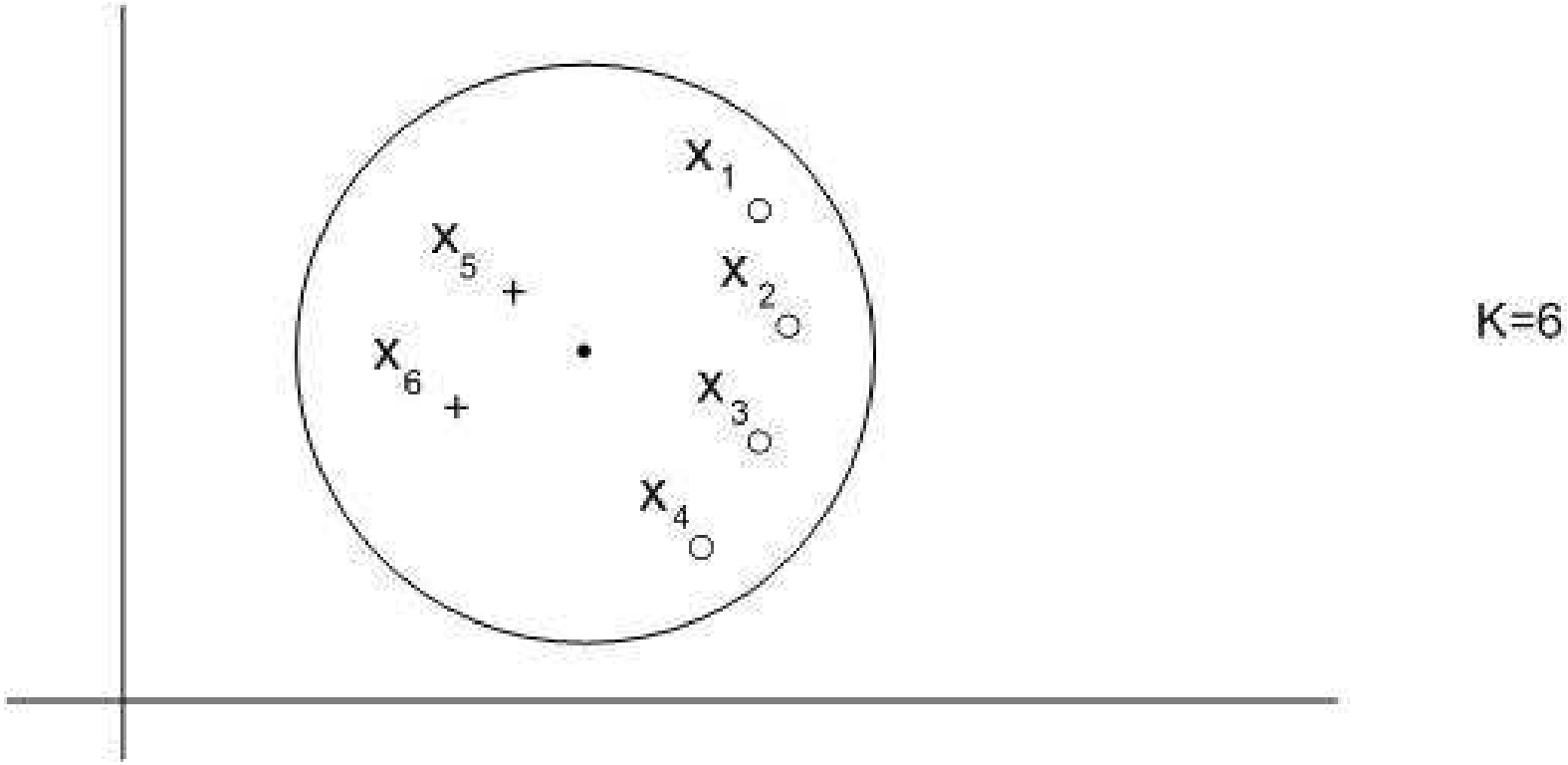


Ejemplo de ilustración del K-NN con distancia media

K-NN con distancia mínima

- Seleccionar un caso por clase (ej. el más cercano al baricentro de la clase)
- Reducción de la dimensión del fichero almacenado de N a m
- Ejecutar un 1-NN a dicho fichero reducido
- Efectividad condicionada a la homogeneidad dentro de las clases. A mayor homogeneidad más efectivo

K-NN con pesado de vecinos



Ejemplo de ilustración del K-NN con pesado de casos seleccionados

K-NN con pesado de vecinos

	$d(\mathbf{x}_i, \mathbf{x})$	w_i
$\mathbf{x}_1$	2	$0{,}5$
$\mathbf{x}_2$	2	$0{,}5$
$\mathbf{x}_3$	2	$0{,}5$
$\mathbf{x}_4$	2	$0{,}5$
$\mathbf{x}_5$	$0{,}7$	$1/0{,}7$
$\mathbf{x}_6$	$0{,}8$	$1/0{,}8$

Peso a asignar a cada uno de los 6 objetos seleccionados

K-NN con pesado de variables

- Mismo peso a todas las variables:

$$d(\mathbf{x}, \mathbf{x}_r) = \sum_{j=1}^{n} (x_j - x_{rj})^2$$

- Distinto peso a cada variable:

$$d(\mathbf{x}, \mathbf{x}_r) = \sum_{j=1}^{n} w_j (x_j - x_{rj})^2$$

- Determinar w_j a partir de $I(X_j, C)$ la cantidad de infomación mutua entre X_j y C

K-NN con pesado de variables

X_1	X_2	C
0	0	1
0	0	1
0	0	1
1	0	1
1	0	1
1	1	1
0	1	0
0	1	0
0	1	0
1	1	0
1	1	0
1	0	0

La variable X_1 no es relevante para C, mientras que la variables X_2 si lo es

K-NN con pesado de variables

$$I(X_1, C) = p_{(X_1,C)}(0,0)\log\frac{p_{(X_1,C)}(0,0)}{p_{X_1}(0)\cdot p_C(0)} + p_{(X_1,C)}(0,1)\log\frac{p_{(X_1,C)}(0,1)}{p_{X_1}(0)\cdot p_C(1)} +$$

$$p_{(X_1,C)}(1,0)\log\frac{p_{(X_1,C)}(1,0)}{p_{X_1}(1)\cdot p_C(0)} + p_{(X_1,C)}(1,1)\log\frac{p_{(X_1,C)}(1,1)}{p_{X_1}(1)\cdot p_C(1)} =$$

$$\frac{3}{12}log\frac{\frac{3}{12}}{\frac{6}{12}\cdot\frac{6}{12}} + \frac{3}{12}log\frac{\frac{3}{12}}{\frac{6}{12}\cdot\frac{6}{12}} + \frac{3}{12}log\frac{\frac{3}{12}}{\frac{6}{12}\cdot\frac{6}{12}} + \frac{3}{12}log\frac{\frac{3}{12}}{\frac{6}{12}\cdot\frac{6}{12}} = 0$$

$$I(X_2, C) = p_{(X_2,C)}(0,0)\log\frac{p_{(X_2,C)}(0,0)}{p_{X_2}(0)\cdot p_C(0)} + p_{(X_2,C)}(0,1)\log\frac{p_{(X_2,C)}(0,1)}{p_{X_2}(0)\cdot p_C(1)} +$$

$$p_{(X_2,C)}(1,0)\log\frac{p_{(X_2,C)}(1,0)}{p_{X_2}(1)\cdot p_C(0)} + p_{(X_2,C)}(1,1)\log\frac{p_{(X_2,C)}(1,1)}{p_{X_2}(1)\cdot p_C(1)} =$$

$$\frac{1}{12}log\frac{\frac{1}{12}}{\frac{6}{12}\cdot\frac{6}{12}} + \frac{5}{12}log\frac{\frac{5}{12}}{\frac{6}{12}\cdot\frac{6}{12}} + \frac{5}{12}log\frac{\frac{5}{12}}{\frac{6}{12}\cdot\frac{6}{12}} + \frac{1}{12}log\frac{\frac{1}{12}}{\frac{6}{12}\cdot\frac{6}{12}}$$

Selección de prototipos

- Edición de Wilson
- Condensación de Hart

Edición de Wilson

- Someter a prueba a cada uno de los elementos del fichero de casos inicial
- Para cada caso se compara su clase verdadera con la que propone un clasificador K-NN obtenido con todos los casos excepto el mismo
- Si ambas clases no coincidan, el caso es eliminado
- Edición de Wilson repetitiva parando el procedimiento cuando en 2 selecciones sucesivas no se produzcan cambios

Condensación de Hart

- Para cada caso, y siguiendo el orden en el que se encuentran almacenados los casos en el fichero, se construye un clasificador K-NN con tan sólo los casos anteriores al caso en cuestión
- Si el caso tiene un valor de la clase distinto al que le asignaría el clasificador K-NN, el caso es seleccionado
- Si por el contrario la clase verdadera del caso coincide con la propuesta por el clasificador K-NN, el caso no se selecciona
- El método es dependiente del orden en que se encuentren almacenados los casos en el fichero