Aprendizaje Basado en Instancias

García Prado, Sergio
sergio@garciparedes.me

9 de abril de 2017

Resumen

[TODO]

1. Introducción

[TODO]

1.1. K-vecinos más cercanos

[TODO]

2. La figura 1 muestra un conjunto de entrenamiento con ejemplos positivos (estrellas) y negativos (círculos). Se desea clasificar la nueva instancia $< 3, 3 >$ mediante el algoritmo K-vecinos más próximos. Obtener la clasificación para los valores de $K = \{1, 3, 5\}$ utilizando las distancias indicadas a continuación

En este ejercicio se realiza una clasificación mediante el algoritmos de clasificación basado en instancias *K-NN*. El conjunto de datos está compuesto por 7 instancias caracterizadas por 2 atributos númericos de caracter entero y la clase de destino de tipo binario (*ESTRELLAS* o *CIRCULOS*). Puesto que el espacio del conjunto de datos está formado por dos dimensiones este se puede representar de forma gráfica tal y como se muestra en la figura 1.

En las ecuaciones (1) y (2) se representan las instancias de cada una de las dos clases en forma de conjuntos de coordenadas. En la ecuación (3) se muestran las coordenadas de la instancia que se desea clasificar.

$$ESTRELLAS = \{(1,1), (1,4), (3,1), (5,3)\} \tag{1}$$
$$CIRCULOS = \{(2,1), (5,2), (6,1)\} \tag{2}$$
$$instancia = (3,3) \tag{3}$$

Puesto que se pretende clasificar una nueva instancia apoyandose en la intuicción del algoritmo de los k-vecinos más cercanos, por tanto, la medida de bondad en que se basará será la distancia. Para ello se utilizará la distancia *Euclídea*(2.1), *Euclídea Ponderada*(2.2), *Manhattan*(2.3) y *Hamming*(2.4).

Algunas de estas medidas de distancia requieren de la necesidad de normalización de los valores para poder ser calculadas de manera apropiada, por tanto el siguiente paso es normalizar los mismos respecto de cada una de las dimensiones de los datos. Para ello se ha utilizado la estrategia de normalización definida por $normalize\colon \mathbb{R}^2 \rightarrow [0,1]^2$ que se describe en la ecuación (4). Nótese que para ello es necesario obtener el máximo y mínimo para cada uno de los atributos, dichos resultados se muestran en las ecuaciones (5) y (6).

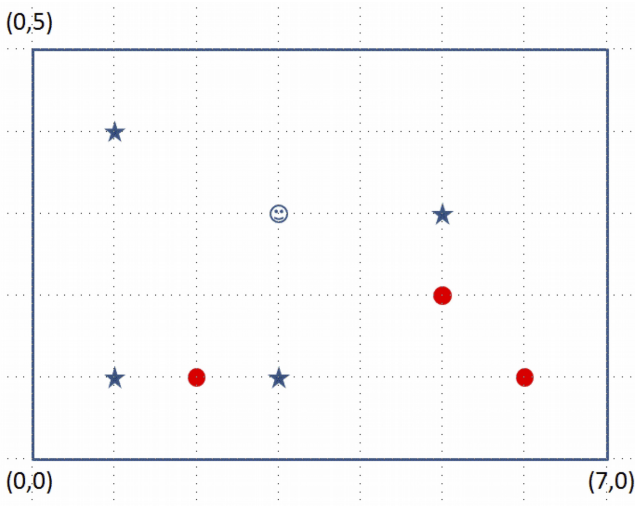

Figura 1: *Representación Gráfica del problema 2*

$$normalize(x, y) = (\frac{x - min_x}{max_x - min_x}, \frac{y - min_y}{max_y - min_y}) \tag{4}$$

$$max_x = 6 \qquad min_x = 1 \tag{5}$$
$$max_y = 4 \qquad min_y = 1 \tag{6}$$

Los valores normalizados se muestran en las ecuaciones (7), (8) y (9). Por tanto, una vez hecho esto, ya se está en condiciones de calcular las distancia de la *instancia* a clasificar con respecto del resto de instancias y aplicar la intuición de cercanía en que se apoya el clasificador *K-NN*.

$$ESTRELLAS_{normalized} = \{(0,0), (0,1), (\tfrac{2}{5}, 0), (\tfrac{4}{5}, \tfrac{2}{3})\} \tag{7}$$
$$CIRCULOS_{normalized} = \{(\tfrac{1}{5}, 0), (\tfrac{4}{5}, \tfrac{1}{3}), (1,0)\} \tag{8}$$
$$instancia_{normalized} = (\tfrac{2}{5}, \tfrac{2}{3}) \tag{9}$$

2.1. Distancia Euclídea

La *Distancia Euclídea* se define tal y como se muestra en la ecuación (10). En este caso es necesario que los valores de entrada hayan sido normalizados previamente. Los resultados de distancia de la *instancia* a clasificar con respecto al resto de instancias se muestran en la ecuación (11).

$$D_{euclidean}(a, b) = \sqrt{(a_x - b_x)^2 + (a_y - b_y)^2} \tag{10}$$

$$R_{euclidean} = \{0.777_e, 0.52_e, 0.666_e, 0.4_e, 0.696_c, 0.52_c, 0.896_c\} \tag{11}$$

El último paso es utilizar la función *min* encargada de obtener los k valores más pequeños de un determinado conjunto. Esto encaja perfectamente con lo que se pretende que haga el clasificador *K-NN*. Por tanto, los valores más cercanos así como la clase en la que se clasifica *instancia* para $k \in \{1, 3, 5\}$ se muestran en las ecuaciones (12), (13) y (14) respectivamente.

$$min(R_{euclidean}, 1) = \{0.4_e\} \implies instancia \in ESTRELLAS \tag{12}$$
$$min(R_{euclidean}, 3) = \{0.4_e, 0.52_e, 0.52_c\} \implies instancia \in ESTRELLAS \tag{13}$$
$$min(R_{euclidean}, 5) = \{0.4_e, 0.52_e, 0.52_c, 0.666_e, 0.696_e\} \implies instancia \in ESTRELLAS \tag{14}$$

2.2. Distancia Euclídea Ponderada: $w_x = 0.2, w_y = 0.8$

La *Distancia Euclídea Ponderada* se define tal y como se muestra en la ecuación (15). En este caso también es necesario que los valores de entrada hayan sido normalizados previamente. Además, la importancia que tienen cada una de las dimensiones no es igual, sino que está ponderada por los valores fijados previamente en el vector w. Los resultados de distancia de la *instancia* a clasificar con respecto al resto de instancias se muestran en la ecuación (16).

$$D_{w_{euclidean}}(a, b) = \sqrt{w_x(a_x - b_x)^2 + w_y(a_y - b_y)^2} = \sqrt{0.2(a_x - b_x)^2 + 0.8(a_y - b_y)^2} \tag{15}$$

$$R_{w_{euclidean}} = \{0.622_e, 0.347_e, 0.596_e, 0.178_e, 0.602_c, 0.347_c, 0.653_c\} \tag{16}$$

El último paso es utilizar la función *min* encargada de obtener los k valores más pequeños de un determinado conjunto. Por tanto, los valores más cercanos así como la clase en la que se clasifica *instancia* para $k \in \{1, 3, 5\}$ se muestran en las ecuaciones (17), (18) y (19) respectivamente.

$$min(R_{w_{euclidean}}, 1) = \{0.178_e\} \implies instancia \in ESTRELLAS \tag{17}$$
$$min(R_{w_{euclidean}}, 3) = \{0.178_e, 0.347_e, 0.347_c, \} \implies instancia \in ESTRELLAS \tag{18}$$
$$min(R_{w_{euclidean}}, 5) = \{0.178_e, 0.347_e, 0.347_c, 0.596_e, 0.602_c\} \implies instancia \in ESTRELLAS \tag{19}$$

2.3. Distancia de Manhattan

La *Distancia de Manhattan* se define tal y como se muestra en la ecuación (20). Es una medida de distancia para valores enteros, por tanto, en este caso no es apropiado utilizar las instancias normalizadas. Los resultados de distancia de la *instancia* a clasificar con respecto al resto de instancias se muestran en la ecuación (21).

$$D_{manhattan}(a, b) = |a_x - b_x| + |a_y - b_y| \tag{20}$$

$$R_{manhattan} = \{4_e, 3_e, 2_e, 2_e, 3_c, 3_c, 5_c\} \tag{21}$$

El último paso es utilizar la función *min* encargada de obtener los k valores más pequeños de un determinado conjunto. Por tanto, los valores más cercanos así como la clase en la que se clasifica *instancia* para $k \in \{1, 3, 5\}$ se muestran en las ecuaciones (22), (23) y (24) respectivamente.

$$min(R_{manhattan}, 1) = \{2_e\} \implies instancia \in ESTRELLAS \tag{22}$$
$$min(R_{manhattan}, 3) = \{2_e, 2_e, 3_e\} \implies instancia \in ESTRELLAS \tag{23}$$
$$min(R_{manhattan}, 5) = \{2_e, 2_e, 3_e, 3_c, 3_c, \} \implies instancia \in ESTRELLAS \tag{24}$$

2.4. Distancia de Hamming

La *Distancia de Hamming* se define tal y como se muestra en la ecuación (25). Es una medida de distancia para valores discretos, a pesar de ello, en este caso la utilizaremos con valores numéricos enteros presuponiendo que cada valor es una categoría diferente. Los resultados de distancia de la *instancia* a clasificar con respecto al resto de instancias se muestran en la ecuación (26).

$$D_{hamming}(a,b) = (a_x \neq b_x) + (a_y \neq b_y) \tag{25}$$

$$R_{hamming} = \{2_e, 2_e, 1_e, 1_e, 2_c, 2_c, 2_c\} \tag{26}$$

El último paso es utilizar la función *min* encargada de obtener los k valores más pequeños de un determinado conjunto. Por tanto, los valores más cercanos así como la clase en la que se clasifica *instancia* para $k \in \{1, 3, 5\}$ se muestran en las ecuaciones (27), (28) y (29) respectivamente.

$$min(R_{hamming}, 1) = \{1_e\} \implies instancia \in ESTRELLAS \tag{27}$$
$$min(R_{hamming}, 3) = \{1_e, 1_e, 2_e\} \implies instancia \in ESTRELLAS \tag{28}$$
$$min(R_{hamming}, 5) = \{1_e, 1_e, 2_e, 3_c, 3_c, \} \implies instancia \in ESTRELLAS \tag{29}$$

3. Dígitos manuscritos

[TODO]

[TODO]

Validación cruzada de 10 particiones — K-Vecinos más Próximos									
Datos	Tasa de Error ($K =$)								
	1	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*
Entrenae	3.448 %	3.524 %	3.065 %	3.141 %	3.371 %	3.218 %	3.295 %	3.295 %	3.371 %

Tabla 1: *Tasa de error obtenida tras realizar un experimento de Validación cruzada de 10 particiones con el clasificador* K-NN *para* $k \in \{1, 2, ..., 9\}$

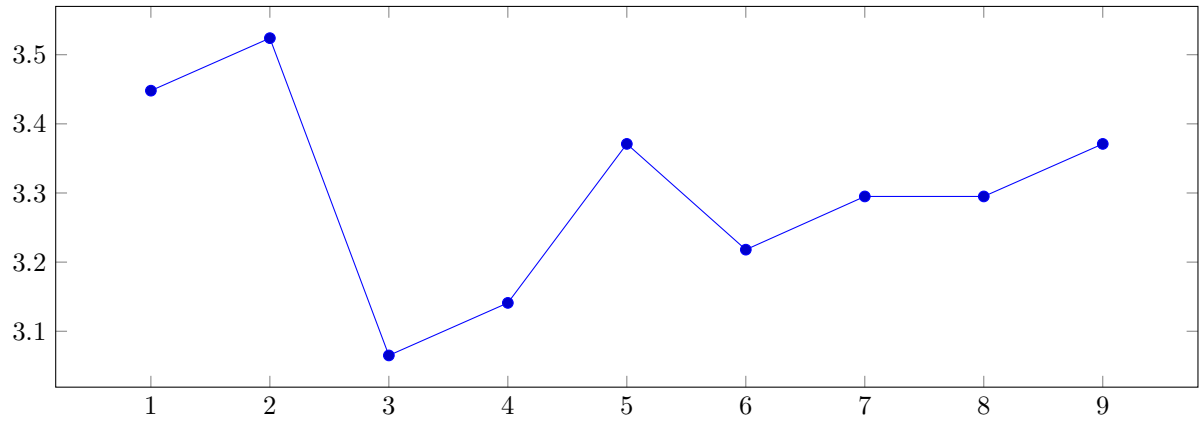

Figura 2: *Representación Gráfica de la tasa de error obtenida tras realizar un experimento de Validación cruzada de 10 particiones con el clasificador* K-NN *para* $k \in \{1, 2, ..., 9\}$

Referencias

[CCAG17] Teodoro Calonge Cano and Carlos Javier Alonso GonzáLez. Técnicas de Aprendizaje Autómatico, 2016/17.

[GP17] Sergio García Prado. Aprendizaje basado en instancias. `https://github.com/garciparedes/machine-learning-instance-based`, 2017.

[too] Weka. `http://www.cs.waikato.ac.nz/ml/weka/`.