

Aprendizaje Basado en Instancias

García Prado, Sergio
sergio@garciparedes.me

9 de abril de 2017

Resumen

En este documento se describe la estrategia de clasificación supervisada basada en instancias del algoritmo K-NN. También se describen las medidas de distancia utilizadas comunmente por los algoritmos de aprendizaje que se apoyan en dicha estrategia. Por último, se realizan dos casos de prueba que demuestran, el funcionamiento de dicho clasificador en la sección 2, y las bondades a nivel de resultados así como la necesidad de ajuste de parámetros en la sección 3.

1. K -VECINOS MÁS CERCANOS

El algoritmo de clasificación denominado *K-vecinos más cercanos* se apoya en la intuición de cercanía de la instancia a clasificar respecto de las instancias aprendidas previamente en la fase de entrenamiento. Por tanto, se enmarca dentro de la categoría de aprendizaje basado en instancias. Esta estrategia tienen un coste muy reducido en la fase de aprendizaje (ya que se limita al simple almacenamiento de las instancias de entrenamiento) pero por contra el coste de en la fase de clasificación es elevado.

La clasificación se basa en el cálculo de la distancia de la nueva instancia a clasificar respecto de cada una de las instancias del conjunto aprendido previamente, lo cual presenta un coste lineal respecto de dicho conjunto. El siguiente paso es la selección de las k instancias más próximas a la instancia que se pretende clasificar. La clase de destino que se designa a la nueva instancia es la mayoritaria en dicho conjunto.

Este tipo de aprendizaje se caracteriza por no estar basado en medidas de error, si no de distancia o cercanía. En las siguientes secciones se muestran distintas medidas de distancia comunmente utilizadas por el algoritmo *K-NN*. A pesar del elevado coste de clasificación y su gran simplicidad algorítmica, mediante la observación de resultados se ha demostrado que es una estrategia muy similar a nivel de tasas de error respecto de otras más sofisticadas en una gran cantidad de casos.

2. LA FIGURA 1 MUESTRA UN CONJUNTO DE ENTRENAMIENTO CON EJEMPLOS POSITIVOS (ESTRELLAS) Y NEGATIVOS (CÍRCULOS). SE DESEA CLASIFICAR LA NUEVA INSTANCIA $\langle 3, 3 \rangle$ 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 numéricos 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.

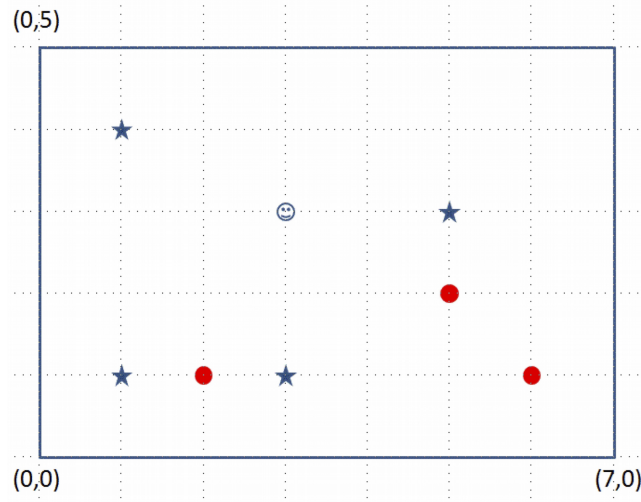


Figura 1: Representación Gráfica del problema 2

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.

$$ESTRELLAS = \{(1, 1), (1, 4), (3, 1), (5, 3)\} \quad (1)$$

$$CIRCULOS = \{(2, 1), (5, 2), (6, 1)\} \quad (2)$$

$$instancia = (3, 3) \quad (3)$$

Puesto que se pretende clasificar una nueva instancia apoyandose en la intuición 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 $normalize: \mathbb{R}^2 \rightarrow [0, 1]^2$ 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).

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

$$max_x = 6 \quad min_x = 1 \quad (5)$$

$$max_y = 4 \quad min_y = 1 \quad (6)$$

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.

$$ESTRELLAS_{normalized} = \{(0, 0), (0, 1), (\frac{2}{5}, 0), (\frac{4}{5}, \frac{2}{3})\} \quad (7)$$

$$CIRCULOS_{normalized} = \{(\frac{1}{5}, 0), (\frac{4}{5}, \frac{1}{3}), (1, 0)\} \quad (8)$$

$$instancia_{normalized} = (\frac{2}{5}, \frac{2}{3}) \quad (9)$$

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).

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

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

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.

$$\min(R_{euclidean}, 1) = \{0.4_e\} \implies instancia \in ESTRELLAS \quad (12)$$

$$\min(R_{euclidean}, 3) = \{0.4_e, 0.52_e, 0.52_c\} \implies instancia \in ESTRELLAS \quad (13)$$

$$\min(R_{euclidean}, 5) = \{0.4_e, 0.52_e, 0.52_c, 0.666_e, 0.696_e\} \implies instancia \in ESTRELLAS \quad (14)$$

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).

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

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

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.

$$\min(R_{w_{euclidean}}, 1) = \{0.178_e\} \implies instancia \in ESTRELLAS \quad (17)$$

$$\min(R_{w_{euclidean}}, 3) = \{0.178_e, 0.347_e, 0.347_c\} \implies instancia \in ESTRELLAS \quad (18)$$

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

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).

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

$$R_{manhattan} = \{4_e, 3_e, 2_e, 2_e, 3_c, 3_c, 5_c\} \quad (21)$$

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.

$$\min(R_{manhattan}, 1) = \{2_e\} \implies \text{instancia} \in ESTRELLAS \quad (22)$$

$$\min(R_{manhattan}, 3) = \{2_e, 2_e, 3_e\} \implies \text{instancia} \in ESTRELLAS \quad (23)$$

$$\min(R_{manhattan}, 5) = \{2_e, 2_e, 3_e, 3_c, 3_c\} \implies \text{instancia} \in ESTRELLAS \quad (24)$$

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).

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

$$R_{hamming} = \{2_e, 2_e, 1_e, 1_e, 2_c, 2_c, 2_c\} \quad (26)$$

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.

$$\min(R_{hamming}, 1) = \{1_e\} \implies \text{instancia} \in ESTRELLAS \quad (27)$$

$$\min(R_{hamming}, 3) = \{1_e, 1_e, 2_e\} \implies \text{instancia} \in ESTRELLAS \quad (28)$$

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

3. DÍGITOS MANUSCRITOS

En esta sección se comparan los resultados para distintos valores de k sobre un conjunto de datos de mayores dimensiones, tanto en cuanto al número de instancias que presenta como atributos de cada una de ellas. El conjunto de datos ha sido adaptado por *D. Barber* para el libro *Bayesian Reasoning and Machine Learning* [Bar12]. En dicha sección se ha utilizado la implementación en *WEKA* [too] del algoritmo de los *K-vecinos más cercanos* denominada *IBK*. Como medida de distancia se ha escogido la seleccionada por defecto (*Distancia Euclídea*).

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 $k \in \{1, 2, \dots, 9\}$

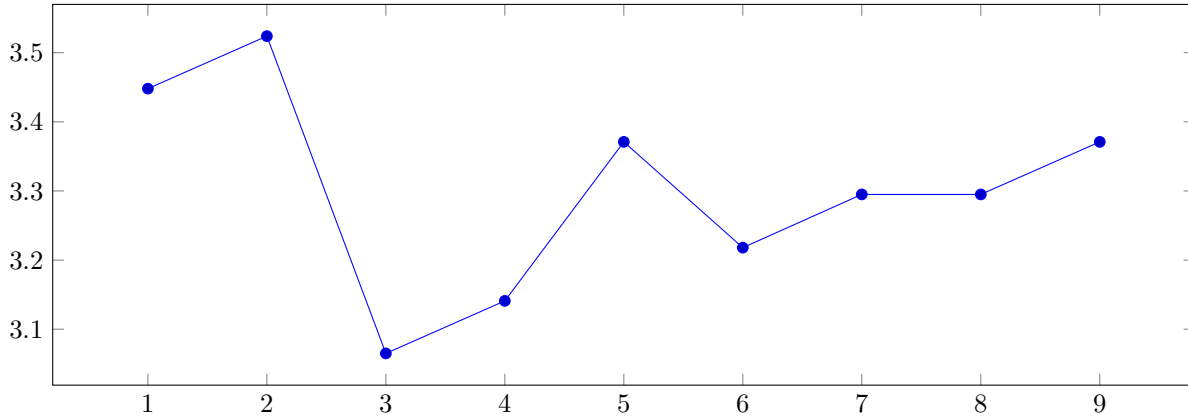


Figura 2: Representación Gráfica de la tasa de error obtenida tras realizar un experimento de Validación cruzada de 10 particiones con el clasificador K-NN para $k \in \{1, 2, \dots, 9\}$

Dicho conjunto de datos se corresponde con representaciones de mapas de bits formadas por $28^2 = 784$ atributos de carácter numérico que se refieren a dígitos numéricos. En concreto, el caso del conjunto de datos que se provee está formado únicamente por representaciones de **5** y **9**, que por su semejanza en cuanto a simbología es necesario escoger tanto un clasificador como parámetros adecuados. El clasificador seleccionado, al igual que en la sección anterior, es k -vecinos más próximos. Para el experimentos se proveen dos conjuntos de datos, el primero de ellos referido a tareas de entrenamiento está formado por 684 instancias que codifican el dígito 5 y 621 instancias que codifican el valor 9, lo cual presenta un total de $684 + 621 = 1305$ instancias. A partir de este conjunto de datos se pretende ajustar el valor óptimo de k para el cuál el clasificador de los mejores resultados. El experimento se ha realizado siguiendo una estrategia de **Validación Cruzada** de 10 particiones para $k \in \{1, 2, \dots, 9\}$.

Los resultados a nivel de *Tasas de Error* obtenidos tras la realización del experimento se muestran de manera tabulada en la tabla 1 y de manera gráfica en la figura 2. Tal y como se puede apreciar, el valor k que minimiza la tasa de error para el experimento realizado es $k = 3$. La siguiente labor es utilizar el segundo conjunto de datos, formado en este caso por 256 instancias referidas al dígito 5 y 352 referidas al 9. Por tanto, el total de instancias en este conjunto de datos es $256 + 352 = 608$. Nótese que el particionamiento de estos dos conjuntos de datos no sigue una distribución estratificada, ya que el ratio del número de instancias de cada clase no es semejante.

El resultado final que se pide en esta sección es la clasificación del conjunto de datos de prueba mediante el entrenamiento del clasificador con el conjunto de entrenamiento utilizado para encontrar el valor k óptimo ($k = 3$). Los resultados se muestran en la tabla 2. Tal y como se puede apreciar la tasa de error en este caso es del 0 %, lo cual significa que todas las instancias han sido clasificadas correctamente.

K -Vecinos más Próximos	
Datos	Tasa de Error ($K =$)
	3
Pruebas	0.0 %

Tabla 2: Tasa de error obtenida tras realizar un experimento con el conjunto de datos de Prueba y el clasificador K-NN para $k = 3$

REFERENCIAS

- [Bar12] D. Barber. *Bayesian Reasoning and Machine Learning*. Cambridge University Press, 2012.
- [CCAG17] Teodoro Calonge Cano and Carlos Javier Alonso González. Técnicas de Aprendizaje Automático, 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/>.