



200 lines (200 loc) · 12.6 KB

---

# MACHINE LEARNING

## TERCERA PARTE

M.SC. LUIS FERNANDO ESPINO BARRIOS  
2020

---

## APRENDIZAJE NO SUPERVISADO

- Es un tipo de ML que busca patrones no detectados en un conjunto de datos sin predicciones con una supervisión humana mínima.

# K-MEANS

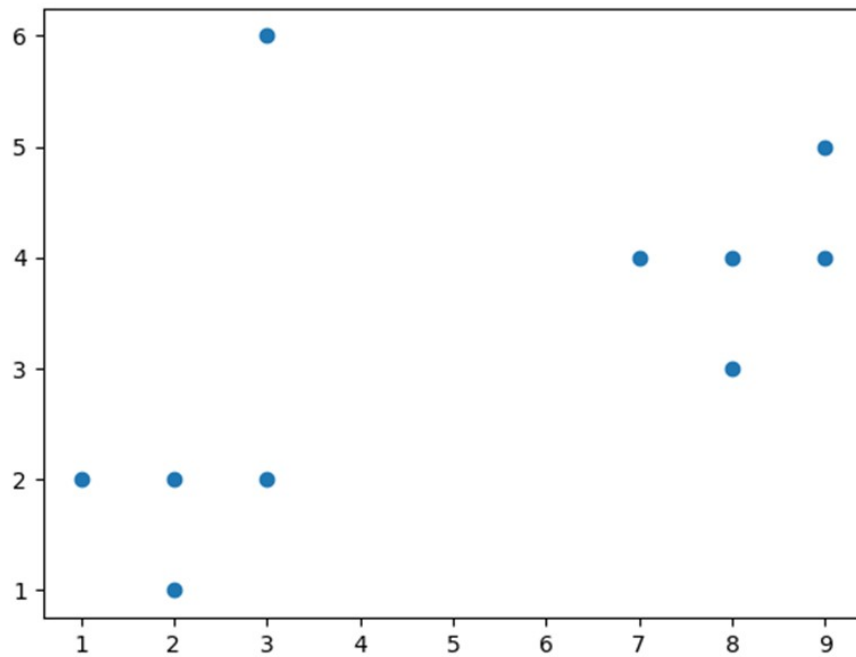
## ALGORITMO

- Seleccionar número de clúster
- Seleccionar centroides iniciales
- Calcular distancias de todos los puntos hacia los centroides para determinar a qué clúster pertenecen
- Recalcular el centroide dependiendo de los puntos que pertenecen a cada clúster
- Iterar los dos incisos anteriores hasta una convergencia del 100% o que no hayan cambios.

# EJEMPLO

- Detectar la anomalía mediante k-means.

Individuo	X	Y
A	3	2
B	2	2
C	2	1
D	1	2
E	3	6
F	7	4
G	8	4
H	8	3
I	9	4
J	9	5





	3	2	2	2	2	1	1	2	3	6	7	4	8	4	8	3	9	4	9	5
8.2 4	7.2	8.2	9.2	9.2	7.2	1.2	0.2	1.2	0.8	1.8										
3 4	2	3	4	4	2	4	5	6	6	7										
1.66 1.66	1.68	0.68	1	1	5.68	7.68	8.68	7.68	9.68	10.7										
	0	0	0	0	0	1	1	1	1	1										
	0	0	0	0	1	0	0	0	0	0										
	1	1	1	1	0	0	0	0	0	0										
8.2 4	7.2	8.2	9.2	9.2	7.2	1.2	0.2	1.2	0.8	1.8										
3 6	4	5	6	6	0	6	7	8	8	7										
2 1.75	1.25	0.25	0.75	1.25	5.25	7.25	8.25	7.25	9.25	10.3										
	0	0	0	0	0	1	1	1	1	1										
	0	0	0	0	1	0	0	0	0	0										
	1	1	1	1	0	0	0	0	0	0										

- Los centroides de cada clúster son:
  - Clúster 1 (8.2, 4)
  - Clúster 2 (3,6)
  - Clúster 3 (2,1.75)
- La pertenencia de los puntos en cada clúster es:
  - Clúster 1 (F, G, H, I, J)
  - Clúster 2 (E)
  - Clúster 3 (A,B,C,D)
- Por lo que la anomalía es el punto E (3,6)

# K-NEAREST NEIGHBOR

- Se calcula la distancia del punto a clasificar contra todos los datos de entrenamiento. El que tenga la menor distancia es el vecino más cercano, por lo que corresponde a dicha clase.

- Decidir a qué clase pertenece el individuo G mediante el método k-nearest neighbor utilizando la distancia Manhattan.

Individuo	X	Y	Clase
A	3	2	N
B	5	3	P
C	2	1	N
D	2	2	N
E	6	3	P
F	6	4	P
G	4	4	?

- La distancia más corta es 2, por lo que G pertenece a la clase P.

Individuo	X	Y	Clase	Distancia
A	3	2	N	3
B	5	3	P	2
C	2	1	N	5
D	2	2	N	4
E	6	3	P	3
F	6	4	P	2
G	4	4	P	

## EJEMPLO

Individuo	V1	V2	V3	CLASE
A	I	9	2	I
B	9	8	8	III
C	2	8	3	I
D	9	I	8	II
E	2	8	2	I
F	9	8	9	III
G	8	2	9	II
H	7	5	6	?



## MANHATTAN $|Y_2 - Y_1| + |X_2 - X_1|$

Individuo	V1	V2	V3	CLASE	Manhattan
A	1	9	2	I	14
B	9	8	8	III	7
C	2	8	3	I	11
D	9	1	8	II	8
E	2	8	2	I	12
F	9	8	9	III	8
G	8	2	9	II	7
H	7	5	6	?	
					II o III

## EUCLIDIANA $\text{RAÍZ}((Y_2 - Y_1)^2 + (X_2 - X_1)^2)$

Individuo	V1	V2	V3	CLASE	Euclidiana
A	1	9	2	I	8.2462113
B	9	8	8	III	4.1231056
C	2	8	3	I	6.5574385
D	9	1	8	II	4.8989795
E	2	8	2	I	7.0710678
F	9	8	9	III	4.6904158
G	8	2	9	II	4.3588989
H	7	5	6	?	
					III

# CONJUNTOS DIFUSOS

- Los conjuntos difusos fueron definidos por Zadeh en 1965.
- Un conjunto difuso  $A$  de un espacio de puntos  $X$  está caracterizado por la función de pertenencia que asocia a cada punto un número real de  $[0, 1]$

# CONJUNTOS DIFUSOS

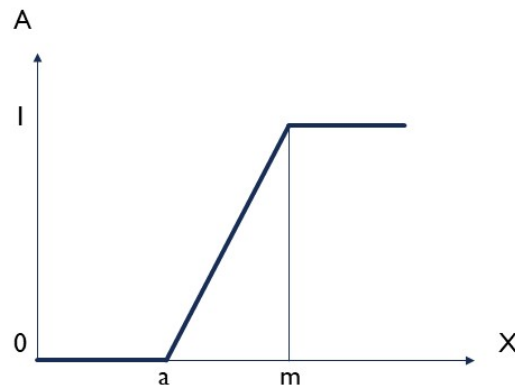
$$A: X \rightarrow [0, 1]$$

- Es decir, el valor  $f_A(x)$  representa un grado de pertenencia de  $x$  en  $A$ .
- Hay muchas aplicaciones para los conjuntos difusos: variables lingüísticas, lógica difusa, control difuso, bases de datos difusas, SQL difuso, etc.

# CONJUNTOS DIFUSOS

- Para aplicar los conjuntos difusos en la clasificación de clústers se toman los tipos de funciones de pertenencia como la triangular, medio trapezoide, función S, función trapezoide, pseudo exponencial, trapezoide extendida.
- Nos interesa la función medio trapezoide, donde:

$$A(x) = \begin{cases} 0 & \text{si } x \leq a \\ \frac{x-a}{m-a} & \text{si } m > x > a \\ 1 & \text{si } x \geq m \end{cases}$$



## NORMALIZACIÓN POR ESCALADA (X-MIN)/(MAX-MIN)

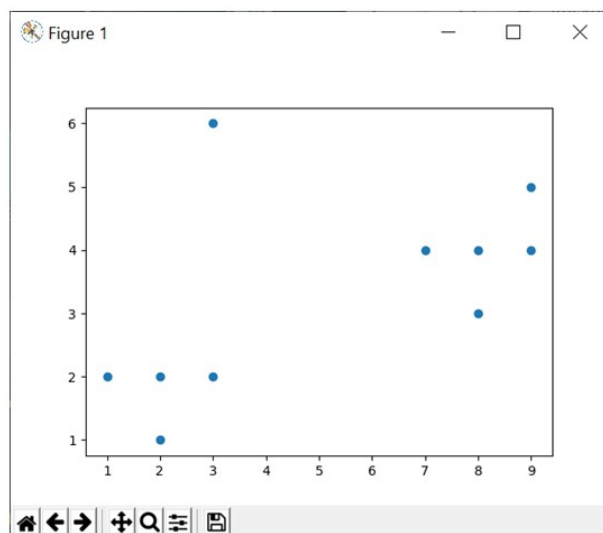
Individuo	V1	V2	V3	CLASE	v.estandar
A	1	9	2	I	
B	9	8	8	III	
C	2	8	3	I	
D	9	1	8	II	
E	2	8	2	I	
F	9	8	9	III	
G	8	2	9	II	
H	7	5	6	?	

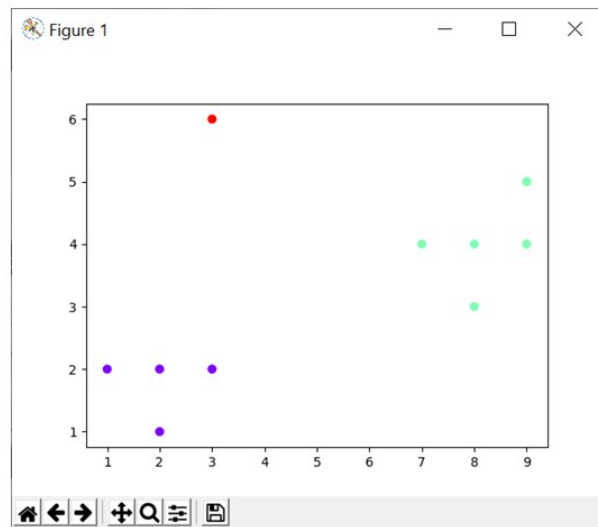
## NORMALIZACIÓN POR ESCALADA (X-MIN)/(MAX-MIN)

Individuo	V1	V2	V3	CLASE
A	1	9	2	I
B	9	8	8	III
C	2	8	3	I
D	9	1	8	II
E	2	8	2	I
F	9	8	9	III
G	8	2	9	II
H	7	5	6	?
min	1	1	2	
max	9	9	9	

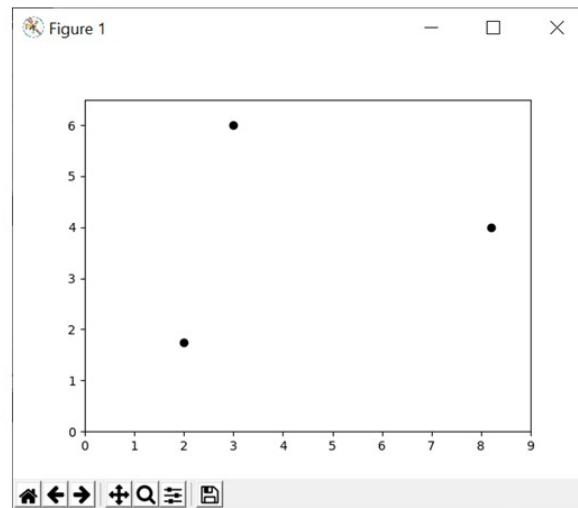
# NORMALIZACIÓN POR ESCALADA (X-MIN)/(MAX-MIN)

Individuo	V1	V2	V3	CLASE	Manhattan	euclidiana
A	0	1	0	I	1.821429	1.067254
B	1	0.875	0.85714286	III	0.910714	0.533627
C	0.125	0.875	0.14285714	I	1.428571	0.845531
D	1	0	0.85714286	II	1.035714	0.6278
E	0.125	0.875	0	I	1.571429	0.926164
F	1	0.875	1	III	1.053571	0.621931
G	0.875	0.125	1	II	0.928571	0.58303
H	0.75	0.5	0.57142857	?		
					III	III
min	1	1	2			
max	9	9	9			





```
Python 3.6.5 (v3.6.5:f59c0932b4, Mar 28 2018, 17:00:18)
4)] on win32
Type "copyright", "credits" or "license()" for more inf
>>>
===== RESTART: D:\python\Machine Learning\05 km
[[2.  1.75]
 [8.2  4.  ]
 [3.  6.  ]]
```



## K-NEAREST NEIGHBOR CON SCIKIT

Individuo	X	Y	Clase
A	3	2	N
B	5	3	P
C	2	1	N
D	2	2	N
E	6	3	P
F	6	4	P
G	4	4	?

```
from sklearn.neighbors import NearestCentroid
import numpy as np
X = np.array([[3, 2], [5, 3], [2, 1],
              [2, 2], [6, 3], [6, 4]])
y = np.array(['N', 'P', 'N',
              'N', 'P', 'P'])
clf = NearestCentroid()
clf.fit(X, y)

print(clf.predict([[4, 4]]))
```

```
===== RESTART: D:\python\Machine
['P']
>>> |
```



# HOJA DE TRABAJO: SCIKIT

Implementar el siguiente ejercicio:

Se desea clasificar las siguientes medicinas según sus atributos para determinar ciertas clases, la importancia radica en tener una efectividad teórica de 10 y un pH balanceado de 7. Para su clasificación se debe utilizar el algoritmo k-means clustering con distancia Manhattan, deduzca cuántas clases debería haber según los datos y cuáles medicinas corresponden a cada clase. Para determinar la primera iteración considere las distancias al centroide (10,7).

Medicina	Efectividad (0-10)	pH (0,14)
A	8	2
B	9	7
C	2	12
D	9	1
E	10	7
F	3	14
G	8	1
H	1	13

## TAREA 9: SCIKIT

Seleccionar un problema donde se pueda utilizar clustering (kmeans y knearest), utilizando los datos desarrolle la solución con scikit explicándola paso a paso.

# REFERENCIAS

- Quinlan, J. R. (1986). Induction of decision Trees. *Machine Learning*, 1(1), 81-106.
- Nilsson, N. (1998). *Artificial Intelligence: A New Synthesis*. United States of America: Morgan Kaufmann Publisher, Inc.
- Russell, S., & Norvig, P. (2003). *Artificial Intelligence: A Modern Approach* (2 ed.). United States of America: Pearson Education, Inc.
- Guía de Usuario de scikit-learn [https://scikit-learn.org/stable/user\\_guide.html](https://scikit-learn.org/stable/user_guide.html)