



Universidad de Alcalá

Escuela Politécnica Superior

Universidad de Alcalá

Fundamentos de Ciencia de Datos

Juan J. Cuadrado Gallego

Martes 10:00 – 12:00

Grado en Ingeniería Informática – Curso 2019/2020

Marcos Barranquero Fernández – 51129104N

TEMA 4 – CLASIFICACIÓN NO SUPERVISADA

INTRODUCCIÓN

La clasificación no supervisada trata de agrupar los datos sin saber previamente si se pueden clasificar. Partimos de una muestra, y los valores se definen en diferentes *clusters*, durante el proceso de clasificación.

Técnicas:

- K-means
- Clustering aglomerativo
- DBScan (no entra)

K-MEANS

El algoritmo consiste en:

1. Se elige un número k de clusters arbitrario.
2. Para cada cluster, se haya su centroide, que es un valor arbitrario de la muestra.
3. Se realiza el cálculo de la distancia eucladiana de cada punto a cada centroide.

Distancia eucladiana
$d(p_1, p_2) = \sqrt{(p_1x - p_2x)^2 + (p_1y - p_2y)^2}$

4. Para cada punto, se le asigna como cluster aquel cuyo centroide esté más cerca de dicho punto.
5. Se recalculan los centroides. El nuevo centroide es la media de x e y de todos los puntos de ese centroide.

Nuevo centroide

$C' = (\bar{x}, \bar{y})$

6. Se realizan iteraciones desde el paso 3 hasta que no varíen los puntos de los centroides.

EJEMPLO

Para la siguiente muestra:

Índice	X	Y
1	4	4
2	3	5
3	1	2
4	5	5
5	0	1
6	2	2
7	4	5
8	2	1

1. Elegimos 2 clusters. $K = 2$.
2. Elegimos centroides:
 - a. $C1 = (0, 1)$
 - b. $C2 = (2, 2)$

3. Calculamos distancias euclidianas:

Punto	Distancia a C1	Distancia a C2
(4, 4)	$\sqrt{16 + 9} = 5$	$\sqrt{4 + 4} = 2'82$
(3, 5)	$\sqrt{9 + 10} = 5$	$\sqrt{1 + 9} = 3'16$
(1, 2)	$\sqrt{11 + 1} = 1'41$	$\sqrt{1 + 0} = 1$
(5, 5)	$\sqrt{25 + 16} = 6'4$	$\sqrt{9 + 9} = 4'24$
(0, 1)	$\sqrt{0 + 0} = 0$	$\sqrt{4 + 1} = 2'23$
(2, 2)	$\sqrt{4 + 1} = 2'23$	$\sqrt{0 + 0} = 0$
(4, 5)	$\sqrt{16 + 16} = 5'66$	$\sqrt{4 + 9} = 3'60$
(2, 1)	$\sqrt{4 + 0} = 2$	$\sqrt{0 + 1} = 1$

4. Para cada punto, le asignamos como cluster aquel cuya distancia a su centroide sea menor:
 $C1 = \{p5\}$; $C2 = \{p1, p2, p3, p4, p6, p7, p8\}$
5. Se recalculan centroides:
 $C1 = (\bar{x}, \bar{y}) = (0, 1)$; $C2 = (\bar{x}, \bar{y}) = (3, 3'43)$
6. Se vuelven a calcular distancias euclidianas:

Punto	Distancia a C1(0,1)	Distancia a C2(3, 3'43)
(4, 4)	$\sqrt{16 + 9} = 5$	$\sqrt{1 + 0'33} = 1'15$
(3, 5)	$\sqrt{9 + 10} = 5$	$\sqrt{0 + 2'46} = 1'57$
(1, 2)	$\sqrt{11 + 1} = 1'41$	$\sqrt{4 + 2'04} = 2'45$
(5, 5)	$\sqrt{25 + 16} = 6'4$	$\sqrt{4 + 2'46} = 2'54$
(0, 1)	$\sqrt{0 + 0} = 0$	$\sqrt{9 + 5'9} = 3'86$
(2, 2)	$\sqrt{4 + 1} = 2'23$	$\sqrt{1 + 2'05} = 1'74$
(4, 5)	$\sqrt{16 + 16} = 5'66$	$\sqrt{1 + 2'46} = 1'86$
(2, 1)	$\sqrt{4 + 0} = 2$	$\sqrt{1 + 5'9} = 2'63$

7. Para cada punto, volvemos a recalcular su cluster asignado:
 $C1 = \{p3, p5, p8\}$; $C2 = \{p1, p2, p4, p6, p7\}$
8. Se recalculan centroides:
 $C1 = (\bar{x}, \bar{y}) = (1, 1'13)$; $C2 = (\bar{x}, \bar{y}) = (3'6, 4'2)$

9. Se recalculan distancias:

Punto	Distancia a C1(0,1)	Distancia a C2(3, 3'43)
(4, 4)	4'04	0'45
(3, 5)	4'2	1
(1, 2)	0'7	3'41
(5, 5)	5'44	1'61
(0, 1)	1'02	4'82
(2, 2)	1'22	2'12
(4, 5)	4'76	0'89
(2, 1)	1'02	3'58

10. Para cada punto, volvemos a calcular su cluster asignado:

$$C1 = \{p3, p5, p6, p8\}; C2 = \{p1, p2, p4, p7\}$$

11. Se recalculan centroides:

$$C1 = (\bar{x}, \bar{y}) = (1'25, 1'5); C2 = (\bar{x}, \bar{y}) = (4, 4'75)$$

12. Se recalculan distancias euclídeas:

Punto	Distancia a C1(1'25, 1'5)	Distancia a C2(4, 4'75)
(4, 4)	3'72	0'75
(3, 5)	3'91	1'03
(1, 2)	0'56	4'07
(5, 5)	5'13	1'03
(0, 1)	1'34	5'48
(2, 2)	0'9	3'4
(4, 5)	4'45	0'25
(2, 1)	0'9	4'25

13. Para cada punto, volvemos a calcular su cluster asignado:

$$C1 = \{p3, p5, p6, p8\}; C2 = \{p1, p2, p4, p7\}$$




Vemos que ningún punto cambia de cluster, por tanto, podemos concluir el algoritmo. Ya tenemos dos conjuntos diferenciados.

CLUSTERING AGLOMERATIVO

Con esta técnica, tomamos los puntos individuales cada uno como un cluster, e iremos juntando clusters hasta obtener la clasificación.

El algoritmo consiste en:

1. Se construye matriz de distancias euclidianas entre clusters. En la primera iteración, cada punto es un cluster.
2. Se unen los clusters más próximos, atendiendo a 3 criterios:

Distancia mínima	Distancia máxima	Media de distancias
Distancia entre los dos puntos más cercanos entre sí de cada cluster.	Distancia mínima entre los dos puntos más lejanos entre sí de cada cluster.	Media de distancias entre todos los puntos de ambos clusters.
		

3. Repetir hasta que solo quede un cluster.

EJEMPLO

Para el siguiente conjunto de datos, clasificar atendiendo a los 3 criterios.

Índice	X	Y
1	0,89	2'94
2	4'36	5'21
3	3'75	1'12
4	6'25	3'14
5	4'1	1'8
6	3'9	4'27

En primer lugar, calculo la matriz de distancias, común a todos los criterios:

	P1	P2	P3	P4	P5	P6
P1	0	4'15	3'39	5'36	3'4	3'3
P2	4'15	0	4'13	2'8	3'42	1'0
P3	3'39	4'14	0	3'2	0'77	3'1
P4	5'36	2'8	3'2	0	2'5	2'6
P5	3'4	3'14	0'7	2'5	0	2'5
P6	3'3	1'04	3'15	2'6	2'5	0

ALGORITMO CON DISTANCIA MÍNIMA

Para hallar los dos clusters más cercanos, encontramos los dos clusters cuya distancia sea menor. En este caso es $(p3, p5) = 0'7$.

$$C1 = \{p3, p5\}; \text{ Resto} = \{p1, p2, p4, p6\}$$

Vuelvo a construir matriz:

	C1	P1	P2	P4	P6
C1	0	3'39	3'14	3'21	2'48
P1	3'39	0	4'15	5'37	3'23
P2	3'14	4'15	0	2'18	1'04
P4	3'21	5'37	2'8	0	2'6
P6	2'48	3'23	1'04	2'6	0

- $d(P1, C1) = \min(d(p1, p3), d(p1, p5)) = 3'39$
- $d(P2, C1) = \min(d(p2, p3), d(p2, p5)) = 3'14$
- $d(P4, C1) = \min(d(p4, p3), d(p4, p5)) = 3'21$
- $d(P6, C1) = \min(d(p6, p3), d(p6, p5)) = 3'48$

Los clusters más cercanos entre si son $(p2, p6) = 1'04$. Actualizo clusters:

$$C1 = \{p3, p5\}; C2 = \{p2, p6\}; \text{ Resto} = \{p1, p4\}$$

Vuelvo a construir la matriz:

	C1	C2	P1	P4
C1	0	2'5	3'39	3'21
C2	2'5	0	3'3	2'6
P1	3'39	3'3	0	5'37
P4	3'21	2'6	5'37	0

- $d(c1, c2) = \min(d(p3, p2), d(p3, p6), d(p5, p2), d(p5, p6)) = 2'5$
- $d(p1, c2) = \min(d(p1, p2), d(p1, p6)) = 3'3$
- $d(p4, c2) = \min(d(p4, p2), d(p4, p6)) = 2'6$

Los clusters más cercanos entre si son $(c1, c2) = 2'5$. Actualizo clusters:

$$C3 = C1 + C2 = \{p2, p3, p5, p6\}; \text{ Resto} = \{p1, p4\}$$

	C3	P1	P4
C3	0	3'3	2'6
P1	3'3	0	5'37
P4	2'6	5'37	0

- $d(p1, c3) = \min(d(p1, p2), d(p1, p3), \dots) = 3'3$
- $d(p4, c3) = \min(d(p4, p2), d(p4, p3), \dots) = 2'6$

Los clusters más cercanos entre si son $(c3, p4) = 2'6$. Actualizo clusters:

$$C3 = \{p2, p3, p4, p5, p6\}; \text{ Resto} = \{p1\}$$

Como ya tengo dos clusters diferenciados, finalizo el algoritmo.

ALGORITMO CON DISTANCIA MÁXIMA

Partiendo de la matriz de distancias entre puntos. Para hallar los dos clusters más cercanos, encontramos los dos clusters cuya distancia sea menor. En este caso es $(p3, p5) = 0'7$.

$$C1 = \{p3, p5\}; \text{ Resto} = \{p1, p2, p4, p6\}$$

Vuelvo a construir matriz:

	C1	P1	P2	P4	P6
C1	0	3'39	4'14	3'21	3'1
P1	3'39	0	4'15	5'37	3'23
P2	4'14	4'15	0	2'18	1'04
P4	3'21	5'37	2'8	0	2'6
P6	3'1	3'23	1'04	2'6	0

- $d(p2, c1) = \max(d(p2, p3), d(p2, p5)) = 4'14$
- $d(p6, c1) = \max(d(p6, p3), d(p6, p5)) = 3'1$

Los clusters más cercanos entre si son $(p2, p6) = 1'04$. Actualizo clusters:

$$C1 = \{p3, p5\}; C2 = \{p2, p6\}; \text{ Resto} = \{p1, p4\}$$

Vuelvo a construir la matriz:

	C1	C2	P1	P4
C1	0	4'14	3'39	3'21
C2	4'14	0	4'15	2'8
P1	3'39	4'15	0	5'37
P4	3'21	2'8	5'37	0

Los clusters más cercanos entre si son $(c2, p4) = 2'8$. Actualizo clusters:

$$C3 = C2 + P4 = \{p2, p4, p6\}; C1 = \{p3, p5\}; \text{ Resto} = \{p1\}$$

Vuelvo a construir la matriz:

	C1	C3	P1
C1	0	4'13	3'4
C3	4'13	0	5'37
P1	3'4	5'37	0

Los clusters más cercanos entre si son $(c1, p1) = 3'4$. Actualizo clusters:

$$C4 = C1 + p1 = \{p1, p3, p5\}; C3 = \{p2, p4, p6\}$$

Como ya tengo dos clusters diferenciados, finalizo el algoritmo.

ALGORITMO CON DISTANCIA MEDIA

Partiendo de la matriz de distancias entre puntos. Para hallar los dos clusters más cercanos, encontramos los dos clusters cuya distancia sea menor. En este caso es $(p3, p5) = 0'7$.

$$C1 = \{p3, p5\}; \text{ Resto} = \{p1, p2, p4, p6\}$$

Vuelvo a construir la matriz, atendiendo al criterio de distancia media:

	C1	P1	P2	P4	P6
C1	0	3'39	3'77	2'85	2'82
P1	3'39	0	4'15	5'37	3'23
P2	3'77	4'15	0	2'18	1'04
P4	2'85	5'37	2'8	0	2'6
P6	2'82	3'23	1'04	2'6	0

- $d(c_1, p_1) = \frac{d(p_1, p_5) - d(p_1, p_5)}{2} = \frac{3'39 + 3'4}{2} = 3'39$
- Igual para el resto de distancias de c1 ...

Los clusters más cercanos entre si son $(p2, p6) = 1'04$. Actualizo clusters:

$$C1 = \{p3, p5\}; C2 = \{p2, p6\}; \text{ Resto} = \{p1, p4\}$$

Vuelvo a construir la matriz:

	C1	C2	P1	P4
C1	0	3'3	3'39	2'82
C2	3'3	0	3'72	2'7
P1	3'39	3'72	0	5'37
P4	2'82	2'7	5'37	0

Los clusters más cercanos entre si son $(c2, p4) = 2'7$. Actualizo clusters:

$$C3 = C2 + P4 = \{p2, p4, p6\}; C1 = \{p3, p5\}; \text{ Resto} = \{p1\}$$

Vuelvo a construir la matriz:

	C1	C3	P1
C1	0	3'1	3'4
C3	3'1	0	4'27
P1	3'4	4'27	0

Los clusters más cercanos entre si son $(c1, c3) = 3'1$. Actualizo clusters:

$$C4 = C1 + C3 = \{p2, p3, p4, p5, p6\}; \text{ Resto} = \{p1\}$$

Como ya tengo dos clusters diferenciados, finalizo el algoritmo.

TEMA 5 – DATOS ANÓMALOS

INTRODUCCIÓN

Un dato anómalo es aquel que no se ajusta al resto de datos. Distinguimos 2 tipos:

- **Erróneos:** errores de medida en el proceso de estudio que debemos eliminar.
- **Correctos:** datos correctos pero que se apartan de lo normal y deben de ser analizados para verificar si nos interesa descartarlos.

A estos casos anómalos se les llama **outliers**.

El analista tiene la decisión de fijar el grado de outlier, que determinará que datos se consideran anómalos y cuáles no.

Podemos determinar los datos anómalos según dos criterios:

- Si tenemos un modelo de clasificación y los datos no se ajustan, podemos considerarlos anómalos y eliminarlos. Hay que verificar que efectivamente es un dato anómalo con el poder del sentido común.
- Según la técnica empleada, podemos ejecutar diferentes algoritmos para descartar los datos anómalos.
 - Técnicas estadísticas:
 - Caja y bigotes
 - Dispersión y desviación típica
 - Regresión y error estándar
 - Técnicas basadas en proximidad.
 - K-vecinos más próximo
 - Técnicas basadas en densidad de datos.
 - Técnicas basadas en clusters.

ALGORITMO CAJA Y BIGOTES

Algoritmo solo valido para datos de 1 variable.

1. Determinamos grado de outlier **d**.
2. Se ordenan datos y se extraen cuantiles.
3. Se calculan límites del intervalo para valores típicos:
 $(Q_1 - d \cdot (Q_3 - Q_1), Q_1 + d \cdot (Q_3 - Q_1))$
4. Los datos que caigan fuera de ese intervalo son datos anómalos y podemos eliminarlos.

EJEMPLO

Índice	X
1	3
2	3'5
3	4'7
4	5'2
5	6'2
6	7'1
7	14

1. $d = \text{grado de anormalidad} = 1'5$
2. $Q_1 = 7 \cdot \frac{1}{4} = 1'75; 1'75 \notin \mathbb{N} = x_{[1'75+1]} = x_2 = 3'5$
3. $Q_3 = 7 \cdot \frac{3}{4} = 5'25; 5'25 \notin \mathbb{N} = x_{[5'25+1]} = x_6 = 7'1$
4. Intervalo = $(3'5 - 1'5 \cdot (3'6), 3'5 + 1'5 \cdot (3'6)) = (-1'9, 12'5)$
5. El dato 7 (14, 5'3) cae fuera del intervalo, podemos eliminarlo.

K-VECINOS

Este algoritmo se aplica a datos con dos variables.

El algoritmo consta de los siguientes pasos:

1. Determinamos grado de outlier **d**.
2. Determinamos k, grado del vecino a comparar.
 - a. Si $k = 1$, compararemos con el 1er vecino más próximo.
 - b. Si $k = 2$, compararemos con el 2do vecino más próximo.
 - c. Etc.
3. Calculo distancias euclídeas entre puntos.
4. Ordenar y calcular distancias hasta llegar al k-vecino.
5. Identificar outliers: aquellos cuya distancia k-vecina sea mayor que el grado de outlier.

EJEMPLO

Índice	X	Y
1	4	4
2	4	3
3	5	5
4	1	1
5	5	4

1. $d = 2.5$
2. $k = 3$ er vecino más próximo.
3. Matriz de distancias euclídeas:

	P1	P2	P3	P4	P5
P1	0	1	1'41	4'24	1
P2	1	0	2'23	3'60	1'41
P3	1'41	2'23	0	5'65	1
P4	4'24	3'60	5'65	0	5
P5	1	1'41	1	5	0

4. Calculamos distancias en función de k. Como $k = 3$, debemos fijarnos en la tercera distancia más próxima de cada punto.

k-vecino(p1) = p3; $d(p1, p3) = 1'41$
k-vecino(p2) = p3; $d(p2, p3) = 2'24$
k-vecino(p3) = p2; $d(p3, p2) = 2'24$
k-vecino(p4) = p2; $d(p4, p2) = 1'41$
k-vecino(p5) = p1; $d(p5, p1) = 5$

5. Identificamos casos cuya distancia a su k-vecino sea mayor que el grado de outlier. En este caso, en único punto cuya distancia es mayor que d es el p5.

DESVIACIÓN TÍPICA

El algoritmo consiste en los siguientes pasos:

1. Determinar el grado de outlier **d**. Por defecto 2, 2'5, etc.
2. Sacar media de \bar{x}
3. Sacar desviación típica σ
4. Se calcula Intervalo de valores aceptables:
$$\text{Intervalo} = (\bar{x} - d \cdot \sigma, \bar{x} + d \cdot \sigma)$$
5. Los valores fuera del intervalo son outliers.

EJEMPLO

Encontrar outliers para el siguiente conjunto de datos:

Índice	X
1	3
2	3'5
3	4'7
4	5'2
5	6'2
6	7'1
7	14

1. Establecemos **d** como 2.
2. Averiguamos $\bar{x} = 5'65$
3. Averiguamos $\sigma = 2'85$
4. $\text{Intervalo} = (5'65 - 2 \cdot 2'85, 5'65 + 2 \cdot 2'85) = (-0'05, 11'35)$.
5. El dato 7 se excede del rango, por lo tanto, es un outlier.