INGENIERÍA DE LA INFORMACIÓN IN4151-2
OTOÑO 2024

AUX 8

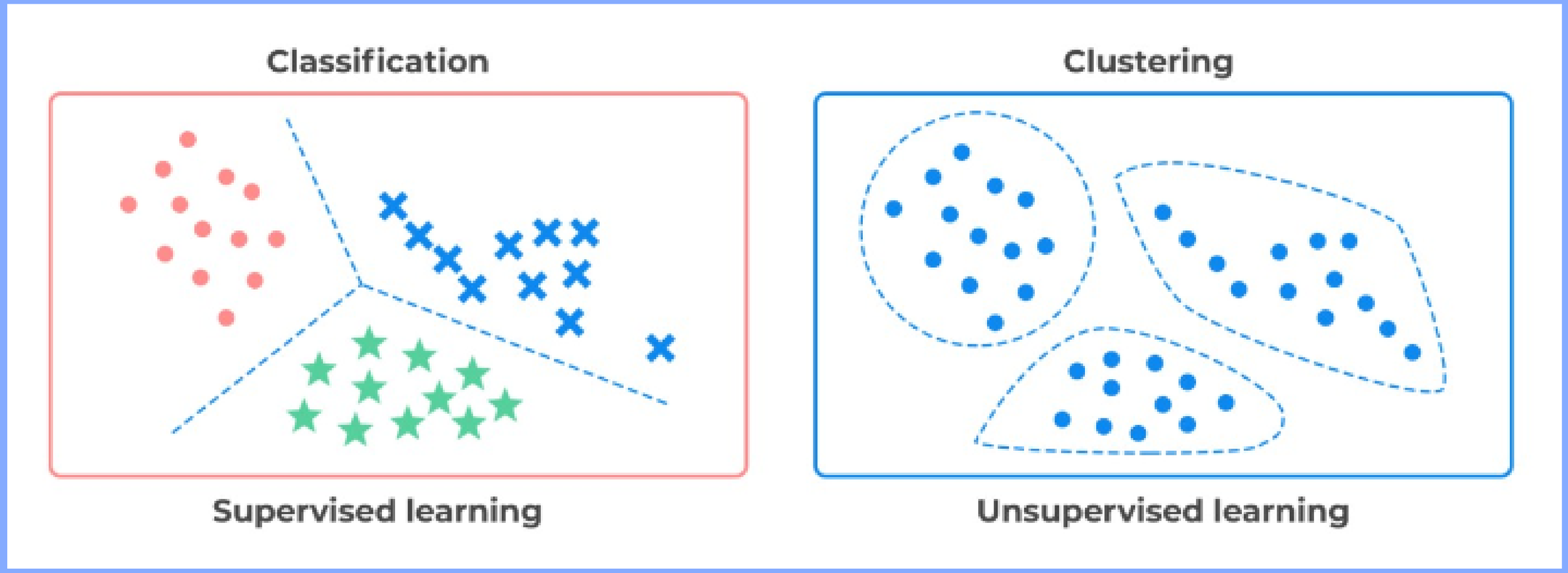
JOSÉ SAFFIE Y JOSÉ SOZA

K'S

AGENDA

- Aprendizaje No Supervisado
- K-means
- K-modes
- K-prototypes

APRENDIZAJE NO SUPERVISADO



K-MEANS

1. First, choose K initial centroids, where K is a user specified parameter, namely, the number of clusters desired.
2. Each point is then assigned to the closest centroid, and each collection of points assigned to a centroid is a cluster.
3. Then, the centroid of each cluster is updated based on the points assigned to the cluster.
4. Repeat the assignment and update steps until *no point changes clusters* (until centroids remain the same).

$$SSE = \sum_{i=1}^{K} \sum_{x \in C_i} dist(c_i, x)^2$$

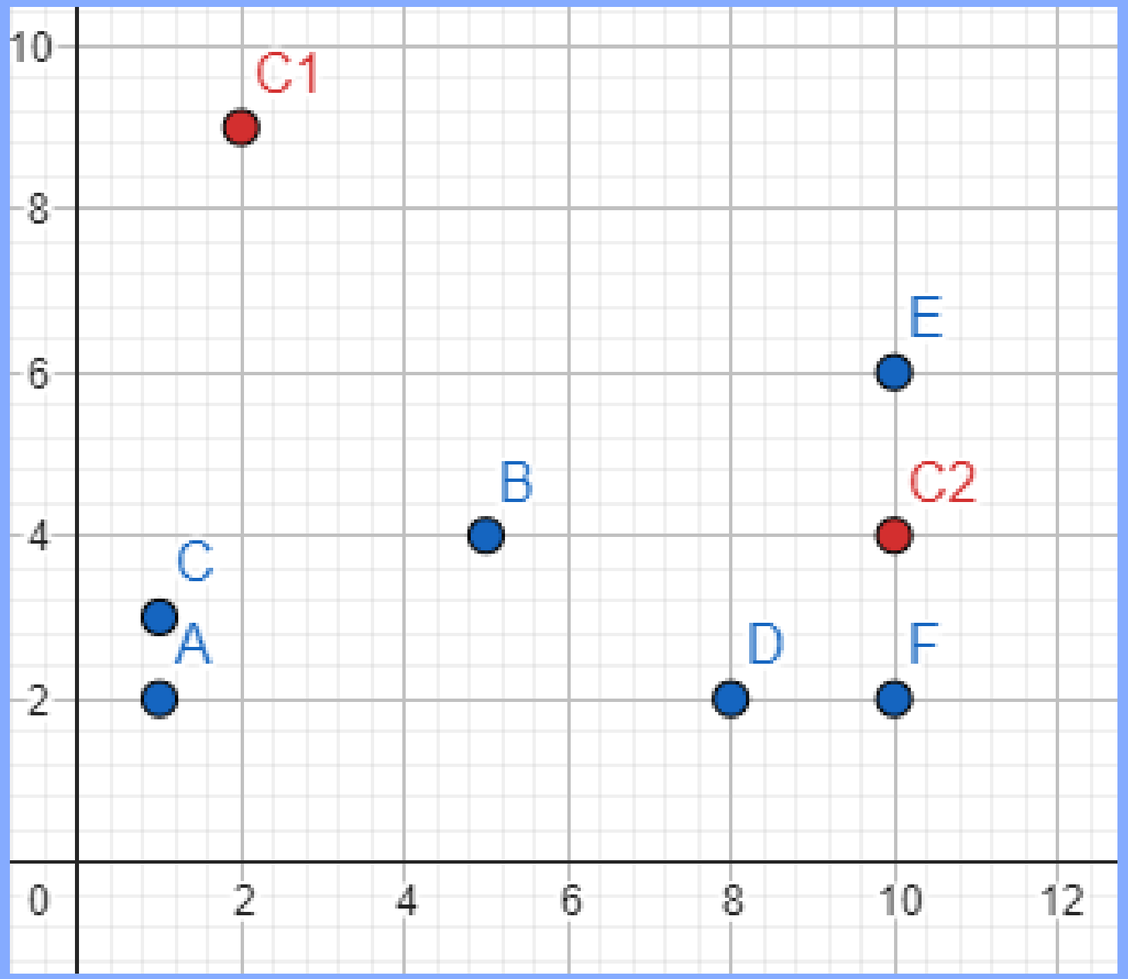
K-MEANS

P1. K-means.- Sea el siguiente dataset

x	1	5	1	8	10	10
y	2	4	3	2	6	2

Construya clusters para los datos utilizando el algoritmo de K-means, indicando claramente sus centroides y los puntos que pertenecen a cada uno. Para ello, considere 2 clusters, y como centroides iniciales $(x = 2, y = 9)$ y $(x = 10, y = 4)$. Luego indique a qué cluster pertenece el punto $(x = 3, y = 7)$.

K-MEANS



K-MEANS

Revisamos a dónde pertenece el punto A = (1,2)

C1 = (2,9)

$$d(x, y) = \sqrt{\sum_{i=1}^{n}(y_i - x_i)^2}$$

Distancia Euclidiana

C2 = (10,4)

$\sqrt{(2-1)^2 + (9-2)^2} = 7.071$

$\sqrt{(10-1)^2 + (4-2)^2} = 9.219$

En C1 posee menos distancia, por lo que A se agrupará con C1

K-MEANS

Revisamos a dónde pertenece el punto B = (5,4)

C1 = (2,9)

$$d(x, y) = \sqrt{\sum_{i=1}^{n}(y_i - x_i)^2}$$

Distancia Euclidiana

C2 = (10,4)

$$\sqrt{(2-5)^2 + (9-4)^2} = 5.830$$

$$\sqrt{(10-5)^2 + (4-4)^2} = 5$$

En C2 posee menos distancia, por lo que B se agrupará con C2

K-MEANS

Repetimos el procedimiento para todos los puntos:

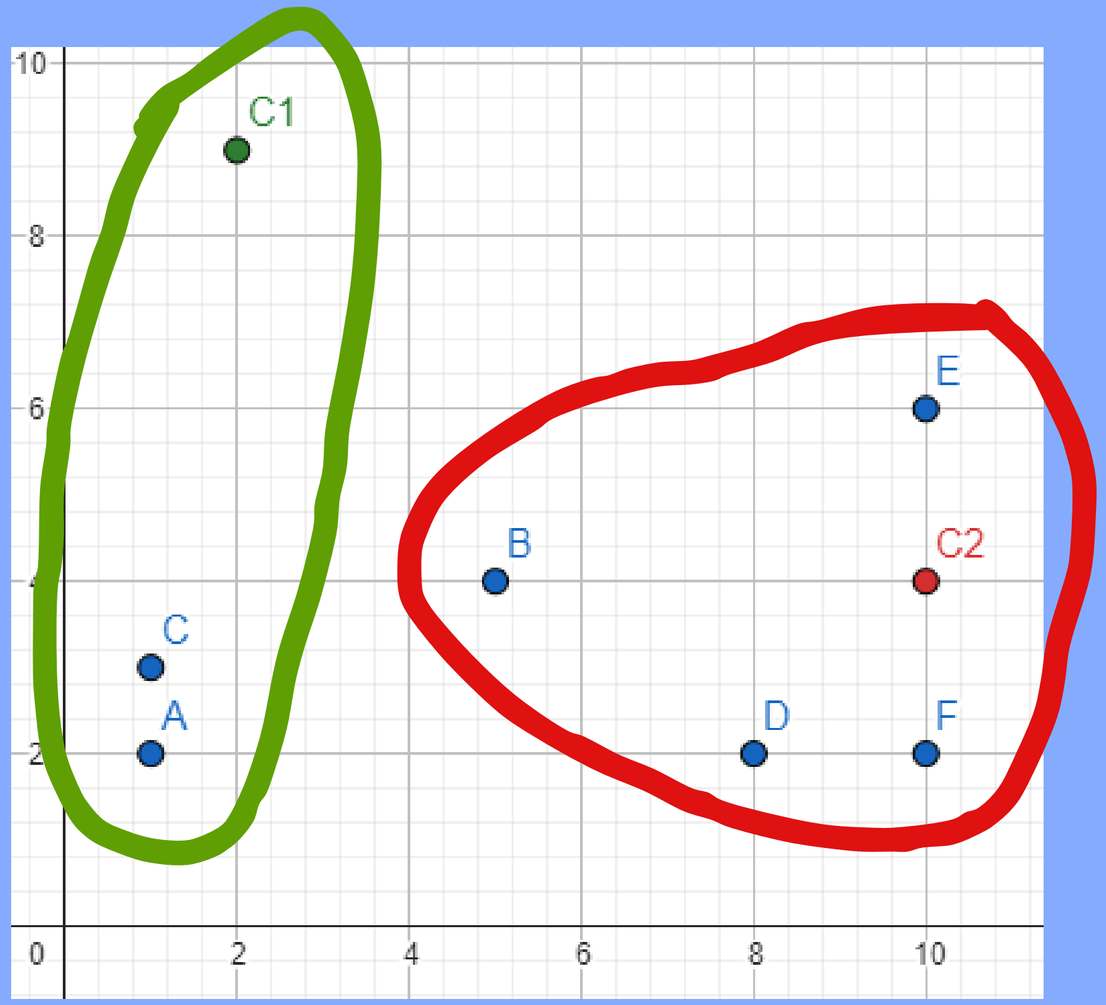
C1 = (2,9)

A,C

C2 = (10,4)

B,D,E,F

K-MEANS



K-MEANS

Actualizamos los centroides con sus respectivos puntos

C1 posee a A = (1,2) y C = (1,3)

C1'(x) = (1+1)/2

C1'(y) = (2+3)/2

C1' = (1,5/2)

K-MEANS

Actualizamos los centroides con sus respectivos puntos

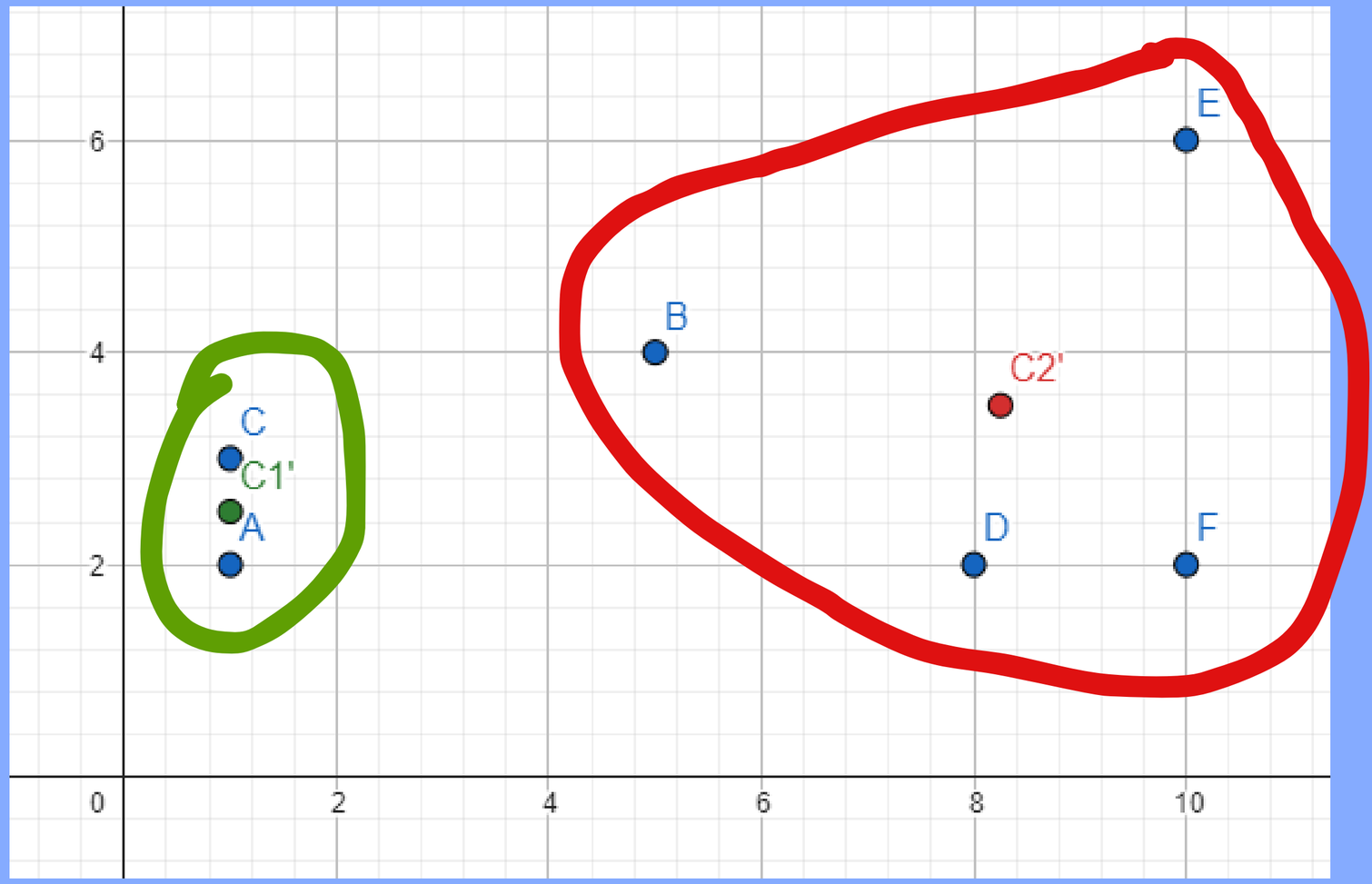
C2 posee a B = (5,4), D = (8,2), E = (10,6) y F = (10,2)

C2'(x) = (5+8+10+10)/4

C2'(y) = (4+2+6+2)/4

C2' = (33/4,14/2)

K-MEANS



K-MEANS

Con la actualización de los centroides, ahora debemos repetir el mismo procedimiento de agrupar los puntos en los nuevos centroides. Si resulta que no hay cambios en la agrupación, se termina el algoritmo.

K-MEANS

Revisamos a dónde pertenece el punto A = (1,2)

$$d(x, y) = \sqrt{\sum_{i=1}^{n}(y_i - x_i)^2}$$

Distancia Euclidiana

C1' = (1,5/2)

$$\sqrt{(1-1)^2 + (5/2-2)^2} = 3/2$$

C2' = (33/4,14/4)

$$\sqrt{(33/4-1)^2 + (14/4-2)^2} = 4.968$$

En C1' posee menos distancia, por lo que A se agrupará con C1'

K-MEANS

SPOILER: Todos los puntos quedan en su misma agrupación anterior, por lo que se termina el algoritmo :)

Hagan igual el algoritmo para comprobarlo

K-MEANS

Una vez terminado el algoritmo,
nos falta identificar dónde
pertenecerá el punto X = (3,7)

Ya tenemos nuestros centroides finales, por lo
que resta calcular dónde posee menor distancia

K-MEANS

Revisamos a dónde pertenece el punto X = (3,7)

C1' = (1,5/2)

$$d(x, y) = \sqrt{\sum_{i=1}^{n}(y_i - x_i)^2}$$

Distancia Euclidiana

C2' = (33/4,14/4)

$$\sqrt{(1-3)^2 + (5/2-7)^2} = 4.924$$

$$\sqrt{(33/4-3)^2 + (14/4-7)^2} = 6.3$$

En C1' posee menos distancia, por lo que X se agrupará con C1'

K-MEANS++ Y K-MEDOIDS

K-Means++: Algoritmo que incluye la elección de los K centroides iniciales

K-Medoids: Algoritmo que disminuye la importancia de outliers, ocupando los mediodes.

K-MODES

The following are the steps for k-modes based clustering (Huang 2008):

1. Select the k initial mode
2. Allocate the observation to the closest cluster based on a simple dissimilarity measure. Update each cluster mode after each allocation.
3. After all the observations have been allocated to a cluster, check the dissimilarity value of each observation against the mode. If an observation turns out that the closest mode is in another cluster, move the observation to the appropriate cluster and update the mode of both clusters.
4. Repeat step 3 until none of the observations change to another cluster.

K-MODES

P2. K-modes.- Sea el siguiente dataset:

Individuo	Comuna	Carrera	Estado Civil	Vehículo
1	Puente Alto	Ingeniería Civil	Soltero	Sedan
2	Maipú	Medicina	Casado	Moto
3	Santiago	Ingeniería Civil	Divorciado	Sedan
4	Independencia	Arquitectura	Soltero	SUV
5	La Florida	Ingeniería Civil	Casado	SUV
6	Maipú	Arquitectura	Casado	SUV
7	Puente Alto	Medicina	Soltero	Sedan
8	Maipú	Arquitectura	Casado	Moto

Aplique el algoritmo de K-modes para segmentar los individuos en 2 clusters, indicando claramente a qué cluster pertenece cada uno y sus respectivos centroides. Para ello, considere como clusters iniciales a los individuos 3 y 8.

K-MODES

Identificamos nuestros centroides:

C1 = {Santiago, Ingeniería Civil, Divorciado, Sedan}

C2 = {Maipú, Arquitectura, Casado, Moto}

K-MODES

Revisamos las disimilitudes del Individuo 1 con ambos centroides:

I1 = {Puente Alto, Ingeniería Civil, Soltero, Sedan}

C1 = {Santiago, Ingeniería Civil, Divorciado, Sedan}

Se diferencian en 2 atributos

C2 = {Maipú, Arquitectura, Casado, Moto}

Se diferencian en 4 atributos

Tiene menos disimilitudes con C1, por lo que I1 se agrupa con C1

K-MODES

Revisamos las disimilitudes del Individuo 5 con ambos centroides:

I5 = {La Florida, Ingeniería Civil, Casado, SUV}

C1 = {Santiago, Ingeniería Civil, Divorciado, Sedan}

Se diferencian en 3 atributos

C2 = {Maipú, Arquitectura, Casado, Moto}

Se diferencian en 3 atributos

Tienen la misma disimilitud en ambos Clusters! D: (lo resolveremos luego...)

K-MODES

Repitiendo el mismo procedimiento para todos los Individuos:

C1 = {Santiago, Ingeniería Civil, Divorciado, Sedan}

C2 = {Maipú, Arquitectura, Casado, Moto}

11,13,17

12,14,16,18

K-MODES

Ahora actualizamos los centroides según la moda:

C1 posee a I1,I3,I7:

I1 = {Puente Alto, Ingeniería Civil, Soltero, Sedan}

I3 = {Santiago, Ingeniería Civil, Divorciado, Sedan}

I7 = {Puente Alto, Medicina, Soltero, Sedan}

C1' = {Puente Alto, Ingeniería Civil, Soltero, Sedan}

K-MODES

Ahora actualizamos los centroides según la moda:

C2 posee a I2,I4,I6,I8:

I2 = {Maipú, Medicina, Casado, Moto}
I4 = {Independencia, Arquitectura, Soltero, SUV}
I6 = {Maipú, Arquitectura, Casado, SUV}
I8 = {Maipú, Arquitectura, Casado, Moto}

C2' = {Maipú, Arquitectura, Casado, Moto o Suv (?)}

K-MODES

Ya que C2' quedó con un atributo doble, podemos usar al I5 que habíamos dejado afuera para forzar el óptimo del centroide:

C2' = {Maipú, Arquitectura, Casado, Moto o Suv (?)}

I5 = {La Florida, Ingeniería Civil, Casado, SUV}

C2' = {Maipú, Arquitectura, Casado, SUV}

K-MODES

Finalmente, los centroides actualizados:

C1' = {Puente Alto, Ingeniería Civil, Soltero, Sedan}

11,13,17

C2' = {Maipú, Arquitectura, Casado, SUV}

12,14,15,16,18

K-MODES

Ahora debemos repetir el algoritmo, es decir, evaluar las disimilitudes de todos los individuos con estos nuevos centroides

SPOILER: Quedan igual :)

K-PROTOTYPES

- k-prototypes is an algorithm to cluster **mixed-type** objects.
- It is straightforward to integrate the k-means and k-modes algorithms into the k-prototypes algorithm.
- Is practically more useful because frequently encountered objects in real world databases are mixed-type objects.
- The dissimilarity between two mixed-type objects X and Y, which are described by attributes $A_1^r, A_2^r, A_p^r, A_{p+1}^c, ..., A_m^c$ can be measured by:

$$d_2(X,Y) = \sum_{j=1}^{p}(x_j - y_j)^2 + \gamma \sum_{j=p+1}^{m} \delta(x_j, y_j)$$

where:

  - The first term is the **squared Euclidean distance** measure on the numeric attributes.
  - The second term is the **simple matching dissimilarity** measure on the categorical attributes.
  - The weight γ is used to avoid favouring either type of attribute.

K-PROTOTYPES

Individuo	Comuna	Carrera	Estado Civil	Vehículo	Altura (cm)
1	Puente Alto	Ingeniería Civil	Soltero	Sedan	173
2	Maipú	Medicina	Casado	Moto	187
3	Santiago	Ingeniería Civil	Divorciado	Sedan	183
4	Independencia	Arquitectura	Soltero	SUV	163
5	La Florida	Ingeniería Civil	Casado	SUV	184
6	Maipú	Arquitectura	Casado	SUV	192
7	Puente Alto	Medicina	Soltero	Sedan	157
8	Maipú	Arquitectura	Casado	Moto	190

a) Aplique el algoritmo de K-prototypes para segmentar los individuos en 2 clusters, indicando claramente a qué cluster pertenece cada uno y sus respectivos centroides. Para ello, considere un valor de γ de 5 y clusters iniciales:

- Cluster 1: (Comuna=Santiago, Carrera=Ingeniería Civil, Estado Civil=Divorciado, Vehículo=Sedan, Altura=150)
- Cluster 2: (Comuna=Maipú, Carrera=Arquitectura, Estado Civil=Casado, Vehículo=Moto, Altura=200)

b) Repita la pregunta anterior, pero usando $\gamma = 20$.

c) En base a las preguntas anteriores, comente: ¿Qué representa el valor de γ en el algoritmo K-prototypes?

K-PROTOTYPES

Identificamos nuestros centroides:

C1 = {Santiago, Ingeniería Civil, Divorciado, Sedan, 150}

C2 = {Maipú, Arquitectura, Casado, Moto, 200}

$y = 5$

K-PROTOTYPES

Revisamos las disimilitudes del Individuo 1 con ambos centroides:

I1 = {Puente Alto, Ingeniería Civil, Soltero, Sedan, 173}

C1 = {Santiago, Ingeniería Civil, Divorciado, Sedan, 150}

Se diferencian en y*2 atributos categóricos + 23 diferencia numérica = 33

C2 = {Maipú, Arquitectura, Casado, Moto, 200}

Se diferencian en y*4 atributos categóricos + 27 diferencia numérica = 47

Se tiene menor distancia en C1, por lo que I1 se agrupa con C1

K-PROTOTYPES

Revisamos las disimilitudes del Individuo 2 con ambos centroides:

I2 = {Maipú, Medicina, Casado, Moto, 187}

C1 = {Santiago, Ingeniería Civil, Divorciado, Sedan, 150}

Se diferencian en y*4 atributos categóricos + 37 diferencia numérica = 57

C2 = {Maipú, Arquitectura, Casado, Moto, 200}

Se diferencian en y*1 atributo categórico + 13 diferencia numérica = 18

Se tiene menor distancia en C2, por lo que I2 se agrupa con C2

K-PROTOTYPES

Y así, evaluamos para todos los individuos, donde finalmente resulta:

C1 = {Santiago, Ingeniería Civil, Divorciado, Sedan, 150}

11,13,14,17

C2 = {Maipú, Arquitectura, Casado, Moto, 200}

12,15,16,18

K-PROTOTYPES

De la misma forma, actualizamos las categóricas usando K-modes, y para las numéricas aplicamos el promedio:

C1 posee a I1,I3,I5,I7:

I1 = {Puente Alto, Ingeniería Civil, Soltero, Sedan, 173}
I3 = {Santiago, Ingeniería Civil, Divorciado, Sedan, 183}
I4 = {Independencia, Arquitectura, Soltero, SUV, 163}
I7 = {Puente Alto, Medicina, Soltero, Sedan, 157 }

Promedio simple

C1' = {Puente Alto, Ingeniería Civil, Soltero, Sedan, 169}

K-PROTOTYPES

Hacemos lo mismo para C2 con tal de calcular C2', donde finalmente resulta:

C1' = {Puente Alto, Ingeniería Civil, Soltero, Sedan, 169}

C2' = {Maipú, Arquitectura, Casado, Moto o SUV, 188.25}

11,13,14,17

12,15,16,18

K-PROTOTYPES

Ya que no tenemos un atributo libre para escoger el óptimo el C2', simplemente elegimos Moto o SUV al azar:

C2' = {Maipú, Arquitectura, Casado, Moto, 188.25}

K-PROTOTYPES

Ahora debemos repetir el algoritmo, es decir, evaluar las disimilitudes de todos los individuos con estos nuevos centroides

SPOILER: Quedan igual :)

K-PROTOTYPES

La pregunta B nos pide lo mismo, pero esta vez cambiando el valor de $y = 20$

¿Qué cambiaría esto?

Propuesto: Hacer la B y ver los cambios en las agrupaciones

K-PROTOTYPES

- A small γ value indicates that the clustering is dominated by numeric attributes while a large γ value implies that categorical attributes dominate the clustering.
- (Huang 1998) and (Huang 1997a) have suggested that the average standard deviation of numeric attributes may be used as a guidance in specifying γ, however, it isn't a general rule.
- User's knowledge about the data is important in specifying γ. If one thinks the clustering should be favoured on numeric attributes, then one needs a small γ. If one believes categorical attributes are important, then one needs a large γ.

¿PREGUNTAS?

INGENIERÍA DE LA INFORMACIÓN IN4151-2
OTOÑO 2024

AUX 8

JOSÉ SAFFIE Y JOSÉ SOZA

K'S