



MODELOS NO SUPERVISADOS

FUZZY-C-MEANS

JAIME MIRANDA

Departamento de Ingeniería Industrial
Universidad de Chile

Aplicaciones

- Marketing
 - Segmentación de clientes
 - Ofertas focalizadas

Fortalezas

- No asume ninguna distribución estadística entre los datos.
- Los resultados son mas intuitivos y fáciles de entender e interpretar

Debilidades

- Son muy sensitivos a los valores fuera de rango (outliers)
- No trabajan bien con variables categóricas



Método no supervisado

A través de la iteración de un algoritmo matemático, determina el grado de pertenencia de un objeto a una clase determinada

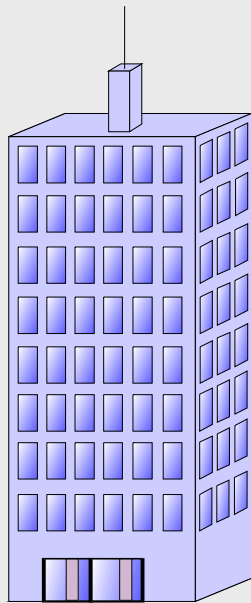
Determina centros o prototipos de clases – elemento ideal de la clase

La base de este algoritmo es la minimización de la suma de las distancias de cada objeto al centro de la clase

Se formula un problema de optimización cuya función objetivo es minimizar la suma de las distancias ponderadas por el grado de pertenencia del objeto

SEGMENTACIÓN DE CLIENTES

Bank



Product 1



Product 2



Product 3



?

?

?

?

?

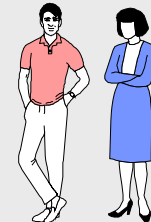
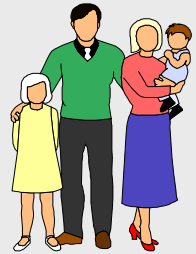
Need

Need

Need

Need

Customer



Which product for which (segment of) customers ?

VARIABLES Y PARÁMETROS

→ n = Número de objetos o muestras

→ c = Número de clases a priori

→ X_i = Vector con valores de los atributos del objeto i , con $i = (1, \dots, n)$

→ C_j = Centro de la clase j con $j = (1, \dots, c)$

→ $d^2(X_i, C_j)$ = Distancia entre el objeto i al centro de la clase j



VARIABLES Y PARÁMETROS

→ u_{ij} = Grado de pertenencia del objeto i a la clase j

→ $u_{i,j} \in [0,1]$

Matriz de pertenencia

$$\sum_{j=1}^c u_{ij} = 1$$

$$U = (u_{ij})_{ij}$$



Función Objetivo:
$$\min \sum_{i=1}^n \sum_{j=1}^c (u_{ij})^m d^2(x_i, c_j)$$

m = Parámetro de difusibilidad

PASO 1

→ Determinar una matriz $U = (u_{ij})_{ij}$ que cumpla :

$$\sum_{j=1}^c u_{ij} = 1$$

PASO 2

→ Determinar los centros de las clases con la siguiente relación:

$$c_j = \frac{\sum_{i=1}^n u_{ij}^m x_i}{\sum_{i=1}^n u_{ij}^m}$$

PASO 3

→ Actualizar los grados de pertenencia:

$$u_{ij} = \frac{1}{\sum_{k=1}^c \left(\frac{d(x_i, c_j)}{d(x_i, c_k)} \right)^{\frac{2}{m-1}}}$$

PASO 4

→ Criterio de detención:

$$\|U^{k+1} - U^k\| < \varepsilon$$

U^k = matriz _en _la _iteración _k

Número de clases

- ¿En cuántos segmentos separo mi muestra?
- “Cluster Analysis” – determinación de cluster
- Regla de los codos

Coeficiente de difusibilidad (m)

- Rango (1 , ∞)

Umbral de detención

- El error de ajuste que el modelo acepta

Valores iniciales de los centros

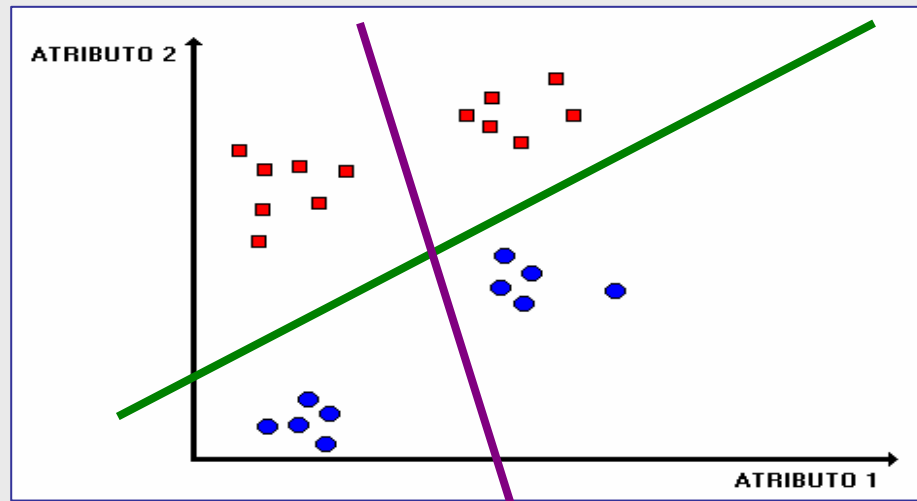


Este método a través de un criterio especial nos indica el número óptimo de cluster's de la muestra

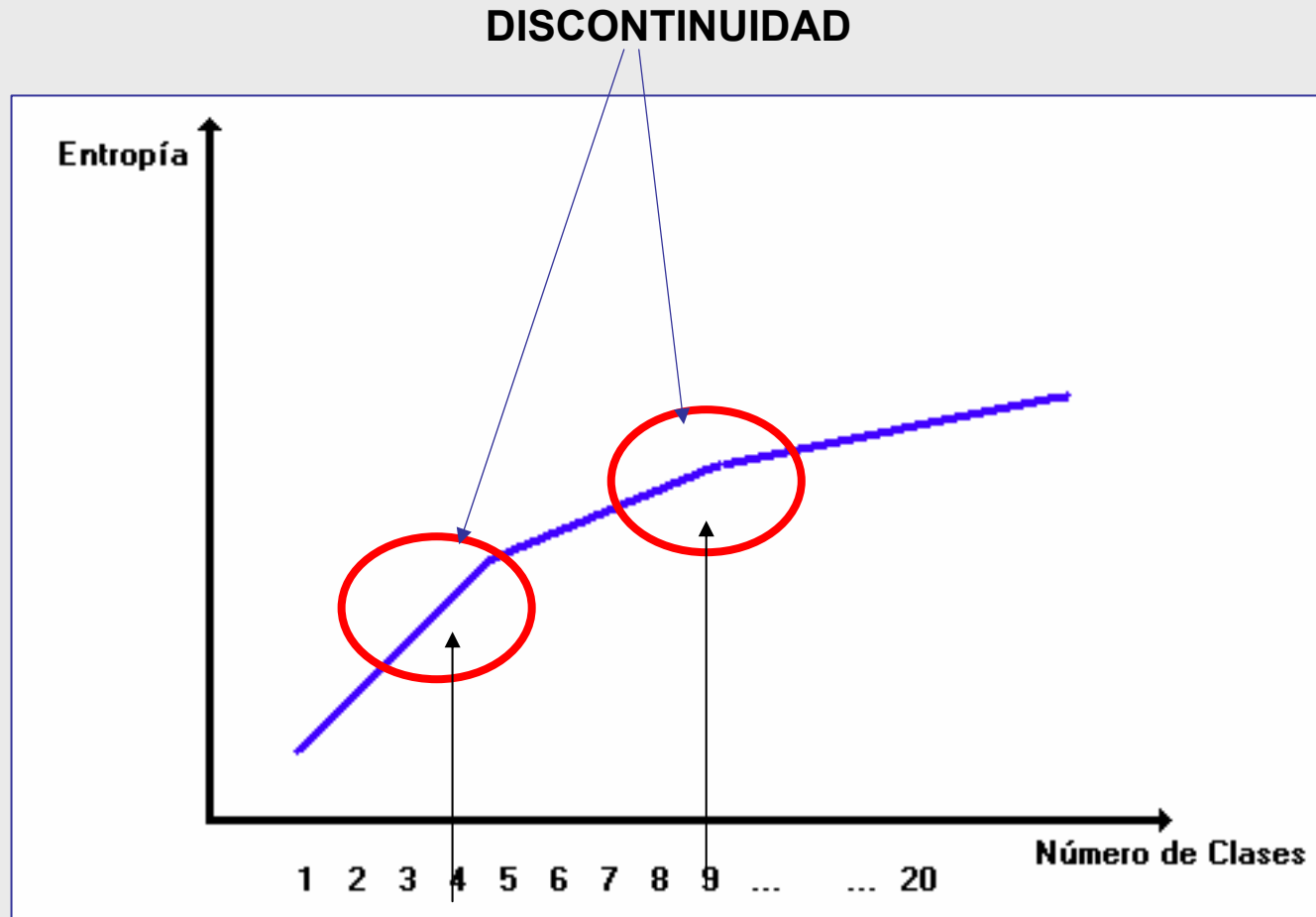
Un indicador posible: Entropía (Classification Entropy):

→ Regla de los codos

- mide la discontinuidad de la entropía respecto al numero de clases



REGLA DE LOS CODOS



PARÁMETRO DE DIFUSIBILIDAD (m)

Muestra que tan difusa será la clasificación

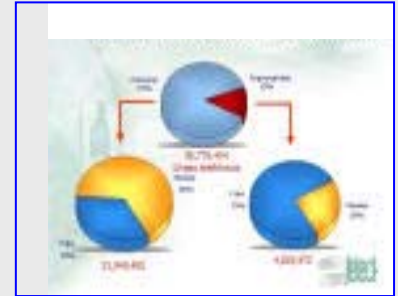
Rango permitido (1 , ∞)

Valor mínimo = 1 , mínima difusibilidad (todos distintos)

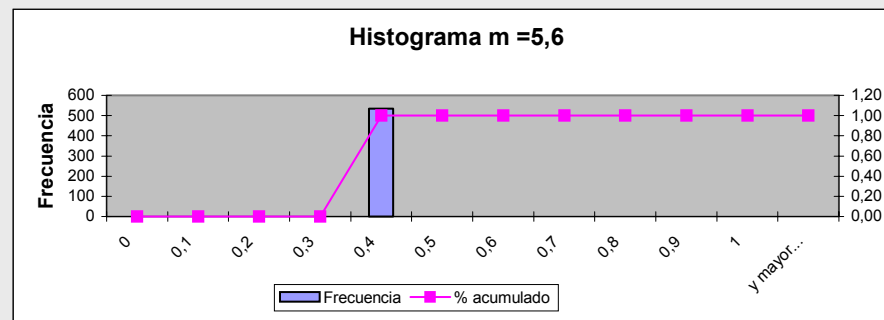
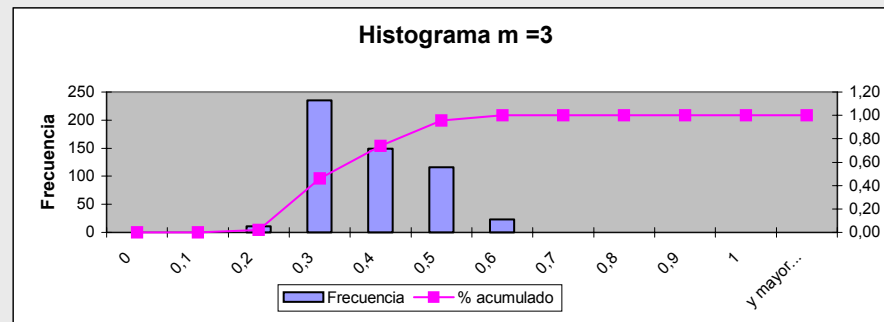
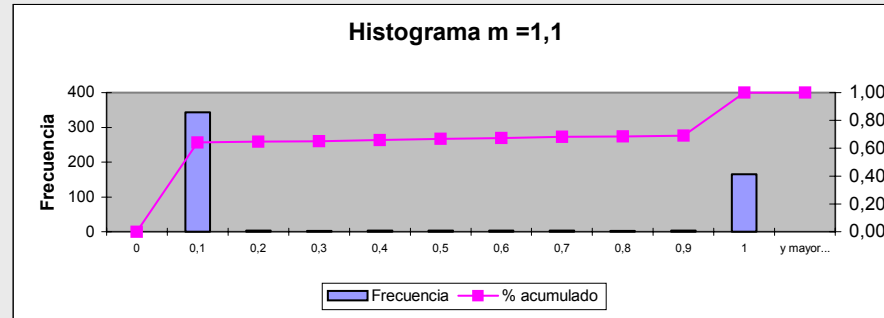
Valor máximo = ∞ máxima difundida (todos iguales)

Valores pequeños dan clasificaciones más discretas

Al aumentar este parámetro la clasificación se hace más difusa



PARÁMETRO DE DIFUSIBILIDAD (m)





MODELOS NO SUPERVISADOS

FUZZY-C-MEANS

JAIME MIRANDA

Departamento de Ingeniería Industrial
Universidad de Chile