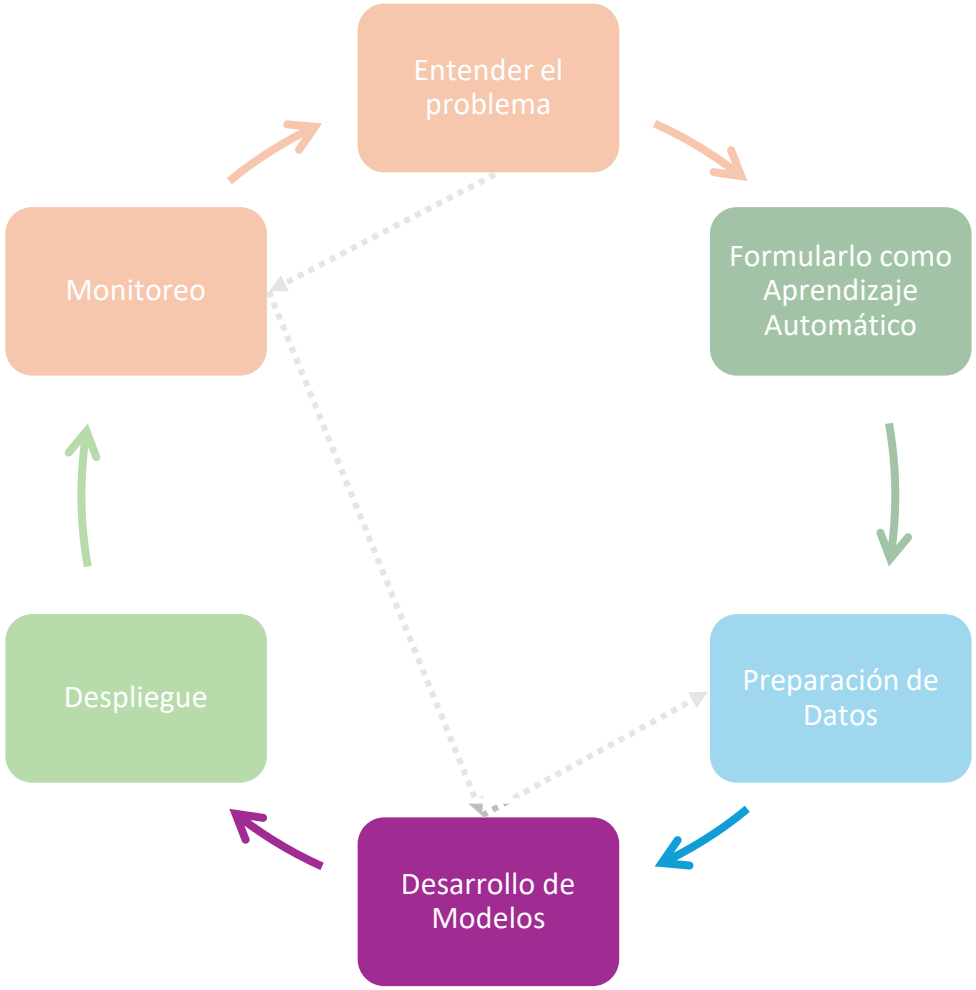
Aprendizaje No Supervisado

Algoritmos de Agrupamiento

experto en procesamiento del lenguaje natural

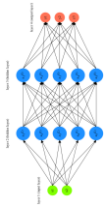
Ciclo de Vida Proyecto de Aprendizaje Automático



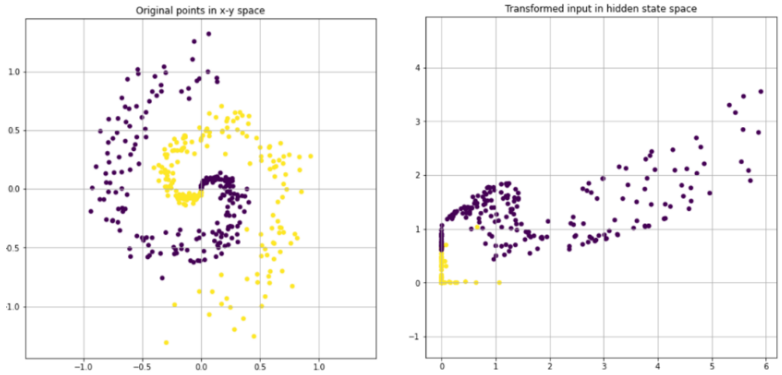
experto en procesamiento del lenguaje natural

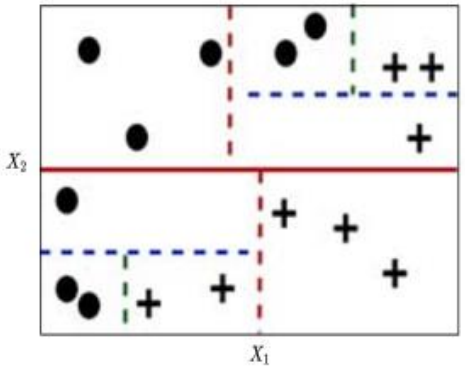
Diferentes algoritmos de Aprendizaje Supervisado

Redes de Neuronales Artificiales y Support Vector Machines que transforman los datos de entrada llevándolos a un nuevo espacio vectorial donde se resuelve el problema.



Árboles de Decisión que particionan recursivamente el espacio de los datos siguiendo algún criterio, por ejemplo, crear el mínimo número de particiones lo más homogéneas posibles.

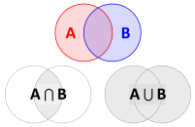




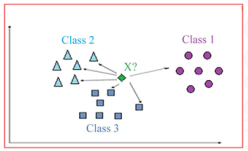
experto en procesamiento del lenguaje natural

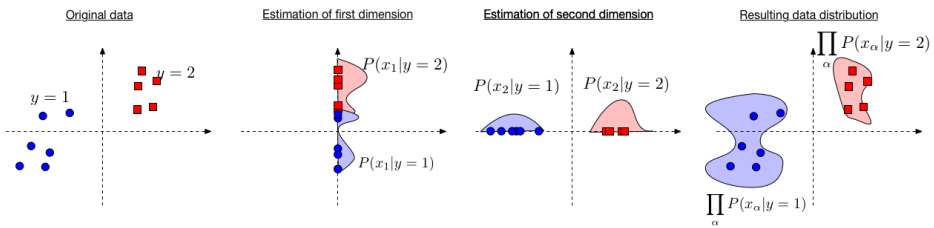
Diferentes algoritmos de Aprendizaje Supervisado

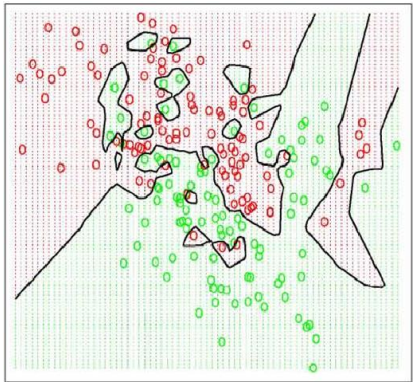
Clasificadores Bayesianos que calculan la probabilidad de cada clase dado el valor de los atributos de una instancia



Basados K-Vecinos más Cercanos: que resuelven el problema a partir de las soluciones de los ejemplos que más se parecen al problema actual.







- A diferencia del aprendizaje supervisado, en este caso no se conoce la salida esperada.

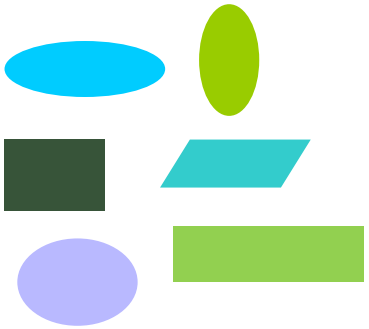
- Los algoritmos se centran en descubrir los patrones subyacentes en los datos.

experto en procesamiento del lenguaje natural

- Se muestra al sistema un conjunto de instancias

$$X = \{x_1, x_2, \dots, x_N\}$$

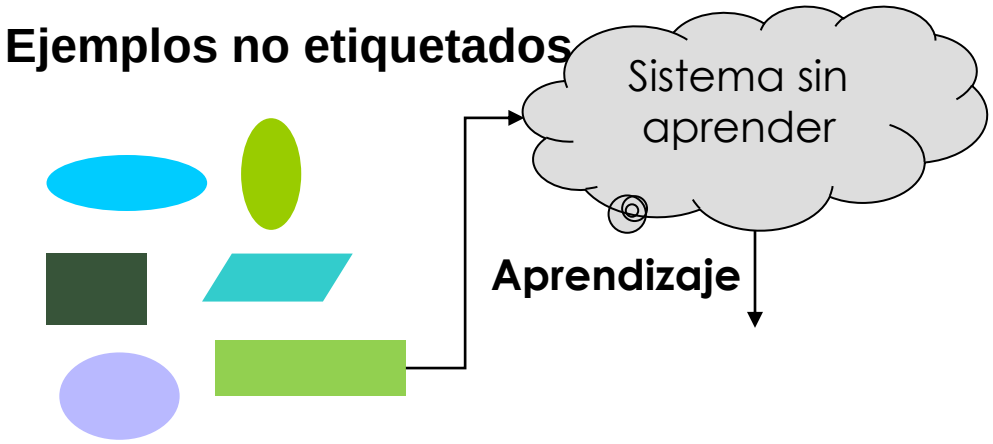
Ejemplos no etiquetados



experto en procesamiento del lenguaje natural

- Se muestra al sistema un conjunto de instancias

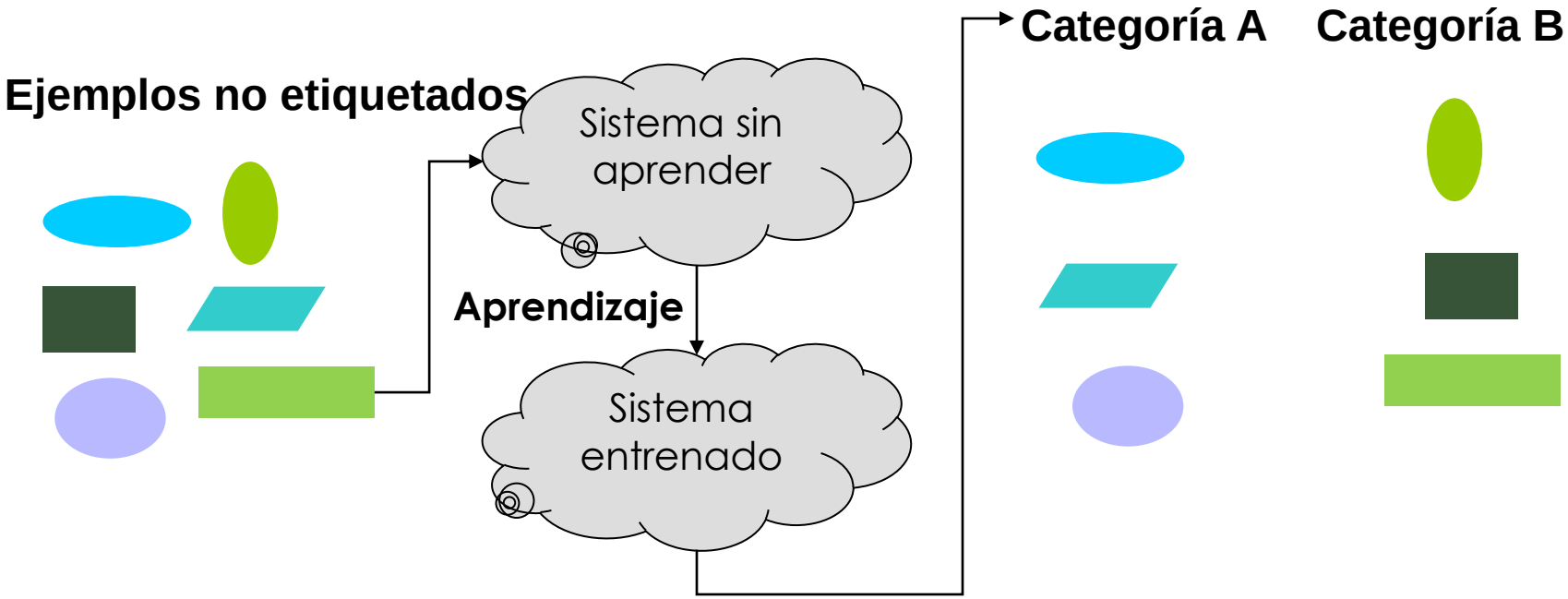
$$X = \{x_1, x_2, \dots, x_N\}$$



experto en procesamiento del lenguaje natural

- Se muestra al sistema un conjunto de instancias

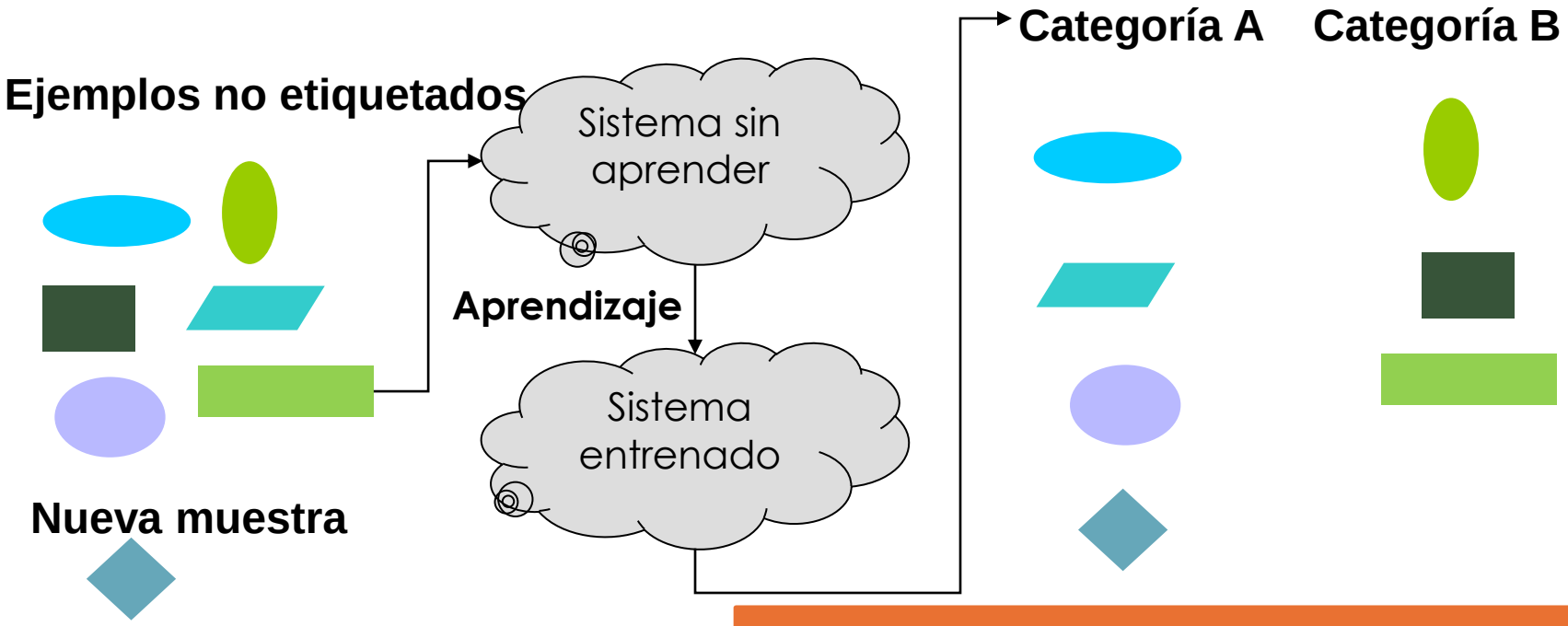
$$X = \{x_1, x_2, \ldots, x_N\}$$



experto en procesamiento del lenguaje natural

- Se muestra al sistema un conjunto de instancias

$$X = \{x_1, x_2, \ldots, x_N\}$$

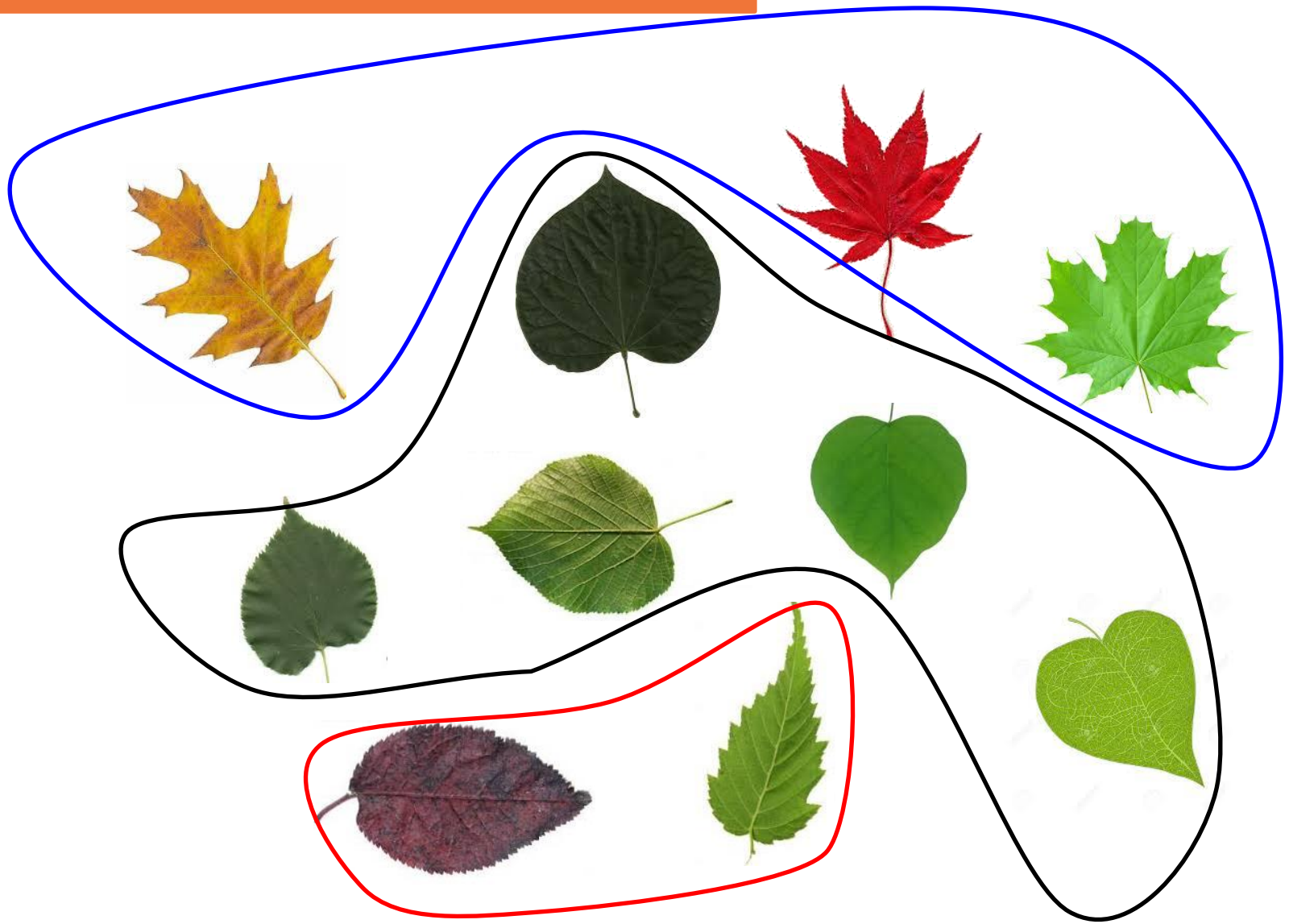


experto en procesamiento del lenguaje natural

Algoritmos de Agrupamiento

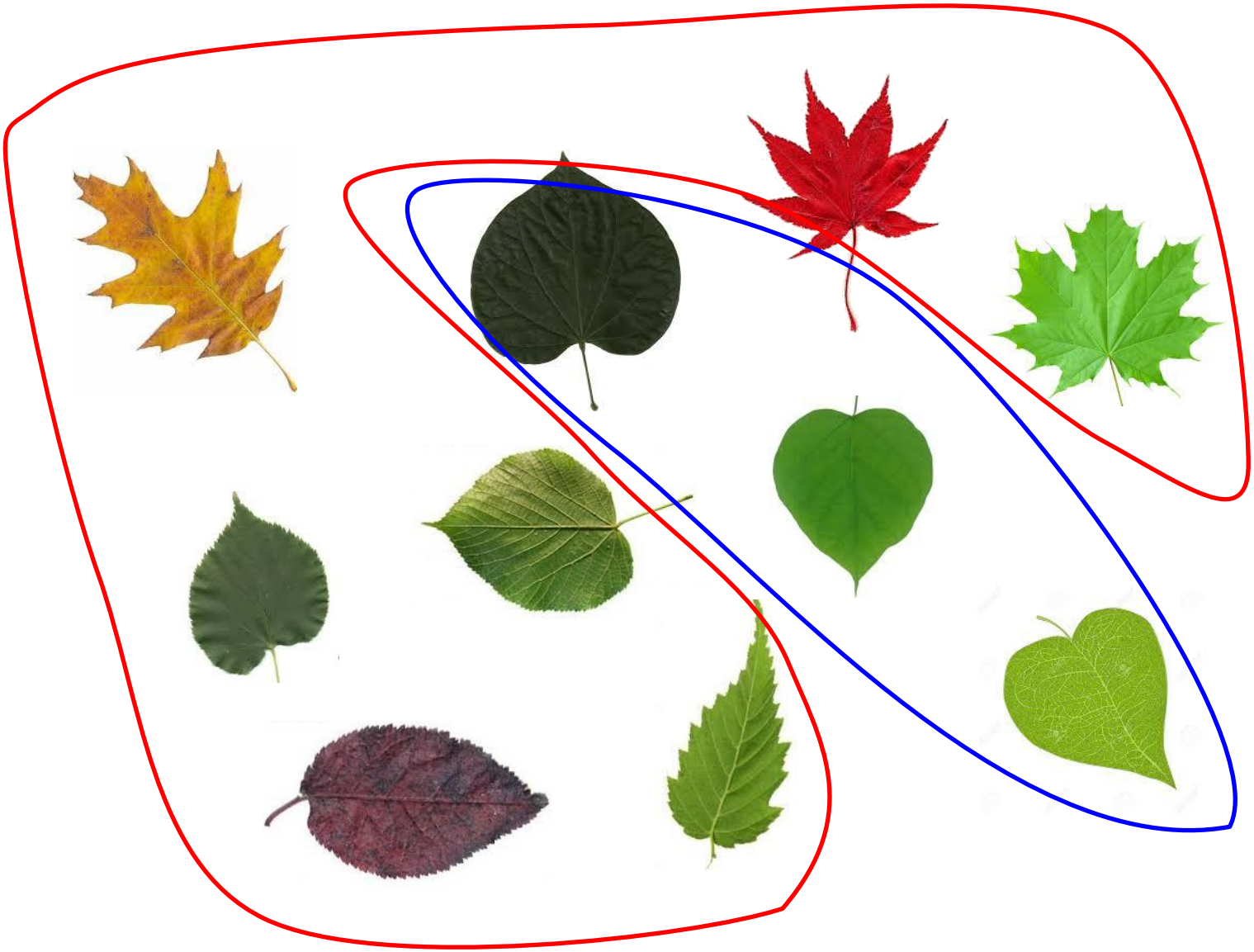
experto en procesamiento del lenguaje natural

experto en procesamiento del lenguaje natural



experto en procesamiento del lenguaje natural

Introducción



experto en procesamiento del lenguaje natural

Algunas definiciones (informales)

- Descubrir alguna estructura en una colección de datos no etiquetados.

- Proceso de agrupar un conjunto de elementos de acuerdo con cierto criterio de similitud de modo que los elementos dentro de un mismo grupo muestren una alta similitud al mismo tiempo que elementos pertenecientes a clúster diferentes sean poco similares.

experto en procesamiento del lenguaje natural

Aplicaciones:

- **Reducción de datos:** sustituir un conjunto de instancias (clúster) por su representante. Se reduce el tamaño del conjunto de entrenamiento por ejemplo al utilizar un clasificador K-NN.

- **Generación de hipótesis:** se puede inferir alguna hipótesis relativa a la naturaleza de los datos. Por ejemplo, si se observan diferencias sustanciales en una de las variables entre los clústeres, se pueden generar hipótesis de las razones.

- **Prueba de hipótesis:** se parte de una hipótesis, por ejemplo, cierta estructura de los datos. Con el análisis, se produce evidencia a favor de la hipótesis.

- **Predicción basada en grupos:** para predecir una nueva instancia, se determina el clúster al que pertenecerá con mayor probabilidad. Se realiza la predicción a partir de las instancias en el clúster.

experto en procesamiento del lenguaje natural

Definiciones

- Sea $X = \{x_1, x_2, \dots x_N\}$ un conjunto de puntos en un espacio L-dimensional, pueden describirse los clústeres como regiones continuas del espacio que contienen una densidad de puntos relativamente alta, separada de otras regiones de alta densidad por regiones de una densidad relativamente baja.

- **Clústeres No Difusos**

  Se define un K-agrupamiento (*K-clustering*) de X en K conjuntos (clústeres) $C_1, C_2, \dots C_K$ tal que se cumplen las siguientes condiciones:

  - $C_k \neq \emptyset, i = 1, \dots K$
  - $\bigcup_{k=1}^{K} C_k = X$
  - $C_i \cap C_i = \emptyset, i \neq j, i, j = 1, \dots, K$

experto en procesamiento del lenguaje natural

Definiciones

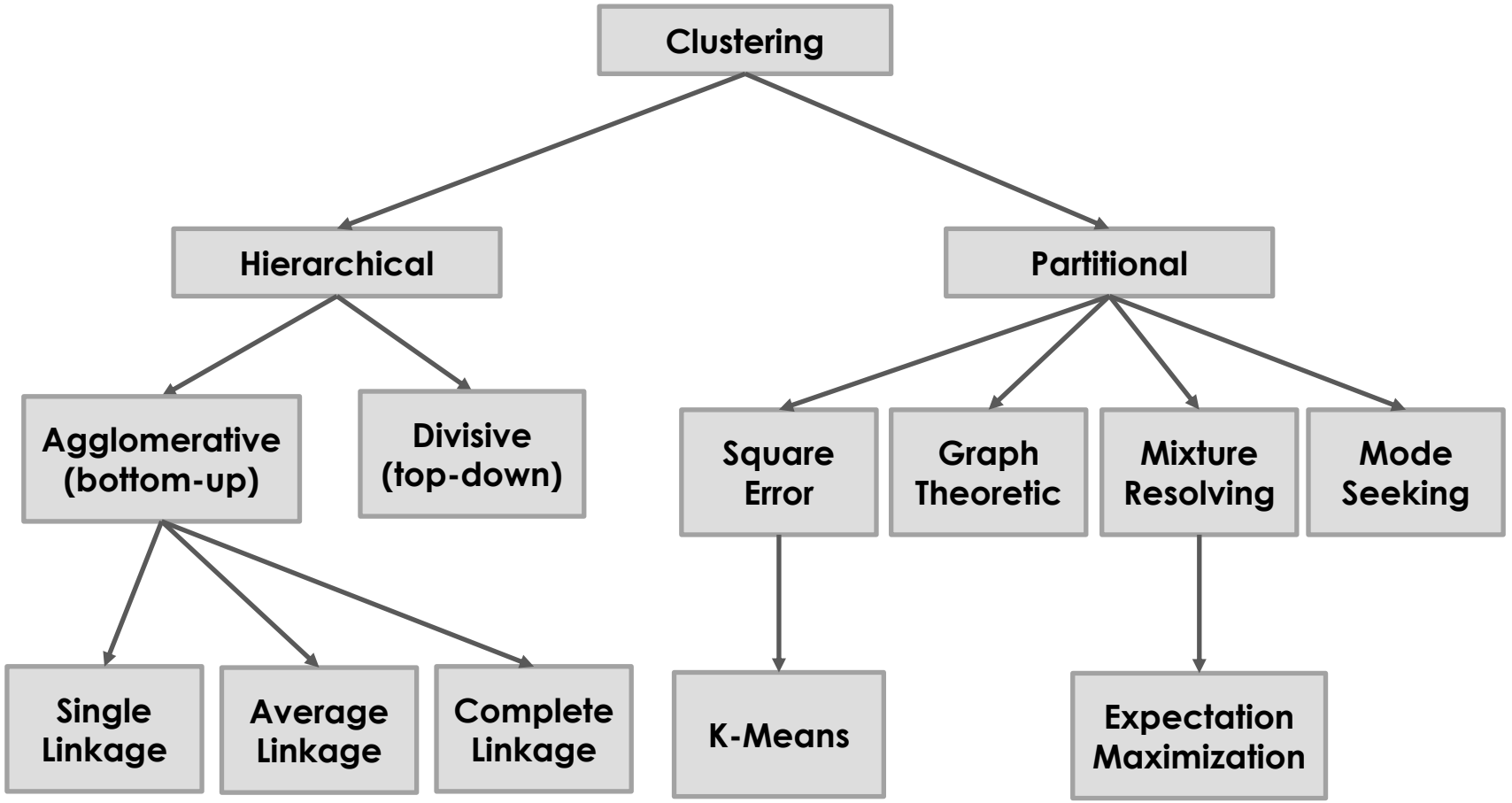
- Sea $X = \{x_1, x_2, \dots x_N\}$ un conjunto de puntos en un espacio L-dimensional.

- **Clústeres Difusos**

  Se define un K-agrupamiento difuso (*K-clustering*) de X mediant K funciones$u_1, u_2, \dots u_K$ tales que:

  - $u_k : X \to [0,1],\ k = 1, \dots, K$
  - $\sum_{k=1}^{K} u_k(x_n) = 1,\ n = 1,2, \dots, N$
  - $0 < \sum_{n=1}^{N} u_k(x_n) < N$

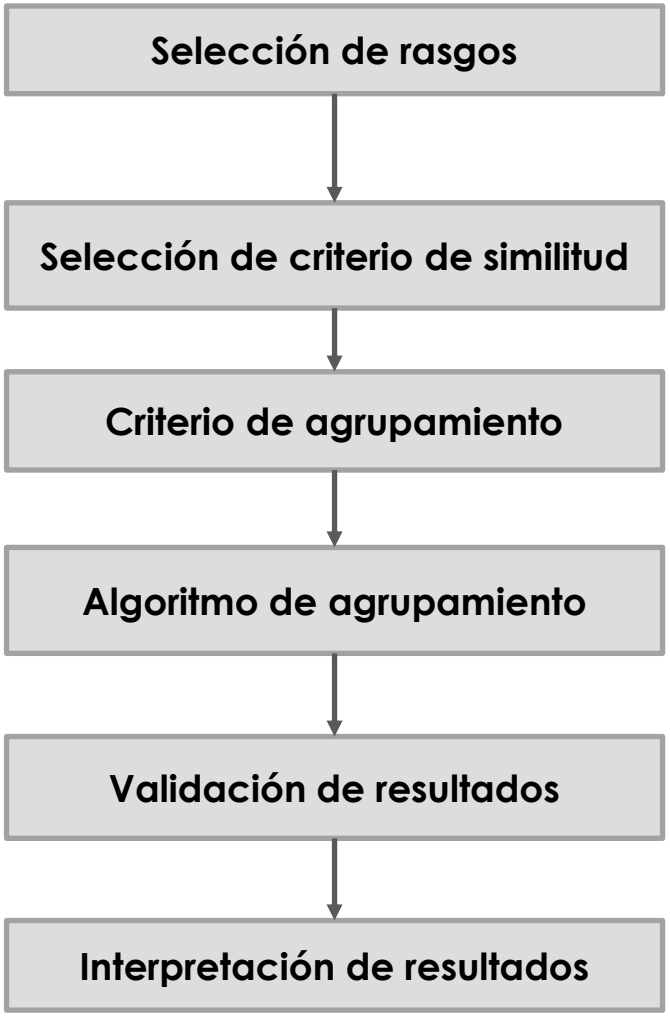
experto en procesamiento del lenguaje natural



Taxonomía de algoritmos de agrupamiento.

Para otra clasificación más exhaustiva, ver Theodoridis, S. & Koutroumbas, K. (2006), *Pattern Recognition. 4th Edition*, sec. 12.2

experto en procesamiento del lenguaje natural



Metodología general para aplicar algoritmos de agrupamiento.

experto en procesamiento del lenguaje natural

Tipos de datos

- **Nominales:** típicamente describen diferentes estados, por ejemplo, el tipo sanguíneo A, B, AB, 0. En general no tiene sentido realizar comparaciones cuantitativas.

- **Ordinales:** se puede establecer un ordenamiento con contenido semántico, por ejemplo, las notas del curso.

- **"Interval-scaled":** existe un orden y la diferencia entre dos valores tiene un significado mientras su proporción puede no tenerlo. Ejemplo la temperatura, ¿qué significa o qué información aporta al problema decir que un sitio es el doble de cálido que otro?, ¿o que en sitio hay 10 grados más que en otro?

- **"Ratio-scaled":** en este caso la proporción tiene significado mientras que su diferencia puede que no. Por ejemplo, la concentración de un producto químico.

experto en procesamiento del lenguaje natural

Medidas de (Di)similitud

- **Propiedades** (usuales) **de una medida**

Sean dos instancias x_i, x_j, d_{ij} una medida de (di)similitud entre estas:

- $d_{ij} \geq 0$
- $d_{ii} = 0$
- $d_{ij} = d_{ji}$

Si además $d_{ij} \leq d_{ik} + d_{kj} \; \forall \, i,k,j$ entonces d_{ij} es una métrica

Se deben atender diferentes casos:

- datos discretos
- datos continuos
- datos heterogéneos
- para grupos de instancias

experto en procesamiento del lenguaje natural

Medidas de (Di)similitud. Datos Discretos

Sean dos ejemplos x_i, x_j donde cada atributo es binario:

a: número de atributos que son 1 en ambas instancias

b: número de atributos que son 1 en x_i y 0 en x_j

c: número de atributos que son 0 en x_i y 1 en x_j

d: número de atributos que son 0 en ambas instancias

		j	
		1	0
i	1	a	b
	0	b	d

- **Simple Matching Coefficient**

$$d_{ij} = \frac{b+c}{a+b+c+d}$$

- **Jaccard´s Coefficient**

$$d_{ij} = \frac{b+c}{a+b+c}$$

experto en procesamiento del lenguaje natural

Medidas de (Di)similitud. Datos Discretos

Sean dos ejemplos x_i, x_j donde cada atributo es binario:

a: número de atributos que son 1 en ambas instancias

b: número de atributos que son 1 en x_i y 0 en x_j

c: número de atributos que son 0 en x_i y 1 en x_j

d: número de atributos que son 0 en ambas instancias

		j	
		1	0
i	1	a	b
	0	b	d

- **Simple Matching Coefficient**

$$d_{ij} = \frac{b+c}{a+b+c+d}$$

- **Jaccard´s Coefficient**

$$d_{ij} = \frac{b+c}{a+b+c}$$

Caso más general

$$d_{ij} = \frac{b+c}{\alpha a+b+c+\delta d}$$

experto en procesamiento del lenguaje natural

Se define en base al coeficiente de macheo δ_{jm}^{l}

$$\delta_{jm}^{l} = \begin{cases} 1, j \neq m \\ 0, j = m \end{cases}$$

donde j, m son los valores del i–$ésimo$ atributo, en el caso de rasgos nominales

Para atributos ordinales discretos debe cumplir:

$\delta_{jm}^{l} < \delta_{jr}^{l}$ si j $> m > r$ o j $< m < r$

- **Generalized Discrete Coefficient**

$d_{ij} = \frac{1}{L}\sum_{l=1}^{L} \delta_{jm}^{l}$ donde L es el número de atributos discretos.

experto en procesamiento del lenguaje natural

Medidas de (Di)similitud. Datos Continuos

Sea L el número de atributos, v_n^l el valor del l-ésimo atributo continuo de la instancia n, y w_l el peso con el que se ponderará el atributo l

- **Minkowski**

$$d_{ij} = \left(\sum_{l=1}^{L}(w^l)^\lambda |v_i^l - v_j^l|^\lambda\right)^{\frac{1}{\lambda}},\ \lambda \geq 1$$

- **Manhattan o City block**

$$d_{ij} = \left(\sum_{l=1}^{L} w_l |v_i^l - v_j^l|\right)$$

- **Euclideana**

$$d_{ij} = \left(\sum_{l=1}^{L}(w^l)^2 |v_i^l - v_j^l|^2\right)^{\frac{1}{2}}$$

- **Canberra**

$$d_{ij} = \begin{cases} \mathbf{0} & v_i^l = \mathbf{0} = v_j^l \\ \sum_{l=1}^{L} \dfrac{|v_i^l - v_j^l|}{|v_i^l| + |v_j^l|} & \boldsymbol{e.o.c} \end{cases}$$

Observar que la distancia de Manhattan y Euclideana son casos particulares de Minkowski con $\lambda = 1$ y $\lambda = 2$ respectivamente.

experto en procesamiento del lenguaje natural

Medidas de (Di)similitud. Datos Continuos

- **Angular Separation**

$$d_{ij} = \frac{1 - \phi_{ij}}{2}$$

$$\phi_{ij} = \frac{\sum_{l=1}^{L} v_i^l v_j^l}{\sqrt{\sum_{l=1}^{L} (v_i^l)^2 \sum_{l=1}^{L} (v_j^l)^2}}$$

- **Coeficiente de Correlación**

$$d_{ij} = \frac{1 - \phi_{ij}}{2}$$

$$\bar{v}_i = \frac{1}{L} \sum_{l=1}^{L} v_i^l$$

$$\phi_{ij} = \frac{\sum_{l=1}^{L} (v_i^l - \bar{v}_i)(v_j^l - \bar{v}_j)}{\sqrt{\sum_{l=1}^{L} (v_i^l - \bar{v}_i)^2 \sum_{l=1}^{L} (v_j^l - \bar{v}_j)^2}}$$

experto en procesamiento del lenguaje natural

Medidas de (Di)similitud. Grupos de Instancias

- **Disimilitud Vecino Más Cercano**

$d(C_h, C_g) = \min d_{ij}, i \in C_h, j \in C_g$

Disimilitud Vecino Más Lejano

$d(C_h, C_g) = \max d_{ij}, i \in C_h, j \in C_g$

- **Euclidiana**

$d(C_h, C_g) = \left(\sum_{l=1L} \left|\bar{v}_h^l - \bar{v}_g^l\right|^2\right)^{\frac{1}{2}}$, $\bar{v}_h^l, \bar{v}_g^l$ valor del l-ésimo atributo del centroide del grupo C_h, C_g respectivamente.

- **Mahalanobis**

$d(C_h, C_g) = (\bar{v}_h - \bar{v}_g)^T W^{-1} (\bar{v}_h - \bar{v}_g)^T$ donde W^{-1}es la matriz de covarianza intra-clúster para ambos grupos.

experto en procesamiento del lenguaje natural

Medidas de (Di)similitud. ¿Cuál emplear?

- ¡No hay una regla absoluta, depende de cada problema en particular!

- Tener en cuenta:
  - Naturaleza de los datos.
  - Rango de valores de los atributos.
  - Estandarización de atributos.
  - Algoritmo de agrupamiento que se utilizará.
  - Pesos de cada atributo.
  - Instancias con valores faltantes para algún atributo.

experto en procesamiento del lenguaje natural

Instancia Representativa

Sea $C = \{x_1, x_2, \dots x_N\}$ un clúster, algunas alternativas para elegir una instancia representativa del clúster son:

- **Punto medio**:

$m = \frac{1}{N}\sum_{x_i \in C} x_i$ de modo que $v_m^l = \frac{1}{N}\sum_{n=1}^{N} v_n^l$

- **Centro medio:**

Definido como $x_m \in C$ tal que $\sum_{n=0}^{N} d_{mn} \leq \sum_{n=0}^{N} d_{jm}$ $\forall j \in C$

- **Mediana**

Definida como $x_m \in C$ tal que $med(d_{mn}) \leq med(d_{jn})$ $\forall j \in C$.

$med(T)$ es la mediana, es decir, el mínimo número en T que es mayor que $\frac{|T|}{2}$ elementos de T.

experto en procesamiento del lenguaje natural

Algoritmos de Particionamiento

experto en procesamiento del lenguaje natural

Consideraciones

- Construyen una partición del conjunto en un número específico de clústeres maximizando o minimizando un determinado criterio.

- Número exponencial de posibles particiones, para N objetos y K grupos se tienen $\left\{ {N \atop K} \right\} = \frac{1}{K!}\sum_{i=0}^{K}(-1)^{K-i}\binom{K}{i}^{i^N} = \sum_{i=0}^{K}\frac{(-1)^{K-i}j^N}{(K-i)!\,i!}$

Criterios de particionamiento

- Densidad de los clústeres (*compactness*): qué tan similares son las instancias dentro de un mismo grupo.

- Grado de aislamiento (*isolation*) : qué tan separado está el clúster de otros o del resto de los ejemplos.

¡Depende del tipo de datos!

experto en procesamiento del lenguaje natural

Criterios de particionamiento

Sea $h(C)$ una medida de la densidad o aislamiento de un clúster y $C = \{C_1, C_2, \ldots C_K\}$ una posible partición, algunos posibles criterios de particionamiento son:

- $c(N, K) = \sum_{k=1K} \frac{h(C_k)}{K}$
- $c(N, K) = \max h(C_k)$
- $c(N, K) = \min h(C_k)$

experto en procesamiento del lenguaje natural

Criterios de particionamiento

Datos continuos

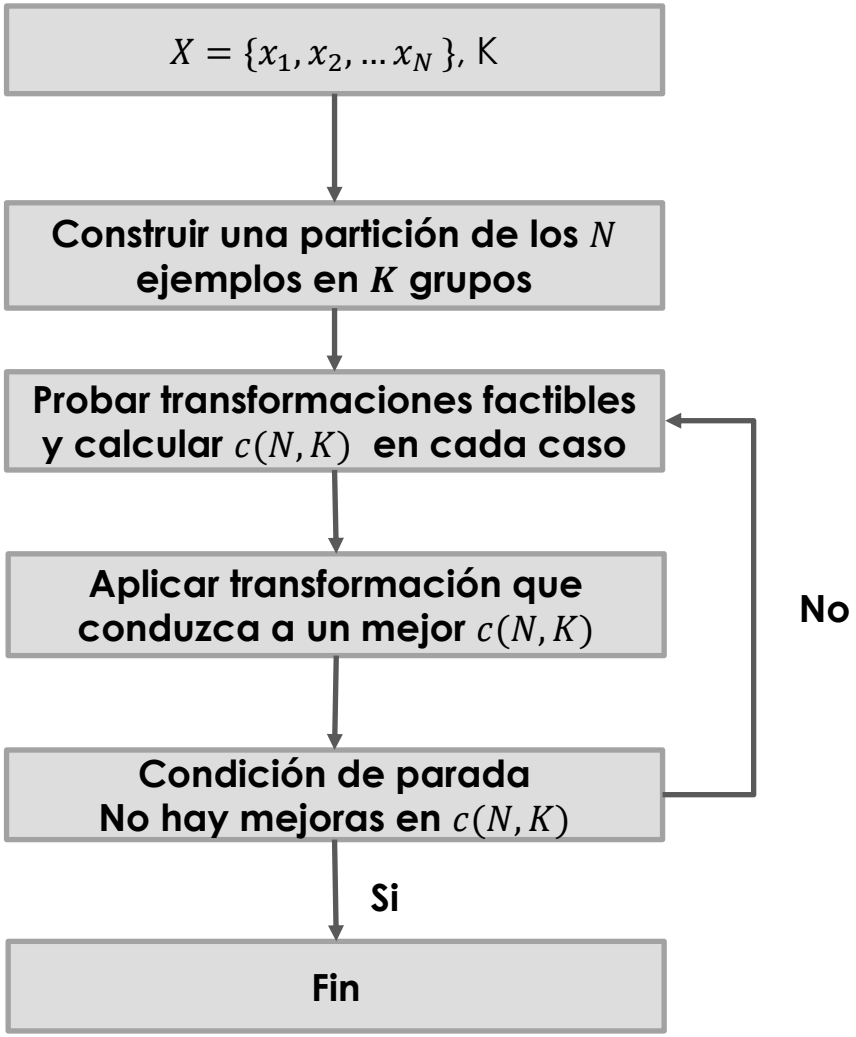
- **Suma de cuadrados** (medida de densidad) $h(C) = \sum_{x_n \in C} \sum_{l=1}^{L} \left(v_n^l - \bar{v}_C^l\right)^2$ donde $\bar{v}_C^l = \frac{1}{|C|} \sum_{x_n \in C} v_n^l$
- $\boldsymbol{L_1}$ (medida de densidad) $h(C) = \sum_{x_n \in C} \sum_{l=1}^{L} \left| v_n^l - \bar{m}_C^l \right|$ donde $\bar{m}_C^l = \text{median}(v_n^l)$

A partir de la matriz de disimilaridad

- **Diámetro** (medida de densidad) $h(C) = \max d_{ij} \;\; x_i, x_j \in C$
- **Star** (medida de densidad) $h(C) = min_{x_i} \sum d_{ij}, x_i, x_j \in C$
- **Suma de Distancias** (medida de densidad) $h(C) = \sum \;\; d_{ij}\, x_i, x_j \in C, j < i$
- **Split** (medida de aislamiento) $h(C) = min_{x_i} d_{ij} \;\; x_i \in C, x_j \notin C$
- **Cut** (medida de aislamiento). $h(C) = \sum_{x_i \in C} \sum_{x_j \notin C} d_{ij}$

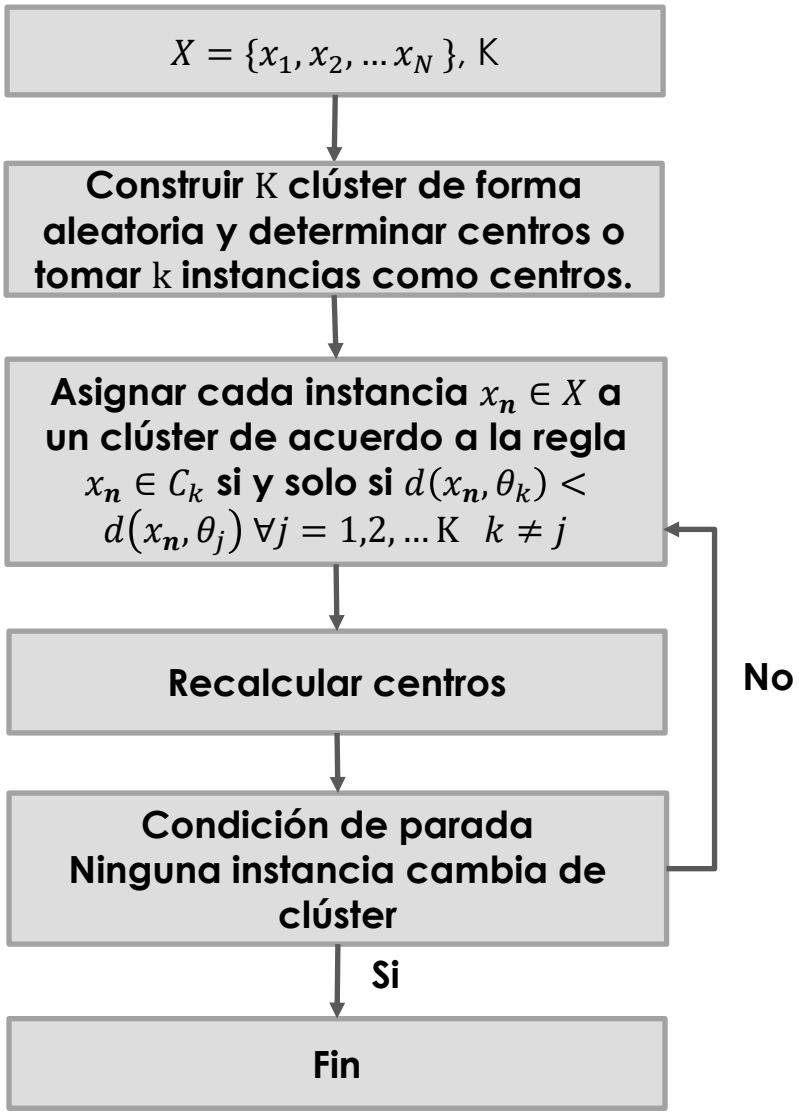
experto en procesamiento del lenguaje natural

Esquema básico



experto en procesamiento del lenguaje natural

K-Medias



- $c(N,K) = \sum_{n=1}^{N} \sum_{k=1}^{K} u_{nk} |x_n - \theta_k|^2$
- θ_k : instancia representativa del clúster C_k
- $u_{nk} = \begin{cases} 1 & x_n \in C_k \\ 0 & x_n \notin C_k \end{cases}$

experto en procesamiento del lenguaje natural

K-Medias

$X = \{x_1, x_2, \ldots x_N\}$, K

Construir K clúster de forma aleatoria y determinar centros o tomar k instancias como centros.

Asignar cada instancia $x_n \in X$ a un clúster de acuerdo a la regla $x_n \in C_k$ si y solo si $d(x_n, \theta_k) < d(x_n, \theta_j)\ \forall j = 1,2, \ldots \mathrm{K} \quad k \neq j$

Recalcular centros

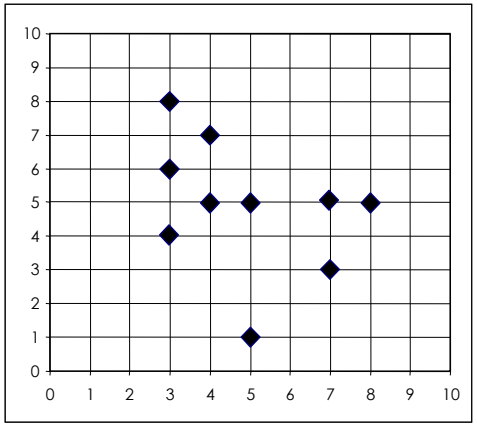
Condición de parada
Ninguna instancia cambia de clúster

No

Si

Fin

a	b	c	d	e	f	g	h	i	j
(3,4)	(3,6)	(3,8)	(4,5)	(4,7)	(5,1)	(5,5)	(7,3)	(7,5)	(8,5)



experto en procesamiento del lenguaje natural

K-Medias

$X = \{x_1, x_2, \ldots x_N\}$, K

Construir K clúster de forma aleatoria y determinar centros o tomar k instancias como centros.

Asignar cada instancia $x_n \in X$ a un clúster de acuerdo a la regla $x_n \in C_k$ si y solo si $d(x_n, \theta_k) < d(x_n, \theta_j)\ \forall j = 1,2, \ldots K \quad k \neq j$

Recalcular centros

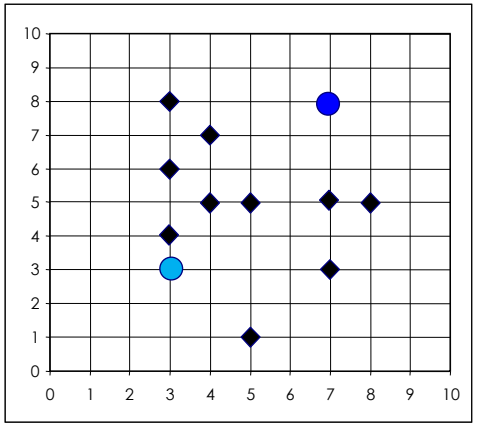
Condición de parada
Ninguna instancia cambia de clúster

No

Si

Fin

	a	b	c	d	e	f	g	h	i	j
	(3,4)	(3,6)	(3,8)	(4,5)	(4,7)	(5,1)	(5,5)	(7,3)	(7,5)	(8,5)
C_1=(3,3)										
C_2=(7,8)										



experto en procesamiento del lenguaje natural

K-Medias

$X = \{x_1, x_2, \ldots x_N\}$, K

Construir K clúster de forma aleatoria y determinar centros o tomar k instancias como centros.

Asignar cada instancia $x_n \in X$ a un clúster de acuerdo a la regla $x_n \in C_k$ si y solo si $d(x_n, \theta_k) < d(x_n, \theta_j)\ \forall j = 1,2, \ldots K \quad k \neq j$

Recalcular centros

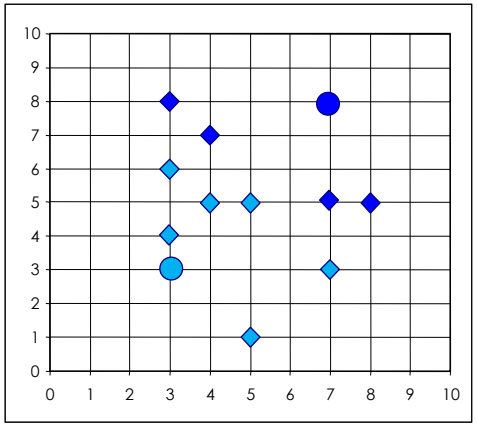
Condición de parada
Ninguna instancia cambia de clúster

No

Si

Fin

	a	b	c	d	e	f	g	h	i	j
	(3,4)	(3,6)	(3,8)	(4,5)	(4,7)	(5,1)	(5,5)	(7,3)	(7,5)	(8,5)
C_1=(3,3)	1.0	3.0	5.0	2.2	4.1	2.8	2.8	4.0	4.5	5.4
C_2=(7,8)	5.7	4.5	4.0	4.2	3.2	7.3	3.6	5.0	3.0	3.2



experto en procesamiento del lenguaje natural

K-Medias

$X = \{x_1, x_2, \ldots x_N\}$, K

Construir K clúster de forma aleatoria y determinar centros o tomar k instancias como centros.

Asignar cada instancia $x_n \in X$ a un clúster de acuerdo a la regla $x_n \in C_k$ si y solo si $d(x_n, \theta_k) < d(x_n, \theta_j)\ \forall j = 1,2, \ldots K \quad k \neq j$

Recalcular centros

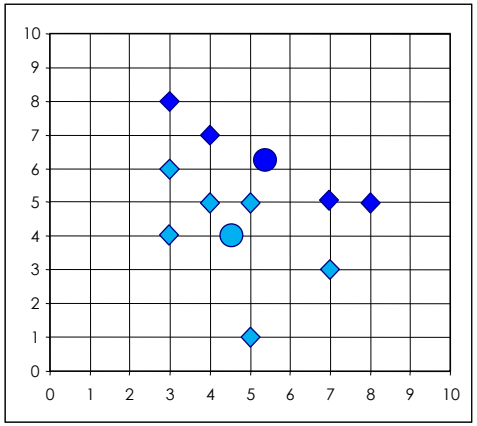
Condición de parada
Ninguna instancia cambia de clúster

No

Si

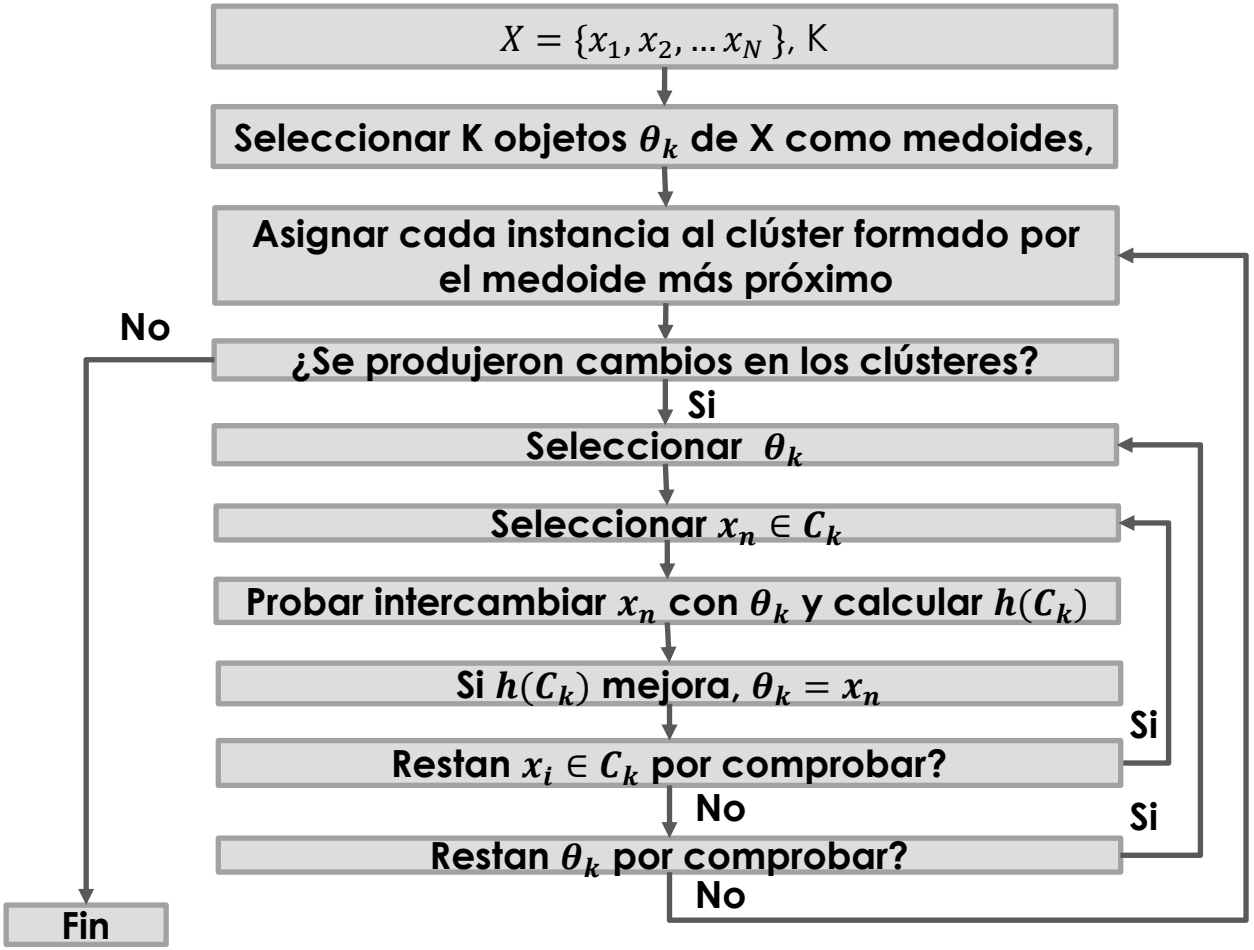
Fin

	a	b	c	d	e	f	g	h	i	j
	(3,4)	(3,6)	(3,8)	(4,5)	(4,7)	(5,1)	(5,5)	(7,3)	(7,5)	(8,5)
C_1=(4.5,4)	1.0	3.0	5.0	2.2	4.1	2.8	2.8	4.0	4.5	5.4
C_2=(5.5,6.3)	5.7	4.5	4.0	4.2	3.2	7.3	3.6	5.0	3.0	3.2



experto en procesamiento del lenguaje natural

Partition Around Medoids (PAM)

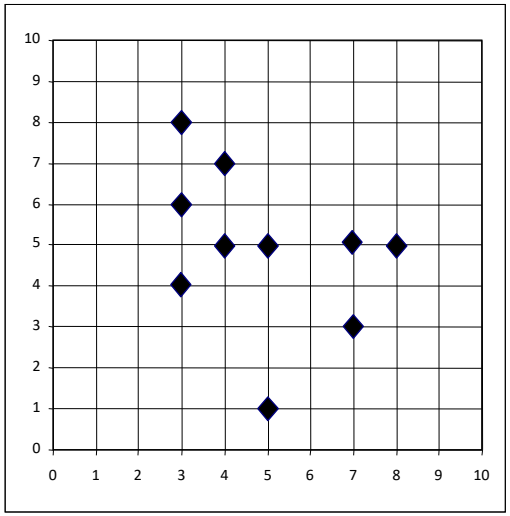


experto en procesamiento del lenguaje natural

PAM (Partition Around Medoids)

Datos los siguientes datos de ejemplo, realizar tres iteraciones del algoritmo PAM.

a	b	c	d	e	f	g	h	i	j
(3,4)	(3,6)	(3,8)	(4,5)	(4,7)	(5,1)	(5,5)	(7,3)	(7,5)	(8,5)



experto en procesamiento del lenguaje natural

Expectation Maximization (EM)

Enfoques basados en medidas de (di)similitud

- Se calcula la (di)similitud entre pares de instancias, agrupando las instancias similares en un clúster.

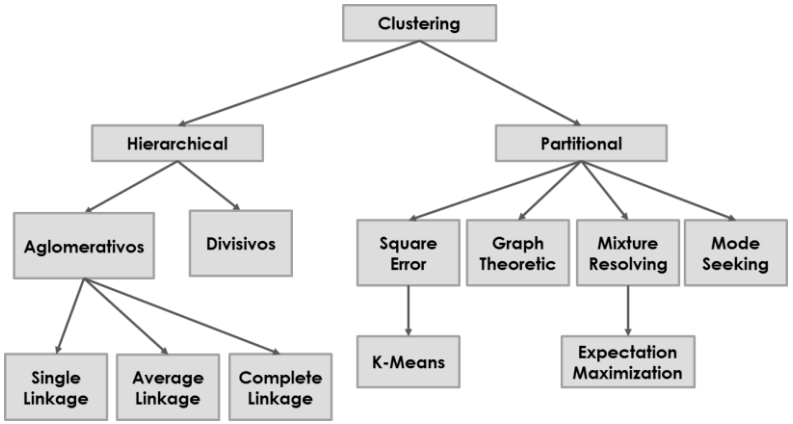
Enfoques generativos o basados en modelos

- Intentan aprender modelos capaces de generar los datos, cada modelo representa un clúster en particular

- En general el número de clústeres se define con antelación.

- El tipo modelo se especifica a priori, ej. Gaussiano.

- El aprendizaje consiste en determinar los parámetros del modelo.



experto en procesamiento del lenguaje natural

Expectation Maximization (EM)

Algoritmo K-Medias

- Cada instancia se asigna a exactamente un clúster.
- ¿Qué ocurre si los clústeres se solapan?
- Difícil de decir cuál es el clúster correcto para cada instancia.
- Emplea distancia Euclidiana.
- ¿Qué ocurre si los clústeres no son circulares?

Un enfoque alternativo es utilizar **Mezclas de Gaussianas**

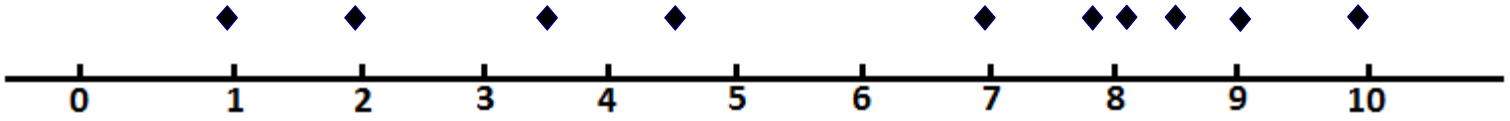
- Los clústeres se modelan como Gaussianas, no solamente por su media.
- Algoritmo EM, asigna cada instancia a un clúster con cierta probabilidad.
- Obtiene un modelo probabilístico de los datos.

experto en procesamiento del lenguaje natural

Expectation Maximization (EM)

Sean $X = \{x_1, x_2, \ldots x_N\}$ $K = 2$, y $N_A(\mu_A, \sigma_A^2)$, $N_B(\mu_B, \sigma_B^2)$ dos distribuciones Gaussianas (normales) con media μ_A, μ_B y varianza σ_A^2, σ_B^2 respectivamente.

¿Cómo estimar μ_A, μ_B y σ_A^2, σ_B^2 de modo que describan lo mejor posible los datos observados?



experto en procesamiento del lenguaje natural

Expectation Maximization (EM)

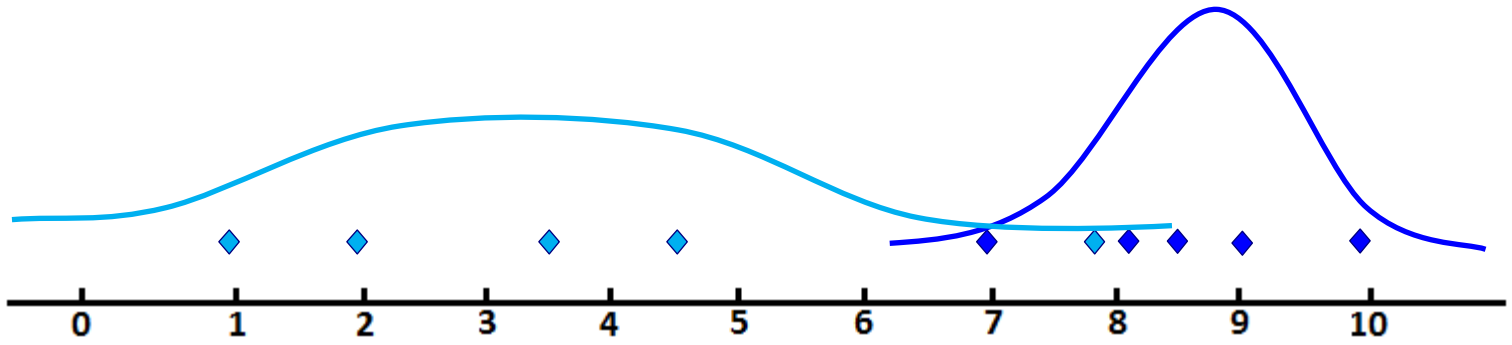
Sean $X = \{x_1, x_2, \ldots x_N\}$ $K = 2$, y $N_A(\mu_A, \sigma_A^2)$, $N_B(\mu_B, \sigma_B^2)$ dos distribuciones Gaussianas (normales) con media μ_A, μ_B y varianza σ_A^2, σ_B^2 respectivamente.

¿Cómo estimar μ_A, μ_B y σ_A^2, σ_B^2 de modo que describan lo mejor posible los datos observados?

!Trivial si se conoce la fuente de cada observación!

$$\mu_A = \frac{x_1 + x_2 + x_3 + x_4 + x_6}{5}$$

$$\mu_B = \frac{x_5 + x_7 + x_8 + x_9 + x_{10}}{5}$$



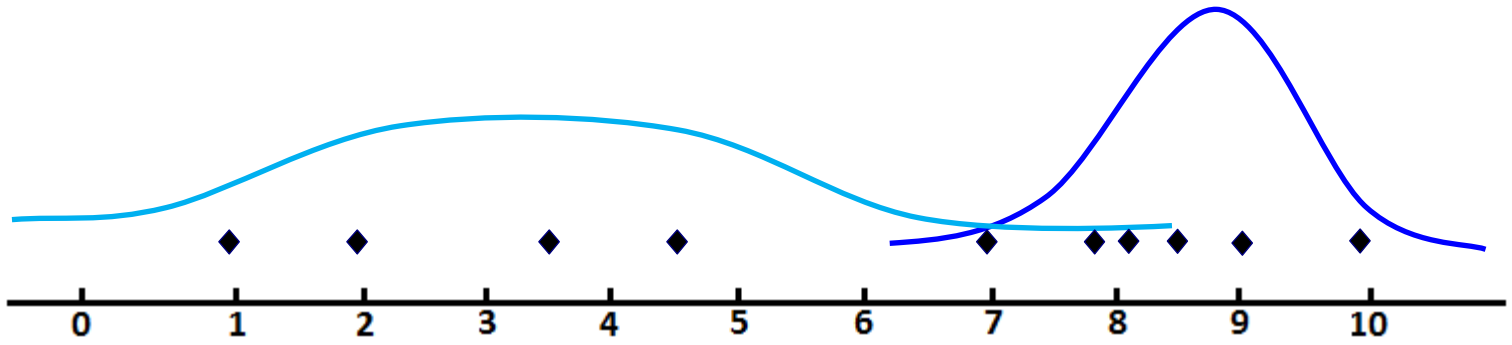
experto en procesamiento del lenguaje natural

Expectation Maximization (EM)

Sean $X = \{x_1, x_2, \ldots x_N\}$ $K = 2$, y $N_A(\mu_A, \sigma_A^2)$, $N_B(\mu_B, \sigma_B^2)$ dos distribuciones Gaussianas (normales) con media μ_A, μ_B y varianza σ_A^2, σ_B^2 respectivamente.

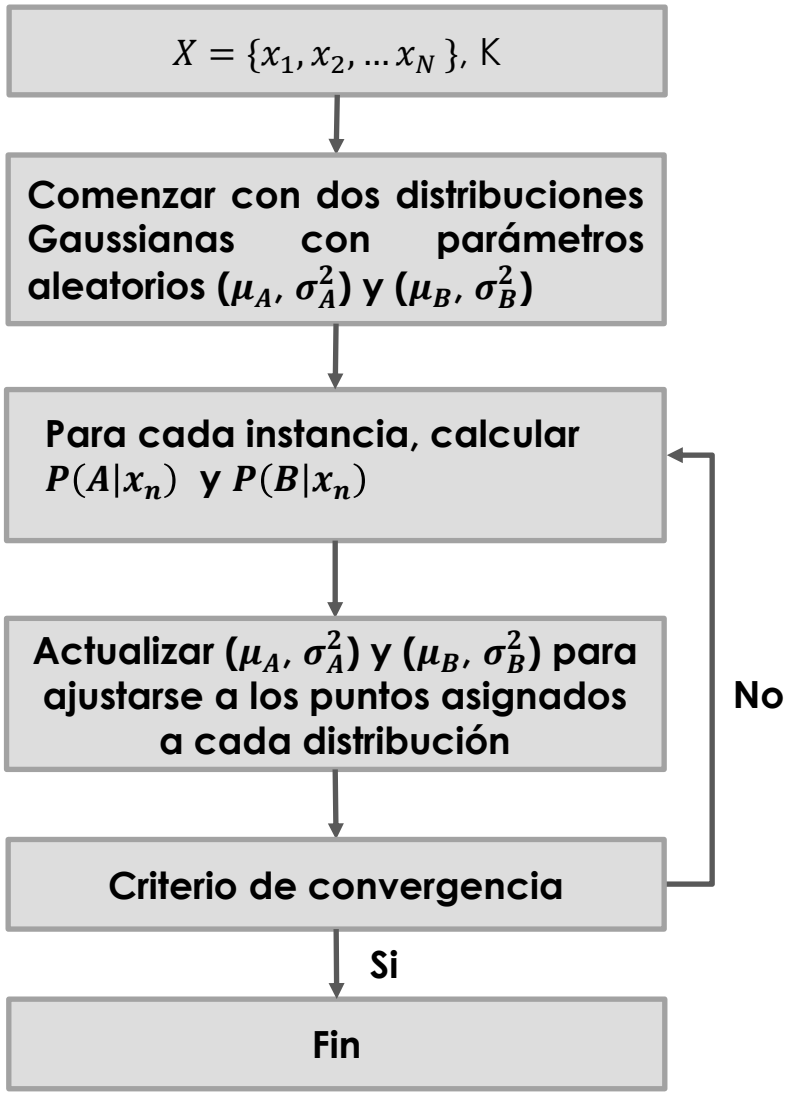
¿Cómo estimar μ_A, μ_B y σ_A^2, σ_B^2 de modo que describan lo mejor posible los datos observados?

!No tan trivial si no se conoce la fuente de cada observación!



experto en procesamiento del lenguaje natural

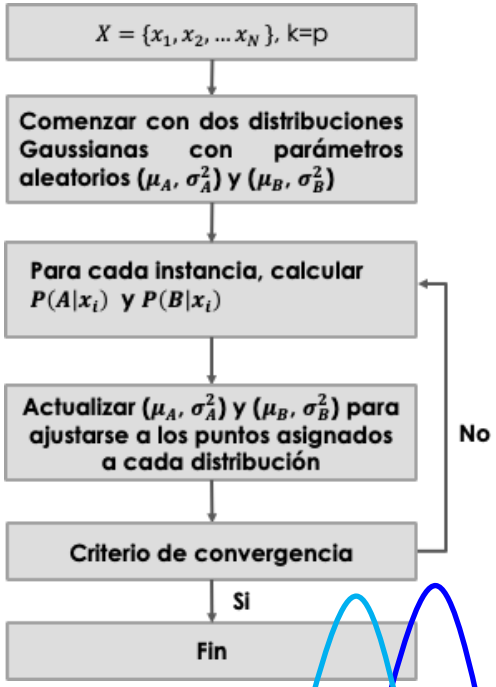
Expectation Maximization (EM)



- Se necesitan μ_A, μ_B y σ_A^2, σ_B^2 para conocer qué distribución es más probable haya generado cada ejemplo.
- Se necesita conocer la distribución que generó cada ejemplo para estimar μ_A, μ_B y σ_A^2, σ_B^2.

experto en procesamiento del lenguaje natural

Expectation Maximization (EM)



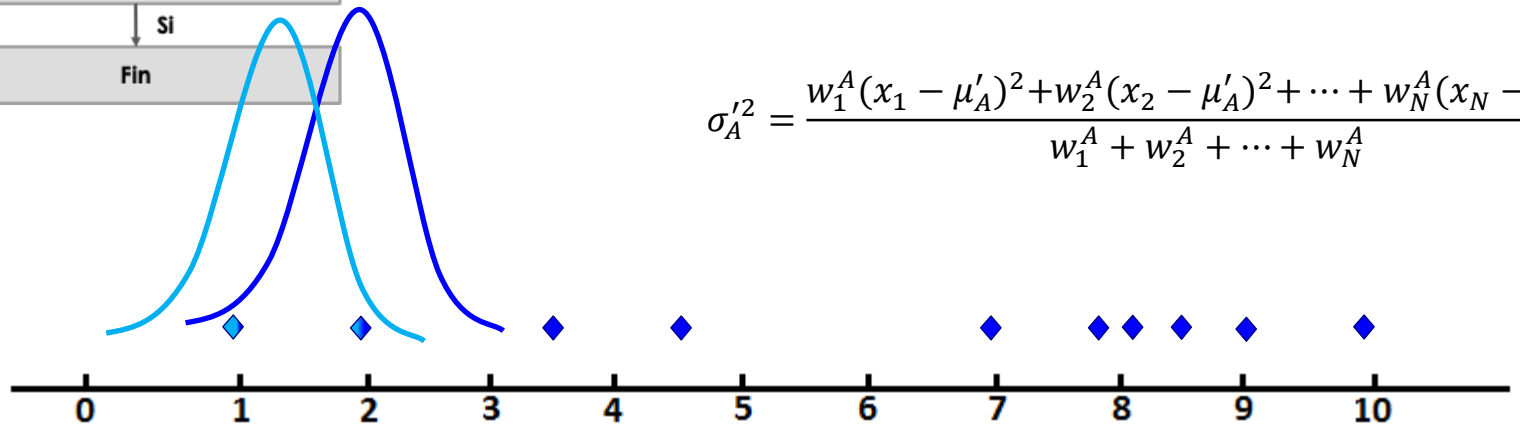
$$P(x_i|A) = \frac{1}{\sqrt{2\pi\sigma_A^2}} \exp\left(-\frac{(x_i - \sigma_A^2)^2}{2\sigma_A^2}\right)$$

$$P(A|x_i) = \frac{P(x_i|A)P(A)}{P(x_i|A)P(A) + P(x_i|B)P(B)}$$

$$P(B|x_i) = 1 - P(A|x_i) \qquad w_i^A = P(A|x_i)$$

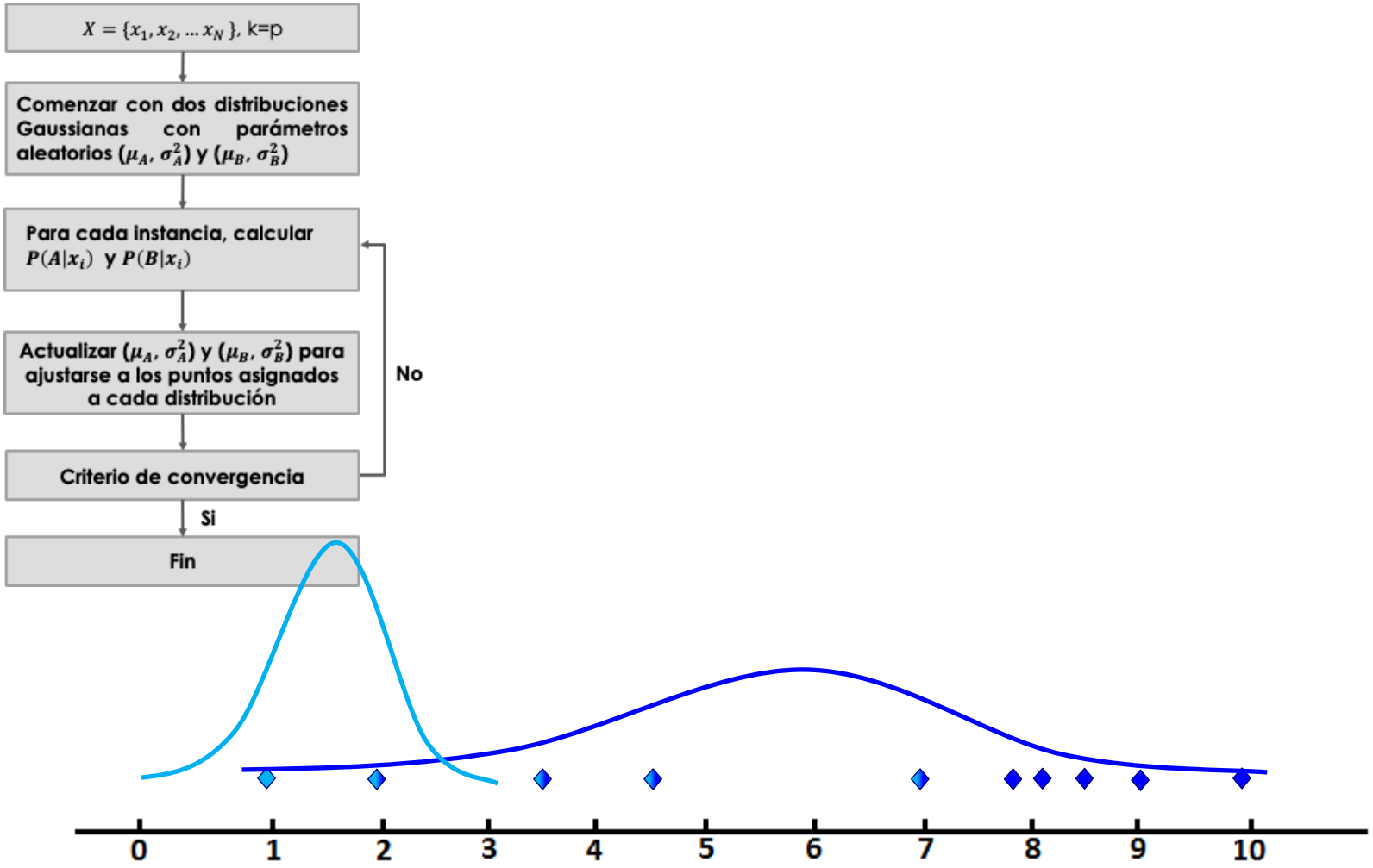
$$\mu'_A = \frac{w_1^A x_1 + w_2^A x_2 + \cdots + w_N^A x_N}{w_1^A + w_2^A + \cdots + w_N^A}$$

$$\sigma_A'^2 = \frac{w_1^A (x_1 - \mu'_A)^2 + w_2^A (x_2 - \mu'_A)^2 + \cdots + w_N^A (x_N - \mu'_A)^2}{w_1^A + w_2^A + \cdots + w_N^A}$$



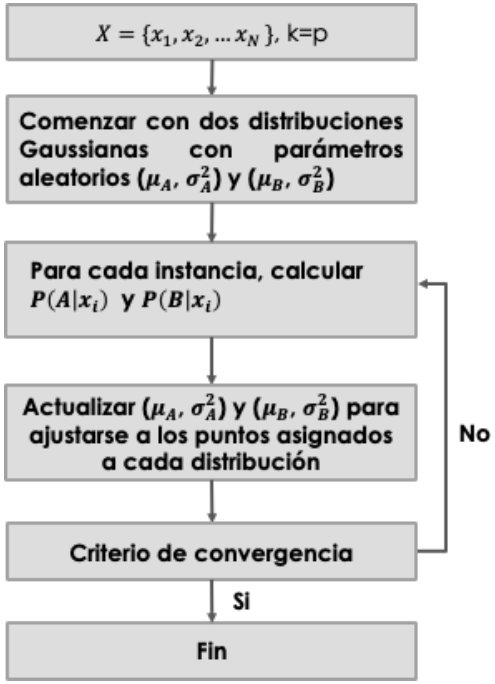
expertos en procesamiento del lenguaje natural

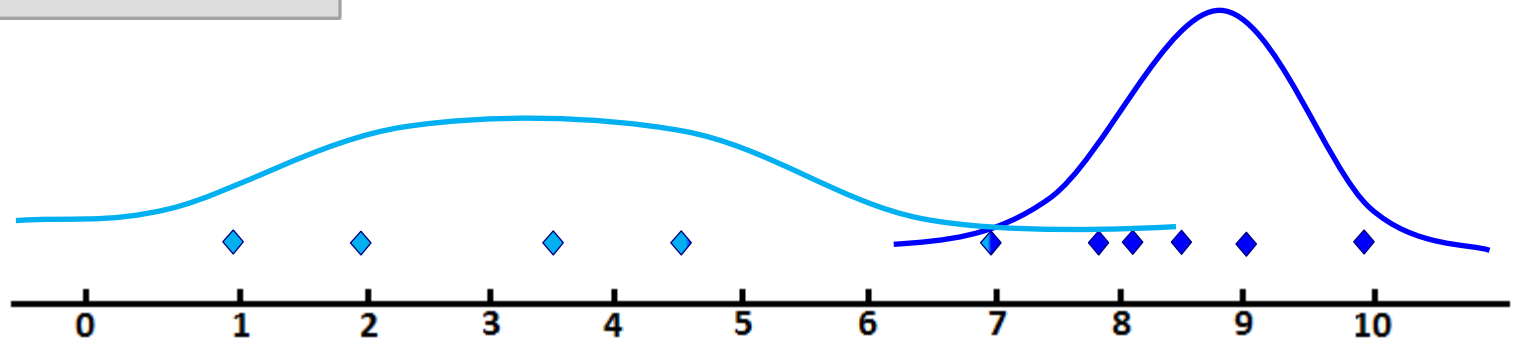
Expectation Maximization



experto en procesamiento del lenguaje natural

Expectation Maximization (EM)



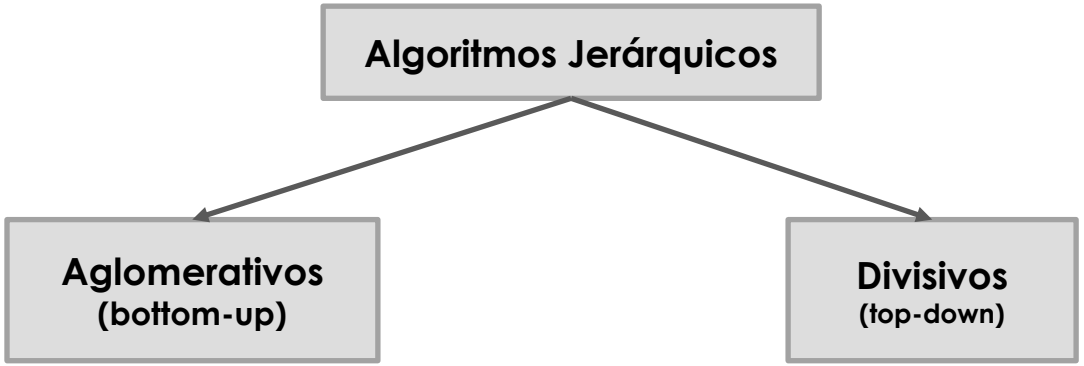


experto en procesamiento del lenguaje natural

Algoritmos Jerárquicos

experto en procesamiento del lenguaje natural

Algoritmos Jerárquicos



experto en procesamiento del lenguaje natural

Criterios de similaridad entre clústeres:

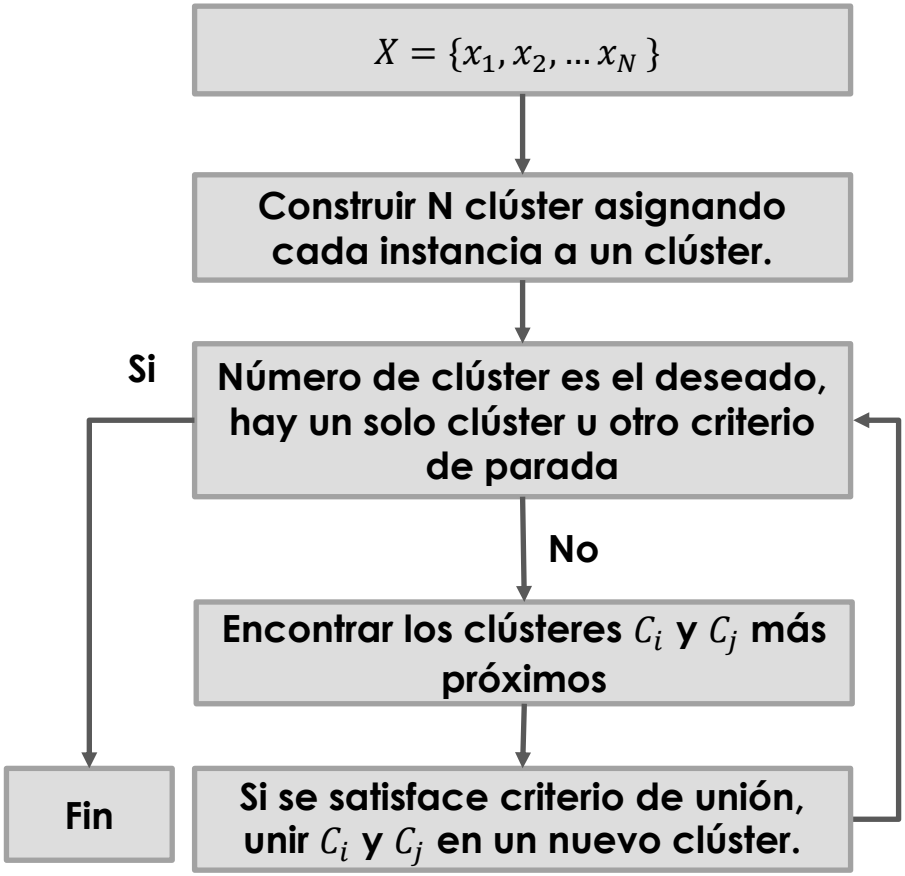
- **Single linkage:** $d(C_h, C_g) = \min d_{ij}, x_i \in C_h, x_j \in C_g$
  - Produce clústeres alargados donde pueden agruparse instancias bastante heterogéneas.

- **Complete linkage:** $d(C_h, C_g) = \max d_{ij}, x_i \in C_h, x_j \in C_g$
  - Clústeres más compactos, de forma esférica donde todas las instancias se encuentran dentro de un diámetro dado respecto al resto de las instancias.

- **Average linkage:** $d(C_h, C_g) = \frac{1}{|C_h| * |C_g|} \sum_{xi \in C_h} \sum_{xj \in C_g} d_{ij}$
  - Criterio de enlace menos dependiente de valores extremos. Clústeres con variabilidad interna aproximadamente igual.

experto en procesamiento del lenguaje natural

Esquema básico

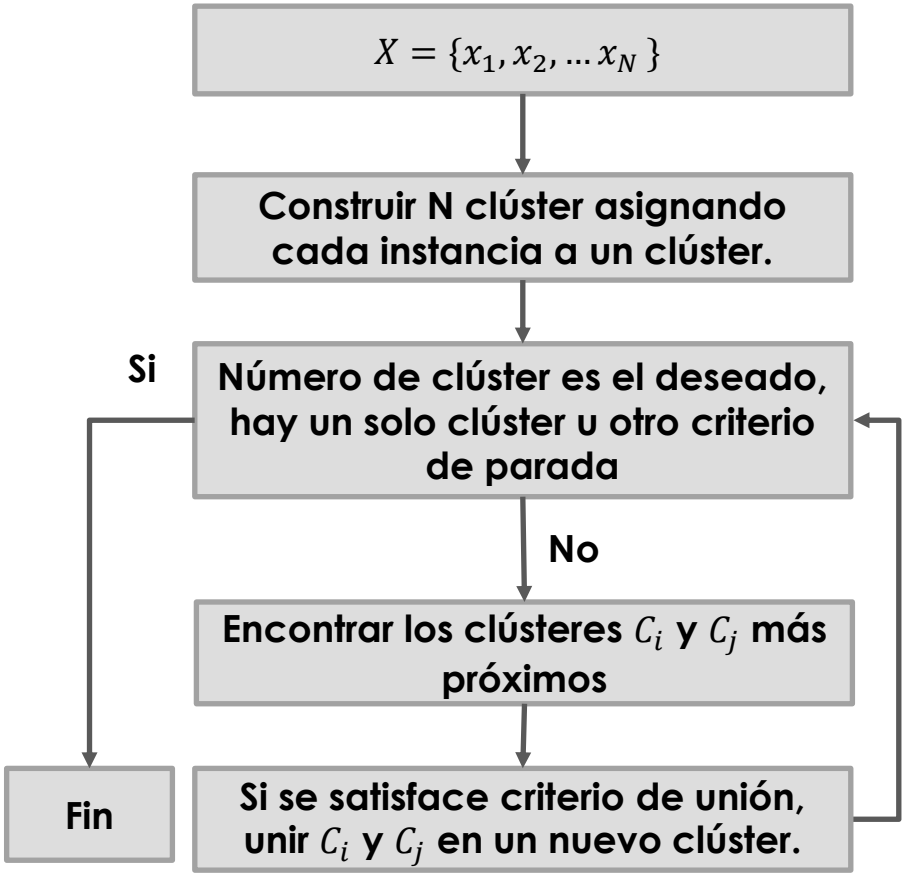


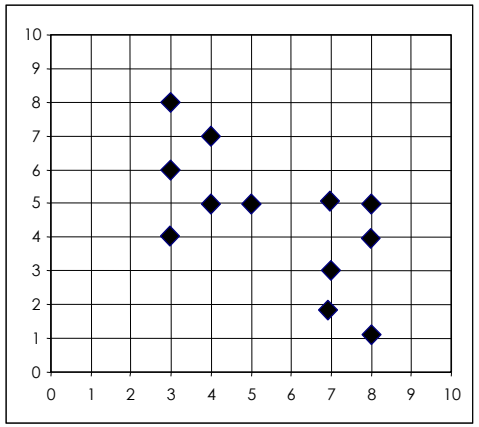
experto en procesamiento del lenguaje natural

Single Linkage

experto en procesamiento del lenguaje natural

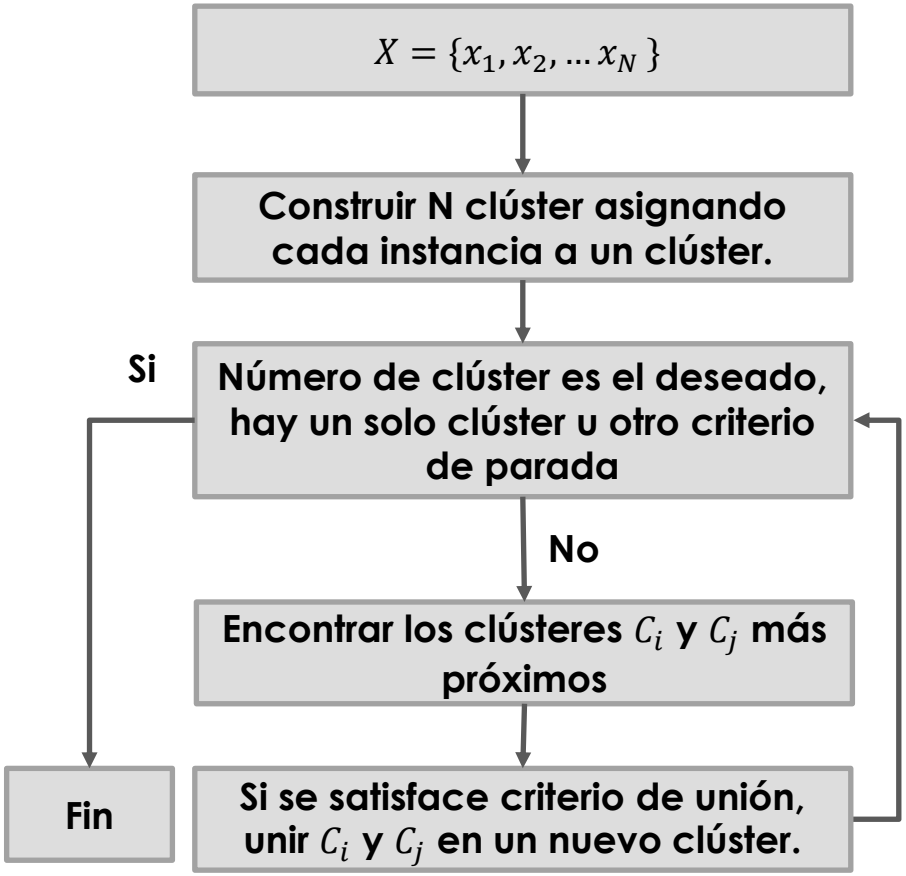
Single Linkage

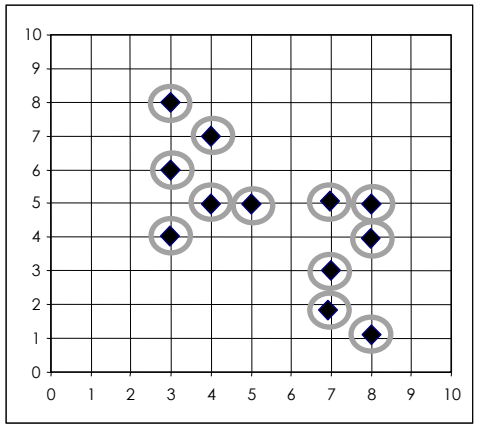




experto en procesamiento del lenguaje natural

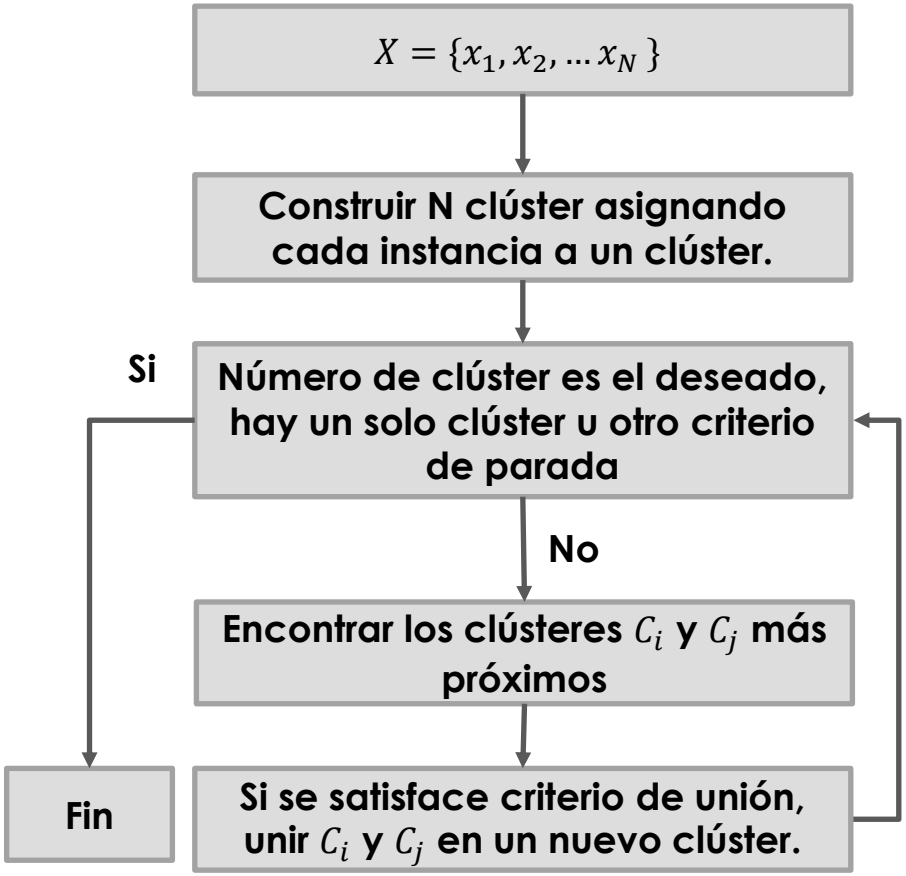
Single Linkage

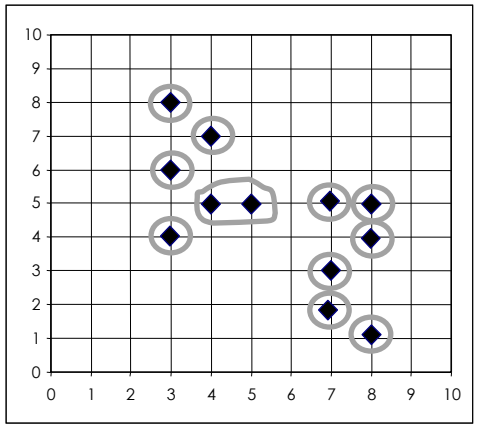




experto en procesamiento del lenguaje natural

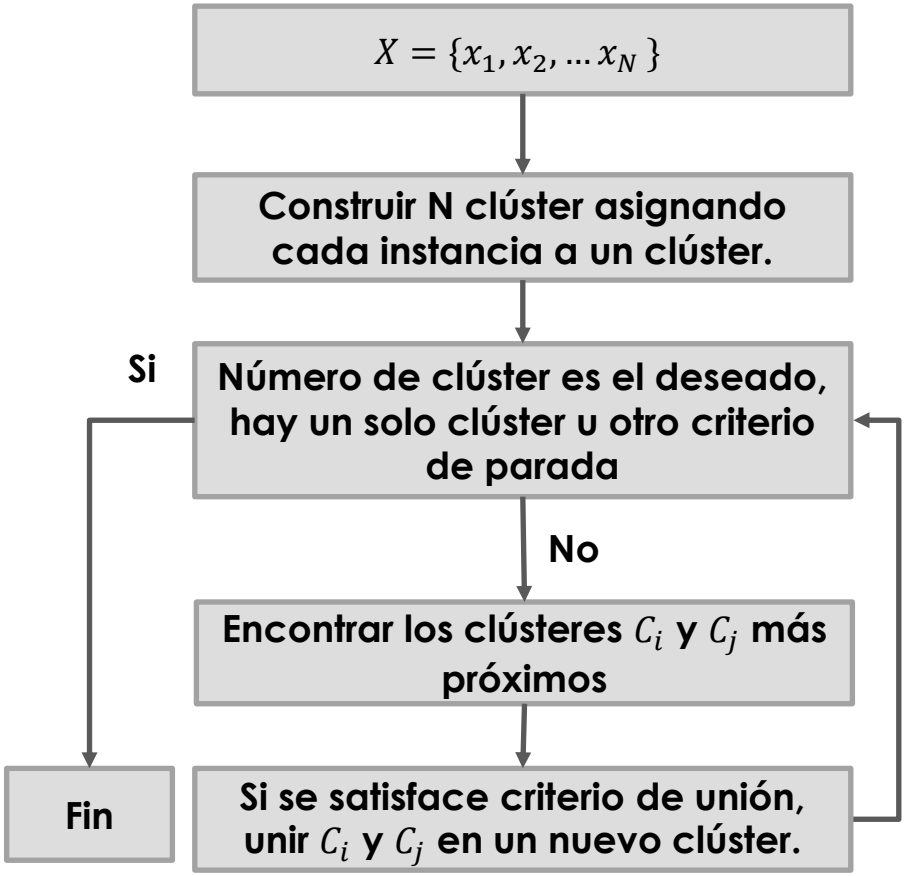
Single Linkage

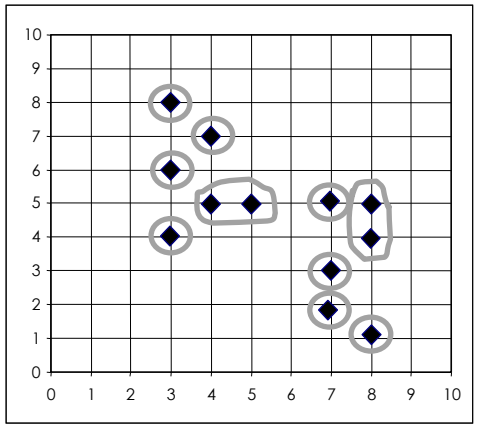




experto en procesamiento del lenguaje natural

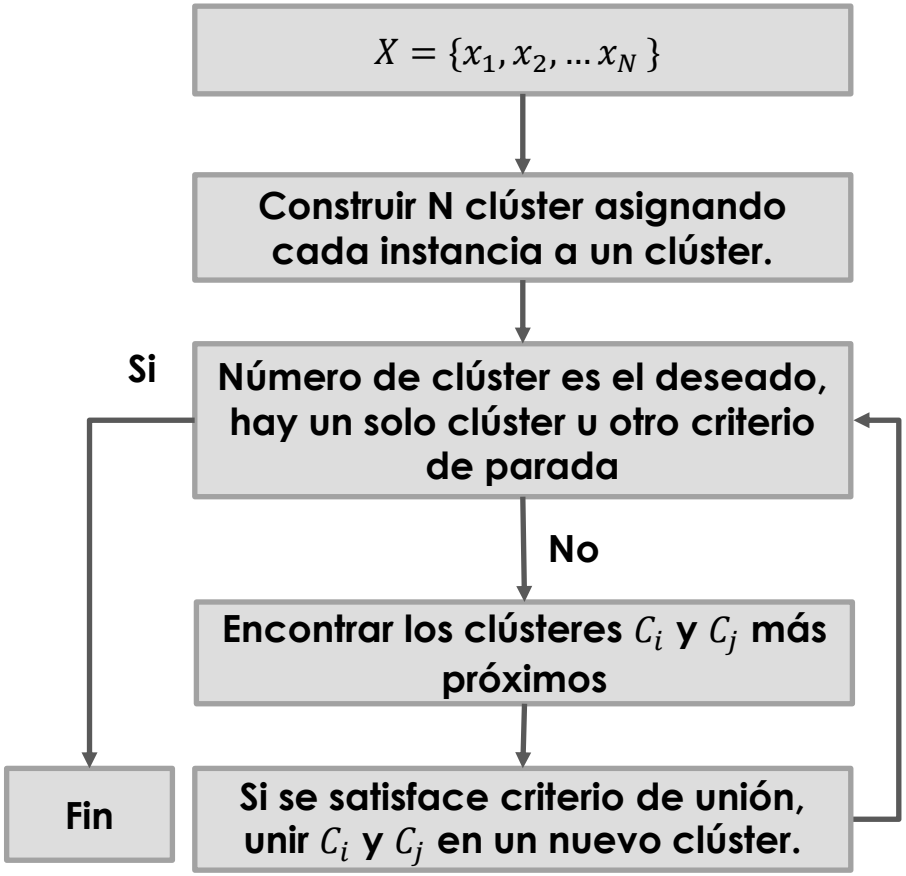
Single Linkage

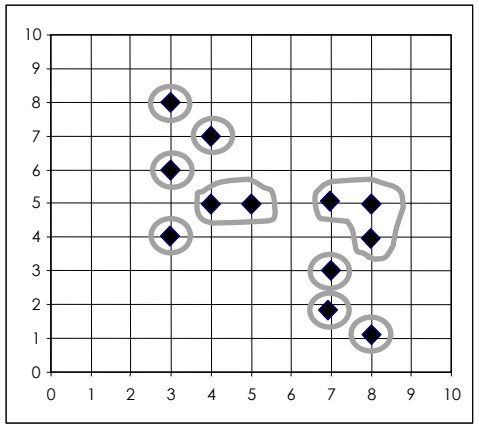




experto en procesamiento del lenguaje natural

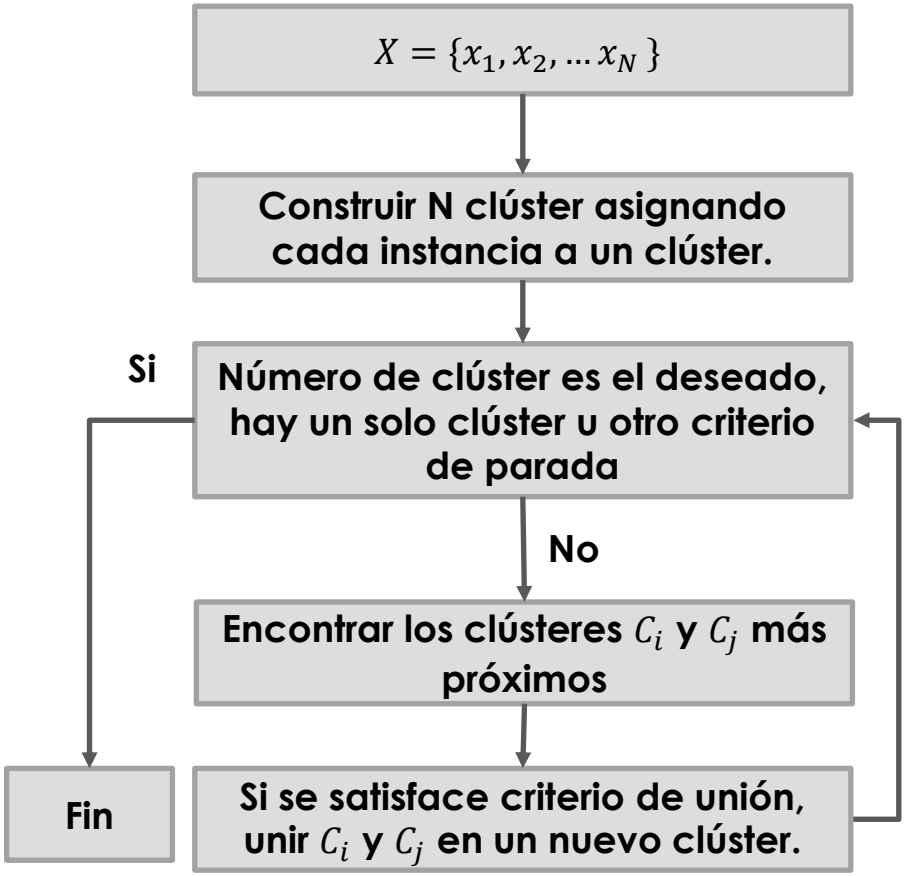
Single Linkage

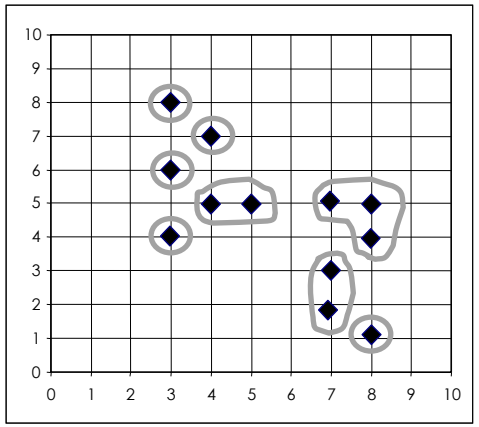




experto en procesamiento del lenguaje natural

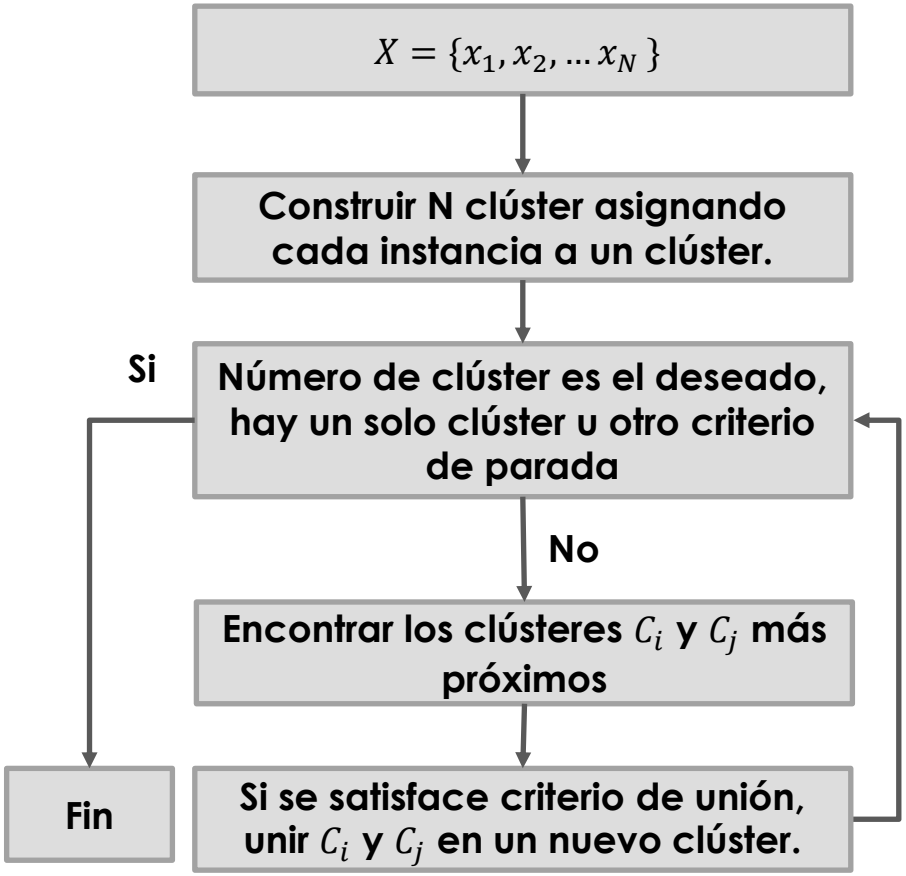
Single Linkage

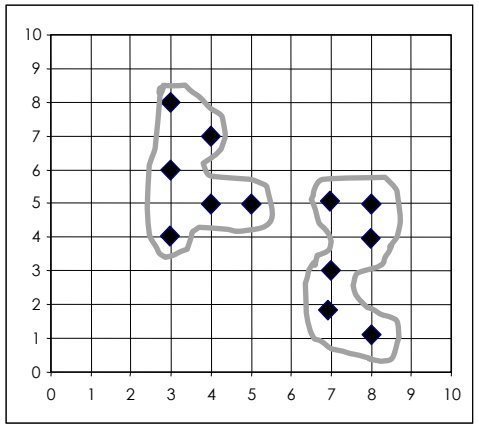




experto en procesamiento del lenguaje natural

Single Linkage



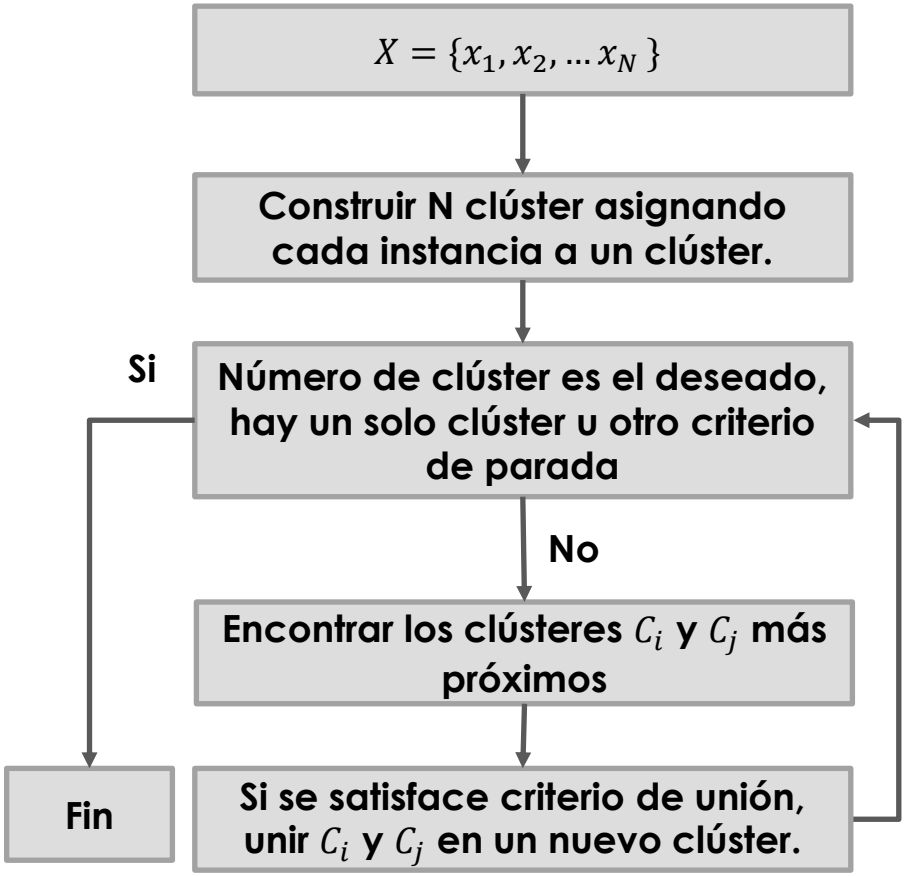


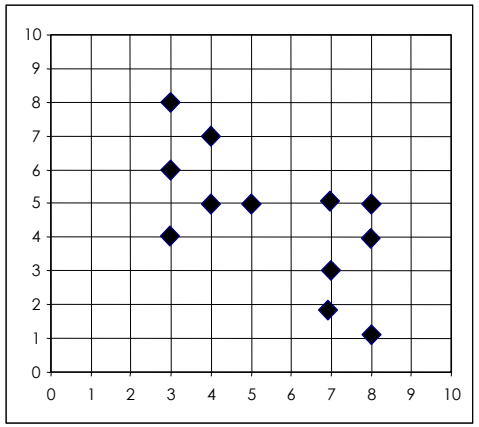
experto en procesamiento del lenguaje natural

Complete Linkage

experto en procesamiento del lenguaje natural

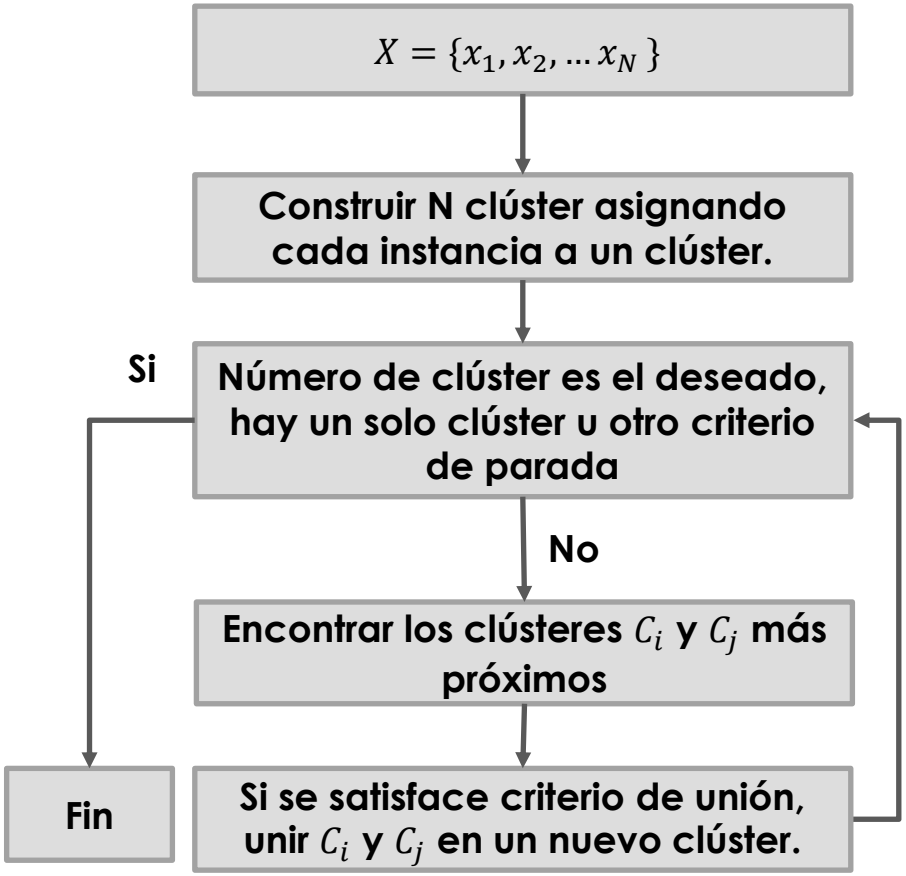
Complete Linkage

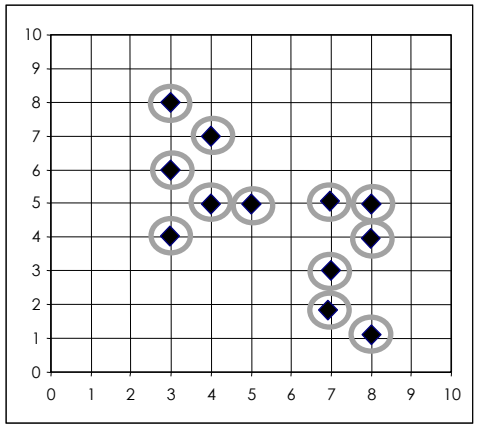




experto en procesamiento del lenguaje natural

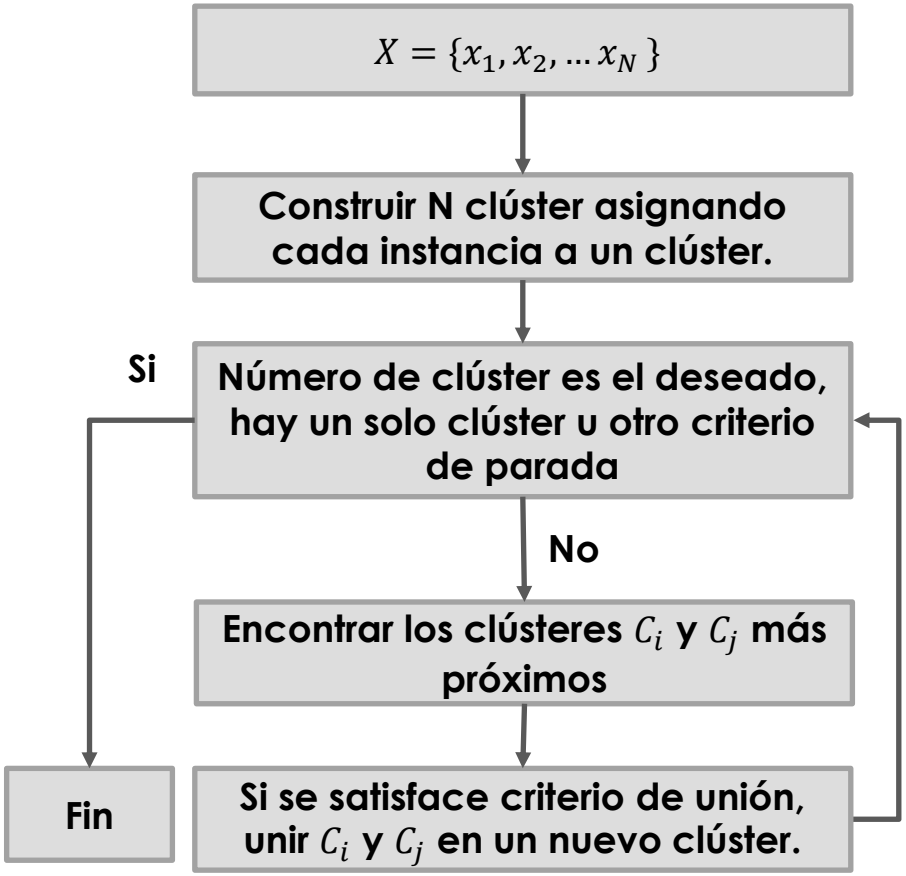
Complete Linkage

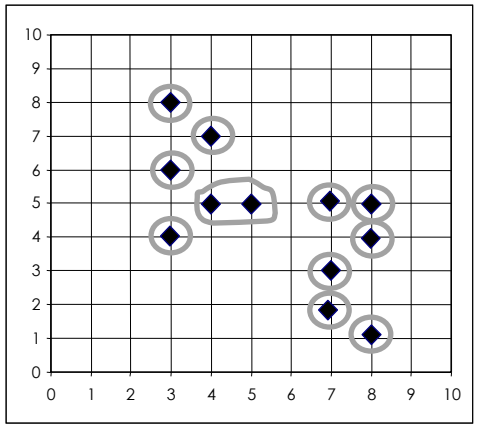




experto en procesamiento del lenguaje natural

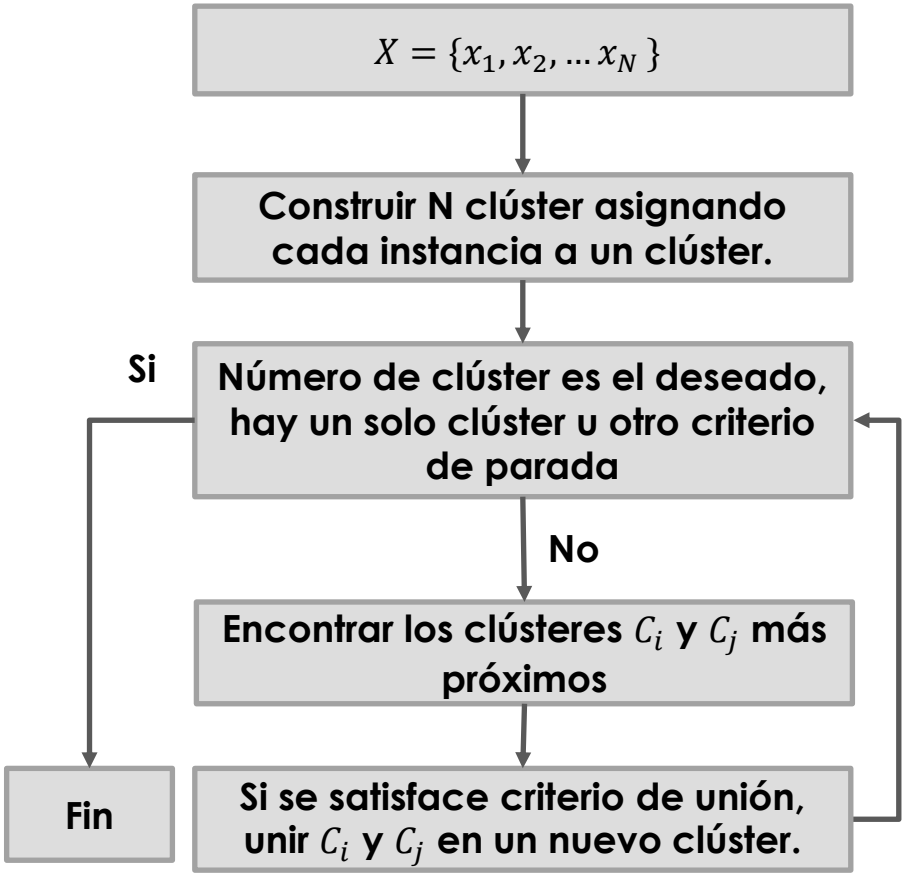
Complete Linkage

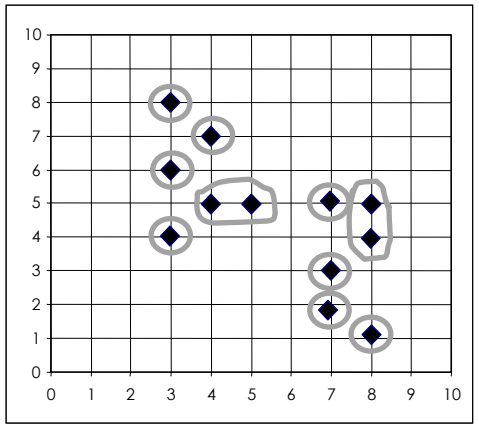




experto en procesamiento del lenguaje natural

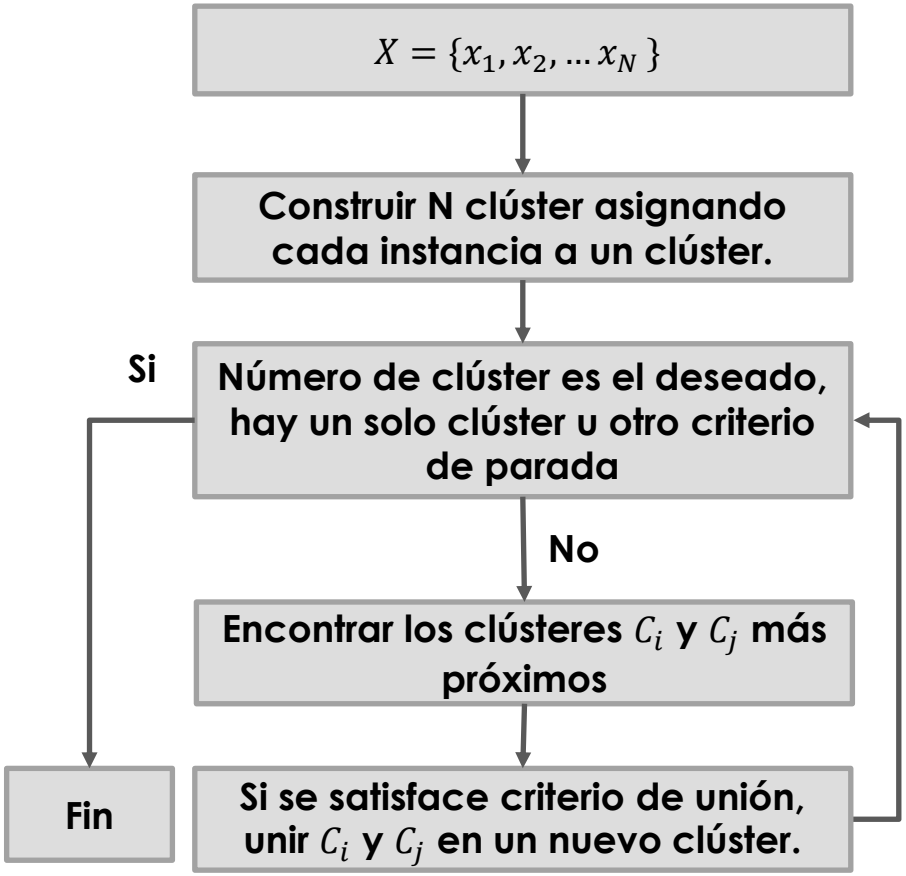
Complete Linkage

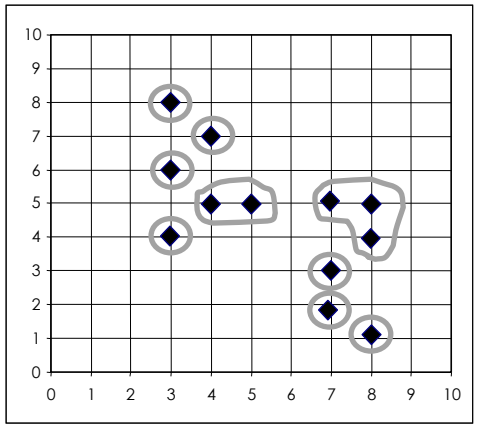




experto en procesamiento del lenguaje natural

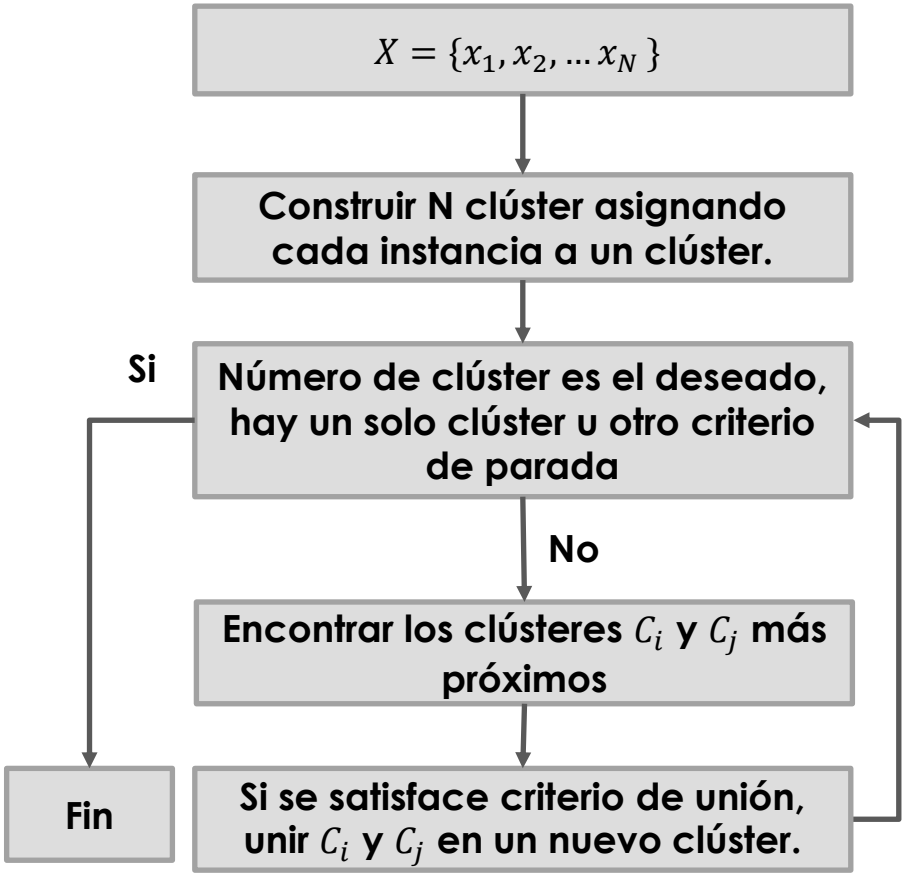
Complete Linkage

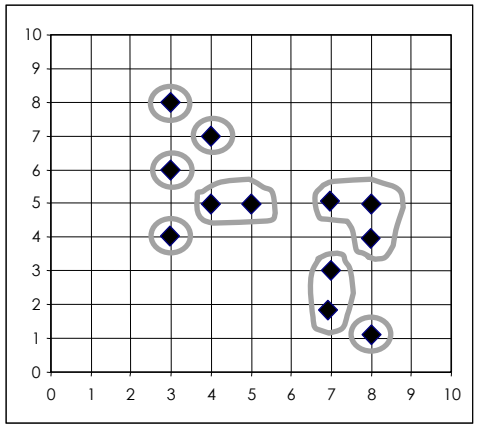




experto en procesamiento del lenguaje natural

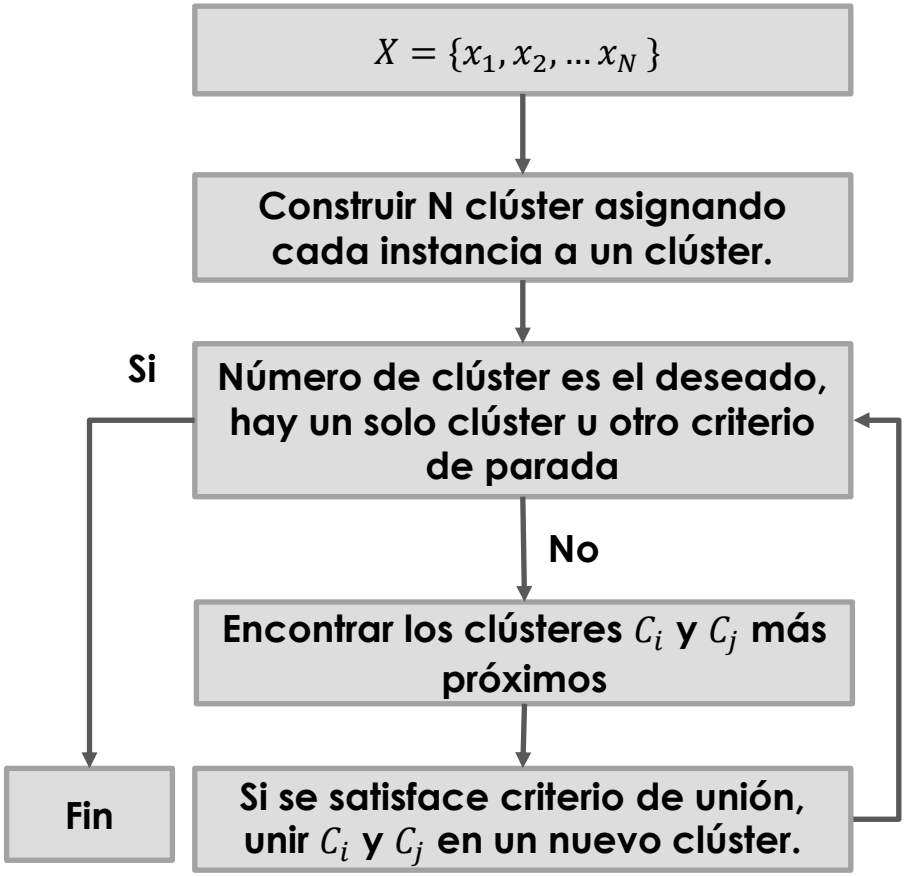
Complete Linkage

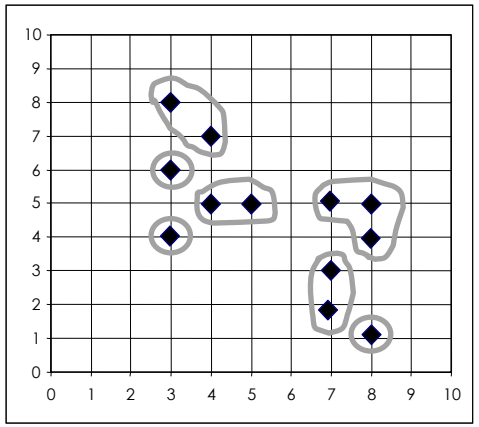




experto en procesamiento del lenguaje natural

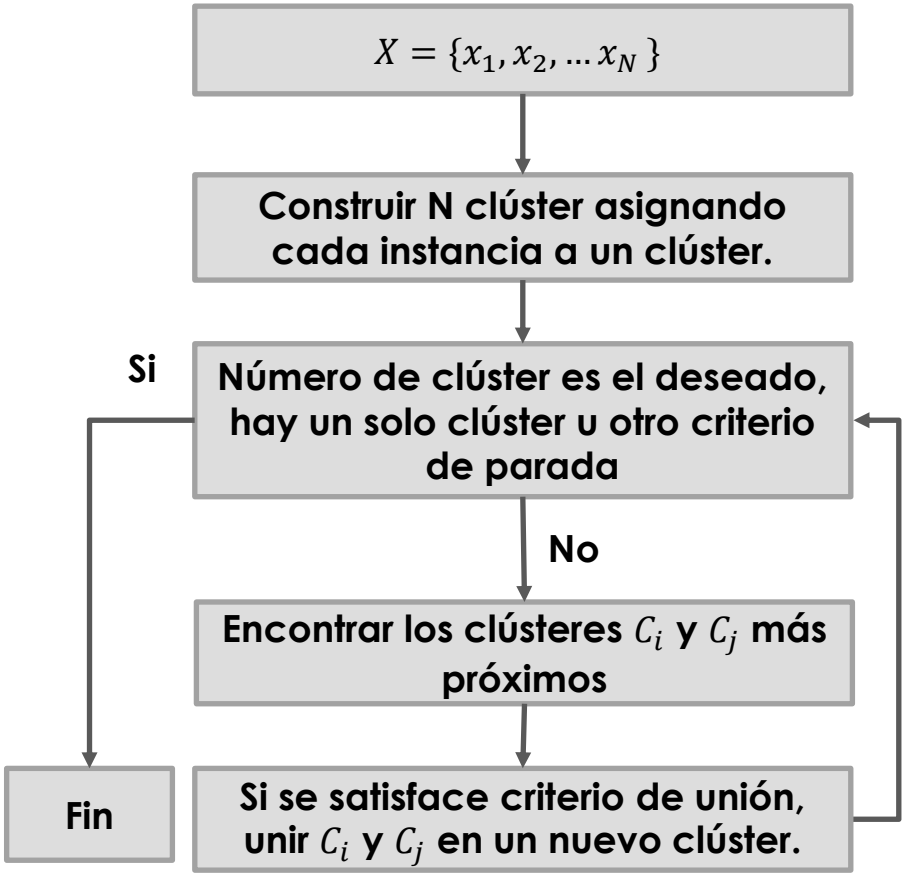
Complete Linkage

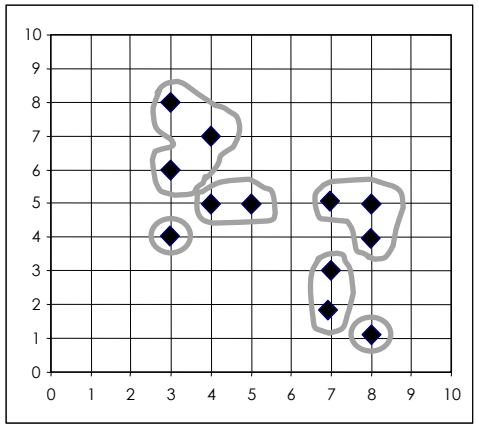




experto en procesamiento del lenguaje natural

Complete Linkage



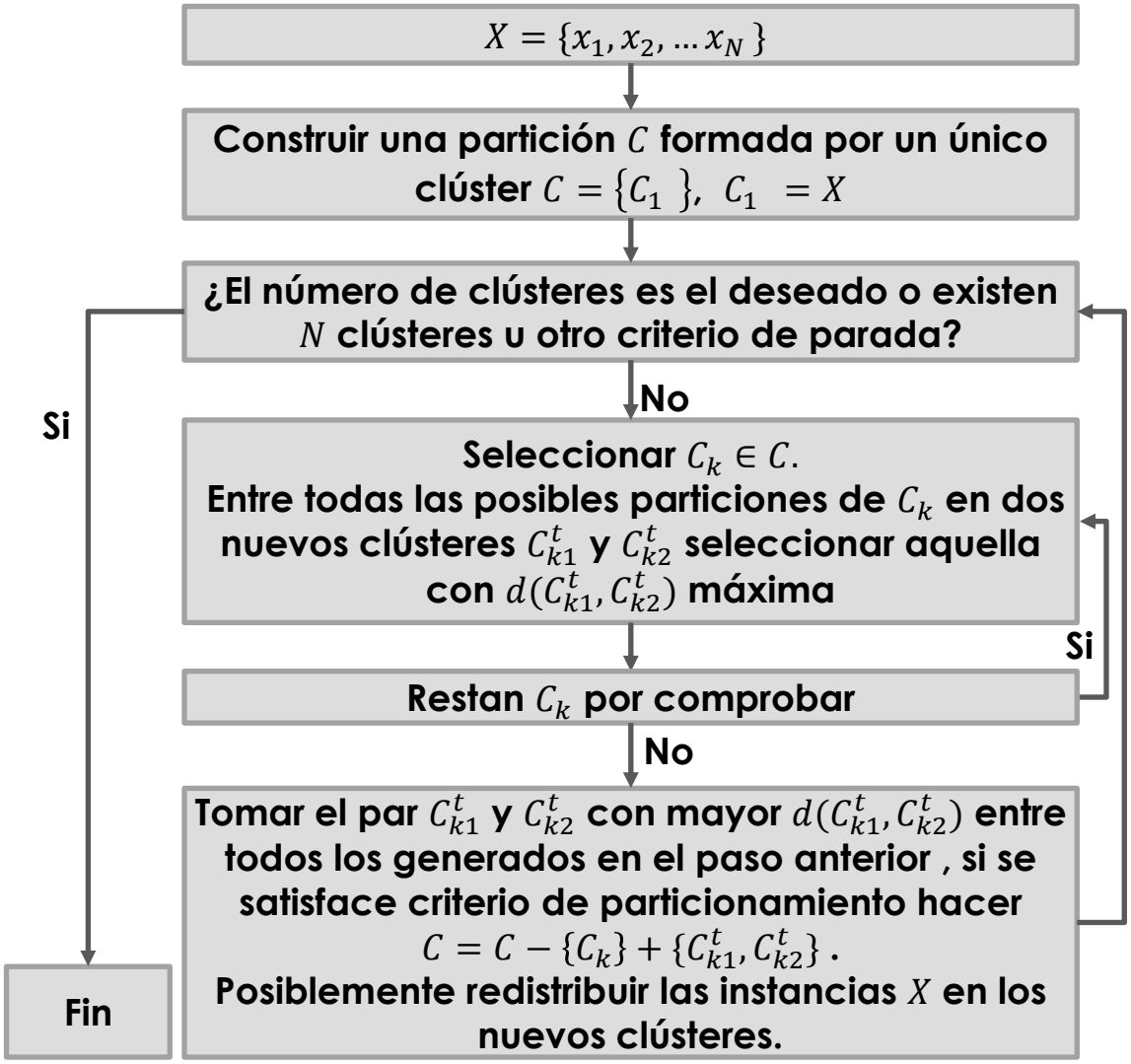


experto en procesamiento del lenguaje natural

Algoritmos Divisivos

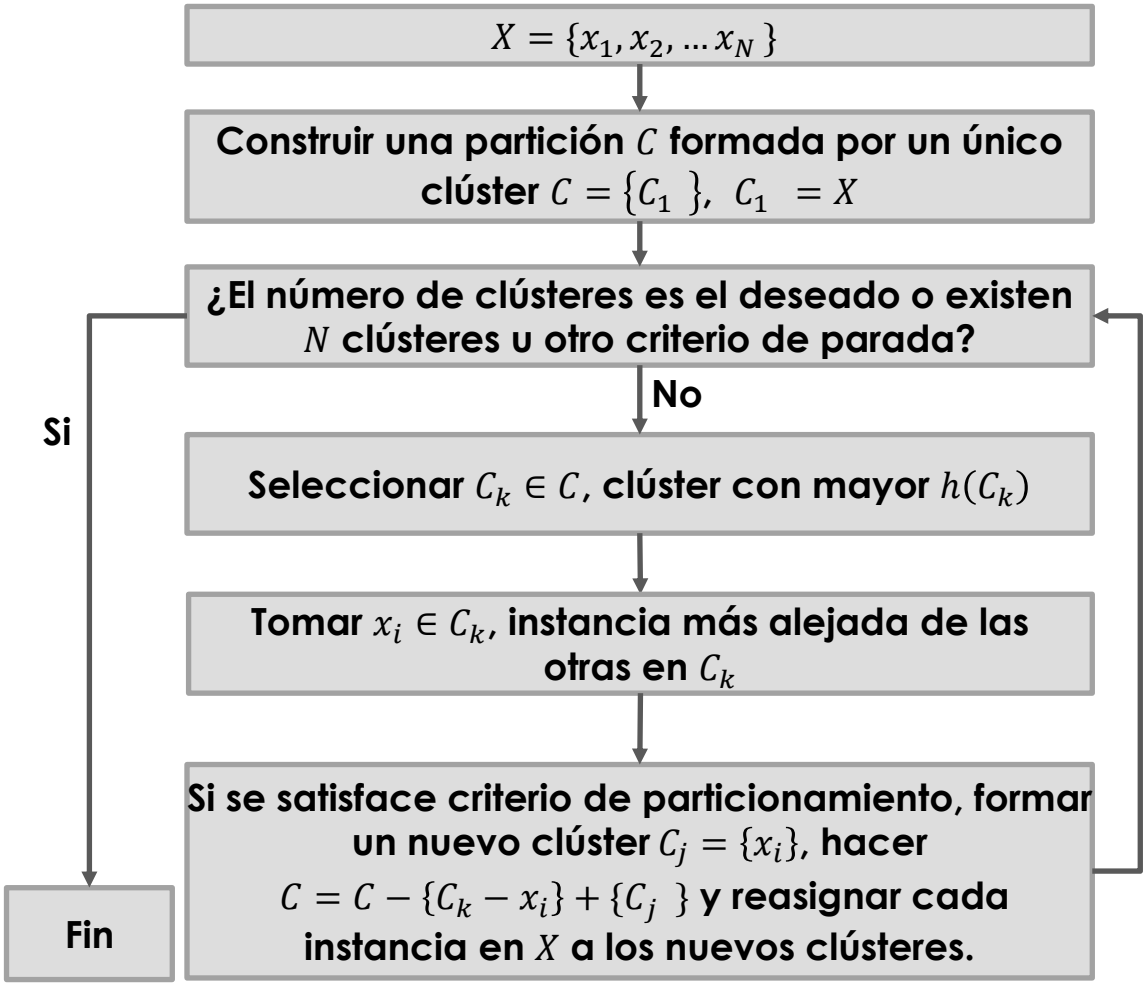
experto en procesamiento del lenguaje natural

Esquema básico



experto en procesamiento del lenguaje natural

Esquema básico (heurístico)



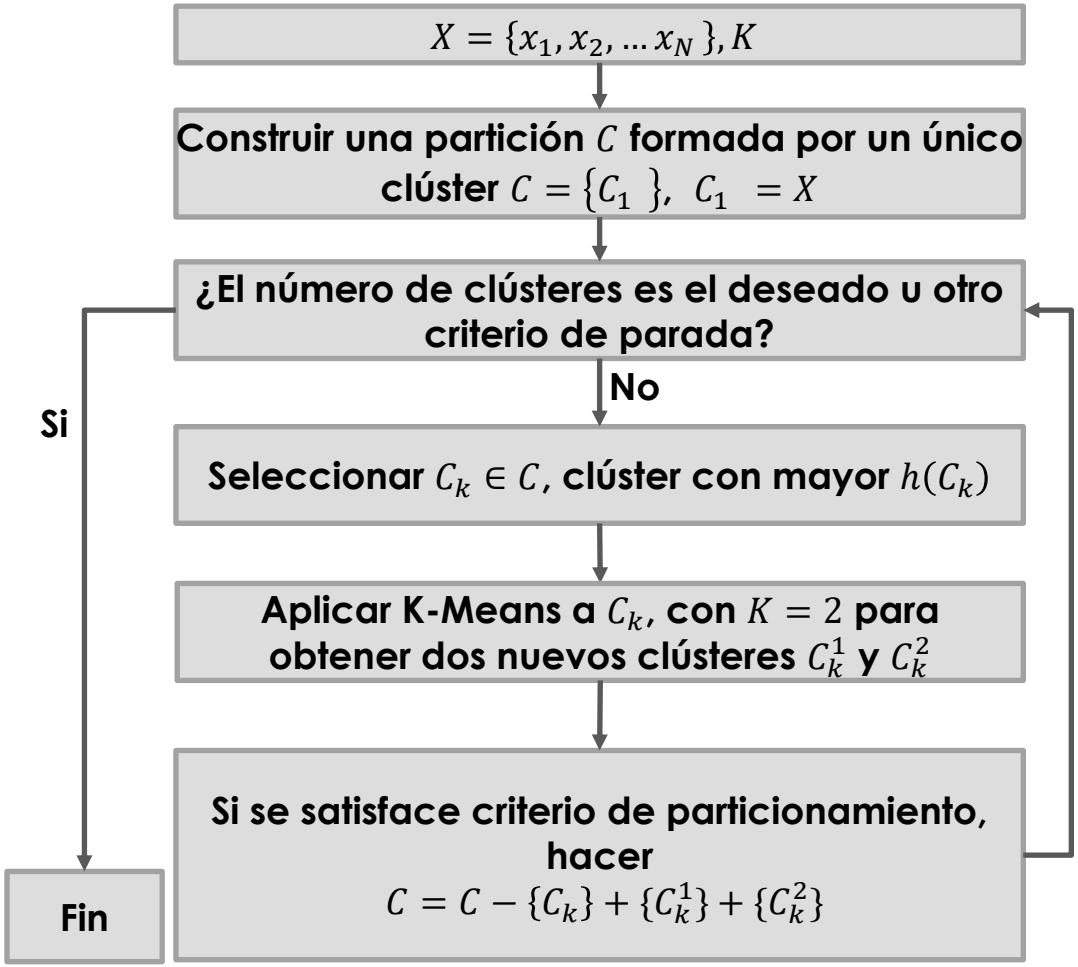
Notar no se verifica cada particionamiento posible

experto en procesamiento del lenguaje natural

Bisecting K-Means

experto en procesamiento del lenguaje natural

Algoritmo Bisecting K-Means



experto en procesamiento del lenguaje natural

Evaluación

experto en procesamiento del lenguaje natural

- ¿Son los clústeres "buenos''?, esto es, ¿la similitud intra-clúster es mayor que la inter-clúster?
- ¿Qué instancias parecen estar bien ubicadas?, ¿cuáles no?, ¿cuáles parecen pertenecer a varios clústeres?
- ¿Cuál es la estructura general de los datos?
- ¿Cuál es el número natural de clústeres?

Existen diferentes enfoques para responder estas preguntas, por ejemplo:

- Silhouette Coefficient
- Elbow Curve

experto en procesamiento del lenguaje natural

Silhouette Coefficient

Sea $X = \{x_1, x_2, \dots x_N\}$ un conjunto de instancias y un particionamiento en clústers $C_1, C_2, \dots, C_K$, y una medida de (di)similitud "ratio scaled", se definen:

- $a(x_n) = \frac{1}{|C_h|}\sum_{j=1}^{|C_h|-1} d_{nj}, n \neq j, x_n, x_j \in C_h$, en otras palabras, $a(x_n)$ es la (di)similitud promedio de x_n a las instancias de su clúster.
- $d(x_n, C_h) = \frac{1}{|C_h|}\sum_{j=1}^{|C_h|-1} d_{nj}, \notin C_h, x_j \in C_h$, en otras palabras, $d(x_n, C_h)$ es la (di)similitud promedio de x_n a las instancias del clúster C_h.
- $b(x_n) = min_C d(x_n, C)$, es decir, es el valor menor $d(x_n, C_h)$ para alguno de los clústeres. Para una medida de similitud, se define como el máximo. El clúster más próximo a x_n según $b(x_n)$ se llama vecino de x_n

Para disimilitud: $s(x_n) = \frac{b(x_n) - a(x_n)}{\max\{a(x_n), b(x_n)\}}$, $-1 \leq s(x_n) \leq 1$ o $s(x_n) = \frac{a(x_n) - b(x_n)}{\max\{a(x_n), b(x_n)\}}$ para similitud.

experto en procesamiento del lenguaje natural

Silhouette Coefficient

Se define el Coeficiente Silhouette $SC = \frac{1}{N}\sum_{n=1}^{N} s(x_n)$, es decir, el promedio de los $s(x_n)$

Los valores de SC pueden interpretarse de acuerdo con el siguiente criterio:

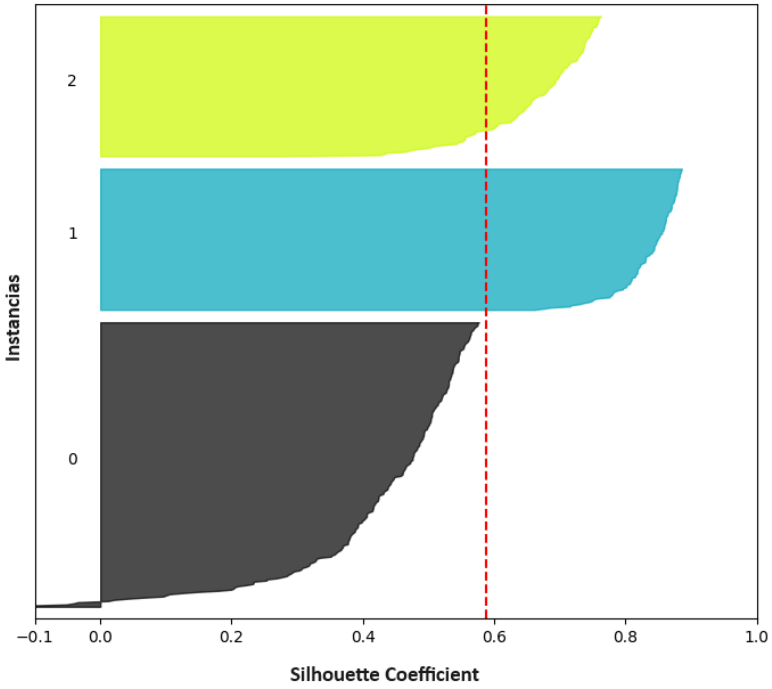
- 0.70 a 1.00: estructura bien definida.
- 0.50 a 0.70: estructura "razonablemente" bien definida.
- 0.25 a 0.50: estructura débil.
- < 0.25: no se descubre una estructura sustancial.

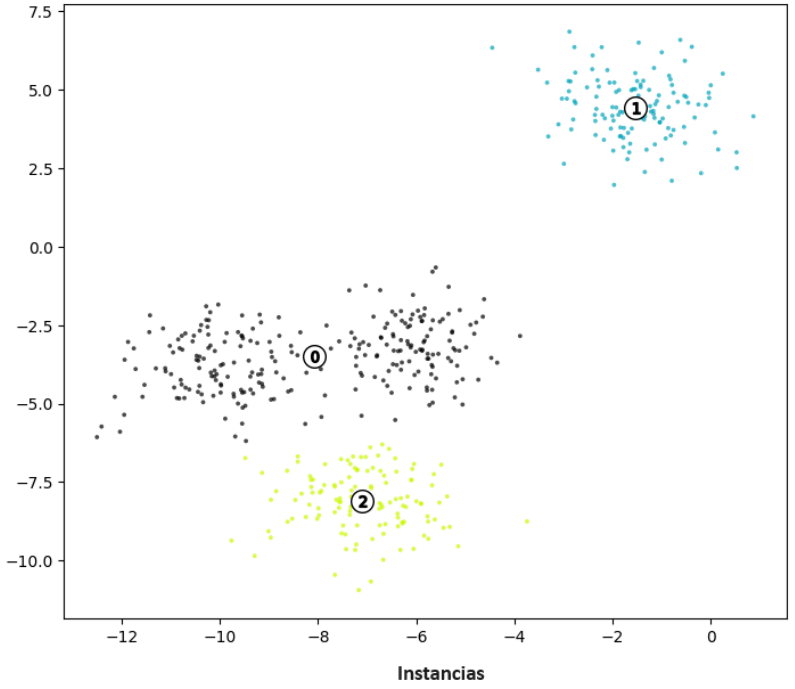
Notar que existe otra definición donde SC es el máximo de los SC para valores de $k = 1, \ldots, K$

experto en procesamiento del lenguaje natural

Silhouette Coefficient

La gráfica Silhouette permite resumir que tan apropiado es el clúster para cada instancia. Se construye como una gráfica de barras horizontales, ordenando las instancias de cada clúster por su valor $s(x_n)$.





experto en procesamiento del lenguaje natural

Conclusiones

Características de un buen algoritmo de agrupamiento:

- Escalabilidad.
- Manejo de atributos de diferente naturaleza.
- Posibilidad de descubrir clúster de forma arbitraria.
- Requerimiento mínimo de conocimiento específico del dominio.
- Tolerancia a ruido, outliers, datos faltantes.
- Insensibilidad al orden de entrada de las instancias.
- Correcto manejo de datos de dimensionalidad alta.
- Interpretable y usable.

experto en procesamiento del lenguaje natural

Fin

experto en procesamiento del lenguaje natural