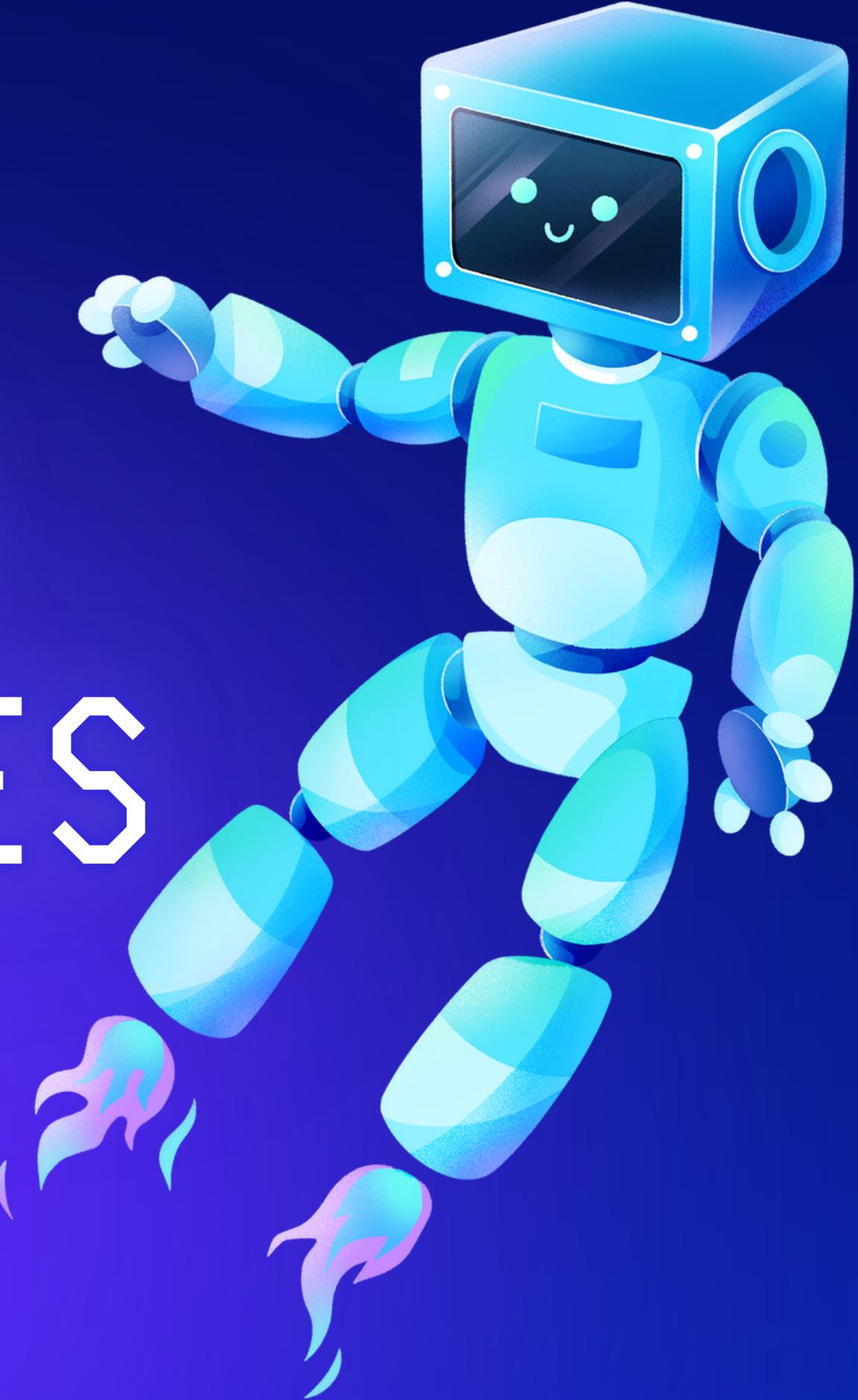


ALGORITMOS PARTICIONALES

Munguía Valadez Miguel Ángel

&

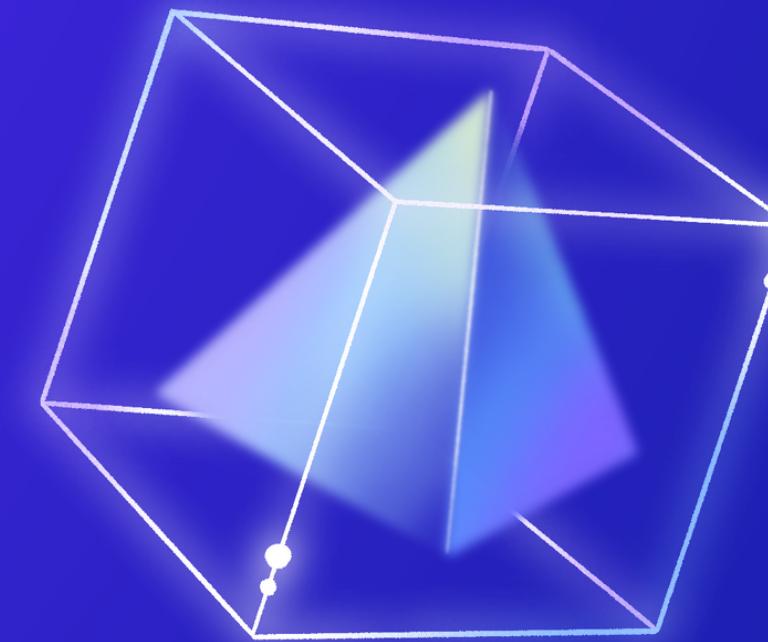
Orozco Solorio Kevin Adrián





CONTENIDO

- 5.5 Algoritmos particionales
 - 5.5.1 Árboles de expansión mínima
 - 5.5.2 Algoritmo de agrupamiento de error cuadrático
 - 5.5.3 Agrupamiento por K-medias (K-means)
 - 5.5.4 Algoritmo de vecinos más cercanos (KNN)



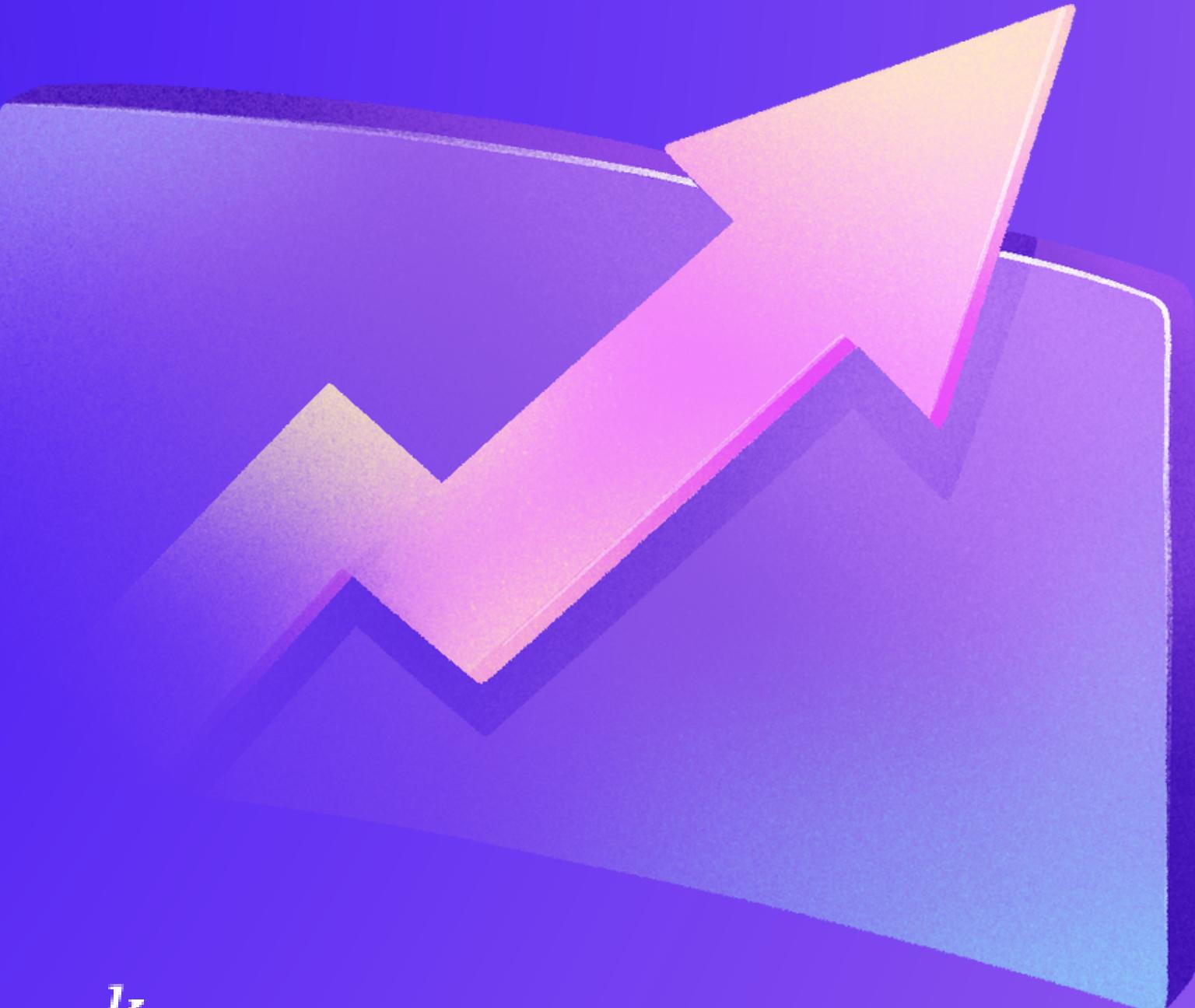
5.5 ALGORITMOS PARTICIONALES

El agrupamiento (clustering) particional, también conocido como no jerárquico, crea grupos en un solo paso, en vez de varios. Esto sucede debido a que se crea un cluster con varios clústeres internos. Por ello, se debe ingresar el número deseado k de clústeres.

De la misma manera, se debe utilizar alguna métrica o criterio para determinar la viabilidad de su utilización.

Una medida común usada es la métrica del error cuadrado, la cual mide la distancia de cada punto al centroide para el agrupamiento asociado.

Un problema de los algoritmos partionales es que sufren de una explosión combinacional debido al número de posibles soluciones.



$$\sum_{m=1}^k \sum_{t_{mi} \in K_m} dis(C_m, t_{mi})^2$$

5.5.1 ÁRBOLES DE EXPANSIÓN MÍNIMA



Ya que tenemos que los algoritmos aglomerativos y divisivos están basados en el uso de árboles de expansión mínima, también se presenta un algoritmo de este tipo para los algoritmos particionales. Sin embargo, primero debemos definir ciertos conceptos.

Zahn propone medidas de inconsistencias más razonables basadas en el peso(distancia) de una arista en comparación con aquellas cercanas a esta.

El problema de esto radica nuevamente en que la complejidad del algoritmo se vuelve $O(n^2)$

ALGORITMO DE ÁRBOL DE EXPANSIÓN MÍNIMA

Input:

$D = \{t_1, t_2, \dots, t_n\}$ //Conjunto de elementos

A //Matriz de adyacencia mostrando la distancia entre elementos

k //Número de clústeres

Output:

f //Conjunto de clústeres

Algoritmo particional de árbol de expansión mínima:

$M = MST(A)$

Identificar la inconsistencia de los extremos en M

Remover los extremos $k-1$ inconsistentes

Crear una representación como salida.

5.5.2 ALGORITMO DE AGRUPAMIENTO DE ERROR CUADRÁTICO



El algoritmo de agrupamiento de error cuadrático minimiza el error cuadrático. Este se define como la suma de las distancias Euclidianas al cuadrado entre cada elemento del conjunto y el centroide del conjunto. Donde un cluster K_i y un conjunto de elementos $\{t_1, t_2, \dots, t_m\}$. El error cuadrático está definido como:

$$se K_i = \sum_{j=1}^m \|t_{ij} - c_k\|^2$$

Dado un grupo de clústeres $K = \{K_1, K_2, \dots, K_m\}$, el error cuadrático está definido por:

$$se K = \sum_{i=1}^k se K_i$$

ALGORITMO DE ERROR CUADRÁTICO

Input:

$D = \{t_1, t_2, \dots, t_n\}$ //Conjunto de elementos
 k //Número de clústeres

Output:

K //Conjunto de clústeres

Algoritmo del error cuadrado:

Asigna cada elemento t_i a un clúster.

Calcula el centro de cada clúster.

Repetir

Asigna cada elemento t_i al clúster que tiene el centro más próximo.

Calcula el nuevo centro del clúster.

Calcula el error cuadrado.

Hasta que la diferencia entre los errores sucesivos sea menor al límite

5.5.3 AGRUPAMIENTO POR K-MEANS



K-means o K-medias es un algoritmo de agrupamiento iterativo en el cual los elementos son movidos entre los diferentes grupos de clústeres hasta que el grupo deseado es alcanzado.

Puede ser visto como un tipo de algoritmo de error cuadrático, a pesar de que el criterio de convergencia necesita ser definido con el error cuadrado.

En este caso, se obtiene un gran grado de semejanza entre los elementos de los clústeres, mientras que un gran grado de disimilitud entre elementos de diferentes clústeres también se obtiene.

El clúster promedio de $K_i = \{t_{i1}, t_{i2}, \dots, t_{im}\}$ está definido como:

$$m_i = \frac{1}{m} \sum_{j=1}^m t_{ij}$$

En esta definición se asume que cada tupla tiene solo valores NUMERICOS. El algoritmo K-means requiere que una definición de clúster promedio exista, pero no tiene que ser uno en específico. Aquí, el promedio se define igual a la definición de centroide.

Este algoritmo asume que el parámetro K es el número deseado de clústeres.

ALGORITMO DE K-MEANS

Input:

$D = \{t_1, t_2, \dots, t_n\}$ //Conjunto de elementos
 k //Número de clústeres

Output:

K //Conjunto de clústeres

Algoritmo K-Means:

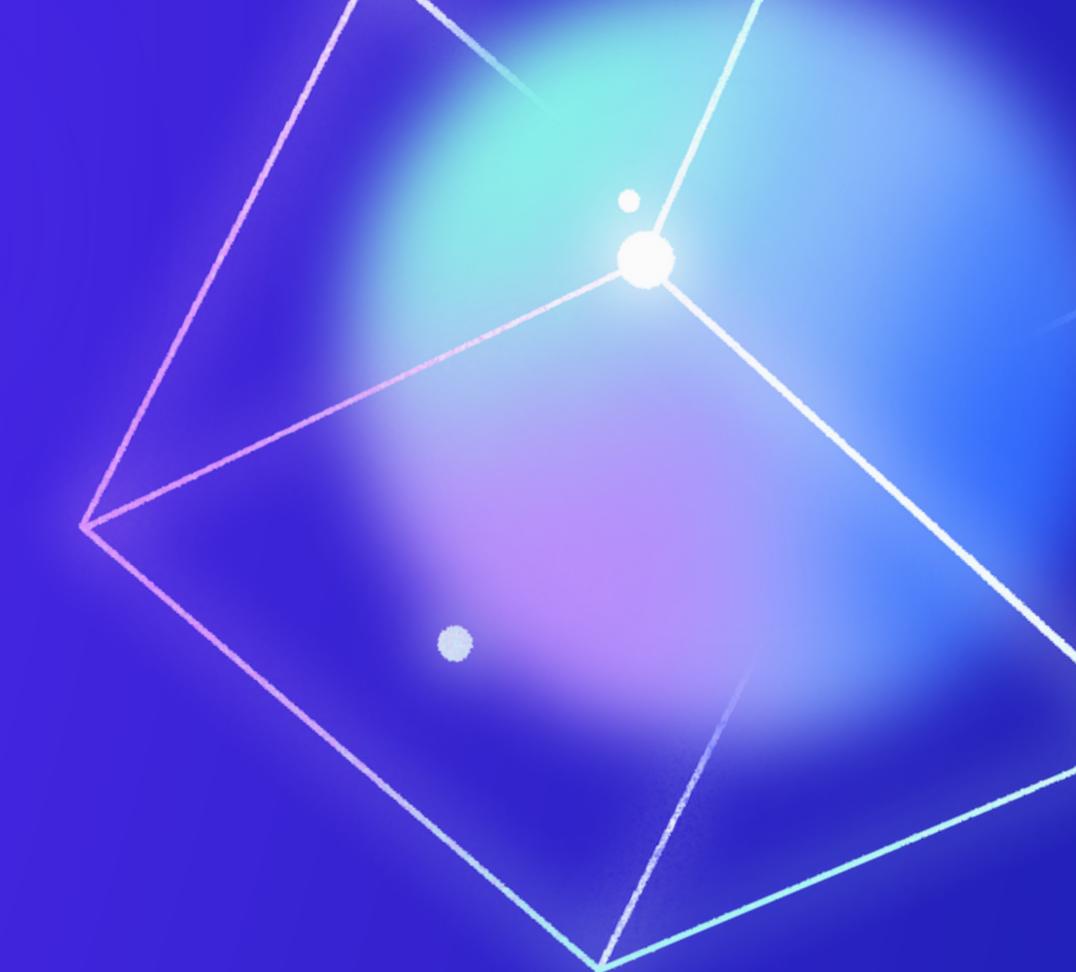
Asigna valores iniciales para los promedios m_1, m_2, \dots, m_k ;

Repetir

Asigna cada elemento t_i al clúster con el promedio más cercano.

Calcula un nuevo promedio para el clúster.

Hasta que el criterio de convergencia es alcanzado.

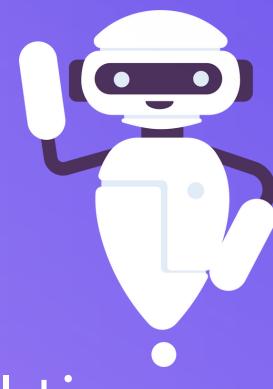


CARACTERÍSTICAS

Un valor típico para K, va de 2-10.

A pesar de dar buenos resultados, no es eficiente en términos de tiempo.

No maneja los valores atípicos correctamente.

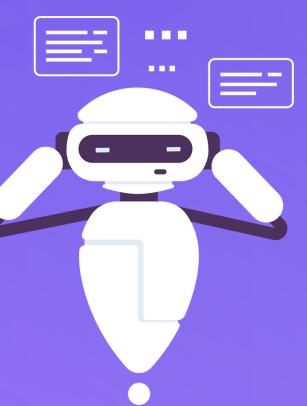


El tiempo de complejidad es de $O(tkn)$ donde t es el número de iteraciones.



K-means encuentra el óptimo local, pero falla al encontrar el óptimo global.

Aunque algunas versiones buscan mejorar este defecto.



K-means trabaja con valores numéricos.

5.5.3 ALGORITMO DE VECINOS MÁS CERCANOS



El KNN (K-Nearest Neighbors) es un algoritmo similar a la técnica del enlace simple.

Con este algoritmo serial, los elementos son fusionados iterativamente en clústeres existentes que estén cerca.

En este algoritmo, se utiliza un límite t el cuál es usado para determinar si los elementos serán añadidos a un clúster existente o a uno nuevo.

En este algoritmo se debe comparar cada elemento con cada elemento ya existente en un clúster. Por ello, su complejidad es de $O(n^2)$

Input:

$D = \{t_1, t_2, \dots, t_n\}$ //Conjunto de elementos

A //Matriz de adyacencia mostrando la distancia entre elementos.

Output:

K //Conjunto de clústeres

Algoritmo KNN:

$K_1 = \{t_1\}$

$K = \{K_1\}$

$k = 1;$

Para $i = 2$ hasta n

 Encontrar el t_m en algún cluster K_m tal que su $\text{dis}(t_i, t_m)$ es
 la más pequeña.

 Si $\text{dist}(t_i, t_m) \leq t$, entonces

$K_m = K_m \cup t_i$

 Si no

$k = k+1;$

$K_k = \{t_i\}$

EJEMPLOS





BIBLIOGRAFÍA

LIBRO

Dunham, Margaret H. Data Mining Introductory and Advanced Tools. 2nd ed., Morgan Kaufmann, 2019.

EJEMPLOS

- Alvarez Estrada José E.. Agrupamiento (clustering) en Knime mediante K-means. Noviembre 17, 2021. [Archivo de video] Disponible: https://www.youtube.com/watch?v=EFLy4Ph_7dl
- Maheswari, R. Data Analytics – Classification – KNN – Knime By Dr. Maheswari.R Noviembre 8, 2022. [Archivo de video] Disponible: <https://www.youtube.com/watch?v=euGjDNyAa7E>

MUCHAS GRACIAS
POR VER ESTA PRESENTACIÓN

