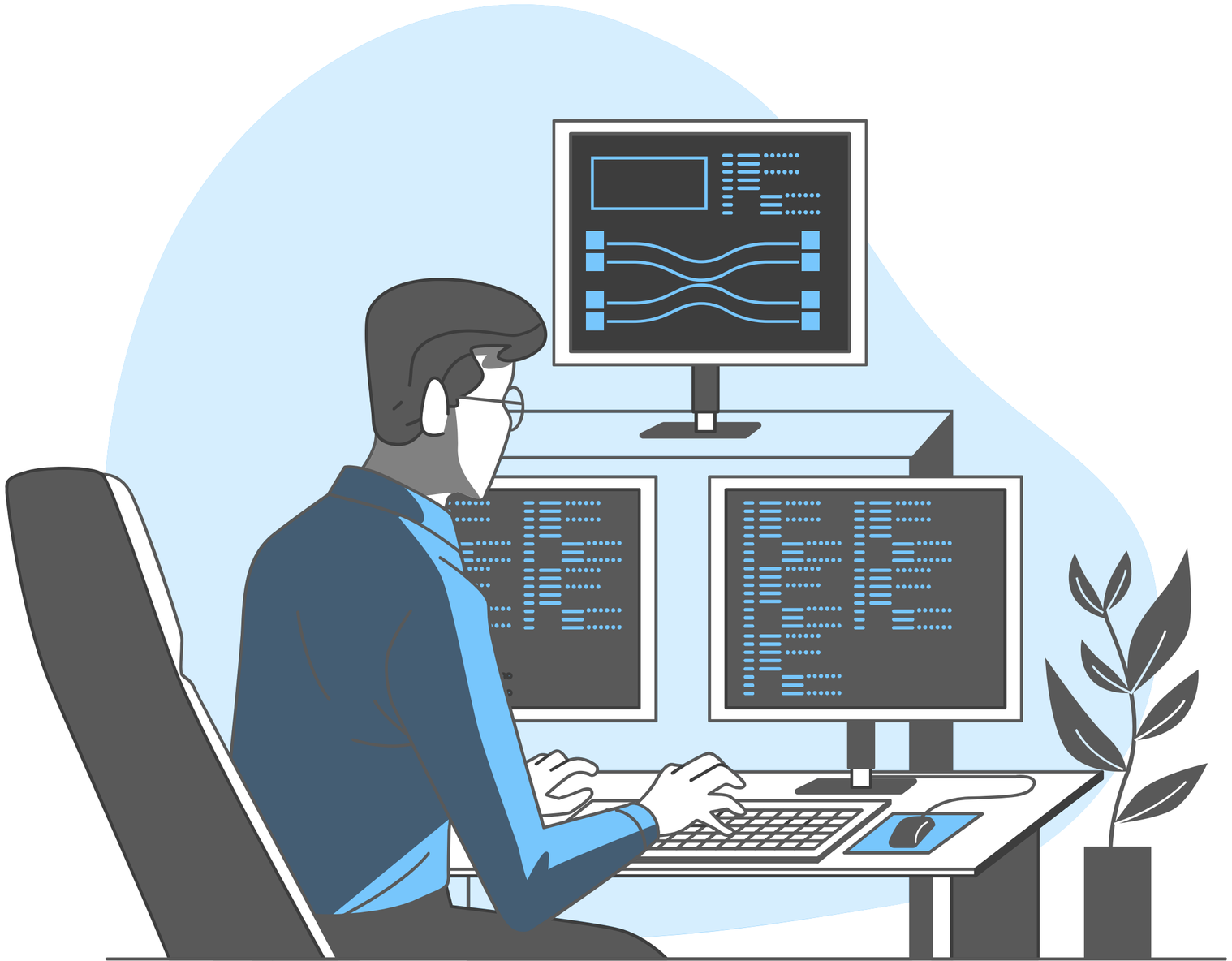


Fuzzy C-Means: Una técnica avanzada de clustering

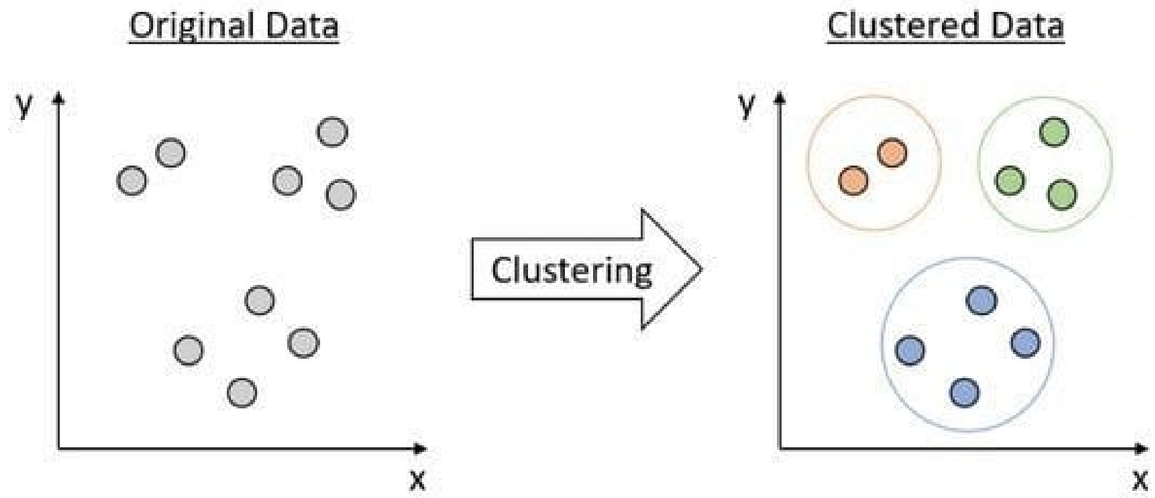
Universidad Politécnica Salesiana
Análisis Multivariado
John Sanmartín
Periodo - 62

Contenido

Introducción

El clustering es una técnica de aprendizaje no supervisado que nos permite agrupar datos similares en conjuntos.

- El clustering nos ayuda a descubrir patrones y estructuras ocultas en los datos sin necesidad de etiquetas predefinidas.
- Es una herramienta esencial en el aprendizaje automático, ya que nos permite explorar grandes conjuntos de datos y encontrar agrupaciones significativas.
- El clustering se aplica en una amplia variedad de campos, como la segmentación de clientes, la detección de anomalías, el análisis de texto y muchas otras áreas donde la comprensión de las agrupaciones de datos es fundamental.



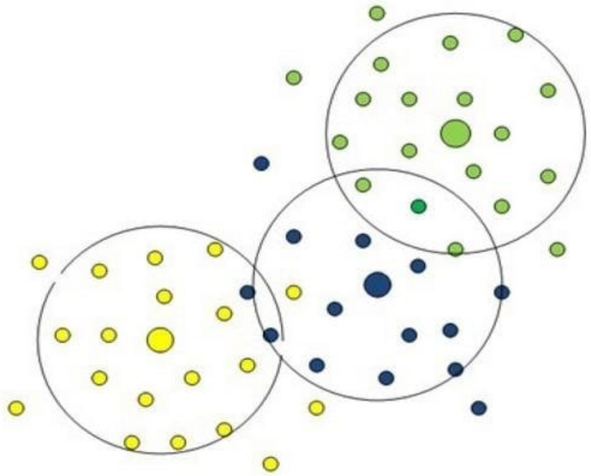
Conceptos clave de Fuzzy C-Means

Fuzzy C-Means es una técnica de clustering ampliamente utilizada y efectiva debido a su capacidad para asignar grados de pertenencia difusos

Pertenencia difusa

Es una característica fundamental de Fuzzy C-Means. En lugar de asignar de manera binaria un punto de datos a un único cluster, Fuzzy C-Means asigna grados de pertenencia difusos a cada punto de datos.

...



Centroides difusos

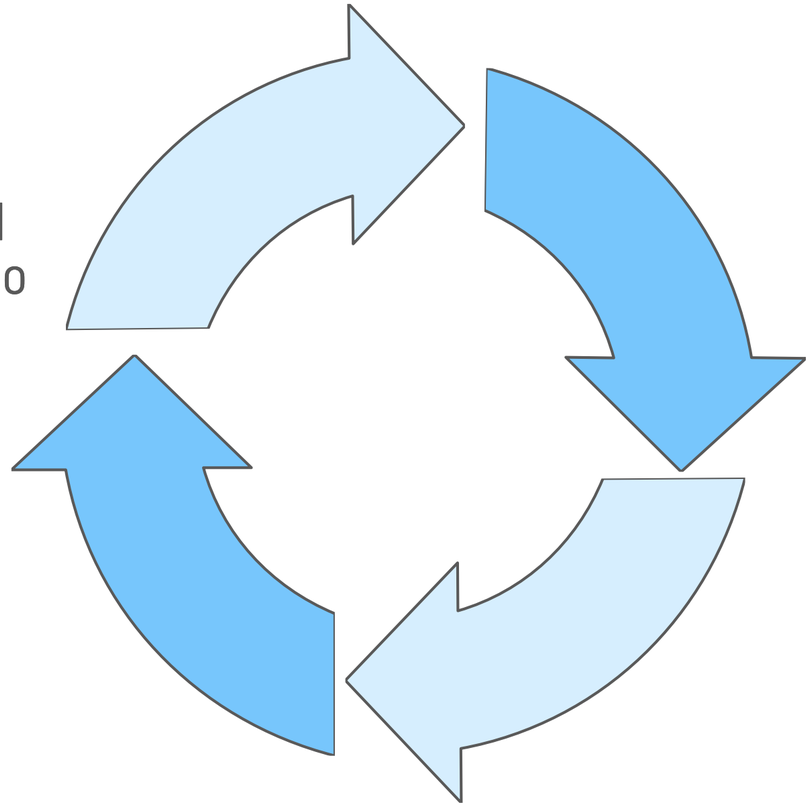
Son puntos representativos de cada cluster y se calculan considerando los grados de pertenencia difusos de los puntos de datos.

...

El algoritmo FCM sigue los siguientes pasos:

Converge

epiten hasta que se alcance un criterio de convergencia, que puede ser una cantidad máxima de iteraciones o una pequeña variación en los centroides.



Inicialización

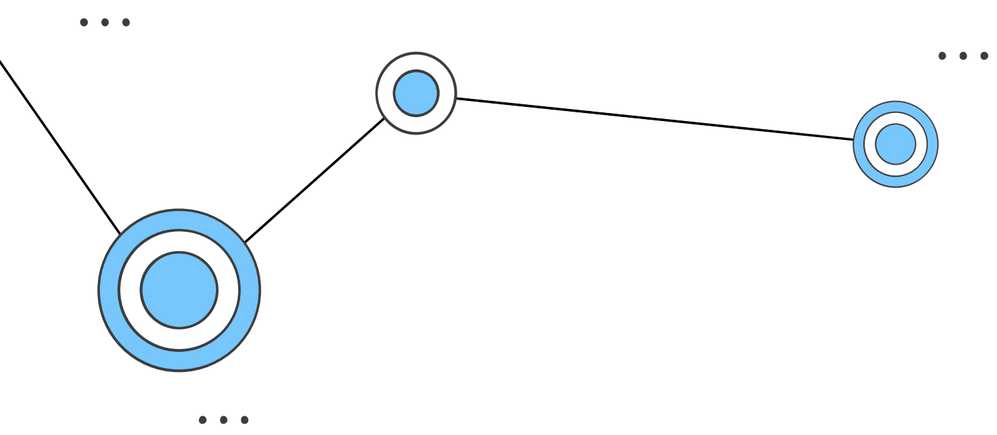
Se selecciona el número de clusters deseados y se eligen aleatoriamente los centroides iniciales para cada cluster

Actualización de centroides

Se recalculan los centroides para cada cluster utilizando los grados de pertenencia difusos de los puntos de datos.

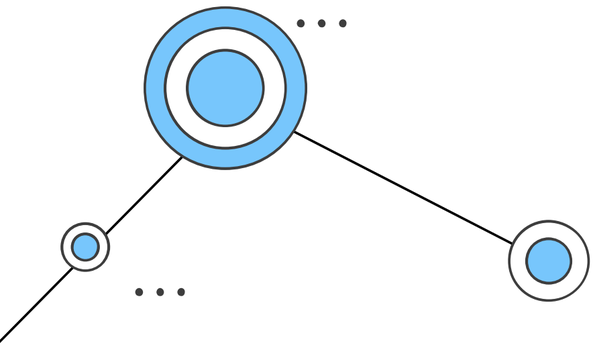
Cálculo de grados de pertenencia

Se calculan los grados de pertenencia difusos para cada punto de datos en relación con cada cluster.

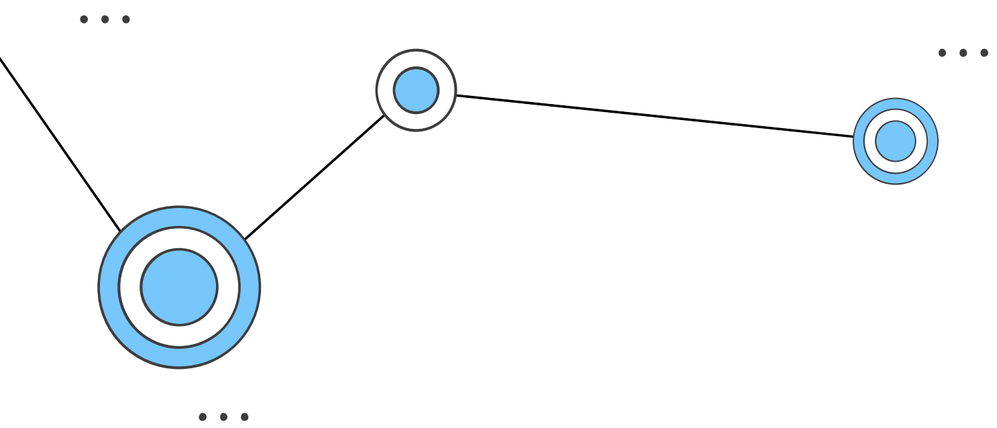


Cálculo de grados de pertenencia difusos (U)

$$U_{ij} = \left(\sum_{k=1}^{K} \left(\frac{\text{distancia}(x_i, C_j)}{\text{distancia}(x_i, C_k)} \right)^{\frac{2}{p-1}} \right)^{-1}$$



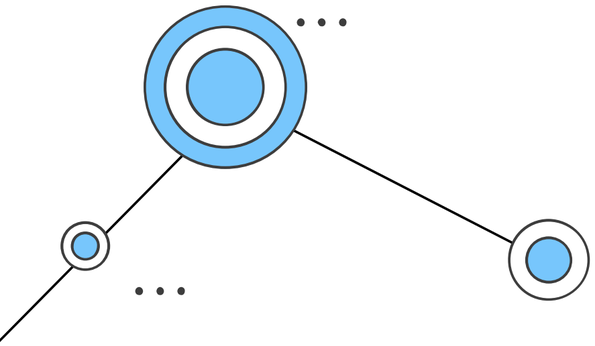
- **Uij** representa el grado de pertenencia difuso del punto de datos **xi** al cluster **Cj**
- Distancia**(a,b)** es una medida de distancia entre los puntos **a** y **b**.
- **K** es el número total de clusters.
- **p** es el valor difuso.



Actualización de centroides (C)

$$C_j = \frac{\sum_{i=1}^{n} \left(U_{ij}^{p} \cdot x_i \right)}{\sum_{i=1}^{n} U_{ij}^{p}}$$

- **Cj** representa las coordenadas del centroide del cluster **j**.
- **Uij** representa el grado de pertenencia difuso del punto de datos **xi** al cluster **Cj**.
- **n** es el número total de puntos de datos en el conjunto **X**.



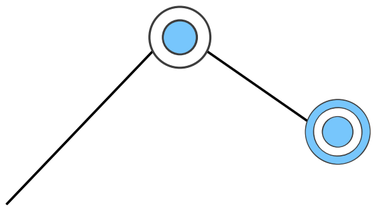
Ventajas

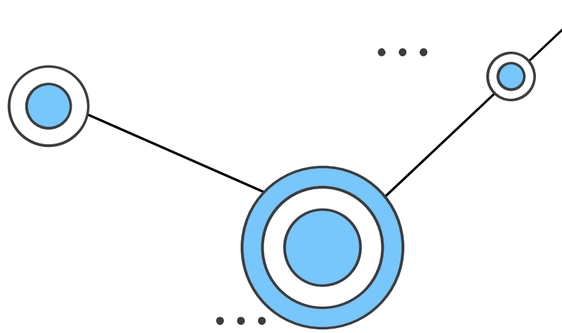
Flexibilidad en la asignación de puntos de datos

Permite asignar grados de pertenencia difusos a cada punto de datos en lugar de asignarlos a un solo cluster. Esto ofrece una mayor flexibilidad y capacidad para lidiar con situaciones en las que los puntos de datos pueden pertenecer a múltiples clusters simultáneamente.

Robustez frente al ruido y datos atípicos

Debido a la asignación de grados de pertenencia difusos, los puntos de datos anómalos o ruidosos no afectan drásticamente los resultados del clustering





Eficiencia computacional

Es computacional mente eficiente en comparación con otros algoritmos de clustering más complejos. Esto lo hace adecuado para aplicaciones en tiempo real o con limitaciones de recursos computacionale s.

Desventajas

Interpretación de los grados de pertenencia

En lugar de asignaciones claras y binarias, los grados de pertenencia difusos requieren una interpretación más cuidadosa y pueden generar cierta ambigüedad en la asignación de puntos de datos a clusters específicos.

Sensibilidad a la elección de parámetros

El rendimiento del algoritmo Fuzzy C-Means puede depender en gran medida de la elección adecuada de los parámetros, como el número de clusters y el valor difuso (p).

Referencias

Kassambara, A. (s/f). Fuzzy C-Means Clustering Algorithm. Datanovia.com. Recuperado de: https://www.datanovia.com/en/lessons/fuzzy-clustering-essentials/fuzzy-c-means-clustering-algorithm/

azminetoushikwasi. (2022, julio 1). Different Clustering Techniques and Algorithms. Kaggle.com; Kaggle. https://kaggle.com/code/azminetoushikwasi/different-clustering-techniques-and-algorithms

¡Gracias!