Ciencia de datos

Santiago Murillo Rendón - Reinel Tabares Soto

UTraining
Entrenamiento en Innovación y Tecnología
UNIVERSIDAD DE ANTIOQUIA
Universidad de Caldas
Ubicua Technology

Agenda

1. Métodos no supervisados y métodos semi-supervisados.
2. Agrupamiento Jerárquico. Hierarchical clustering.
3. Agrupamiento por k-medias. K-means clustering.

Tipos de Aprendizaje

- **Aprendizaje semisupervisado**: Aprendizaje usando tanto datos etiquetados como no etiquetados, en general una pequeña cantidad de datos etiquetados junto a muchos datos no etiquetados.

- **Aprendizaje no supervisado**: Aprendizaje donde un modelo se ajusta a las observaciones sin variable objetivo. Utilizado para agrupar, asociar o detectar anomalías.

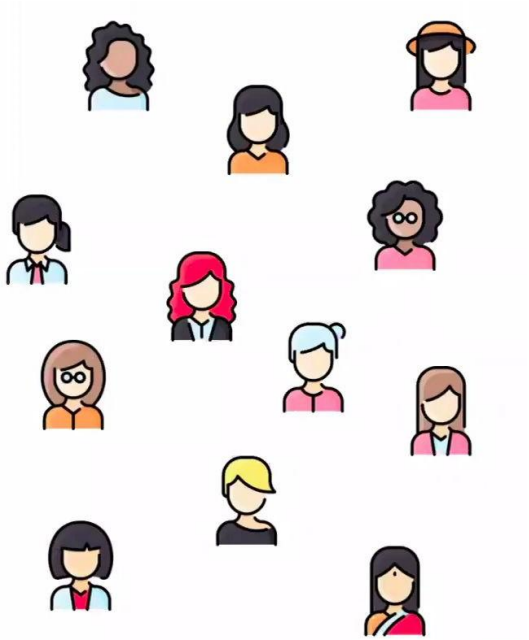
Aprendizaje no supervisado

Encontrar estructura

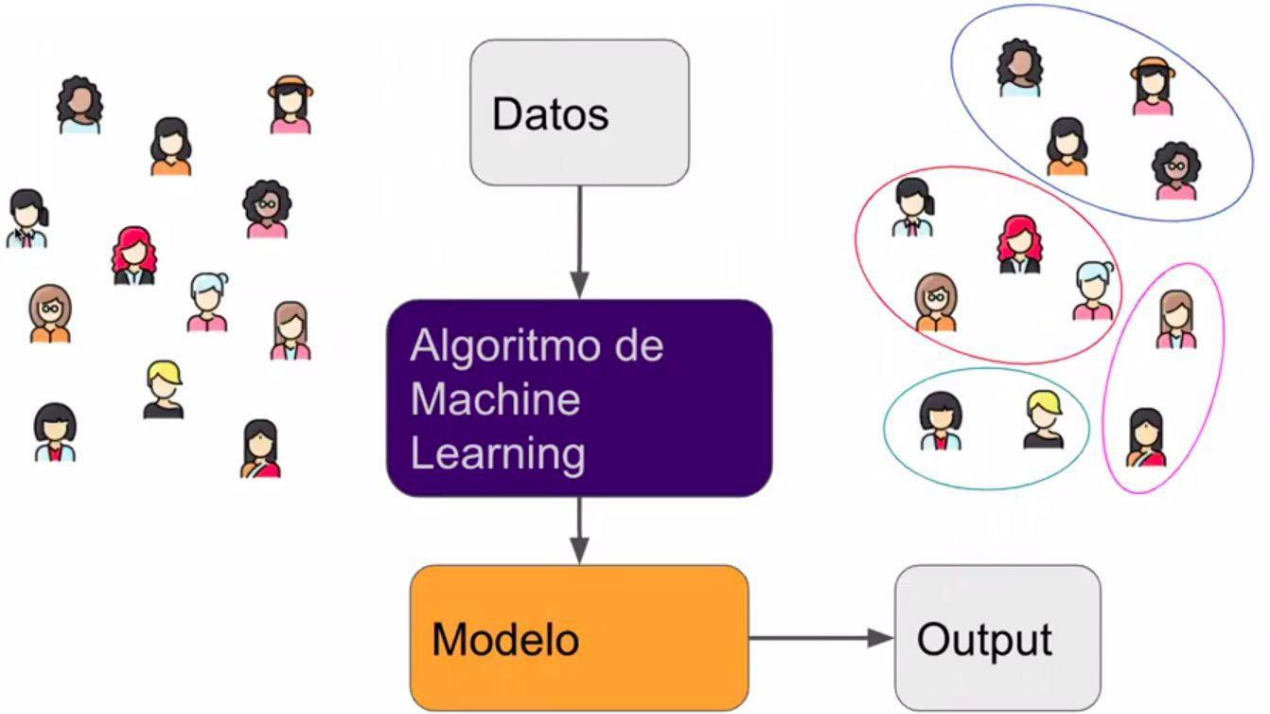
- Clustering (agrupación)
- Detección de anomalías

Datos de entrenamiento

- Muestras no se encuentran etiquetadas



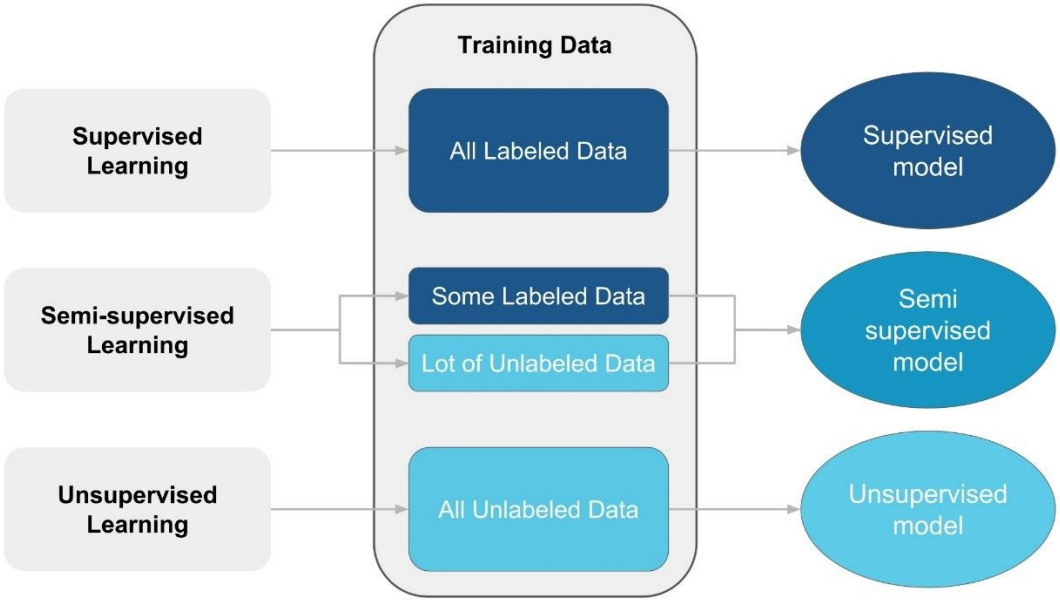
Aprendizaje no supervisado



Aprendizaje no supervisado

- No contamos con las clases o valores observados (reales) de la variable objetivo
- Intentamos entender los datos
- Buscamos estructuras o patrones
- Evaluación indirecta o cualitativa
  - ¿Puedo hacer algo útil con esto?
  - ¿Tiene sentido esto?

Comparando el aprendizaje supervisado, semisupervisado y no supervisado

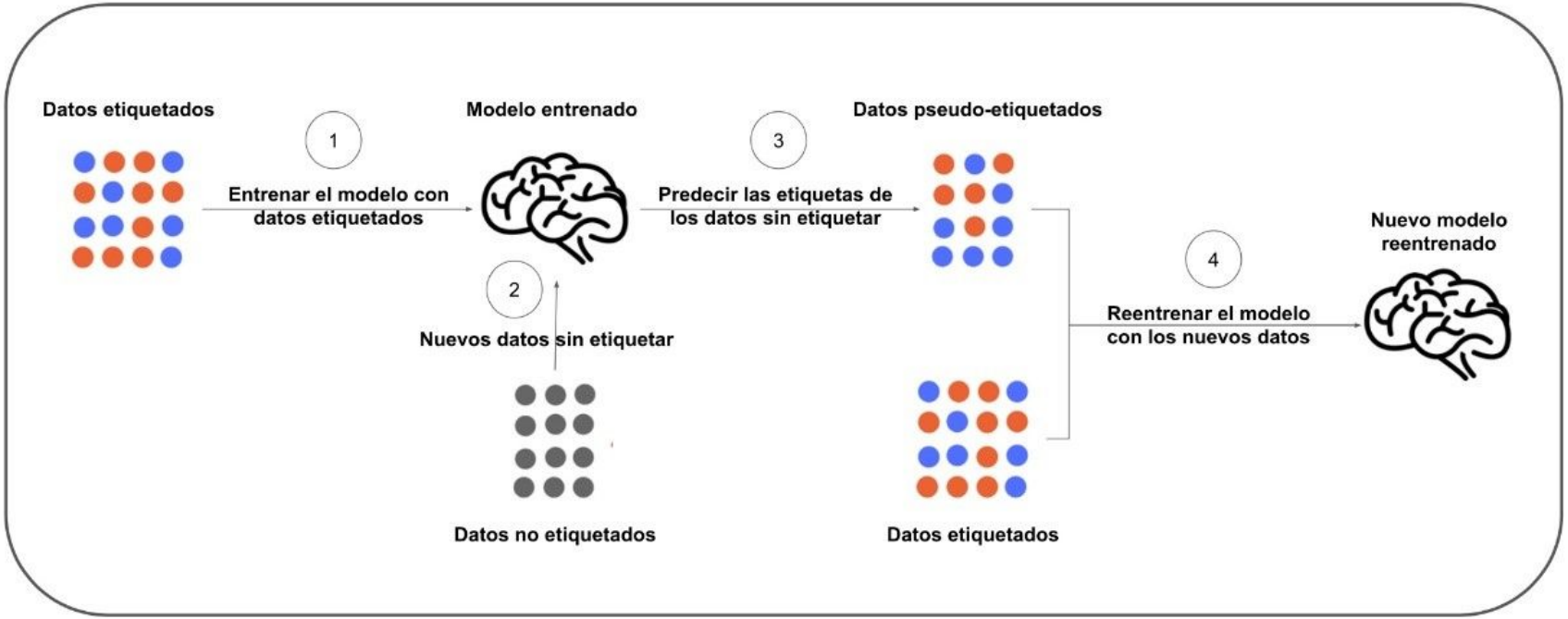


https://empresas.blogthinkbig.com/semi-supervised-learningel-gran-desconocido/

Aprendizaje semisupervisado

- Útil en problemas de clasificación
- Conjunto de datos:
  - CON las etiquetas (pocos)
  - SIN las etiquetas (muchos)
- Combinar datos etiquetados y no etiquetados para lograr tener más datos etiquetados
- Un camino medio entre aprendizaje supervisado y no supervisado

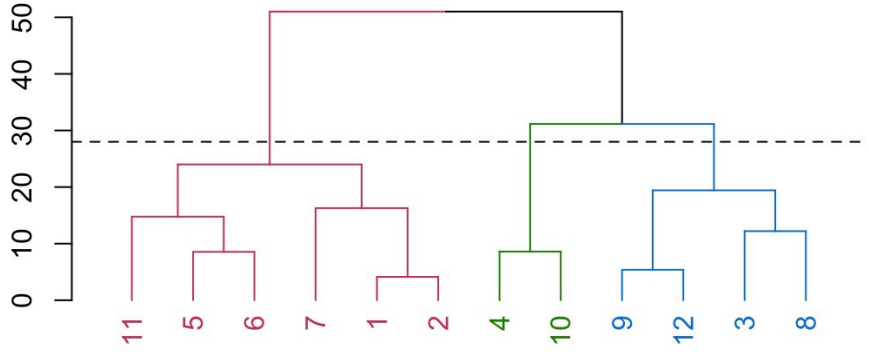
Aprendizaje semisupervisado



https://empresas.blogthinkbig.com/semi-supervised-learningel-gran-desconocido/

Técnicas de Aprendizaje no supervisado:

Análisis Jerárquico



Características

- El agrupamiento jerárquico es un método de aprendizaje no supervisado para agrupar puntos de datos.
- El algoritmo crea grupos midiendo las diferencias entre los datos.
- Este método se puede utilizar en cualquier dato para visualizar e interpretar la relación entre puntos de datos individuales.

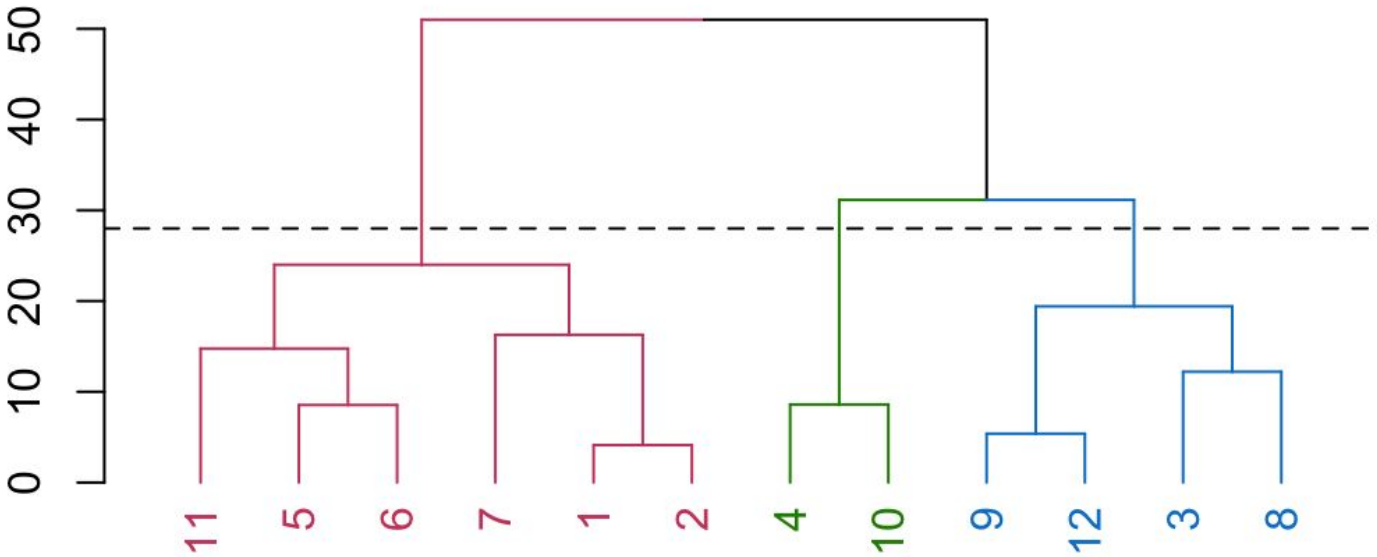
Funcionamiento

- El clustering jerárquico sigue un enfoque de abajo hacia arriba.
- Trata cada punto de los datos como su propio grupo.
- Luego, une los grupos que tienen la distancia más corta entre ellos para crear grupos más grandes.
- Este paso se repite hasta que se forma un grupo grande que contiene todos los puntos de datos.

Funcionamiento

- Una vez se ha creado el Dendograma, se puede decidir según su gráfico la cantidad de Clusters más apropiada.
- Con la cantidad de clusters, se hace la separación de los datos según corresponda.
- Graficación para visualizar los clusters de datos encontrados

El dendrograma

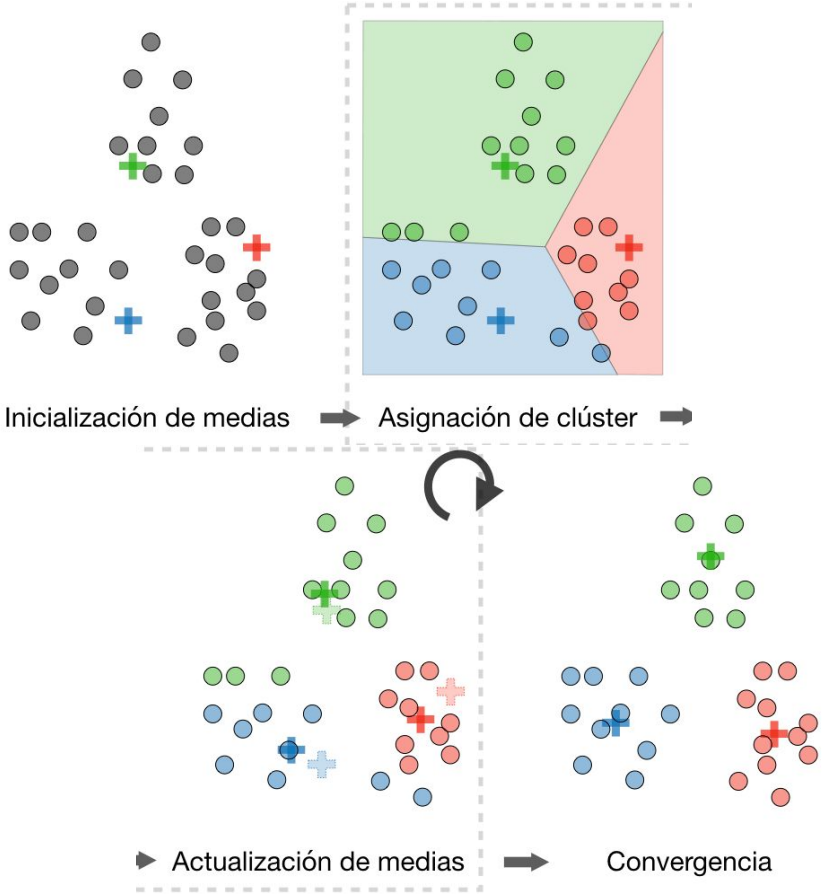


Aprendizaje semisupervisado

- Utilizado en problemas de clasificación.
- Se irán generando nuevas etiquetas sólo con las predicciones más confiables. Y se repite el proceso (incluyendo las nuevas etiquetas) hasta que ya no queden datos sin etiquetar (self-training).
- Lograremos pasar de tener pocos datos etiquetados, a muchos datos etiquetados.

Técnicas de Aprendizaje no supervisado:

K-means Clustering



Técnicas de Aprendizaje no supervisado: K-means Clustering. Conceptos base

- Algoritmo para aprendizaje no supervisado
- Sirve para particionar un dataset en K grupos o Clusters
- El valor “final” de K lo determina el analista o científico de datos
- Cada uno de los Clusters será representado por un centroide
- Un centroide es la media de los puntos asignados al Cluster

Técnicas de Aprendizaje no supervisado: K-means Clustering. Pasos de ejecución del algoritmo

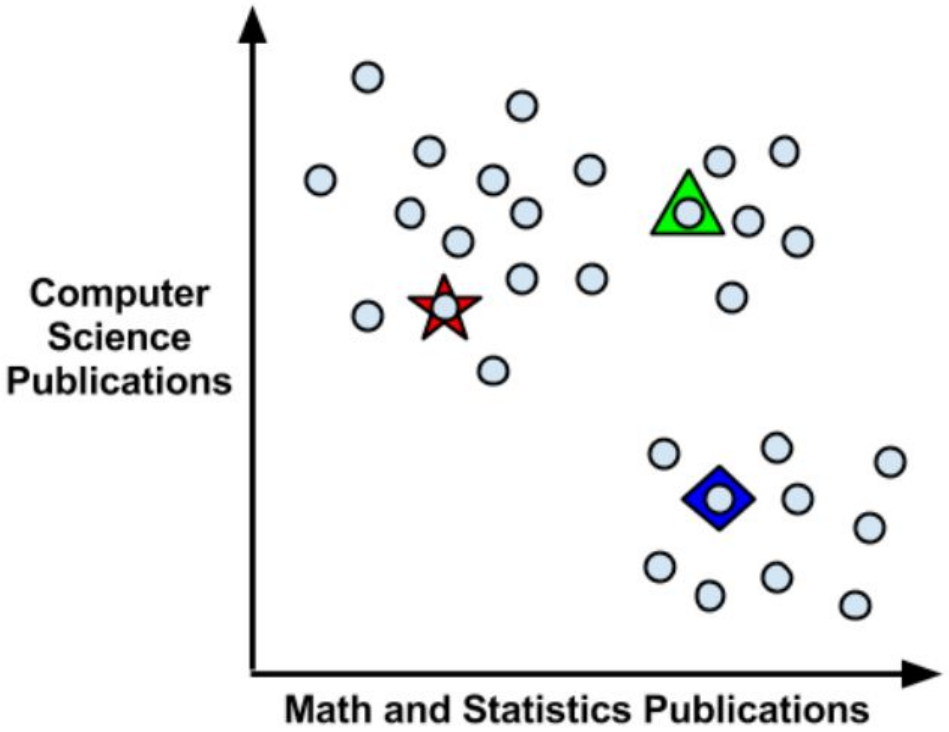
- Decidir la cantidad de K clusters a realizar (analista o científico de datos)

- Seleccionar de forma arbitraria K puntos iniciales (primeras posiciones de los centroides)

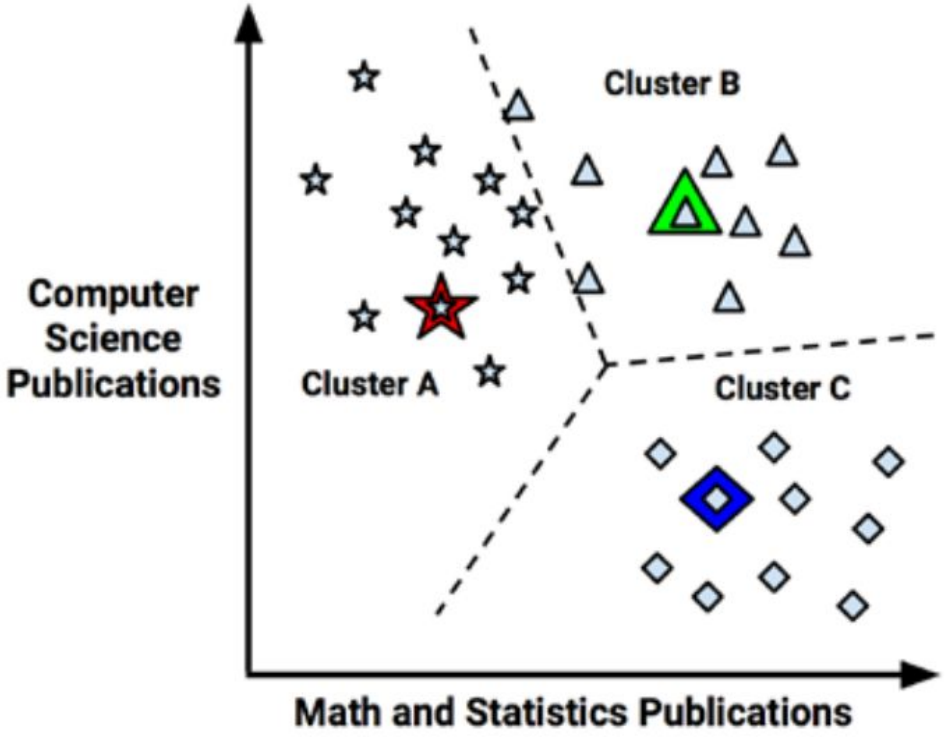
Técnicas de Aprendizaje no supervisado: K-means Clustering. Pasos de ejecución del algoritmo

- Asignar cada una de las muestras de los datos al centroide más cercano (distancia euclidiana entre cada muestra y los centroides)
- Actualizar las posiciones de los centroides según la media de las distribuciones de las muestras actuales
- Repetir los dos pasos anteriores hasta alcanzar la cantidad máxima de iteraciones, o hasta que las asignaciones de clústeres dejen de cambiar

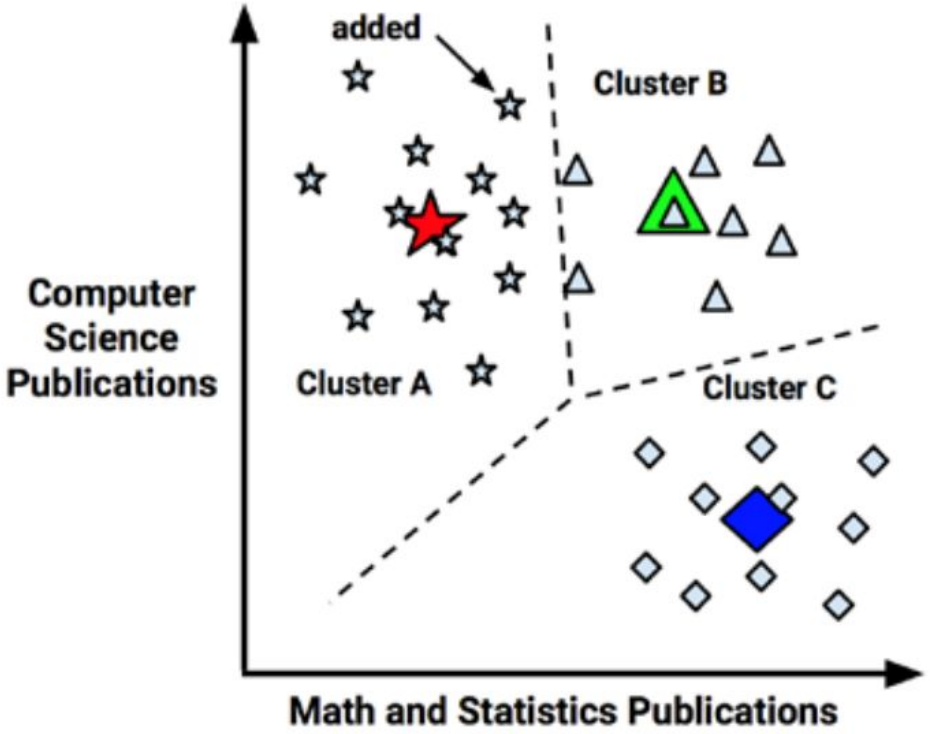
Explicación gráfica



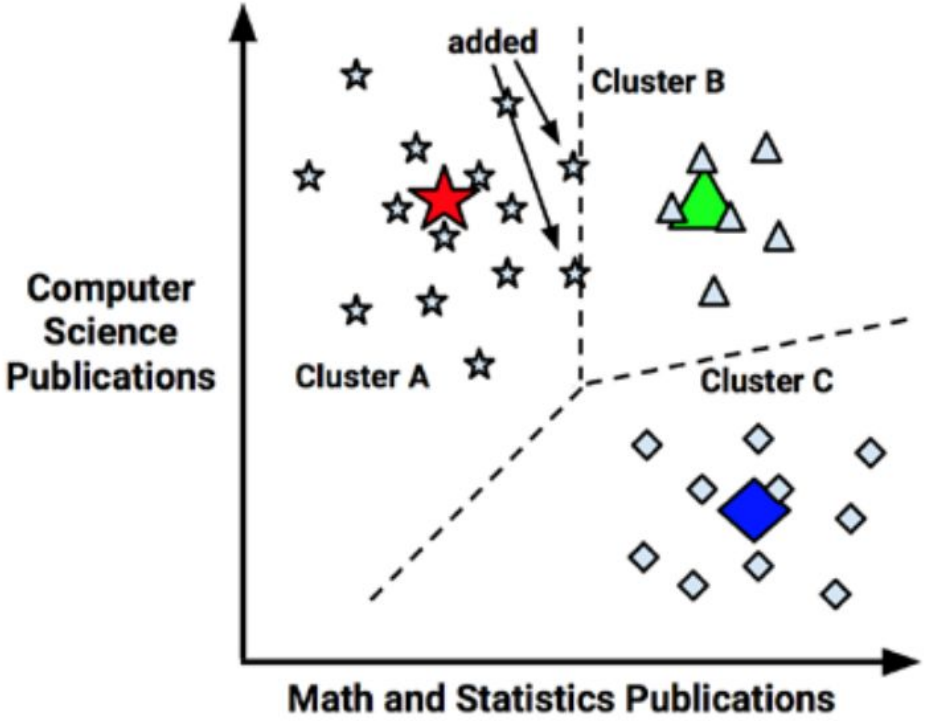
Explicación gráfica



Explicación gráfica



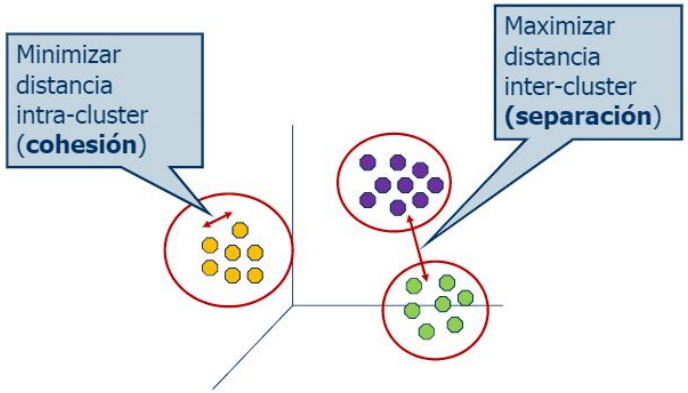
Explicación gráfica

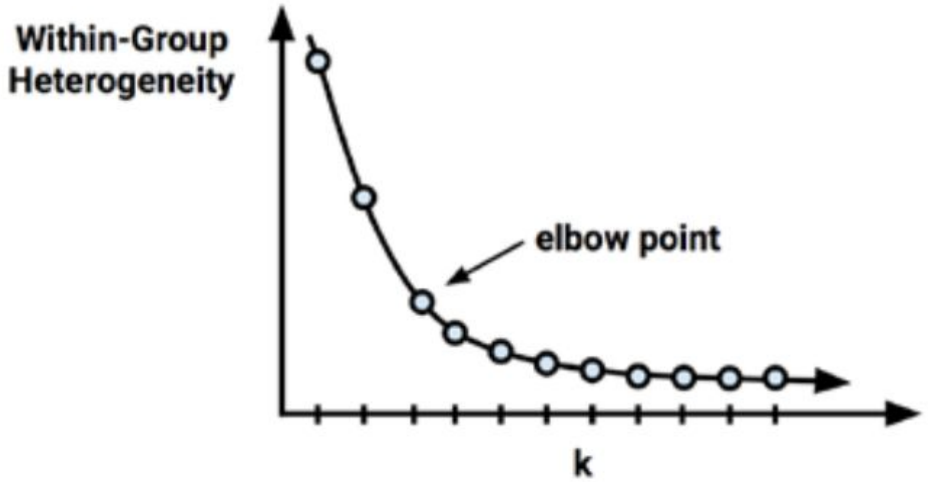


Selección de K clusters

- Forma arbitraria
- Conocimiento del dominio - requerimientos especiales de negocio
- Método del codo

Método del codo





Ventajas

- Usa principios simples explicables
- Flexible, se puede ajustar con facilidad
- Buen desempeño en muchos casos
- Visible en 2 y 3 dimensiones

Desventajas

- No es tan sofisticado como otros algoritmos más modernos
- No garantiza encontrar los clusters óptimos
- Podría requerir "adivinar" cuántos clusters hay en los datos
- No es ideal para clusters no circulares