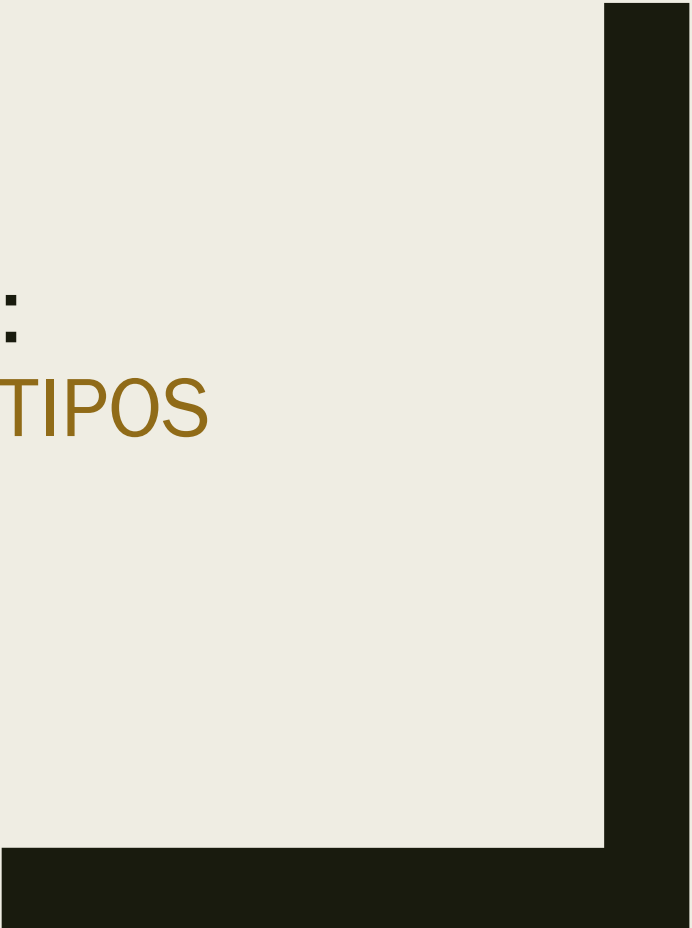




ANALÍTICA AVANZADA DE DATOS: AGRUPAMIENTO BASADO EN PROTOTIPOS

A. Alejandra Sánchez Manilla
asanchezm.q@gmail.com



Tipos de algoritmos de agrupamiento

Agrupamiento basado en densidad:

- En este enfoque, los datos se agrupan en áreas con una alta concentración de puntos rodeadas por áreas con baja concentración de puntos.
- El algoritmo identifica las regiones densas y las considera como grupos. Una ventaja de este tipo de algoritmo es que los grupos pueden tener cualquier forma, no están limitados por suposiciones preestablecidas.
- Sin embargo, los valores atípicos suelen ser ignorados en este tipo de agrupamiento.

Agrupamiento basado en la distribución:

- Este enfoque considera que cada punto de datos tiene una probabilidad de pertenecer a un determinado grupo.
- Se establece un punto central y a medida que la distancia de un punto de datos al centro aumenta, disminuye la probabilidad de que forme parte de ese grupo.
- Este tipo de agrupamiento es útil cuando no se conoce la distribución exacta de los datos.

Tipos de algoritmos de agrupamiento

Agrupamiento basado en centroides:

- El agrupamiento basado en centroides es uno de los más populares
- En este enfoque, los puntos de datos se asignan a grupos en función de la distancia al cuadrado entre cada punto y los centroides
- Los centroides son puntos representativos que se encuentran en el centro de cada grupo
- Este tipo de agrupamiento es rápido y eficiente, pero puede ser sensible a los parámetros iniciales que se le proporcionen

Agrupamiento jerárquico:

- El agrupamiento jerárquico se utiliza en conjuntos de datos jerárquicos, como bases de datos empresariales o taxonomías.
- Este enfoque construye un árbol de grupos, organizando los datos de arriba hacia abajo en una estructura jerárquica.
- Si bien es más restrictivo en comparación con otros enfoques, es adecuado para conjuntos de datos específicos que tienen una estructura jerárquica clara.

Cuando usar agrupamiento



Detección de anomalías: Detectar valores atípicos o anomalías en los datos. Al identificar grupos y establecer límites, se puede determinar si un punto de datos es inusual o no.



Selección de características: Si no estás seguro de qué características utilizar en tu modelo de aprendizaje automático, el agrupamiento puede ayudarte a descubrir patrones y tendencias en los datos. Estos patrones pueden proporcionar información sobre qué características son más relevantes para tu problema.



Exploración de datos desconocidos: A través del agrupamiento, se pueden descubrir conexiones y relaciones en los datos que podrían no ser evidentes de otra manera. Esto puede conducir a nuevos conocimientos y perspectivas en el análisis de los datos.



Aplicaciones en el mundo real: Se puede utilizar para detectar fraudes en seguros, categorizar libros en una biblioteca, segmentar clientes en el campo del marketing, analizar patrones sísmicos en el estudio de terremotos, o en la planificación urbana para identificar patrones de uso del suelo y necesidades de infraestructura.

Hyper-Sphere Clustering (HSC)

El algoritmo de agrupamiento Hyper-Sphere Clustering (HSC) es un enfoque específico de agrupamiento basado en prototipos hiperesféricos

Este algoritmo se utiliza para dividir un conjunto de datos en clústeres utilizando hiperesferas como representantes de cada clúster

Hyper-Sphere Clustering (HSC)

Una descripción general del HSC:

1. Inicialización:

- Se selecciona un número de prototipos hiperesféricos (centros) aleatoriamente o utilizando algún otro método de inicialización, como k-means
- Estos prototipos representarán los clústeres iniciales

2. Asignación:

- Cada objeto del conjunto de datos se asigna al clúster cuyo prototipo hiperesférico está más cercano en términos de distancia euclidiana
- La distancia se calcula desde el centro de la hiperesfera hasta el objeto
- Un objeto se asigna al clúster con el prototipo más cercano

Hyper-Sphere Clustering (HSC)

3. Actualización:

- Después de la asignación inicial, los prototipos hiperesféricos se actualizan recalculando los nuevos centros de cada hiperesfera
- Esto se hace tomando la media de las coordenadas de los objetos asignados a cada clúster

4. Reasignación:

- Una vez que los prototipos se han actualizado, los objetos se reasignan a los clústeres en función de los nuevos prototipos hiperesféricos
- Se repite el proceso de asignación y actualización hasta que se cumpla un criterio de convergencia, como un número máximo de iteraciones o una estabilidad en la asignación

Hyper-Sphere Clustering (HSC)

5. Resultado:

- Una vez que el algoritmo ha convergido, se obtiene el resultado final del agrupamiento
- Esto incluye los prototipos hiperesféricos, que representan los clústeres, y la asignación de los objetos a cada clúster

Hyper-Sphere Clustering (HSC)



El HSC se diferencia de otros enfoques de agrupamiento en que permite la formación de clústeres no convexos utilizando hiperesferas como representantes



Esto significa que puede encontrar clústeres con formas más complejas que las formas convexas



La elección del número óptimo de clústeres y la inicialización adecuada de los prototipos son consideraciones importantes al utilizar el algoritmo HSC

Agrupamiento basado en prototipos elipsoidales

El agrupamiento basado en prototipos elipsoidales es un enfoque de agrupamiento en el cual los grupos se representan mediante elipses o elipsoides en lugar de puntos individuales

A diferencia de otros métodos de agrupamiento que utilizan centroides o medoides como prototipos, este enfoque busca encontrar elipsoides que encapsulen y representen la forma y distribución de los datos en cada grupo

Agrupamiento basado en prototipos elipsoidales

El proceso de agrupamiento basado en prototipos elipsoidales generalmente sigue los siguientes pasos:

1. Inicialización:

- Se seleccionan aleatoriamente K elipsoides iniciales como prototipos para representar los grupos
- Estos prototipos pueden tener diferentes formas, tamaños y orientaciones

2. Asignación:

- Para cada objeto en el conjunto de datos, se calcula la distancia entre el objeto y cada prototipo elipsoidal
- El objeto se asigna al grupo representado por el elipsoide más cercano

Agrupamiento basado en prototipos elipsoidales

3. Actualización:

- Una vez que todos los objetos han sido asignados a sus respectivos grupos, se actualizan los prototipos elipsoidales
- Esto implica ajustar la forma, tamaño y orientación de cada elipsoide para que se ajuste mejor a los objetos asignados a su grupo

4. Convergencia:

- Los pasos de asignación y actualización se repiten hasta que se alcance un criterio de convergencia, como una cantidad máxima de iteraciones o una mejora mínima en la función de costo

Agrupamiento basado en prototipos elipsoidales



La asignación y actualización se realizan iterativamente para optimizar los prototipos elipsoidales y mejorar la representación de los grupos



El objetivo es encontrar elipsoides que maximicen la coherencia intra-cluster y minimicen la dispersión inter-cluster

Agrupamiento basado en prototipo


En resumen:

Hiperesféricos



- Utiliza hiperesferas como representantes de los clústeres
- El proceso implica la asignación inicial, la actualización iterativa de los prototipos y la reasignación hasta que se alcance la convergencia
- Esto permite la formación de clústeres no convexos y puede ser útil en el análisis de datos con estructuras complejas

Elipsoidales

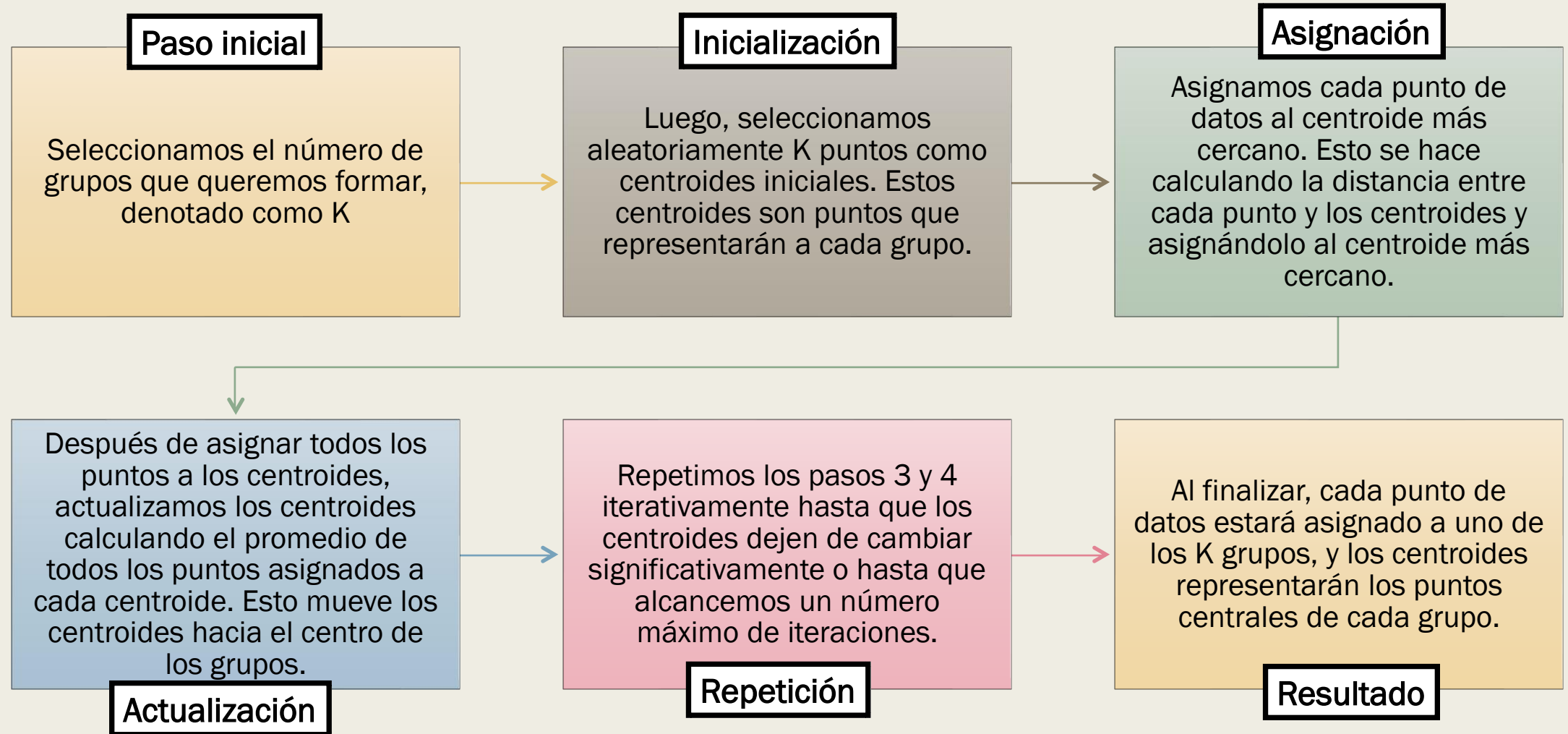
- Busca encontrar elipsoides que representen la forma y distribución de los datos en cada grupo, asignando objetos a los grupos en función de la proximidad a los elipsoides y actualizando los elipsoides para mejorar su ajuste a los objetos asignados.
- Este enfoque permite una mayor flexibilidad en la forma y tamaño de los grupos, ya que los elipsoides pueden capturar estructuras más complejas que los prototipos simples



ALGUNOS ALGORITMOS DE AGRUPAMIENTO



K-means



DBSCAN (Density-Based Spatial Clustering of Applications with Noise)

Es un método de agrupamiento basado en la densidad de los puntos de datos en un espacio

A diferencia del algoritmo K-means, DBSCAN no requiere que se especifique el número de grupos de antemano, sino que puede descubrir automáticamente la cantidad óptima de grupos en función de la distribución de los datos

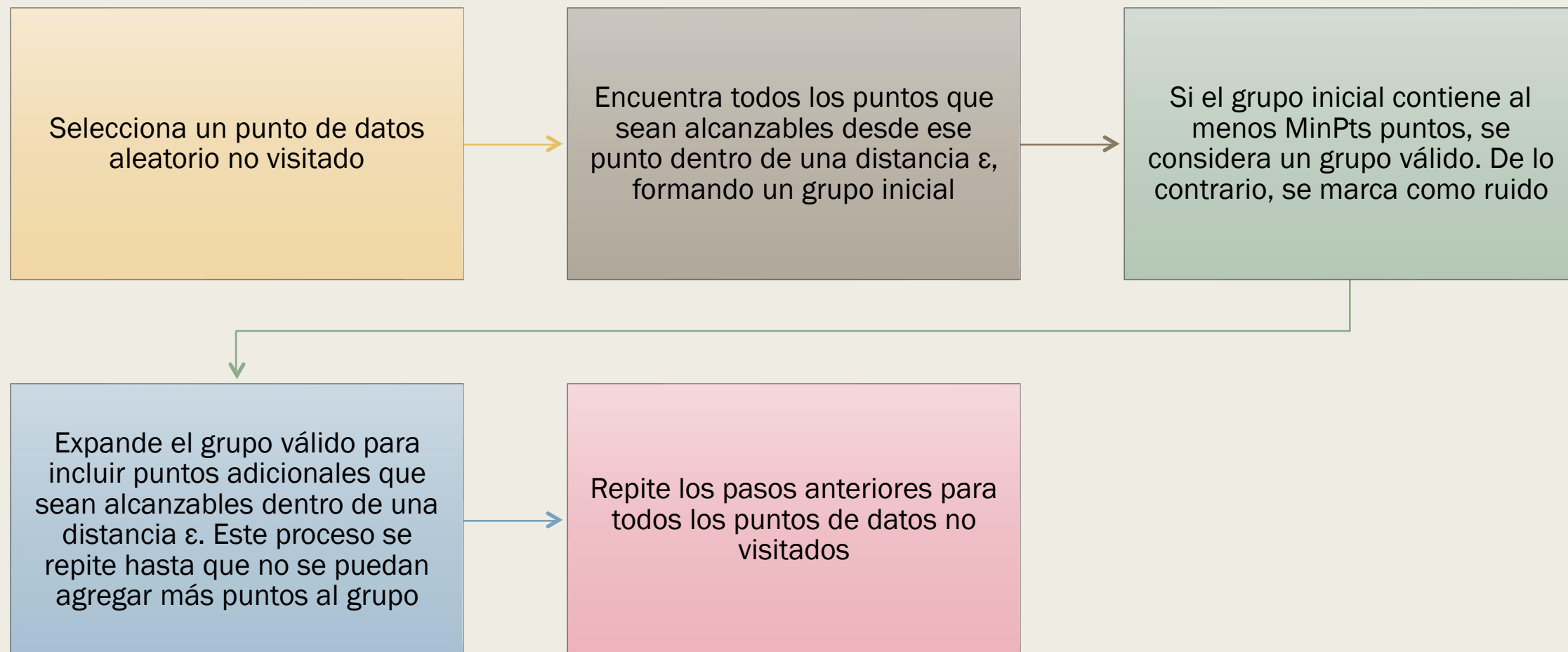
La idea principal detrás de DBSCAN es agrupar puntos de datos que estén cerca unos de otros en términos de distancia y densidad

DBSCAN

Para ello, el algoritmo define dos parámetros principales:

- **Épsilon (ϵ):** Es la distancia máxima que determina la vecindad de un punto. Todos los puntos que se encuentren dentro de una distancia ϵ de otro punto se consideran vecinos
- **Mínimo número de puntos (MinPts):** Es la cantidad mínima de puntos vecinos que deben encontrarse dentro de una distancia ϵ para que un punto sea considerado central

DBSCAN



DBSCAN

DBSCAN clasifica los puntos de datos en tres categorías:

1. **Puntos núcleo:** Son puntos que tienen al menos MinPts puntos dentro de una distancia ϵ
2. **Puntos frontera:** Son puntos que tienen menos de MinPts puntos dentro de una distancia ϵ , pero se encuentran dentro de la vecindad de un punto núcleo
3. **Puntos ruido:** Son puntos que no pertenecen a ningún grupo y no tienen suficientes puntos vecinos

DBSCAN



Es especialmente útil para descubrir grupos de diferentes formas y tamaños, y es robusto ante valores atípicos



Además, no requiere especificar el número de grupos de antemano, lo que lo hace muy adecuado cuando no se tiene información previa sobre la estructura de los datos

Algoritmo de Mezcla Gaussiana

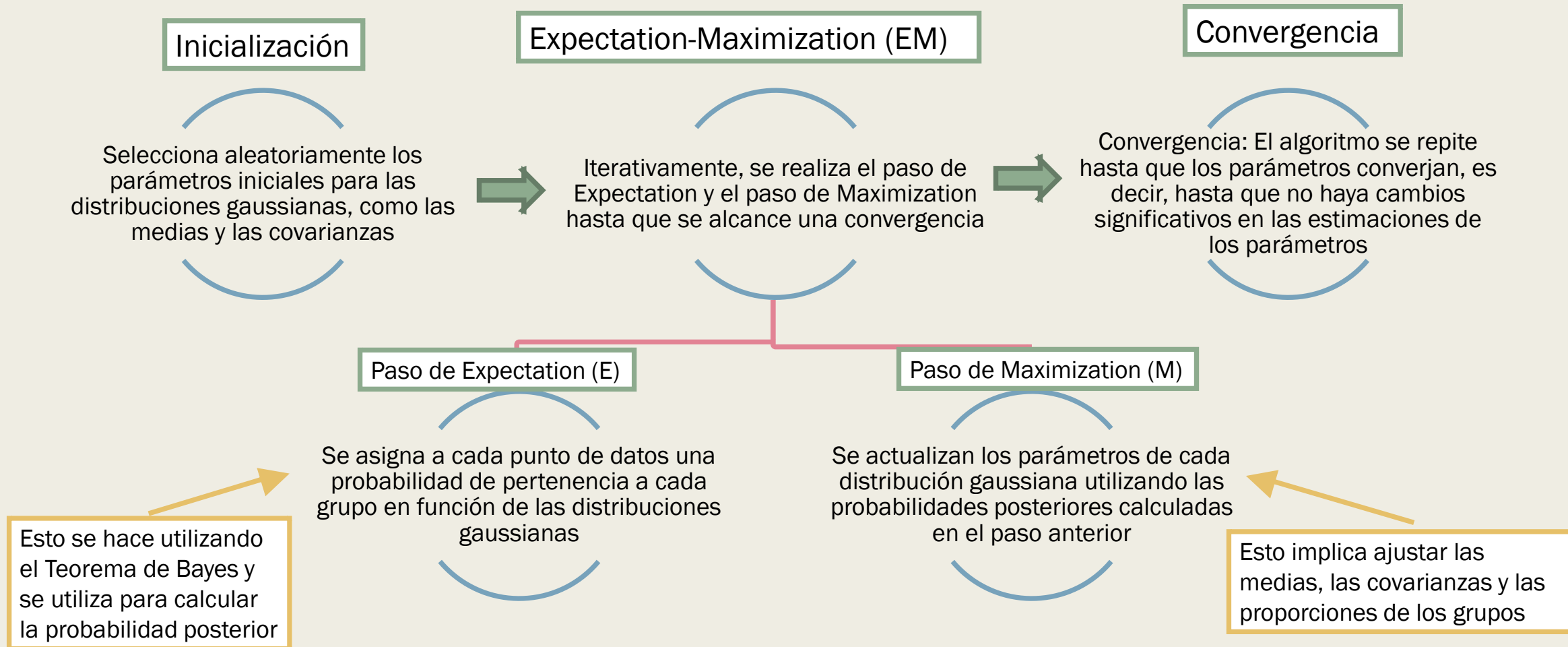
Se basa en el concepto de modelos probabilísticos y asume que los datos provienen de una combinación de varias distribuciones gaussianas o normales

La idea principal del algoritmo es encontrar la mejor manera de representar los datos como una mezcla de distribuciones gaussianas

Cada distribución gaussiana representa un grupo o cluster en el conjunto de datos

El objetivo del algoritmo es estimar los parámetros de cada distribución gaussiana, como la media y la covarianza, y determinar la proporción de datos que pertenecen a cada grupo

Algoritmo de Mezcla Gaussiana



Algoritmo de Mezcla Gaussiana



Al finalizar el algoritmo, cada punto de datos se asigna al grupo con la probabilidad más alta de pertenencia



Esto permite obtener una segmentación o agrupamiento de los datos en función de las distribuciones gaussianas estimadas



Es muy útil cuando los datos no tienen una estructura de agrupamiento clara y pueden pertenecer a múltiples grupos simultáneamente.



Permite modelar grupos con formas y tamaños diferentes al permitir distribuciones gaussianas con diferentes parámetros

Balanced Iterative Reducing and Clustering using Hierarchies

- El algoritmo BIRCH es un algoritmo de agrupamiento jerárquico diseñado para manejar grandes conjuntos de datos de manera eficiente
- A diferencia de otros algoritmos jerárquicos que requieren realizar múltiples pasadas sobre los datos, BIRCH utiliza una estructura de árbol llamada CF Tree (Clustering Feature Tree) para reducir el número de operaciones necesarias

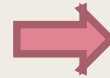
Algoritmo BIRCH

- La idea principal detrás de BIRCH es agrupar los datos en una estructura de árbol que representa una visión compacta de los datos
- El árbol CF tiene nodos internos que representan subgrupos y hojas que contienen los datos individuales o pequeños subgrupos
- Cada nodo del árbol CF mantiene información resumida sobre los subgrupos, como su centroide y el número de puntos en el subgrupo

Algoritmo BIRCH

Construcción del árbol CF

Los puntos de datos se insertan en los nodos del árbol de manera incremental



Refinamiento de los grupos

Se realiza un refinamiento de los grupos utilizando un algoritmo más preciso, como k-means, dentro de cada subgrupo identificado por el árbol CF.

Se actualizan los valores de resumen en cada nodo visitado

Algoritmo BIRCH



Puede manejar grandes volúmenes de datos de manera eficiente, ya que solo requiere un recorrido inicial de los datos para construir la estructura del árbol CF



Esto lo hace especialmente útil en aplicaciones donde los datos se generan en tiempo real o donde el almacenamiento y el procesamiento son limitados