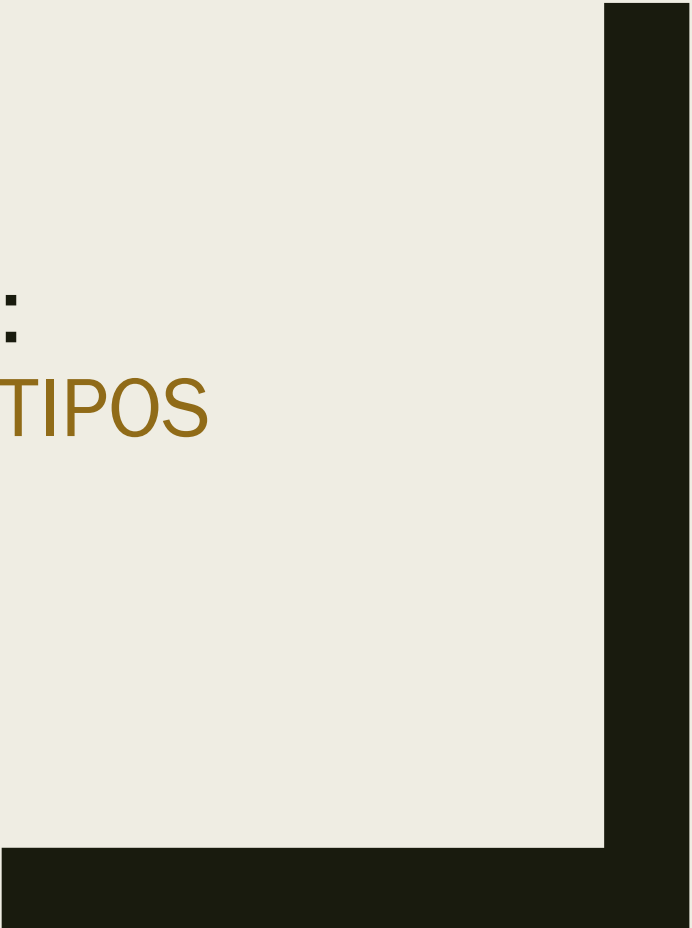




ANALÍTICA AVANZADA DE DATOS: AGRUPAMIENTO BASADO EN PROTOTIPOS

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





¿QUÉ ES EL AGRUPAMIENTO BASADO EN PROTOTIPOS?



Agrupamiento basado en prototipo



El **agrupamiento** es una técnica del aprendizaje automático que permite encontrar estructuras y patrones subyacentes en los datos sin la necesidad de etiquetas o categorías predefinidas

En lugar de tener un conjunto de datos etiquetados, el agrupamiento busca concentrar objetos similares en conjuntos o grupos, de manera que los objetos dentro de un mismo grupo sean más similares entre sí que con aquellos de otros grupos

Agrupamiento basado en prototipo



El agrupamiento es una técnica esencial en el análisis de datos, porque permite descubrir patrones y estructuras ocultas en conjuntos de datos no etiquetados



Su capacidad para organizar, explorar y extraer conocimientos de grandes volúmenes de datos sin supervisión lo convierte en una herramienta valiosa dado que la cantidad de datos disponibles sigue creciendo exponencialmente



Al comprender y aplicar el agrupamiento, podemos obtener información significativa y tomar decisiones más informadas en diversas áreas de estudio y práctica

Agrupamiento basado en prototipo

En esta técnica, se crean prototipos que representan a cada grupo y los objetos se asignan al grupo cuyo prototipo sea más cercano en términos de similitud

El **objetivo** es descubrir la estructura inherente de los datos, identificando grupos homogéneos de objetos y separándolos de otros grupos

Estos prototipos pueden ser representados por puntos o centroides en el espacio de características, y se convierten en las representaciones centrales de cada grupo

Se basa en la idea de que los objetos dentro de un mismo grupo son más similares entre sí que con aquellos de otros grupos.

Agrupamiento basado en prototipo

Hay dos elementos:

Prototipos:

- Son representaciones o puntos de referencia que caracterizan a cada grupo en el proceso de agrupamiento
- Cada grupo está asociado con un prototipo que actúa como una especie de representante central o promedio de los objetos en ese grupo
- Pueden ser puntos en el espacio de características o centroides, que se encuentran en el centro geométrico de los objetos del grupo

Similitud entre objetos:

- Es una medida que cuantifica la cercanía o la afinidad entre dos objetos en el espacio de características
- Se utiliza para calcular la distancia o la similitud entre un objeto y un prototipo o entre dos objetos. Puede ser la distancia euclidiana, la similitud del coseno o la distancia de Mahalanobis

Agrupamiento basado en prototipo

La similitud entre objetos es un factor crítico en el proceso de agrupamiento

- Cuanto más similares sean dos objetos, mayor será su similitud y mayor la probabilidad de que se asignen al mismo grupo
- Por otro lado, si dos objetos son muy diferentes, su similitud será menor y es más probable que se asignen a grupos diferentes

La relación entre los prototipos y la similitud entre objetos es fundamental

- Los prototipos actúan como puntos de referencia para medir la similitud entre un objeto y los grupos existentes. Al comparar la similitud entre un objeto y los prototipos de cada grupo, se determina a qué grupo pertenece más adecuadamente ese objeto

Agrupamiento basado en prototipo

En conjunto, los prototipos y la similitud entre objetos son elementos clave en el agrupamiento basado en prototipos

- Los prototipos representan los grupos y su posición se actualiza en función de los objetos asignados a cada grupo
- La similitud entre objetos determina cómo se asignan los objetos a los grupos y cómo se actualizan los prototipos
- Estos elementos trabajan en conjunto para organizar y estructurar los datos en grupos coherentes y facilitar el análisis y la interpretación posterior

Agrupamiento basado en prototipo

Existen varias medidas de similitud que se utilizan para calcular la distancia o la similitud entre objetos

Estas medidas son fundamentales para determinar qué tan cerca o similar es un objeto con respecto a otros objetos en el espacio de características

Algunas de las principales medidas de similitud utilizadas en el agrupamiento:

- ***Distancia euclidiana:*** Es una medida comúnmente utilizada para calcular la similitud entre dos objetos en un espacio multidimensional. La distancia euclidiana mide la longitud del segmento que conecta dos puntos en el espacio de características y se calcula utilizando la fórmula matemática de la distancia euclidiana

Agrupamiento basado en prototipo

- ***Similitud del coseno:*** Esta medida se utiliza principalmente para **datos vectoriales**. Mide el ángulo entre dos vectores en relación con el origen y calcula la similitud en función de la similitud del coseno entre los vectores. La similitud del coseno es útil cuando se quiere medir la similitud entre objetos independientemente de su magnitud o escala
- ***Distancia de Mahalanobis:*** Esta medida de similitud tiene en cuenta la correlación y la covarianza entre las variables en el espacio de características. A diferencia de la distancia euclidiana, la distancia de Mahalanobis puede tratar con **datos multivariados** y tiene en cuenta las relaciones entre las variables
- ***Coeficiente de correlación:*** Es una medida que cuantifica la relación lineal entre dos variables. Se utiliza para medir la similitud entre objetos en función de su correlación y varía entre -1 y 1. Un coeficiente de correlación cercano a 1 indica una alta similitud, mientras que un valor cercano a -1 indica una similitud negativa

Agrupamiento basado en prototipo

- ***Distancia de Manhattan:*** También conocida como distancia de la ciudad, mide la distancia entre dos puntos en un espacio de características sumando las diferencias absolutas de sus coordenadas. Es especialmente útil cuando **las dimensiones del espacio de características son categóricas o discretas**

La elección de la medida de similitud depende del tipo de datos, la naturaleza del problema y el algoritmo de agrupamiento específico utilizado. Cada medida tiene sus ventajas y desventajas, y es importante seleccionar la más apropiada para el contexto del problema y los datos disponibles

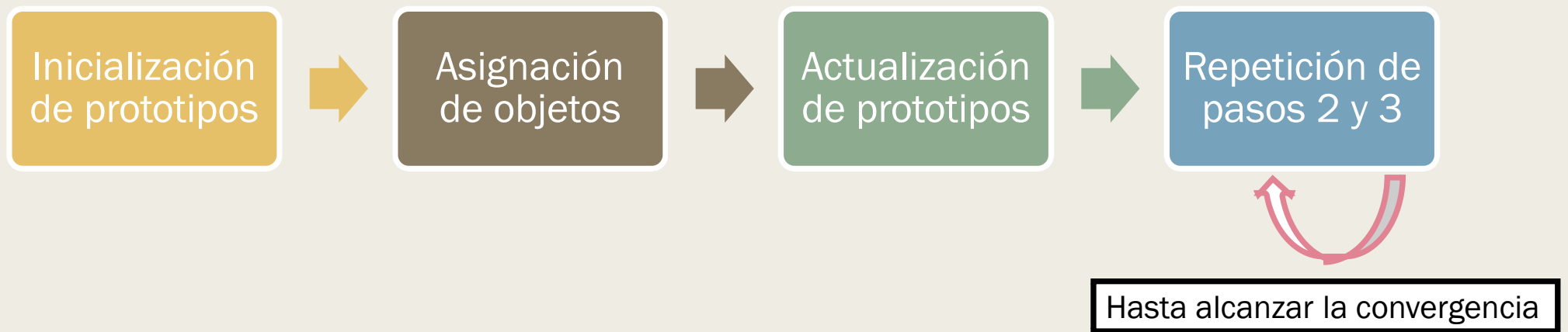


PROCESO DEL AGRUPAMIENTO BASADO EN PROTOTIPOS



Agrupamiento basado en prototipo

Para lograr esto, el proceso de agrupamiento consta de varios pasos fundamentales:



Inicialización de prototipos

- La etapa de inicialización en el proceso de agrupamiento basado en prototipos se refiere a la selección o generación de los prototipos iniciales que representarán a cada grupo en el proceso de agrupamiento
- Esta etapa es crucial, ya que los prototipos iniciales determinarán el punto de partida del algoritmo de agrupamiento y pueden afectar significativamente los resultados finales
- Existen diferentes enfoques para la inicialización de los prototipos, y la elección depende del contexto y los datos específicos:

Inicialización de prototipos

- **Selección aleatoria:**
 - En este enfoque, los prototipos se eligen de manera aleatoria del conjunto de datos
 - Se seleccionan K objetos de manera aleatoria y se asignan como prototipos iniciales, donde K es el número de grupos o clústeres deseados
 - Este método es simple y fácil de implementar, pero puede ser sensible a la selección inicial y puede generar resultados inconsistentes en diferentes ejecuciones
- **Selección basada en muestras representativas:**
 - En este enfoque, se seleccionan objetos representativos del conjunto de datos para ser los prototipos iniciales
 - Puede involucrar técnicas como muestreo estratificado, donde se asegura que cada grupo esté representado en la selección inicial de prototipos

Inicialización de prototipos

- **Generación aleatoria:**
 - En algunos casos, los prototipos pueden generarse de manera aleatoria en el espacio de características, en lugar de seleccionarse directamente de los datos.
 - Esto se utiliza cuando no se dispone de datos reales o cuando se desea explorar la estructura del espacio de características de manera más amplia
- **Inicialización basada en algoritmos:**
 - También es posible utilizar algoritmos específicos para inicializar los prototipos.
 - Por ejemplo, el algoritmo *K-means++* utiliza un enfoque de inicialización inteligente para seleccionar prototipos iniciales que estén lo más alejados posible entre sí, lo que ayuda a mejorar la convergencia y la calidad de los resultados finales

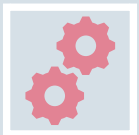
Inicialización de prototipos



Es importante tener en cuenta que la elección de los prototipos iniciales puede influir en la calidad y la eficacia del agrupamiento



Una mala selección inicial puede llevar a una convergencia prematura o a resultados subóptimos



Por lo tanto, es recomendable experimentar con diferentes métodos de inicialización y evaluar su impacto en los resultados

Asignación de objetos

La asignación de objetos es un paso fundamental que implica calcular la similitud entre los objetos y asignarlos a los grupos correspondientes. Este paso se repite iterativamente en cada iteración del algoritmo de agrupamiento hasta alcanzar la convergencia

A continuación, se explica cómo se lleva a cabo:

1. Cálculo de la similitud:

- El primer paso en la asignación de objetos es calcular la similitud entre cada objeto y los prototipos de cada grupo
- La similitud se calcula utilizando una medida de similitud adecuada, como la distancia euclidiana, la similitud del coseno o alguna otra medida específica según el problema
- Para cada objeto, se calcula su similitud con todos los prototipos existentes

Asignación de objetos

2. Asignación al grupo más cercano:

- Una vez que se han calculado las similitudes, se asigna cada objeto al grupo cuyo prototipo esté más cercano en términos de similitud
- El objeto se asigna al grupo cuyo prototipo tenga la similitud más alta con el objeto
- Esto implica encontrar el prototipo más cercano utilizando la medida de similitud previamente calculada

3. Actualización de la asignación:

- Después de asignar un objeto a un grupo, se actualiza la asignación y se registra que el objeto pertenece a ese grupo
- Esta información se utiliza posteriormente en el proceso de actualización de prototipos y en las iteraciones posteriores del algoritmo

Asignación de objetos



Se basa en la similitud entre los objetos y los prototipos existentes en ese momento



Se repite para todos los objetos en el conjunto de datos en cada iteración del algoritmo de agrupamiento



Es un paso esencial en el agrupamiento basado en prototipos, ya que permite organizar los objetos en grupos coherentes y capturar las similitudes entre ellos

Actualización de prototipos

Después de asignar los objetos a los grupos correspondientes, se procede a recalcular la posición de los prototipos en función de los objetos asignados a cada grupo.

Este paso tiene como **objetivo** ajustar los prototipos para que representen de manera efectiva las características de cada grupo

1. Recopilación de objetos asignados:

- El primer paso es recopilar todos los objetos que han sido asignados a cada grupo
- Para cada grupo, se recopilan los objetos que se han asignado a ese grupo en la etapa de asignación de objetos

Actualización de prototipos

2. Cálculo de los nuevos prototipos:

- Una vez que se han recopilado los objetos asignados a cada grupo, se procede a calcular los nuevos prototipos
- La forma en que se calculan los nuevos prototipos depende del enfoque específico del agrupamiento basado en prototipos que se esté utilizando
- Algunas técnicas comunes incluyen:
 - **Promedio:** Se calcula el promedio de las características de los objetos asignados a cada grupo. Para cada dimensión o característica, se calcula el promedio de los valores correspondientes de los objetos asignados al grupo. Esto resulta en un nuevo prototipo que representa el centro del grupo en el espacio de características

Actualización de prototipos

- **Mediana:** En lugar de calcular el promedio, se puede calcular la mediana de las características de los objetos asignados a cada grupo. La mediana es el valor que se encuentra en el medio cuando los valores se ordenan en orden ascendente. Esto puede ser útil cuando los datos contienen valores atípicos que pueden afectar negativamente el cálculo del promedio
- **Centroides ponderados:** Si se desea asignar pesos diferentes a los objetos en la actualización de prototipos, se pueden utilizar centroides ponderados. Cada objeto puede tener un peso asignado según su importancia o relevancia, y se utiliza este peso en el cálculo del prototipo

Actualización de prototipos

3. Reasignación de prototipos:

- Después de calcular los nuevos prototipos, se reemplazan los prototipos anteriores con los nuevos valores
- Esto implica actualizar la posición de los prototipos en el espacio de características para que reflejen las características de los objetos asignados al grupo
- Los nuevos prototipos se utilizarán en la próxima iteración del proceso de agrupamiento

Actualización de prototipos

4. Repetición de asignación y actualización:

- Una vez que se han actualizado los prototipos, se repite el proceso de asignación de objetos utilizando los nuevos prototipos
- Los objetos se asignarán nuevamente a los grupos en función de la similitud con los nuevos prototipos, y luego se actualizarán los prototipos nuevamente
- Este ciclo de asignación y actualización se repite iterativamente hasta alcanzar la convergencia

Actualización de prototipos



Permite refinar la representación de cada grupo y ajustar los prototipos para que sean más representativos de los objetos asignados a cada grupo



A medida que se repite el proceso de asignación y actualización, los prototipos tienden a converger hacia posiciones que mejor representan las características de cada grupo



Esto resulta en una agrupación más precisa y coherente de los objetos en función de sus similitudes

Repetición de los pasos de asignación y actualización

Se lleva a cabo iterativamente hasta alcanzar la convergencia

Este enfoque iterativo permite mejorar gradualmente la asignación de objetos a los grupos y la posición de los prototipos.

1. Asignación de objetos:

- En cada iteración, se realiza la asignación de objetos a los grupos existentes utilizando los prototipos actuales
- Se calcula la similitud entre cada objeto y los prototipos utilizando una medida de similitud apropiada
- Luego, se asigna cada objeto al grupo cuyo prototipo tenga la similitud más alta con el objeto
- Esta asignación se basa en la similitud calculada en la etapa de asignación anterior

Actualización de prototipos

2. Actualización de prototipos:

- Después de realizar la asignación de objetos, se actualizan los prototipos en función de los objetos asignados a cada grupo
- Los prototipos se recalculan utilizando los objetos asignados en la iteración actual
- Se pueden utilizar diferentes métodos de actualización de prototipos, como el cálculo del promedio de las características de los objetos asignados a cada grupo
- El **objetivo** es ajustar los prototipos para que sean más representativos de los objetos asignados a cada grupo

Actualización de prototipos

3. Evaluación de la convergencia:

- Después de actualizar los prototipos, se evalúa si se ha alcanzado la convergencia
- Esto implica verificar si ha habido cambios significativos en la asignación de objetos o en la posición de los prototipos en comparación con la iteración anterior
- Si no hay cambios significativos, se considera que se ha alcanzado la convergencia y se finaliza el proceso de agrupamiento

Actualización de prototipos

4. Iteración del proceso:

- Si no se ha alcanzado la convergencia, se repiten los pasos de asignación y actualización
- Los prototipos actualizados se utilizan en la siguiente iteración para asignar objetos y recalcular los prototipos nuevamente
- Este ciclo de iteración continúa hasta que se alcance la convergencia

Actualización de prototipos



El número de iteraciones puede variar y puede ser definido de antemano o establecido en función de un criterio de convergencia específico



Es posible incluir criterios adicionales para detener el proceso de agrupamiento, como un límite máximo de iteraciones o un umbral de cambio mínimo en la asignación o los prototipos



En cada iteración del proceso de agrupamiento basado en prototipos permite que los grupos se afinen gradualmente y se mejore la calidad de la agrupación



APLICACIONES DEL AGRUPAMIENTO BASADO EN PROTOTIPOS



Agrupamiento basado en prototipo

Minería de datos:

Se utiliza para descubrir patrones y estructuras ocultas en conjuntos de datos. Se aplica en la segmentación de clientes, detección de anomalías, análisis de texto, recomendación de productos y muchas otras tareas relacionadas con el descubrimiento de conocimiento a partir de datos

Reconocimiento de patrones:

Se utiliza para agrupar objetos o instancias similares en clases o categorías. Por ejemplo, en el reconocimiento de imágenes, se pueden utilizar prototipos para agrupar imágenes similares en categorías como paisajes, animales, objetos, entre otros

Bioinformática:

Se aplica para clasificar y agrupar secuencias de ADN o proteínas con el fin de identificar relaciones genéticas, identificar enfermedades genéticas o descubrir nuevos medicamentos. Los prototipos pueden representar patrones de secuencias que permiten agrupar y analizar datos biológicos complejos

Agrupamiento basado en prototipo

Procesamiento de imágenes y visión por computadora:

Se utiliza para segmentar imágenes en regiones u objetos similares. Puede ayudar en tareas como la detección de bordes, la segmentación de imágenes médicas, la clasificación de objetos y la recuperación de imágenes basada en contenido

Análisis de redes sociales:

Se utiliza para descubrir comunidades o grupos de usuarios con intereses similares, comportamientos similares o relaciones sociales cercanas. Puede ayudar en la recomendación de amigos, la detección de comunidades influyentes y el análisis de tendencias en las redes sociales

Optimización de la cadena de suministro:

En el campo de la logística y la gestión de la cadena de suministro, el agrupamiento basado en prototipos se utiliza para clasificar productos, identificar grupos de clientes con necesidades similares, optimizar rutas de entrega y gestionar el inventario de manera eficiente

Aplicaciones del agrupamiento basado en prototipo

Al aplicar el agrupamiento basado en prototipos a un conjunto de datos, se pueden obtener varios beneficios para el análisis y la comprensión de los datos:

1. Descubrimiento de grupos naturales:

- Permite identificar grupos o clústeres naturales en los datos, es decir, agrupaciones de objetos que comparten características similares
- Estos grupos pueden no ser conocidos de antemano y el algoritmo de agrupamiento ayuda a revelarlos de forma automática.

Actualización de prototipos

2. Segmentación de datos:

- Es posible segmentar grandes conjuntos de datos en grupos más pequeños y homogéneos
- Esta segmentación facilita el análisis y la comprensión de los datos, ya que se pueden estudiar las características y los patrones específicos de cada grupo de forma independiente

3. Detección de anomalías:

- Además de identificar grupos similares, también puede ayudar a detectar objetos o instancias anómalas que no se ajustan bien a ninguno de los grupos existentes
- Estas anomalías pueden indicar puntos de datos atípicos o inusuales que requieren una atención especial

Actualización de prototipos

4. Visualización de datos:

- Puede facilitar la visualización de datos complejos en espacios de alta dimensión
- Al asignar colores o etiquetas a los grupos obtenidos, se pueden representar los datos en un espacio de menor dimensión para una mejor comprensión visual

5. Segmentación de clientes y personalización:

- En el ámbito del análisis de clientes, puede ayudar a identificar segmentos de clientes con características y preferencias similares
- Esto permite una personalización más efectiva de los productos o servicios ofrecidos, así como una mejor comprensión del comportamiento del cliente

Importancia del agrupamiento basado en prototipos

También puede ser útil para tareas adicionales como:

1. Clasificación:

- Una vez que se han formado los grupos mediante el agrupamiento basado en prototipos, se puede utilizar esta información para clasificar nuevos objetos o instancias
- Cada prototipo representa un grupo específico con características similares, por lo que, al asignar un nuevo objeto al prototipo más cercano, se puede predecir su clase o categoría
- Esto permite construir modelos de clasificación basados en prototipos, que pueden ser rápidos y eficientes en la asignación de nuevos objetos a grupos predefinidos

Importancia del agrupamiento basado en prototipos

2. Detección de anomalías:

- Al realizar el agrupamiento basado en prototipos, también se puede identificar la presencia de anomalías o puntos de datos atípicos
- Los objetos que no se ajustan bien a ninguno de los grupos existentes pueden considerarse anomalías
- Estas anomalías pueden indicar datos inusuales o potencialmente errores en los datos
- Por lo tanto, la información de agrupamiento puede utilizarse como base para desarrollar métodos de detección de anomalías, identificando aquellos objetos que no se ajustan a los grupos esperados

Importancia del agrupamiento basado en prototipos

3. Etiquetado de datos:

- Una vez que se han formado los grupos, se puede asignar una etiqueta o categoría a cada grupo en función de los objetos que contiene y su interpretación semántica
- Estas etiquetas pueden ser utilizadas posteriormente para clasificar nuevos objetos o instancias en tareas de aprendizaje supervisado
- Además, el etiquetado de grupos puede ayudar en la interpretación y comprensión de los patrones y características que definen cada grupo