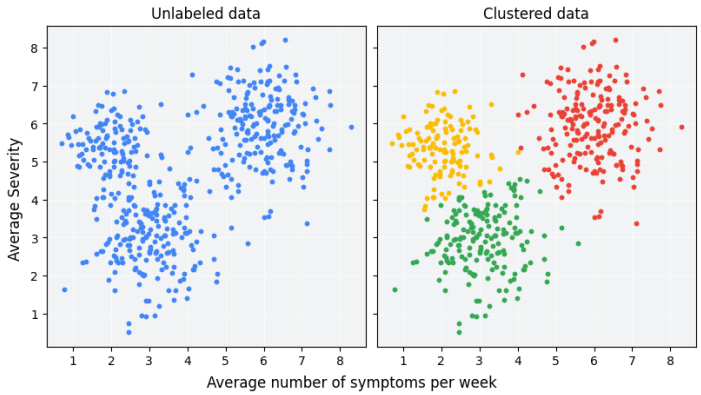
Centro Universitario de Ciencias Exactas e Ingenierías

Universidad de Guadalajara

Actividad 7: Agrupamiento

Aprendizaje Máquina



Alumno: Samuel David Pérez Brambila

Código: 222966286

Profesora: Karla Ávila Cárdenas

Sección: D01

Fecha de Entrega: 27 de Octubre de 2024

Introducción

El clustering o agrupamiento es “un algoritmo de machine learning no supervisado que organiza y clasifica diferentes objetos, puntos de datos u observaciones en grupos o clusters basados en similitudes o patrones.” (IBM, s.f.)

Pero ¿qué es el aprendizaje no supervisado? Según INESDI (2022), es un tipo de aprendizaje automático o machine learning en el que los modelos aprenden a partir de conjuntos de datos sin etiquetar sobre el que se les permite actuar sin supervisión.

Teniendo entendido que es el aprendizaje no supervisado, continuemos hablando acerca del clustering. Existen múltiples algoritmos de clustering, cada uno con distintas formas de definir un cluster. Los distintos enfoques funcionarán bien para diferentes tipos de modelos en función del tamaño de los datos de entrada, la dimensionalidad de los datos, la rigidez de las categorías y el número de conglomerados dentro del conjunto de datos. Vale la pena señalar que un algoritmo puede funcionar muy bien para un conjunto de datos y muy mal en otro. Según IBM (s.f.), cinco de los enfoques más utilizados son:

- Clustering basado en centroides: El clustering basado en centroides es un tipo de método de clustering que divide o divide un conjunto de datos en grupos similares en función de la distancia entre sus centroides. El centroide de cada clúster es la media o la mediana de todos los puntos del clúster, en función de los datos.
- Clustering jerárquico: El clustering jerárquico, a veces denominado clustering basado en la conectividad, agrupa los puntos de datos en función de la proximidad y la conectividad de sus atributos. Este método determina los clústeres en función de la proximidad de los puntos de datos entre sí en todas las dimensiones. La idea es que los objetos que están más cerca están más estrechamente relacionados que los que están lejos unos de otros.
- Clustering basado en distribución: El clustering basado en la distribución, a veces denominado clustering probabilístico, agrupa los puntos de datos en función de su distribución de probabilidad. Este enfoque supone que hay un proceso que genera distribuciones normales para cada dimensión de los datos que crea los centros de clústeres. Se diferencia del clustering basado en centroides en que no utiliza una métrica de distancia como una distancia euclidiana o de Manhattan. En su lugar, los enfoques basados en la distribución buscan una distribución bien definida que aparezca en cada dimensión.
- Clustering basado en densidad: El clustering basado en la densidad funciona mediante la detección de áreas donde se concentran puntos y donde están separados por áreas que están vacías o escasas. A diferencia de los

enfoques basados en centroides, como las medias K, o los enfoques basados en la distribución, como la maximización de expectativas el clustering basado en la densidad puede detectar clústeres de una forma arbitraria. Esto puede ser extremadamente útil cuando los clústeres no están definidos en torno a una ubicación o distribución específica.

- Clustering basado en cuadrícula: Los algoritmos de clustering basados en cuadrículas no se utilizan con tanta frecuencia como en los cuatro enfoques anteriores, pero pueden resultar útiles en clústeres de alta dimensión, ya que otros algoritmos de agrupamiento pueden no tener el mismo rendimiento. En este enfoque, el algoritmo divide un conjunto de datos de alta dimensión en celdas. A cada celda se le asigna un identificador único llamado ID de celda y todos los puntos de datos que se encuentran dentro de una celda se consideran parte del mismo clúster.

Hay muchas áreas de aplicación donde el clustering es una herramienta valiosa para la minería de datos o el análisis exploratorio de datos, acorde con IBM (s.f.) encontramos:

- Detección de anomalías: El clustering puede ayudar a descubrir anomalías mediante la medición de los puntos de datos que no se incluyen en la estructura de clustering definido por el análisis de clústeres. Los puntos de datos que pertenecen a clústeres pequeños o muy dispersos o que están lejos de su clúster asignado se pueden considerar anomalías.
- Investigación de mercado: Al tratar de comprender qué perfiles de clientes o subconjuntos de mercados podrían existir, el clustering puede ser una herramienta poderosa para ayudar a realizar la segmentación de clientes. Es posible que pueda combinar datos demográficos con datos de comportamiento del cliente para encontrar qué tipos de características y patrones de compra se correlacionan con mayor frecuencia.
- Segmentación de imágenes: Las imágenes pueden tener sus píxeles agrupados en una variedad de formas que pueden ayudar a cortar la imagen en diferentes secciones para separar un primer plano de un fondo, detectar objetos usando similitudes en color y brillo, o dividir imágenes en regiones de interés para su posterior procesamiento. Con las imágenes, los métodos de clustering procesan los píxeles de la imagen y definen áreas dentro de la imagen que representan el clúster.

Es así que en el siguiente trabajo, se elaborará un cuadro sinóptico que sintetiza la información clave del capítulo sobre clustering de un libro proporcionado por la profesora. A través de este esquema, se resaltarán los conceptos teóricos más importantes sobre el análisis de clusters. El objetivo es facilitar la comprensión de este tema, organizando las ideas principales de manera clara y estructurada.

Contenido de la Actividad

Introducción al clustering

Aprendizaje no supervisado

El aprendizaje no supervisado es un enfoque que no requiere ejemplos con etiquetas de clase. Busca identificar patrones útiles en los datos sin clasificaciones previas.

La principal tarea del aprendizaje no supervisado es el análisis de clusters. Este busca identificar grupos de ejemplos similares (clusters) en los datos.

Identificación de grupos de ejemplos similares

- **Visualización**: En datos bidimensionales es fácil visualizar los clusters, pero en más dimensiones, los humanos no pueden hacerlo. Es necesario utilizar algoritmos de análisis de clusters.
- **Representación por centroides**: Un cluster puede describirse por su centroide, que es el promedio de los atributos de los ejemplos del cluster. Para atributos numéricos, es un promedio directo, y para atributos discretos, se pueden convertir en variables booleanas.

Propiedades de los clusters

- **No Superposición**: Los clusters no deben solaparse; cada ejemplo pertenece a un solo cluster.
- **Proximidad**: Los ejemplos dentro de un cluster deben estar mucho más cerca entre sí que de los ejemplos en otros clusters.
- **Número de clusters**: El número de clusters puede definirse manualmente o determinarse automáticamente por el algoritmo.

Problemas con la medición de distancias

La distancia euclidiana es la medida más común para evaluar la cercanía entre un ejemplo y un cluster, pero tiene limitaciones.

- No es adecuada para atributos discretos.
- Problemas de escala: Diferencias en las unidades o magnitudes entre atributos pueden alterar las distancias. Solución: normalización de los atributos.

Fórmula general para la medición de distancias

La fórmula para atributos mixtos (numéricos y discretos es:

$$d_M(\mathbf{x}, \mathbf{y}) = \sqrt{\sum_{i=1}^{n} d(x_i, y_i)}$$

- Para atributos continuos: $d(x_i, y_i) = (x_i - y_i)^2$
- Para atributos discretos (booleanos): $d(x_i, y_i) = 0$ si $x_i = y_i$, y $d(x_i, y_i) = 1$ si $x_i \neq y_i$

Asignación de un ejemplo a un cluster

- **Proceso de asignación**:
  - Se calcula la distancia de un ejemplo **x** a cada centroide ci de los clusters.
  - El ejemplo se asigna al cluster cuyo centroide esté más cerca.
- **Ejemplo numérico**:
  - Dado un ejemplo **x** = (4,4) y tres centroides: c1 = (3,4), c2 = (4,6), c3 = (5,7):
  - Distancias euclidianas = d(x,c1) = 1, d(x,c2) = 2, d(x,c3) = $\sqrt{10}$
  - El ejemplo se asigna al cluster c1 ya que tiene la **menor distancia**.

Beneficios del análisis de clusters

- **Estimación de Valores Faltantes**: Si se desconoce un valor de atributo, se puede estimar utilizando el promedio o el valor más frecuente dentro del cluster al que pertenece el ejemplo. Esto es más preciso que usar simplemente el valor promedio de todo el conjunto de datos.
- **Reducción del tamaño de redes y clasificadores**: El análisis de clusters puede reducir el tamaño de clasificadores como las redes de función de base radial (RBF) y los clasificadores bayesianos. En lugar de usar todos los ejemplos, se pueden usar los centroides de los clusters como "centros gaussianos".
- **Clasificación simple**: Si casi todos los ejemplos en un cluster pertenecen a la misma clase, se puede etiquetar el cluster con esa clase, lo que simplifica la clasificación supervisada.

Conclusiones

En conclusión, comprender el funcionamiento del clustering es fundamental para aprovechar al máximo su capacidad de análisis y agrupamiento de datos en machine learning. Desde métodos basados en centroides hasta enfoques más complejos como el clustering jerárquico o basado en densidad, cada técnica ofrece ventajas específicas que permiten abordar diferentes tipos de problemas. La correcta aplicación de estos algoritmos no solo facilita el descubrimiento de patrones ocultos en los datos, sino que también juega un papel crucial en áreas como la detección de anomalías, la segmentación de mercados y el procesamiento de imágenes.

Realizar un cuadro sinóptico que sintetice los puntos clave de estas técnicas ayuda a estructurar y clarificar los conceptos esenciales del clustering. Esta herramienta visual permite una comprensión más profunda de puntos como los beneficios del análisis de clusters, los posibles problemas al realizar medición de distancias en un algoritmo de clustering, entre otros. Además, al reducir la complejidad del tema a ideas concretas, facilita la identificación de los aspectos más relevantes y prepara el terreno para una aplicación efectiva de estos métodos en escenarios reales.

Por lo tanto, este trabajo no solo resalta la relevancia del clustering en el análisis de datos, sino también la importancia de contar con una herramienta como el cuadro sinóptico para condensar la información. Al organizar de manera clara los conceptos principales, se facilita la comprensión y permite abordar estos temas con mayor precisión. Esto es esencial para dominar la aplicación de algoritmos de clustering y aprovechar su potencial en la resolución de problemas complejos en diferentes campos.

Referencias

- IBM (s.f.). *¿Qué es el clustering?*. IBM. Recuperado el 16 de Octubre de 2024 de: https://www.ibm.com/es-es/topics/clustering
- INESDI (2022, 2 de noviembre). *¿Qué es el aprendizaje no supervisado y cuándo usarlo?*. INESDI Business School. Recuperado el 16 de Octubre de 2024 de: https://www.inesdi.com/blog/que-es-aprendizaje-no-supervisado/