
Universidad
Tecmilenio®

Universidad Tecmilenio

Aprendizaje Automático No Supervisado

Introducción al aprendizaje automático no supervisado

Introducción

Las últimas dos décadas se han caracterizado por los avances rápidos y significativos en la recopilación, almacenamiento, transmisión y procesamiento de datos.

Debido a que se han acumulado datos masivos que cubren todos los aspectos de la vida humana, se necesitan métodos efectivos y eficientes para utilizarlos y proporcionar soluciones para liberar el poder de los datos masivos.

Por consiguiente, el aprendizaje automático y las estadísticas proporcionaron la facilidad de dicho análisis de datos.

Entre las técnicas de aprendizaje no supervisado la más investigada y aplicada es la agrupación (*clustering*), la cual conocerás a lo largo de este tema. Otras aplicaciones del paradigma no supervisado incluyen la visualización de información, la reducción de dimensionalidad, el descubrimiento de las reglas de asociación y la detección de anomalías.

Explicación

Es posible dividir los problemas de aprendizaje en dos clases si los datos de entrenamiento están etiquetados o no:

- Aprendizaje supervisado (por ejemplo, clasificación y regresión).
- Aprendizaje no supervisado (por ejemplo, agrupamiento).

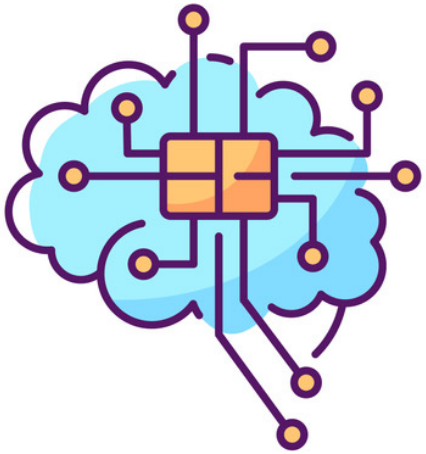
Funcionamiento de un sistema no supervisado

Un sistema basa sus predicciones en una muestra de entrenamiento de casos previamente resueltos en donde se conocen los valores conjuntos de todas las variables. A esto se le llama aprendizaje supervisado o aprendizaje con un maestro. Bajo esta metáfora, el aprendiz presenta una respuesta en la muestra de capacitación, mientras que el supervisor o maestro proporciona la respuesta correcta o un error asociado con la respuesta del aprendiz.

Explicación

El objetivo del aprendizaje automático consiste en aprender sobre modelos que puedan funcionar bien con nuevas muestras, en lugar de ejemplos de entrenamiento y el mismo objetivo también se aplica al paradigma no supervisado, ya que se busca que los grupos aprendidos funcionen bien en las muestras fuera del conjunto de entrenamiento.

La capacidad para trabajar en las nuevas muestras se denomina capacidad de generalización, por lo que un modelo bien generalizado debería funcionar correctamente en todo el espacio muestral.



Explicación

La agrupación en clústeres se puede utilizar por sí misma para identificar la estructura inherente de los datos, por lo que también puede servir como una técnica de preprocesamiento para otras tareas de aprendizaje como la clasificación.

Por ejemplo, una empresa puede querer clasificar a los nuevos usuarios en diferentes categorías, pero esto no es tan fácil.

En tal caso, la agrupación en clústeres puede aplicarse para agrupar a todos los usuarios (donde cada clúster representa una categoría de usuario), por lo que posteriormente se podrá construir un modelo de clasificación sobre los grupos para clasificar la categoría de nuevos usuarios.

Explicación

Hay que tener en cuenta que los algoritmos de agrupación en clústeres desconocen dichas características de la agrupación, por lo que solo son responsables de crear los clústeres.

Entonces, la característica asignada a cada grupo es interpretada por el usuario.

El problema de agrupación puede definirse formalmente de la siguiente manera:

Definición del problema de agrupación (Zhou, 2021)

Dado un conjunto de datos

$$D = \{x_{i1}; x_{i2}; \ldots; x_{in}\}$$

que contiene m muestras sin etiquetar, donde cada muestra

$$x_i = (x_{i1}; x_{i2}; \ldots; x_{in})$$

es un vector n-dimensional. Entonces el algoritmo de agrupamiento divide el conjunto de datos D en k grupos disjuntos

$$\{C_l | l = 1, 2, \ldots, k\}.$$

En consecuencia, se denota

$$\lambda_j \epsilon \{1, 2, \ldots, k\}$$

como la etiqueta de grupo de la muestra

$$x_j \text{ (es decir, } x_i \epsilon C_{\lambda l}).$$

Entonces, el resultado de la agrupación se puede representar como un vector de etiqueta de agrupación

$$\lambda = (\lambda_1; \lambda_2; \ldots; \lambda_m)$$

con m elementos.

Explicación

Aunque con el aprendizaje supervisado hay una medida clara de éxito o falta del mismo, en el no supervisado no existe tal medida directa de éxito, ya que es difícil determinar la validez de las inferencias extraídas de los resultados de la mayoría de sus algoritmos.

Por tal motivo, a veces es necesario recurrir a argumentos heurísticos para juzgar la calidad de los resultados. Esta situación ha provocado una amplia proliferación de diferentes métodos, ya que la eficacia se vuelve subjetiva y no puede verificarse directamente.

Explicación

Dado lo anterior, una pregunta importante que debes plantearte en este punto es la siguiente:

¿Cuál sería una correcta agrupación en clústeres?

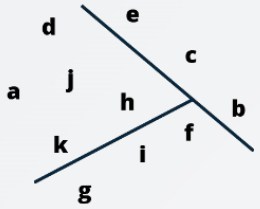
Intuitivamente el objetivo consiste en que cosas de cierto tipo se agrupen, es decir, las muestras de un mismo grupo deben ser lo más similares posible, mientras que las muestras de diferentes grupos deben ser lo más diferentes posible. En otras palabras, se busca que los elementos de cada clúster tengan una alta similitud entre ellos, al mismo tiempo que cada clúster tenga baja similitud con respecto a los demás.

Explicación

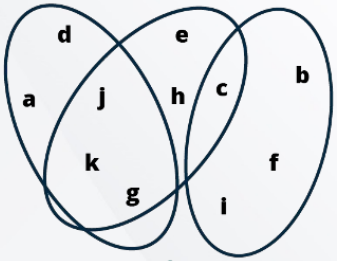
Al aplicar la técnica de agrupamiento, la salida toma la forma de un diagrama que muestra cómo las instancias se dividen en clústeres (a).

Sin embargo, se podrían disponer las instancias en dos dimensiones y generar subconjuntos superpuestos que representan cada agrupación, resultando en un diagrama de Venn (b).

En otros casos, los algoritmos asocian las instancias con clústeres de forma probabilística en lugar de categórica (c).



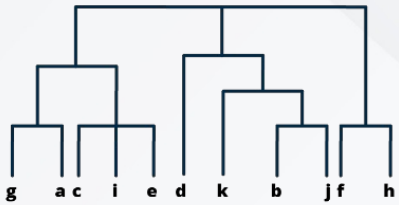
(a)



(b)

	1	2	3
a	0,3	0,2	0,5
b	0,2	0,6	0,2
c	0,4	0,3	0,3
d	0,8	0,1	0,1
e	0,5	0,3	0,2
f	0,3	0,4	0,3
g	0,7	0,1	0,2
h	0,3	0,1	0,6
i	0,3	0,3	0,4

(c)



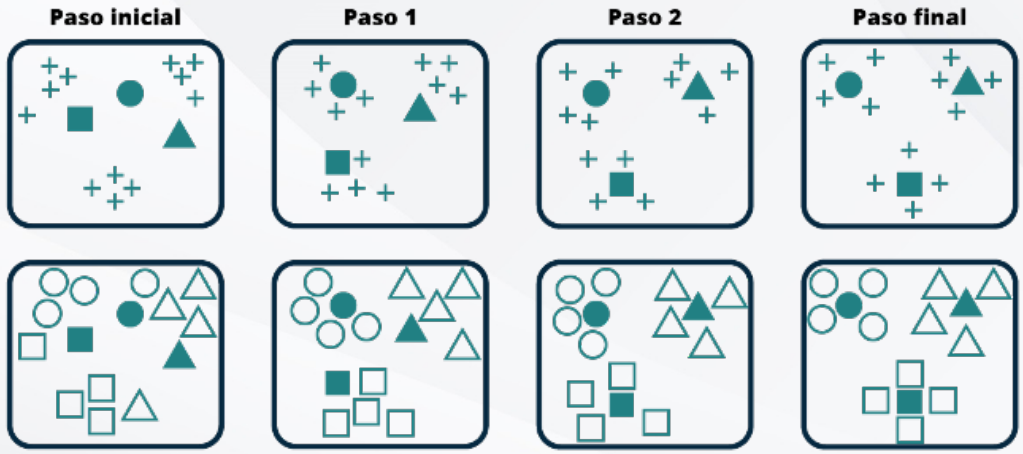
(d)

Además, otros algoritmos producen una estructura jerárquica de clústeres, de modo que en el nivel superior el espacio de instancias se divide en solo unos pocos clústeres, en donde cada uno se divide en sus propios subgrupos en el siguiente nivel (d).

Explicación

La técnica más conocida de agrupamiento se conoce como k-medias. En la figura se muestra un ejemplo de su funcionamiento.

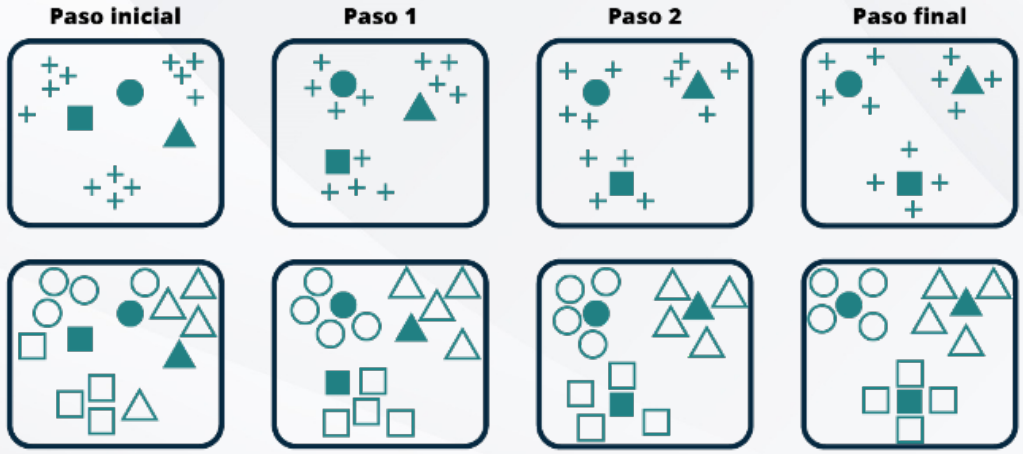
Cada una de las cuatro columnas corresponde a una iteración del algoritmo de k-medias. Este ejemplo asume que se observan tres grupos, por lo que así se establece k = 3.



Inicialmente en la parte superior izquierda se colocan aleatoriamente tres centros de conglomerados, que están representados por diferentes formas geométricas. Por consiguiente, las instancias se asignan tentativamente a los conglomerados al encontrar el centro del conglomerado más cercano para cada instancia, por lo que esto completa la primera iteración del algoritmo.

Explicación

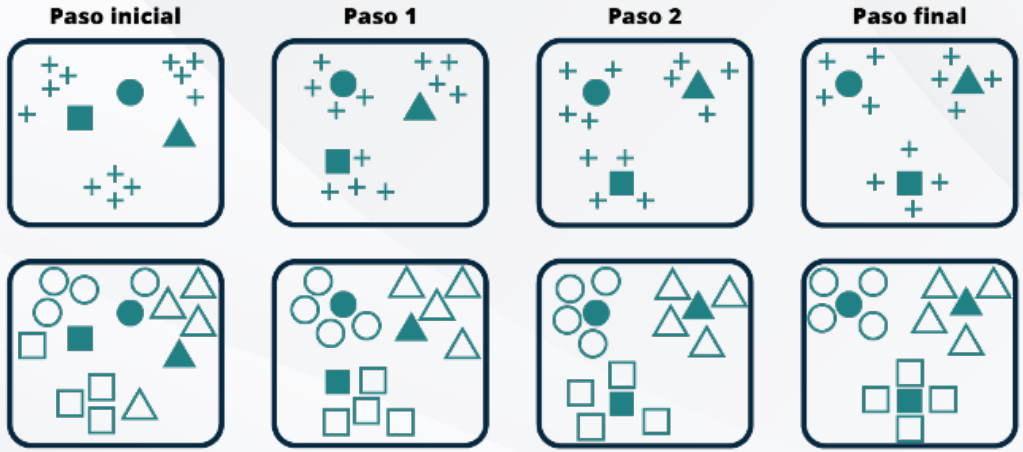
Hasta ese momento la agrupación parece desordenada, pero no es sorprendente ya que los centros de agrupación iniciales eran aleatorios. La clave es actualizar los centros en función de la asignación que se acaba de crear.



En la siguiente iteración los centros de los conglomerados se recalculan en función de las instancias que se han asignado a cada conglomerado para obtener el gráfico superior en la segunda columna.

Explicación

Posteriormente las instancias se vuelven a asignar a estos nuevos centros para obtener la siguiente etapa, lo cual produce un conjunto de clústeres más aceptable. Sin embargo, los centros aún no se encuentran en medio de sus agrupaciones. Por lo tanto, es necesario repetir los dos pasos: recálculo del centro y reasignación de instancias.



Esto da como resultado el paso 2, en donde los grupos parecen muy plausibles, pero los dos centros de clústeres más importantes aún deben actualizarse.

Al volver a calcular las asignaciones en la siguiente y última iteración, se muestra que todas las instancias permanecen asignadas a los mismos centros de clúster, por lo que el algoritmo ha terminado de converger.

Explicación

Selección del número de clústeres para cada problema

Se han desarrollado muchas variantes del algoritmo básico de k-medias. Entonces, ¿cómo se elige *K*? A menudo no se sabe nada sobre el número probable de agrupaciones, por lo que el objetivo de la agrupación es averiguarlo.

Una forma es probar diferentes valores y elegir el mejor. Para hacer esto se necesita saber cómo evaluar el éxito del aprendizaje automático.

Explicación

Selección del número de clústeres para cada problema

Otros autores indican que la elección del número de conglomerados *K* depende del objetivo. A saber, para la segmentación de datos, generalmente *K* se define como parte del problema.

Por ejemplo, una empresa puede emplear a *K* vendedores, por lo que el objetivo es dividir una base de datos de clientes en *K* segmentos (uno para cada vendedor), de modo que los clientes asignados a cada uno sean lo más similares posible.

Actividad

Después de haber estudiado el tema puedes abordar las siguientes cuestiones:

- ¿Cuál consideras que es la principal diferencia entre el aprendizaje supervisado y el no supervisado?
- De acuerdo con la literatura, cuando aplicamos la técnica de agrupamiento (clustering), ¿es posible determinar un número correcto de grupos (clusters)?

Cierre

En este tema se abordaron las diferencias entre el aprendizaje supervisado y no supervisado, resaltando que el segundo puede ser útil para llevar a cabo la tarea de agrupación (clustering), misma que también puede aplicarse en áreas importantes como marketing, medicina, bioinformática y en tecnologías basadas en el aprendizaje automático como la visión artificial.

Asimismo, se explicó en qué consiste el problema de agrupación y de qué manera abordarlo, señalando sus elementos principales. También se mencionaron y detallaron cuáles son las diferentes maneras en que los grupos (clústeres) pueden llegar a representarse.

Además, se explicaron algunos criterios importantes al momento de seleccionar el número de grupos (clústeres) para un problema, basándose en la forma más usual del algoritmo de agrupación: k-medias.

Finalmente resultaría conveniente preguntarte lo siguiente: ¿qué aplicaciones basadas en aprendizaje no supervisado y agrupación podrían desarrollarse?

Universidad Tecmilenio

Aprendizaje Automático No Supervisado

Principios de la clasificación

Introducción

La clasificación es el proceso de ordenar elementos en una o algunas categorías predefinidas. Los algoritmos que realizan esta tarea se aplican sobre categorías que el ser humano separaría usando el sentido común. Sin embargo, el valor de esta tecnología está en la cantidad de datos que puede manejar, en comparación con una persona.

En este tema se incluye un resumen sobre las técnicas de aprendizaje supervisado, el cual representa al primer tipo de aprendizaje automático. Es importante recordar que los datos de entrenamiento se encuentran etiquetados en este tipo de aprendizaje, por lo que suele aplicarse a la clasificación convencional de datos y regresiones.

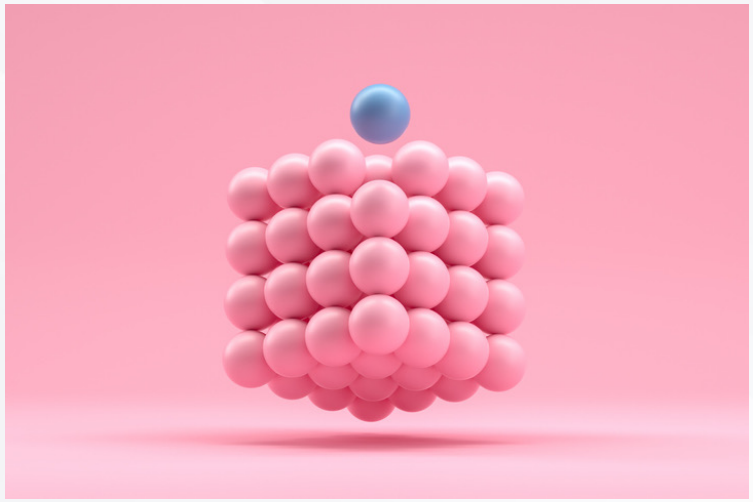
Por su parte, el aprendizaje no supervisado consiste en aquel enfoque en donde los datos de entrenamiento no están etiquetados. Por ende, el modelo aprende en función de similitudes. Esta característica le permite realizar la agrupación de datos con efectividad, por lo que también es posible que un algoritmo de aprendizaje transite entre ambos paradigmas.

Explicación

Un algoritmo de clasificación se realiza definiendo los datos no visibles en grupos como segmentos, mientras que en la técnica de agrupación se hace con la segmentación de ejemplos convertidos en grupos.

La principal diferencia entre ambos problemas es que en el agrupamiento la segmentación se realiza con base en similitudes sin tener conocimiento previo o comprensión de la estructura de los grupos, mientras que en el caso de la clasificación la segmentación se realiza en función del conjunto de datos de entrenamiento, sabiendo cuál es la estructura del grupo.

Es así como al resultado del problema de clasificación se denomina como aprendizaje supervisado, mientras que la agrupación se denomina como aprendizaje no supervisado.



Explicación

Nombre	Descripción	Fundamentos o representación	Ventajas	Desventajas
Regresión logística	La regresión logística es uno de los métodos más poderosos de la estadística que se utilizan para analizar los conjuntos de datos que tienen al menos una variable independiente para determinar el resultado.	➢ Se mide la relación entre una o más variables independientes y la variable dependiente categórica mediante la estimación de sus probabilidades. El resultado tiene dos valores de naturaleza categórica. ➢ Se utiliza para predecir la probabilidad de que ocurra o no ocurra un evento.	• Puede evitar un ajuste excesivo. • Se puede realizar la selección de funciones. • La salida se puede interpretar como una medida de probabilidad. • El sistema es robusto al ruido.	• Para lograr resultados significativos y estables se requieren más datos. • Sobreajuste.
Redes neuronales artificiales	Es un modelo computacional inspirado en el funcionamiento del sistema neuronal biológico. El modelo se entrena ajustando los pesos a la espera de que se ajusten a una relación entrada-salida con los datos.	Se compone de elementos interconectados conocidos como neuronas o nodos que trabajan para producir la función de salida.	• Aprendizaje adaptativo. • Autoorganización. • Predicción rápida. • Identifica fácilmente las relaciones complejas. • Puede manejar los datos ruidosos.	• Todas las entradas deben convertirse en numéricas. • Proporciona un óptimo local. • Sobreajuste. • Difícil de interpretar. • El tiempo de procesamiento es elevado. • No se puede inicializar con conocimientos previos.

Explicación

Después de haber realizado un repaso a las técnicas de aprendizaje supervisado, es necesario comprender que existe otro paradigma de aprendizaje ilustrado en el diagrama de bloques de la figura: no supervisado.

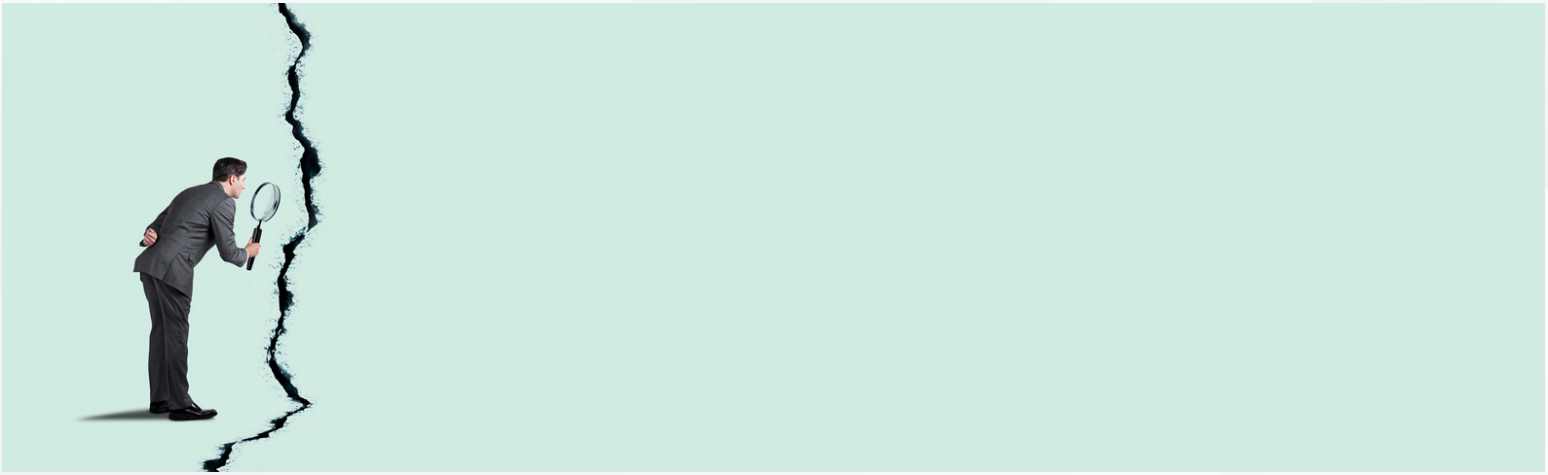
La agrupación de elementos de datos es la tarea a la que se aplican los algoritmos de aprendizaje no supervisados.

Para aquellos casos en donde las etiquetas no están disponibles en lo absoluto será necesario realizar un análisis exploratorio sin supervisión para comprender la estructura y composición de los datos. En general, el aprendizaje no supervisado marca un pilar importante del aprendizaje automático moderno.

Explicación

Estas son las principales aplicaciones del aprendizaje no supervisado:

- Segmentación de conjuntos de datos que comparten atributos.
- Detección de anomalías.
- Simplificación de conjuntos de datos.



Explicación

En la siguiente tabla se resumen los principales algoritmos de **aprendizaje no supervisado.**

Nombre	Descripción	Fundamentos o representación	Ventajas	Desventajas
K-medias	Tiene como objetivo encontrar y agrupar en clases los puntos de datos con alta similitud. En los términos del algoritmo, esta similitud se entiende como lo opuesto de la distancia entre puntos de datos.	La distancia más comúnmente utilizada en K-means, que es la distancia cuadrada de euclidiana: $d(x,y) = \sum_{j=1}^{m} (x_j - y_j)^2$ Por ende, la inercia del clúster será la suma de errores cuadrados dentro del contexto del clustering.	• Muy fácil de implementar. • Eficiente desde el punto de vista computacional.	• El resultado de cualquier set de entrenamiento fijo no siempre será el mismo, ya que los centroides iniciales se fijan al azar. • Debido a la naturaleza de la distancia euclídea, no es un algoritmo adecuado cuando se trata con clústeres que adoptan formas no esféricas.
Clusterización jerárquica	Se utilizan para agrupar datos jerárquicamente.	Divisivo: engloba todos los puntos de datos en un solo grupo para después dividirlo en otros más pequeños hasta que cada uno de ellos contenga solo una muestra. Aglomerativo: cada muestra inicia siendo un grupo diferente, para luego ser fusionadas por las que están más cerca de otras hasta que queda un solo grupo.	• No es necesario especificar el número de agrupaciones. • Permite el trazado de dendogramas (visualizaciones de una agrupación jerárquica binaria). • Es potente cuando el conjunto de datos contiene relaciones jerárquicas reales.	• Altamente sensible a los valores atípicos, disminuyendo así el rendimiento del modelo significativamente. • Costoso desde el punto de vista informático y computacional.

Explicación

Nombre	Descripción	Fundamentos o representación	Ventajas	Desventajas
Agrupamiento basado en densidad (DBSCAN)	Es un algoritmo útil para identificar correctamente el ruido en los datos.	Se basa en un número de puntos con un radio especificado ε y hay una etiqueta especial asignada a cada punto de datos siguiendo este proceso: 1. Se especifica un número de puntos vecinos, MinPts. Si estos existen y caen en el radio ε, se asigna un punto central 2. Un punto fronterizo cae en el radio de ε de un punto central, pero tiene menos vecinos que el número de MinPts.	• No es necesario especificar el número de grupos. • Es flexible con las formas y tamaños que los grupos pueden adoptar. • Es muy útil para identificar y tratar con datos con ruido y valores atípicos.	• Tiene dificultades para tratar puntos de encuentro alcanzables por dos grupos. • No encuentra racimos de densidades variables correctamente.
Modelo de Agrupamiento Gaussiano	Modelo probabilístico en donde se asume que todas las muestras son generadas a partir de una mezcla de un número finito de distribuciones gaussianas con parámetros desconocidos.	Cada punto de datos pertenece a un grupo del conjunto de datos, pero con diferentes niveles de pertenencia. Se asignará cada dato como miembro de acuerdo con una probabilidad que oscila entre 0 y 1.	• Es un método de agrupación en grupos blandos que asigna etiquetas de pertenencia a varios grupos. Esta característica lo convierte en el algoritmo más rápido para aprender modelos de mezclas. • Existe una gran flexibilidad en el número y la forma de los grupos.	• Es muy sensible a los valores iniciales que condicionarán en gran medida su rendimiento. • El MMG puede converger a un mínimo local, lo que constituiría una solución que no es óptima. • Es necesario regularizar artificialmente las covarianzas entre los puntos de datos.

Explicación

Después de haber abordado ambos paradigmas surge la siguiente pregunta: ¿qué tipos de plataformas para ciencia de datos y aprendizaje automático existen?

Plataformas locales

La nube no siempre es la respuesta, ya que no siempre es una solución viable. No todos los expertos en datos pueden darse el lujo de trabajar en la nube por varias razones, incluida la seguridad de los datos y problemas relacionados con la latencia. En casos como el cuidado de la salud o regulaciones estrictas se requiere que los datos estén seguros.

Plataformas en la nube

Estas plataformas brindan la capacidad de entrenar e implementar los modelos en la nube, lo cual también ayuda cuando estos modelos se están integrando en varias aplicaciones, ya que proporciona un acceso más fácil para cambiar y ajustar los modelos que se han implementado.

Plataformas de borde

Algunas plataformas permiten la puesta en marcha de algoritmos en el borde, lo cual consiste en una red en malla de centros de datos que procesan y almacenan datos localmente antes de enviarlos a un centro de almacenamiento centralizado o en la nube.

Actividad

Después de haber estudiado el tema puedes abordar las siguientes cuestiones:

- ¿En qué casos debe usarse el aprendizaje supervisado y en cuáles el no supervisado?
- Si tuvieras que diseñar un programa que clasificara imágenes de frutas, ¿qué paradigma emplearías?
- Si necesitaras encontrar similitudes en un conjunto de diversos datos con el propósito de agruparlos, ¿usarías el paradigma supervisado o el no supervisado?

Cierre

Después de resumir y comparar las principales técnicas de clasificación del aprendizaje supervisado y no supervisado, ahora puedes comprender a qué tipo de problemas y datos se adapta mejor cada enfoque. Asimismo, se consideraron las ventajas y desventajas de cada algoritmo para ayudarte a decidir cuál de ellos es el más conveniente para el proyecto que estés realizando.

Además, se presentaron los tipos de plataformas digitales empleadas en la ciencia de datos y en el aprendizaje automático, las cuales ayudan al desarrollo de soluciones comerciales, remarcando su importancia, categorización y beneficios. Del mismo modo, se mencionaron los desafíos que enfrentan al aplicarse en cualquier ámbito.

Finalmente, valdría la pena hacerse las siguientes preguntas: ¿cuáles otras diferencias identificas entre el aprendizaje supervisado y el no supervisado?, ¿qué herramientas digitales usarías para abordar un problema de aprendizaje automático?, ¿cuál enfoque crees que sea más costoso?, ¿por qué?

Tecmilenio no guarda relación alguna con las marcas mencionadas como ejemplo. Las marcas son propiedad de sus titulares conforme a la legislación aplicable, estas se utilizan con fines académicos y didácticos, por lo que no existen fines de lucro, relación publicitaria o de patrocinio.
