Módulo 1

Lección 3

Aprendizaje Automático No Supervisado

Contenido

Haz clic sobre los títulos para navegar en cada tema.

Tema 1. Introducción

El aprendizaje no supervisado es una rama del aprendizaje automático, donde el modelo se entrena sin la guía de etiquetas o respuestas conocidas. En lugar de aprender a predecir resultados específicos, el objetivo principal es descubrir:

Patrones

Estructuras

Relaciones intrínsecas en los datos

Volver a Contenido

Tipos principales de aprendizaje no supervisado

Agrupamiento (Clustering)

Agrupa datos similares en conjuntos o clústeres sin conocimiento previo de las categorías.

Ejemplo: **segmentación de clientes en grupos basados en patrones de compra similares.**

Reducción de dimensionalidad

Reduce la cantidad de variables en un conjunto de datos manteniendo la información esencial.

Ejemplo: **análisis de componentes principales (PCA) para reducir dimensiones manteniendo la variabilidad**

Asociación

Rescubre patrones de asociación entre variables en conjuntos de datos.

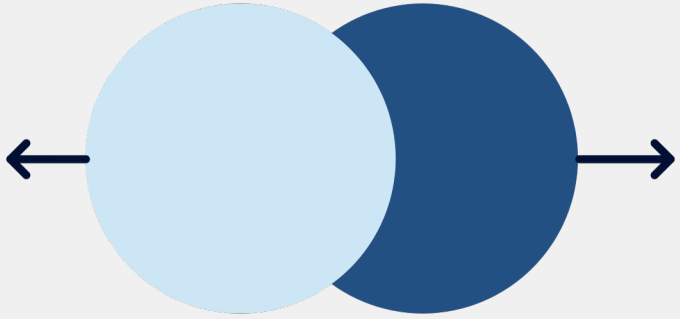
Ejemplo: **análisis de la cesta de la compra para identificar productos frecuentemente comprados juntos.**

Papel del aprendizaje no supervisado

- Descubrimiento de estructuras ocultas: identificación de patrones y relaciones que pueden no ser evidentes en los datos sin etiquetas.
- Generación de representaciones latentes: creación de representaciones compactas y significativas de los datos, facilitando análisis y visualización.
- Exploración y entendimiento de datos: ayuda a revelar la complejidad y las interconexiones dentro de grandes conjuntos de datos.

Desafíos y consideraciones

Evaluación subjetiva: la evaluación del rendimiento en el aprendizaje no supervisado puede ser subjetiva ya que no hay respuestas conocidas.

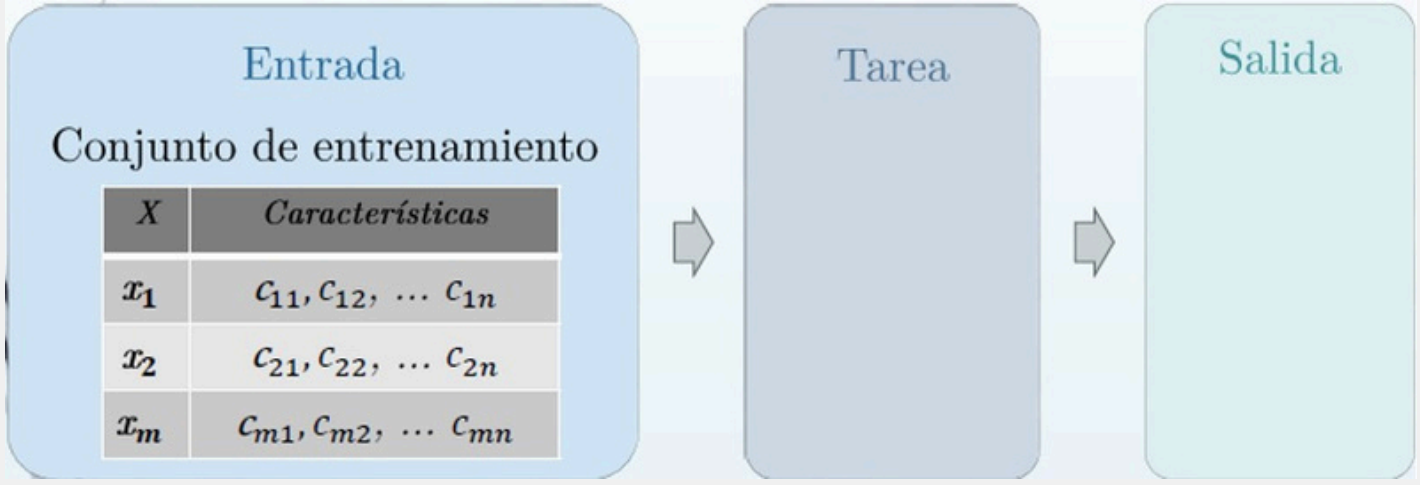


Selección de métodos: la elección de técnicas de aprendizaje no supervisado adecuadas depende de la naturaleza de los datos y los objetivos específicos.

Tendencias y futuras direcciones

- Aprendizaje No Supervisado Profundo: la aplicación de técnicas de aprendizaje profundo en el ámbito no supervisado para descubrir patrones más complejos.
- Integración con aprendizaje supervisado: enfoques que combinan aprendizaje no supervisado y supervisado para mejorar la generalización y el rendimiento.
- Su evolución hacia enfoques más profundos: promete un mayor entendimiento de la complejidad inherente a grandes volúmenes de datos no etiquetados.

El aprendizaje no supervisado desempeña un papel crucial en la exploración y comprensión de datos sin etiquetas claras. Desde la agrupación hasta la reducción de dimensionalidad, estas técnicas son fundamentales para revelar estructuras y patrones subyacentes en conjuntos de datos complejos. En el aprendizaje no supervisado, el agente aprende la distribución de un conjunto de entrenamiento X sin etiquetas.



Tema 2. Clustering

Clustering es una técnica de aprendizaje no supervisado que agrupa conjuntos de datos similares en categorías o clústeres. El objetivo es que los elementos dentro de un clúster sean más similares entre sí que con aquellos en otros clústeres, sin conocimiento previo de las categorías.

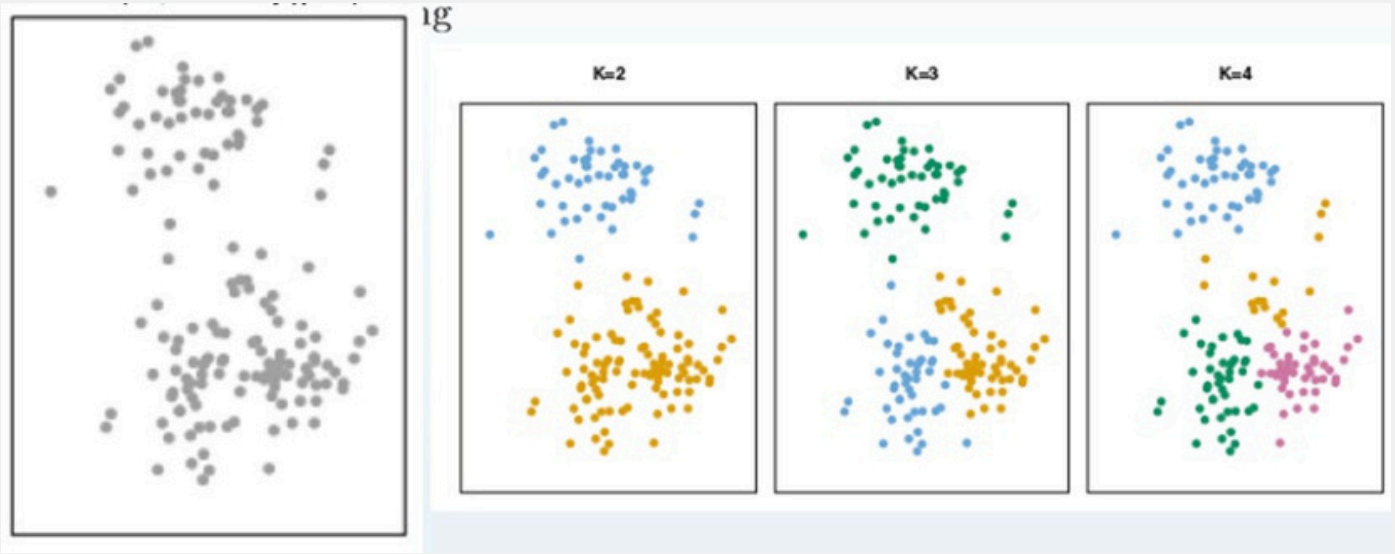
Algunos tipos de algoritmos de clustering son:

- K-Means
- Aglomerativo (Hierarchical Agglomerative Clustering – HAC)
- DBSCAN (Density-Based Spatial Clustering of Applications with Noise)
- Meanshift

Volver a Contenido

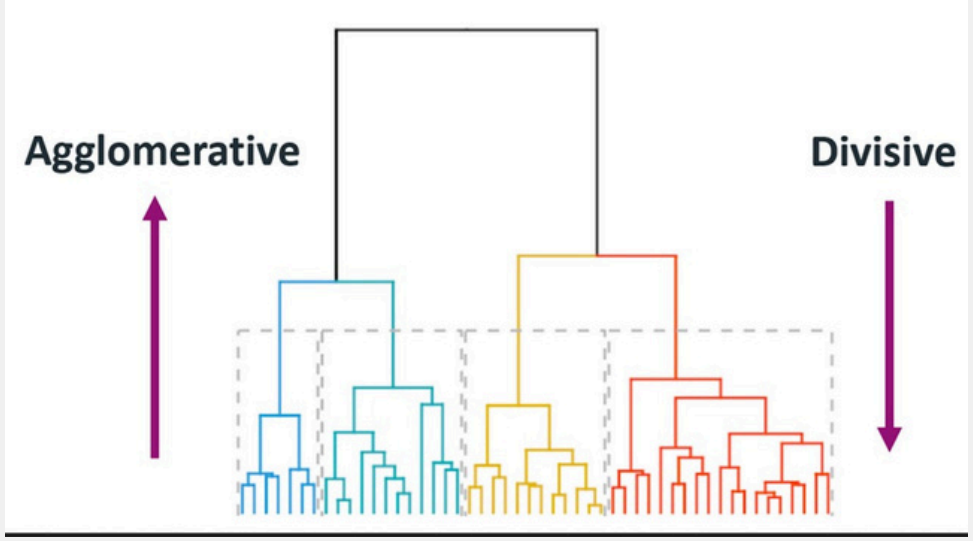
K-Means

Agrupa datos en k clústeres asignando cada punto al clúster cuyo centroide está más cercano. Por ejemplo: la segmentación de clientes basada en comportamientos de compra.



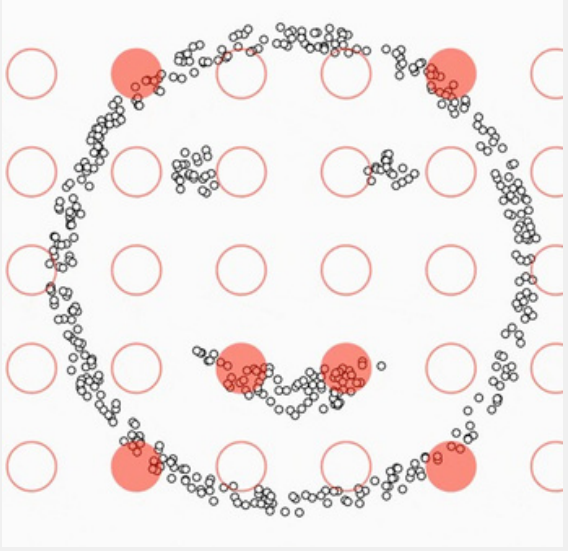
Aglomerativo (Hierarchical Agglomerative Clustering – HAC)

Comienza con cada punto como un clúster individual y fusiona gradualmente los clústeres más cercanos. Por ejemplo: el análisis de similitudes genéticas entre diferentes especies.



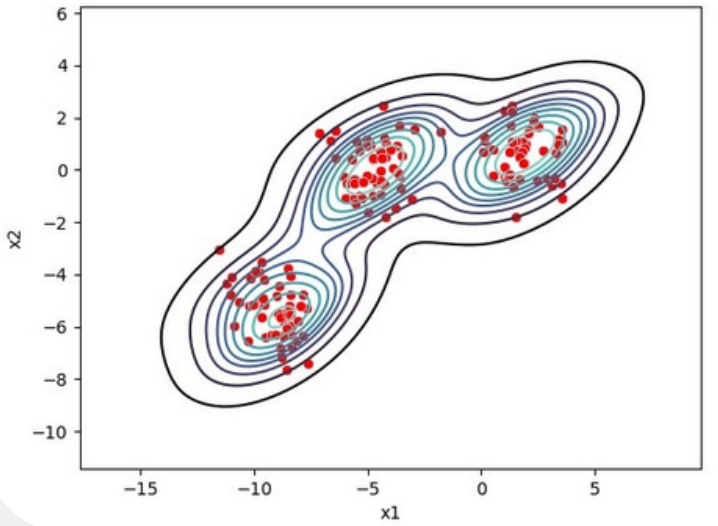
DBSCAN (Density-Based Spatial Clustering of Applications with Noise)

Define clústeres basándose en la densidad de puntos, identificando áreas densas y áreas dispersas. Por ejemplo: la detección de anomalías en conjuntos de datos con ruido.



Meanshift

encuentra clústeres maximizando la densidad de puntos en el espacio de características. Por ejemplo: el seguimiento de objetos en vídeos basado en sus características.



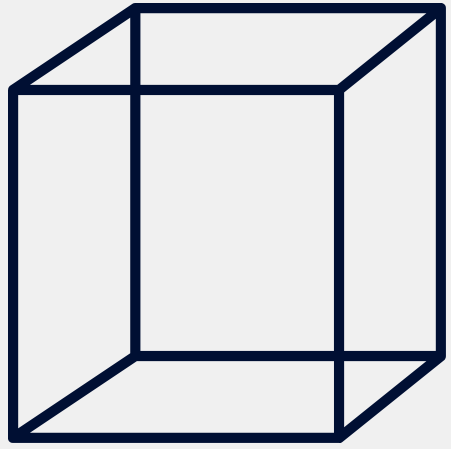
Conozcamos algunos ejemplos prácticos de aplicación:

Marketing y segmentación de clientes	Biología y genómica	Reconocimiento de patrones en imágenes	Detección de fraudes en finanzas
Utilizando técnicas de clustering para identificar grupos de clientes con patrones de comportamiento similares, permitiendo estrategias de marketing más efectivas.	Agrupación de datos genéticos para identificar similitudes y relaciones evolutivas entre especies.	Aplicación de algoritmos de clustering para segmentar y reconocer patrones en imágenes, facilitando la clasificación automática.	Identificación de patrones de transacciones fraudulentas mediante la agrupación de comportamientos anómalos.

Clustering es una herramienta poderosa en diversas disciplinas, desde el análisis de datos hasta la toma de decisiones en negocios y ciencia. La variedad de algoritmos disponibles permite adaptarse a diferentes tipos de datos y objetivos, proporcionando una amplia gama de aplicaciones prácticas en el mundo real.

Tema 3. Reducción de dimensionalidad

La reducción de dimensionalidad, es una técnica que se utiliza para reducir el número de variables o características en un conjunto de datos manteniendo la información esencial. En otras palabras, busca representar la información de un conjunto de datos en un espacio de menor dimensión sin perder demasiada información crucial.



Volver a Contenido

Importancia de reducir la dimensionalidad en conjuntos de datos grandes

1. Eficiencia Computacional: conjuntos de datos con muchas variables pueden requerir más recursos computacionales. Reducir la dimensionalidad ayuda a gestionar la carga computacional.

2. Evitar la Maldición de la Dimensionalidad: en altas dimensiones, la distancia entre puntos aumenta, lo que puede afectar negativamente la calidad de los modelos. Reducir la dimensionalidad contrarresta este problema.

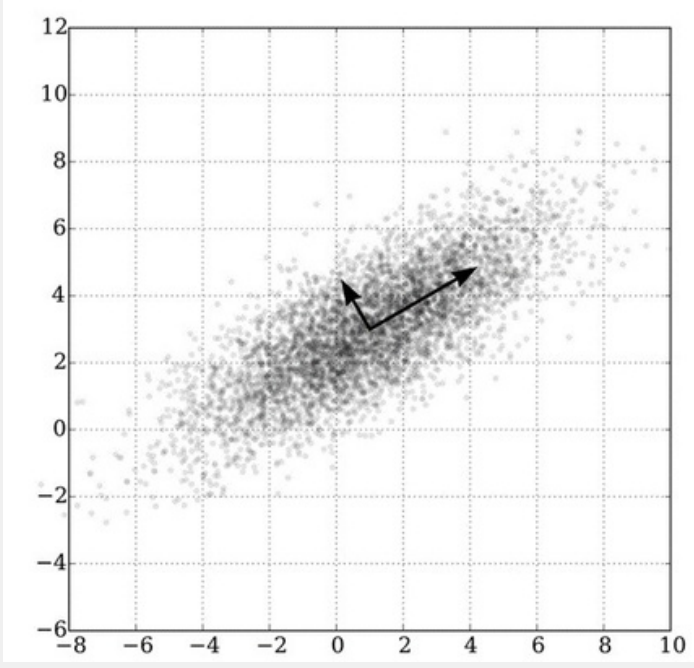
3. Visualización: la representación en espacios de menor dimensión facilita la visualización y comprensión de patrones y relaciones en los datos.

4. Mejora del Rendimiento de Modelos: al reducir la dimensionalidad, los modelos tienden a generalizar mejor, evitando el sobreajuste y mejorando su capacidad predictiva.

Métodos comunes de reducción de dimensionalidad: PCA (Análisis de Componentes Principales)

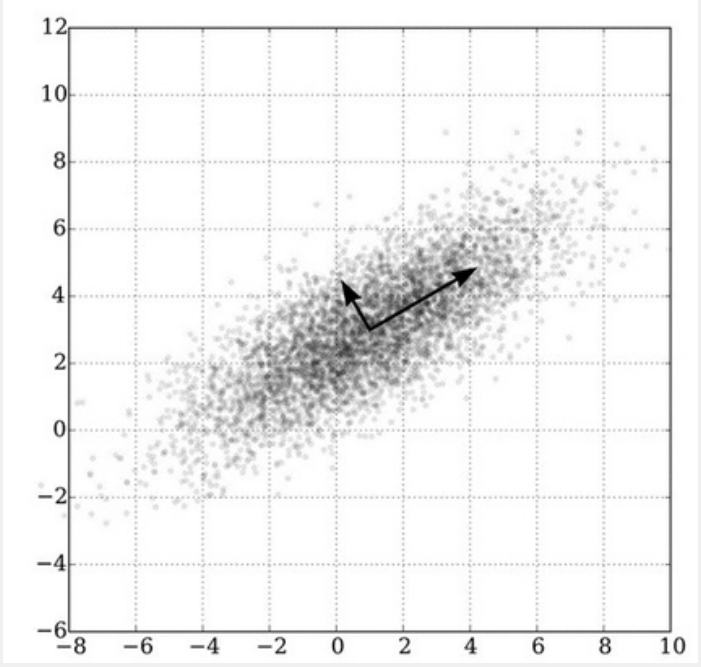
1. Definición:

El Análisis de Componentes Principales (PCA), es una técnica de reducción de dimensionalidad que busca transformar un conjunto de datos originales en un nuevo conjunto de variables, llamadas componentes principales. Estos componentes principales son combinaciones lineales de las variables originales y están ordenados de manera que la primera componente principal captura la mayor varianza en los datos, la segunda la segunda mayor, y así sucesivamente.



2. Cómo Funciona:

- Identificación de Direcciones Principales de Variabilidad: PCA identifica las direcciones en las cuales los datos varían más. Estas direcciones son las que maximizan la varianza en el conjunto de datos.
- Proyección en Nuevas Direcciones: una vez identificadas las direcciones principales, los datos se proyectan en estas nuevas direcciones. Esto implica expresar cada observación del conjunto de datos en términos de las componentes principales.

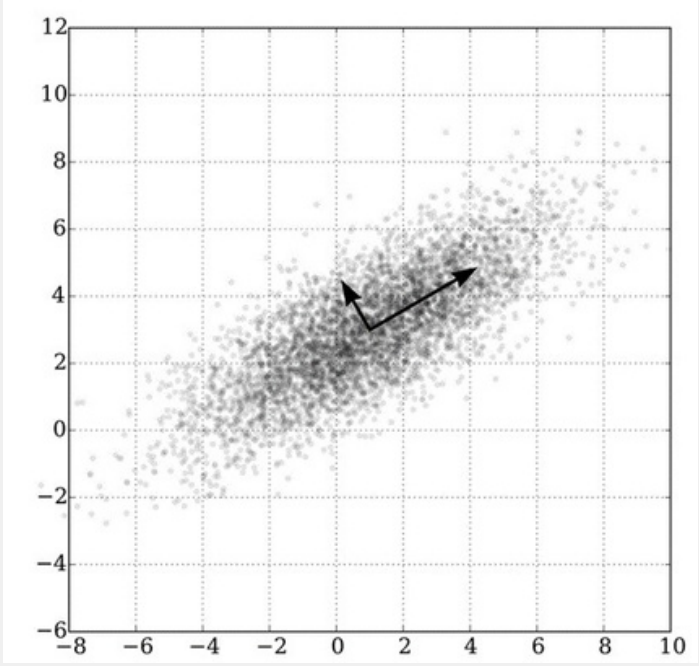


3. Importancia de PCA:

- Maximiza la Varianza: La primera componente principal captura la mayor cantidad de variabilidad presente en los datos. Esto significa que la información más importante se conserva en esta nueva representación.
- Preserva la Información Relevante: al retener las primeras k componentes principales, donde k es un número menor que el número original de variables, se conserva la mayor parte de la información importante mientras se reduce la dimensionalidad.
- Elimina la Correlación entre Variables: las componentes principales son ortogonales entre sí, lo que elimina la correlación entre las variables originales y simplifica la interpretación de las relaciones en los datos.

4. Ejemplo Práctico de PCA:

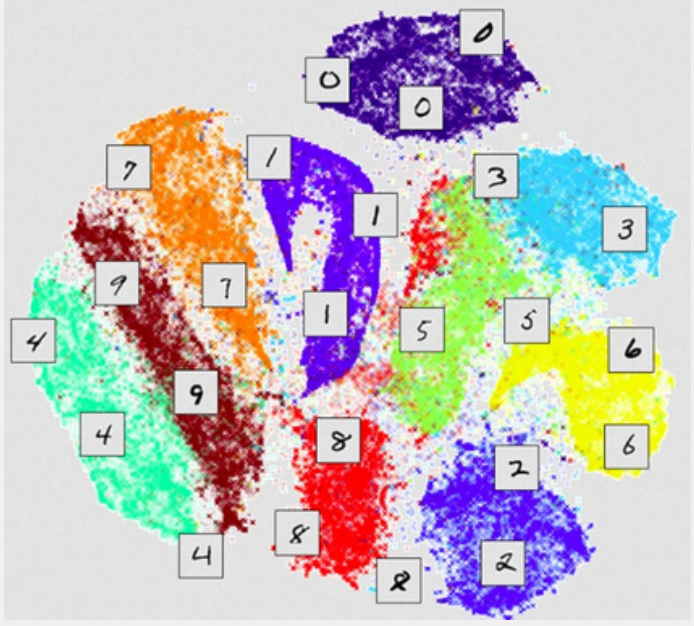
Supongamos que tenemos un conjunto de datos en dos dimensiones (x, y), que representa la variación en la altura y el peso de una muestra de individuos. Aplicar PCA nos dará nuevas dimensiones (componentes principales) que representan las direcciones de máxima variabilidad en los datos; la primera componente principal podría estar relacionada con la altura, mientras que la segunda podría representar la variabilidad en el peso.



Métodos comunes de reducción de dimensionalidad: t-SNE (T-distributed Stochastic Neighbor Embedding)

1. **Definición:**

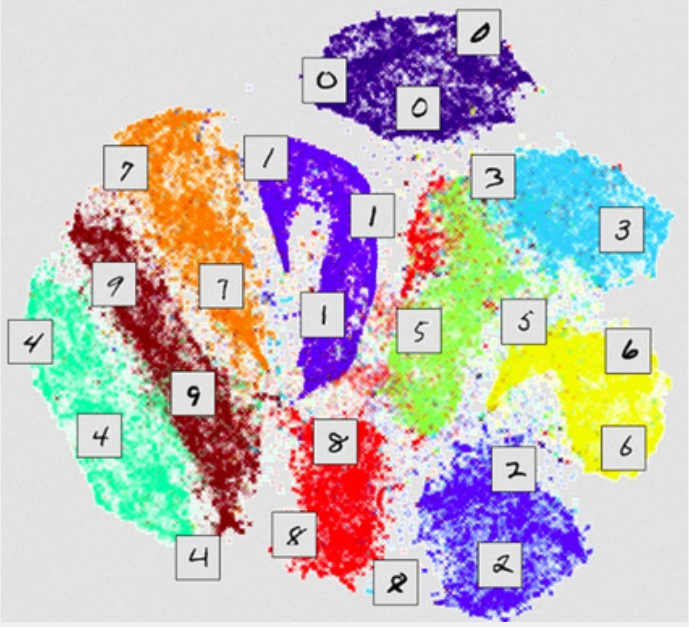
Es una técnica de reducción de dimensionalidad utilizada para visualizar conjuntos de datos complejos en un espacio de menor dimensión. A diferencia de PCA, t-SNE se centra en la preservación de las relaciones de similitud entre puntos en el espacio original al representarlos en un espacio de menor dimensión.



2. Cómo Funciona:

- **Similitud entre Puntos:** t-SNE calcula la similitud entre pares de puntos en el espacio original. La similitud se mide en términos de probabilidad condicional.

- **Representación en un Nuevo Espacio:** a partir de las similitudes, t-SNE genera una representación en un nuevo espacio de menor dimensión, donde las similitudes entre puntos se conservan tanto como sea posible.



3. Importancia de t-SNE:

- **Preservación de Relaciones de Similitud:** la principal fortaleza de t-SNE radica en su capacidad para preservar las relaciones de similitud entre puntos; puntos similares en el espacio original tienden a estar próximos en el espacio de menor dimensión.

- **Visualización Efectiva:** es especialmente útil para la visualización de conjuntos de datos complejos en dos o tres dimensiones, permitiendo la identificación de agrupaciones y patrones de manera más efectiva.

4. Ejemplo Práctico de t-SNE:

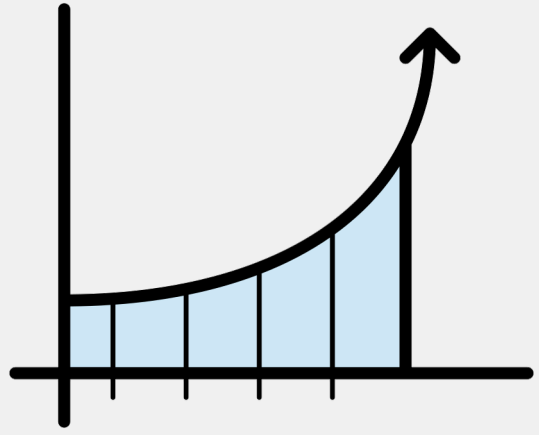
Supongamos que tenemos un conjunto de datos que representa características genéticas de individuos. La aplicación de t-SNE podría ayudar a visualizar la similitud entre individuos en función de estas características en un espacio de menor dimensión, facilitando la identificación de grupos genéticos relacionados.

El t-SNE es una herramienta valiosa para la visualización de datos complejos, especialmente cuando la preservación de las relaciones de similitud es crucial. Su capacidad para representar puntos similares próximos en el espacio de menor dimensión lo convierte en una elección efectiva para la exploración visual de conjuntos de datos multidimensionales.

La reducción de dimensionalidad es una herramienta valiosa para manejar conjuntos de datos grandes y complejos. Permite simplificar la información mientras se mantiene la esencia de los datos, mejorando la eficiencia computacional y facilitando la interpretación y visualización de patrones.

Tema 4. Asociación

Las reglas de asociación, son un componente clave en el análisis de datos que busca identificar patrones de co-ocurrencia entre diferentes variables en un conjunto de datos. Esta técnica es particularmente útil en la identificación de relaciones entre elementos en conjuntos de datos transaccionales.



Volver a Contenido

Cómo funcionan las reglas de asociación

Soporte

El soporte mide la frecuencia con la que una asociación particular aparece en el conjunto de datos. Cuanto mayor sea el soporte, más frecuente es la asociación.

Confianza

La confianza mide la probabilidad de que la ocurrencia de un elemento lleve a la ocurrencia de otro. Indica la fuerza de la relación entre los elementos.

Aplicación en análisis de patrones:

Recomendación de Productos

En el comercio electrónico, las reglas de asociación pueden utilizarse para identificar productos que tienden a comprarse juntos. Por ejemplo, si los clientes que compran laptops también tienden a comprar mochilas, y se pueden generar recomendaciones personalizadas.

Marketing y Estrategias de Ventas

Las reglas de asociación son valiosas en marketing para comprender las relaciones entre diferentes productos o servicios. Esto permite a las empresas diseñar estrategias de venta cruzada y promociones efectivas.

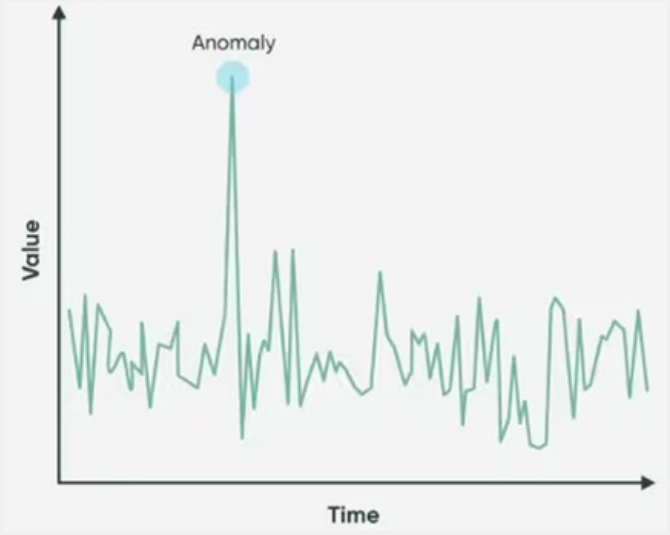
Conozcamos algunos ejemplos prácticos de aplicación:

Recomendación de Productos	Marketing Personalizado
Si un análisis de reglas de asociación revela que los clientes que compran cámaras tambiénindican que los clientes que compran cereales tienden a comprar trípodes, un sitio web de comercio electrónico podría ofrecer paquetes de productos o descuentos en trípodes al comprar una cámara.	En un supermercado, si las reglas de asociación a menudo también compran leche, el supermercado podría enviar cupones personalizados para cereales a clientes que han comprado leche recientemente.

Las reglas de asociación son herramientas poderosas para descubrir patrones significativos en grandes conjuntos de datos transaccionales. Su aplicación en estas áreas del marketing puede impulsar la personalización y la eficacia de las decisiones comerciales.

Tema 5. Detección de anomalías

En el contexto del aprendizaje automático, las anomalías son patrones o eventos que difieren significativamente del comportamiento normal o esperado en un conjunto de datos. La detección de anomalías se centra en identificar estas instancias inusuales, que podrían indicar problemas, fraudes o situaciones atípicas.



Volver a Contenido

- Técnicas para detectar anomalías

Métodos Estadísticos: utilizan medidas como la desviación estándar para identificar puntos que se desvían significativamente de la media.

Aprendizaje No Supervisado: modelos de clustering y reducción de dimensionalidad pueden revelar instancias que no se ajustan al comportamiento general del conjunto de datos.

Aprendizaje Supervisado: utiliza modelos entrenados con datos normales para identificar instancias que difieren notablemente durante la evaluación.

Volver a Contenido

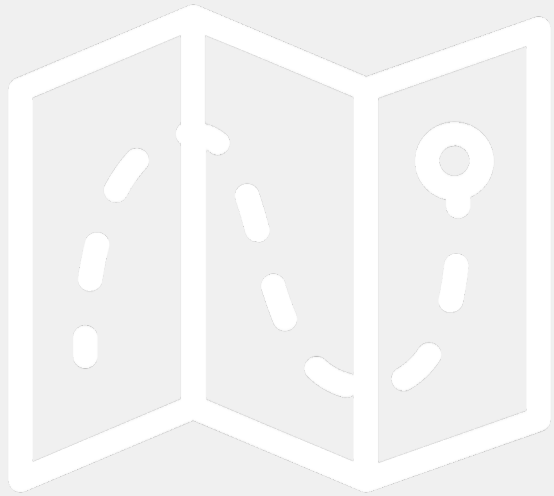
Conozcamos algunos ejemplos prácticos de aplicación:

Ciberseguridad	Mantenimiento predictivo	Servicios financieros
La detección de anomalías es esencial para identificar actividades sospechosas en redes informáticas, patrones inusuales de tráfico o comportamientos pueden indicar intentos de intrusión o ataques. Estaejemplo, en una fábrica, la detección puede identificar patrones de tráfico de red inusuales, como intentos de acceso no autorizado o malware.	En la industria, la detección de anomalías se utiliza para monitorear el estado de la maquinaria; cambios abruptos en el rendimiento pueden indicar la necesidad de mantenimiento preventivo. Por de anomalías puede alertar sobre cambios en la vibración de una máquina, indicando posibles problemas mecánicos.	En transacciones financieras, la detección de anomalías ayuda a identificar posibles fraudes, actividades inusuales en tarjetas de crédito o patrones de retiro atípicos pueden ser señales de alerta. Esta podría alertar sobre compras inusuales o retiros significativos que podrían ser fraudulentos.

La detección de anomalías es crucial en diversas industrias para prevenir riesgos, garantizar la seguridad y mantener un rendimiento eficiente de los sistemas y procesos. Su aplicación práctica abarca desde la ciberseguridad hasta el mantenimiento de maquinaria y la detección de fraudes financieros.

Tema 6. Mapas Autoorganizativos (SOM)

Los Mapas Autoorganizativos (Self-Organizing Maps, SOM), son una técnica de aprendizaje no supervisado que visualiza relaciones complejas y estructuras en datos multidimensionales, funcionan mediante la proyección de datos de alta dimensión en un mapa bidimensional o tridimensional, preservando las relaciones topológicas entre los datos originales. El proceso de autoorganización implica que regiones adyacentes en el mapa corresponden a patrones similares en los datos originales.



Volver a Contenido

Elementos de los mapas organizativos (SOM)

- Inicialización: **se asignan aleatoriamente pesos a los nodos del mapa.**
- Competencia: **los datos se presentan al SOM, y los nodos compiten para representar los datos. En cada iteración, selecciona k-1 pliegues para entrenar el modelo.**
- Cooperación: **los nodos vecinos también se ajustan para representar características similares, favoreciendo la organización topológica.**
- Adaptación Continua: **a medida que se presentan más datos, los pesos de los nodos se ajustan continuamente para reflejar la estructura subyacente de los datos.**

Casos de uso y ejemplos de implementación:

Análisis de Datos Geoespaciales	Segmentación de Clientes en Comercio Electrónico	Procesamiento de Imágenes	Biología Computacional
Los SOM se utilizan para visualizar y entender patrones en datos geoespaciales, como la distribución de recursos naturales o la clasificación de paisajes.	Los SOM pueden identificar segmentos de clientes basados en patrones de comportamiento de compra, lo que ayuda en estrategias de marketing	En el campo de la visión por computadora, los SOM se aplican para organizar y clasificar imágenes en función de características visuales.	Los SOM se utilizan para analizar datos biológicos como expresión génica, ayudando a identificar patrones y relaciones en grandes conjuntos de datos biológicos.

- Ejemplos prácticos de los SOM

1. Ciencias Ambientales: visualizar patrones climáticos complejos al representar datos meteorológicos multidimensionales en un mapa bidimensional.

2. Retail y Segmentación de Clientes: identificar grupos de clientes con comportamientos de compra similares, facilitando estrategias de marketing específicas para cada segmento.

3. Análisis de Datos de Imágenes Médicas: organizar y clasificar imágenes médicas en función de características visuales, facilitando diagnósticos más precisos.

4. Exploración de Expresión Génica: se utilizan para analizar grandes conjuntos de datos de expresión génica, revelando patrones subyacentes en la regulación genética.

Volver a Contenido

Los Mapas Autoorganizativos son herramientas versátiles que encuentran aplicaciones en diversas disciplinas, desde la segmentación de clientes en negocios hasta la exploración de patrones en datos biológicos. Su capacidad para preservar estructuras topológicas en datos complejos los hace valiosos en la visualización y comprensión de conjuntos de datos multidimensionales.

Tema 7. Aplicaciones prácticas

- Estudio de casos reales que demuestren la utilidad de técnicas no supervisadas en problemas del mundo real. Discusión sobre cómo estas técnicas pueden mejorar la toma de decisiones.

Volver a Contenido

Tema 8. Avances y tendencias

Exploración de avances recientes en Aprendizaje No Supervisado.

Discusión sobre las tendencias emergentes y su impacto en diversas industrias.

Volver a Contenido

Tema 9. Aprendizaje Automático SemiSupervisado

El aprendizaje automático semi supervisado es un enfoque que combina elementos de los paradigmas supervisado y no supervisado. En este escenario, el modelo se entrena utilizando un conjunto de datos que contiene tanto ejemplos etiquetados como no etiquetados; la presencia de datos no etiquetados permite al modelo aprender patrones y estructuras subyacentes adicionales, mejorando su capacidad de generalización.

Volver a Contenido

- Diferencias con enfoques supervisado y no supervisado

Supervisado: utiliza un conjunto de datos completamente etiquetado para entrenar el modelo, donde cada entrada tiene una etiqueta correspondiente.

No Supervisado: se basa en datos no etiquetados y busca patrones o estructuras intrínsecas en los datos sin la guía de etiquetas previas.

Semi-supervisado: combina datos etiquetados y no etiquetados, permitiendo que el modelo aprenda tanto de la información explícita proporcionada por las etiquetas como de la estructura inherente en los datos no etiquetados.

Volver a Contenido

- Beneficios del aprendizaje semi-supervisado:

Problemas de clasificación con pocos datos etiquetados	Mejora de la Generalización	Adaptabilidad a cambios en los datos	Reducción de la dependencia de etiquetas	Mejora en problemas de agrupación (Clustering)
En situaciones donde la obtención de datos etiquetados es costosa o difícil, el aprendizaje semi supervisado permite aprovechar datos no etiquetados para mejorar el rendimiento del modelo.	Al incluir datos no etiquetados, el modelo puede aprender patrones más generales y robustos, mejorando su capacidad para generalizar a nuevas instancias,	En entornos dinámicos, donde la distribución de datos puede cambiar con el tiempo, el aprendizaje semi supervisado puede adaptarse mejor a estas variaciones al incorporar datos no etiquetados actualizados.	En casos donde es difícil obtener etiquetas precisas para todos los datos, el aprendizaje semi supervisado reduce la dependencia de una gran cantidad de datos etiquetados.	Al integrar datos no etiquetados, el modelo puede descubrir naturalmente grupos o patrones en los datos, mejorando la eficacia del agrupamiento.

El aprendizaje automático semi-supervisado, representa una estrategia valiosa cuando se enfrenta a limitaciones en la disponibilidad de datos etiquetados; al combinar información supervisada y no supervisada, se logra un equilibrio que puede mejorar la capacidad de generalización y adaptación del modelo en diversas situaciones. Este enfoque, es particularmente beneficioso en escenarios donde la recopilación de datos etiquetados es un desafío.