

SESIÓN APRENDIZAJE DE MÁQUINA NO SUPERVISADO

CONTENIDOS:

- Qué es el aprendizaje de máquina no supervisado.
- Tareas que resuelve.
- Tipos de aprendizaje no supervisado:
 - Clusterización.
 - Reducción de dimensionalidad.
- Aplicaciones de aprendizaje no supervisado.

QUÉ ES EL APRENDIZAJE DE MÁQUINA NO SUPERVISADO

El **aprendizaje de máquina no supervisado** es una rama del aprendizaje automático que no depende de datos etiquetados (es decir, datos que tienen una "respuesta" o clasificación predefinida). En lugar de eso, trabaja con datos "en bruto" para encontrar patrones, relaciones o estructuras subyacentes en ellos.

En el aprendizaje no supervisado, el modelo no tiene una guía específica sobre qué buscar en los datos. En lugar de aprender de pares de entrada y salida (como en el aprendizaje supervisado), analiza las características intrínsecas de los datos, de esta forma puede:

1. **Descubrir Agrupamientos o Grupos:** Identificar subconjuntos o categorías dentro de los datos que tienen características similares.
2. **Reducir Complejidad:** Simplificar grandes volúmenes de datos eliminando redundancias (reducción de dimensionalidad).
3. **Detectar Anomalías:** Encontrar datos que se desvían significativamente del patrón general, lo cual es útil, por ejemplo, en la detección de fraudes.

¿Por qué se utiliza el aprendizaje de máquina no supervisado?

El aprendizaje de máquina no supervisado se utiliza porque es una herramienta poderosa para explorar y analizar datos sin la necesidad de etiquetas o respuestas predefinidas. Esto es especialmente útil en contextos donde etiquetar los datos es costoso, impráctico o simplemente imposible.



Ilustración 1 Estructura de Datos en Aprendizaje Automático: Supervisado vs No Supervisado

Ventajas del aprendizaje no supervisado

- Se pueden descubrir patrones desconocidos en los datos.
- No requiere datos etiquetados (ahorra costos y esfuerzo).
- Es adaptable a una gran variedad de dominios (biología, marketing, tecnología, etc.).
- Requiere menos intervención humana.

Desventajas del aprendizaje no supervisado

- Puede ser difícil interpretar resultados.
- No siempre se garantiza que los patrones sean útiles o relevantes.
- Puede ser difícil validar si los resultados obtenidos son correctos o significativos al no haber etiquetas.

Ejemplo del uso del aprendizaje no supervisado

En plataformas de streaming como Netflix o Spotify, se emplean algoritmos de aprendizaje no supervisado, como la clusterización, para agrupar usuarios con preferencias similares basándose en su historial de interacciones. Esto permite ofrecer recomendaciones personalizadas sin necesidad de etiquetas explícitas, como géneros o categorías previamente definidas.



¿QUÉ TAREAS RESUELVE EL APRENDIZAJE NO SUPERVISADO?

1. **Análisis de Ventas o Finanzas**

- Ayudan a calcular totales, promedios, máximos, y mínimos en informes financieros o de ventas. Por ejemplo, sumar los ingresos mensuales de una tienda, promediar salarios de empleados, o identificar el producto más vendido.

2. **Control de Inventarios**

- Útiles para agrupar productos por categoría y calcular cantidades disponibles. Por ejemplo, saber cuántos productos hay de cada tipo o identificar el producto con menor stock.

3. **Evaluación de Rendimiento Académico o Profesional**

- Permiten evaluar grupos de estudiantes o empleados. Por ejemplo, calcular el promedio de calificaciones por curso, o agrupar empleados por departamento para analizar promedios salariales.

4. **Monitorización de Producción o Procesos**

- En un entorno industrial, permiten agrupar y analizar datos de producción. Por ejemplo, sumar productos fabricados por día o identificar el turno con mayor productividad.

5. **Segmentación de Clientes**

- Sirven para agrupar clientes por categoría (edad, ubicación, consumo) y analizar tendencias. Por ejemplo, identificar al cliente que más compra o calcular el gasto promedio por región.

6. **Análisis de Tendencias Históricas**

- Ayudan a analizar datos históricos agrupados por periodos de tiempo para detectar patrones o cambios. Por ejemplo, analizar las ventas trimestrales o promediar las temperaturas de cada mes.

TIPOS DE APRENDIZAJE NO SUPERVISADO: CLUSTERIZACIÓN Y REDUCCIÓN DE DIMENSIONALIDAD

El aprendizaje no supervisado se divide principalmente en diferentes tipos según los objetivos que se busquen en el análisis de los datos. Veamos los principales tipos, sus funciones y sus diferencias:

1. Clusterización o Agrupamiento

Consiste en dividir un conjunto de datos en grupos (clústeres) basados en similitudes. Los datos dentro de un grupo son más similares entre sí que con los de otros grupos.

- **Funciones:**
 - Segmentación de mercado (agrupar clientes según comportamiento de compra).
 - Agrupación de imágenes similares en bases de datos visuales.
 - Descubrimiento de patrones en datos médicos (por ejemplo, agrupar pacientes por síntomas comunes).
- **Algoritmos populares:**
 - *K-means*
 - *DBSCAN* (algoritmo basado en densidad)
 - *Hierarchical Clustering* (clusterización jerárquica).
- **Ejemplo:**

```
from sklearn.cluster import KMeans
import numpy as np

# Datos ficticios: compras de clientes (ingreso, gasto promedio mensual)
datos = np.array([[30, 300], [40, 400], [25, 250], [35, 350], [50, 500]])

# Aplicar K-Means con 2 clusters
modelo = KMeans(n_clusters=2, random_state=42)
modelo.fit(datos)

# Ver los clusters asignados
print("Etiquetas de cluster asignadas:", modelo.labels_)
```

Ilustración 2 código K-Means en Python

2. Reducción de Dimensionalidad

Busca simplificar grandes conjuntos de datos eliminando variables redundantes, mientras se conserva la mayor cantidad de información posible.

- **Funciones:**
 - Preprocesamiento en sistemas de aprendizaje supervisado.
 - Visualización de datos en gráficos comprensibles (por ejemplo, reducir de 100 variables a 2 dimensiones para generar gráficos 2D).
 - Identificación de características más relevantes en datos masivos, como en el análisis genómico.
- **Métodos populares:**
 - *Principal Component Analysis (PCA)*.
 - *t-SNE* (T-distributed Stochastic Neighbor Embedding).
 - *UMAP* (Uniform Manifold Approximation and Projection).
- **Ejemplo:**

```
from sklearn.decomposition import PCA
import numpy as np

# Datos ficticios con 5 variables
datos = np.random.rand(10, 5)

# Reducimos a 2 dimensiones con PCA
pca = PCA(n_components=2)
datos_reducidos = pca.fit_transform(datos)

print("Datos originales:", datos.shape)
print("Datos reducidos:", datos_reducidos.shape)
```

Ilustración 3 código PCA en Python

Otros tipos de aprendizaje no supervisado

3. Detección de Anomalías

Identifica datos que se desvían significativamente de los patrones comunes. Es ideal para encontrar eventos raros o comportamientos anómalos.

- **Funciones:**
 - Detección de fraudes en transacciones bancarias.
 - Mantenimiento predictivo (predecir fallos en máquinas antes de que ocurran).
 - Identificación de ataques cibernéticos.
- **Ejemplo de algoritmos:**
 - Modelos de mezcla gaussiana (Gaussian Mixture Models).
 - Autoencoders (una técnica de redes neuronales).

4. Sistemas de Recomendación (Filtrado Colaborativo No Supervisado)

Agrupar a usuarios o productos según patrones de comportamiento (sin etiquetas explícitas) para recomendar contenido.

- **Funciones:**
 - Recomendación de series, películas o canciones en plataformas como Netflix o Spotify.
 - Sugerencias de productos en sitios de comercio electrónico.
- **Ejemplo de métodos:**
 - Descomposición matricial.
 - Modelos basados en similitud entre usuarios o entre ítems.

5. Modelado de Distribuciones

Busca modelar la distribución probabilística de los datos y entender cómo están distribuidos en un espacio.

- **Funciones:**
 - Generación de datos sintéticos basados en distribuciones reales.
 - Identificación de características estructurales de los datos.
- **Ejemplo de algoritmos:**
 - Modelos de mezcla gaussiana (*Gaussian Mixture Models*).
 - Mapas autoorganizativos (*Self-Organizing Maps*).

Diferencias clave entre los tipos de aprendizaje no supervisado

Objetivo	Método	Aplicaciones
Algunos buscan agrupar (clusterización), otros reducir (reducción de dimensionalidad) y otros identificar anomalías.	Cada tipo usa algoritmos específicos adaptados a la tarea (como K-means para clusterización vs PCS para reducción de dimensionalidad).	Aunque hay traslape en aplicaciones, como en la detección de fraudes o segmentación de datos, cada tipo resuelve problemas específicos.

APLICACIONES DE APRENDIZAJE NO SUPERVISADO

El aprendizaje no supervisado tiene usos amplios y diversos en la vida real. Algunas aplicaciones destacadas incluyen:

1. **Marketing y Negocios: Segmentación de Clientes**
 - Identificar grupos de consumidores con comportamientos similares para personalizar campañas de marketing y aumentar la efectividad de las estrategias.



2. Finanzas: Detección de Fraudes

- Encontrar patrones de transacciones anómalas en bases de datos financieros para prevenir actividades fraudulentas.

3. Industria y Manufactura: Mantenimiento Predictivo

- Analizar datos de sensores en máquinas para predecir fallos antes de que ocurran, reduciendo costes y tiempos de inactividad.

4. Salud: Análisis Médico

- Agrupar pacientes con síntomas similares para mejorar diagnósticos, desarrollar tratamientos personalizados o analizar datos genómicos.

5. Tecnología: Sistemas de Recomendación

- Ofrecer contenido relevante a usuarios en plataformas como Netflix, YouTube o Spotify, agrupando usuarios con intereses similares sin etiquetado explícito.

6. Ciencias: Investigación y Análisis de Datos Complejos

- Analizar datos astronómicos para identificar galaxias con características comunes o encontrar patrones en grandes volúmenes de datos biológicos.