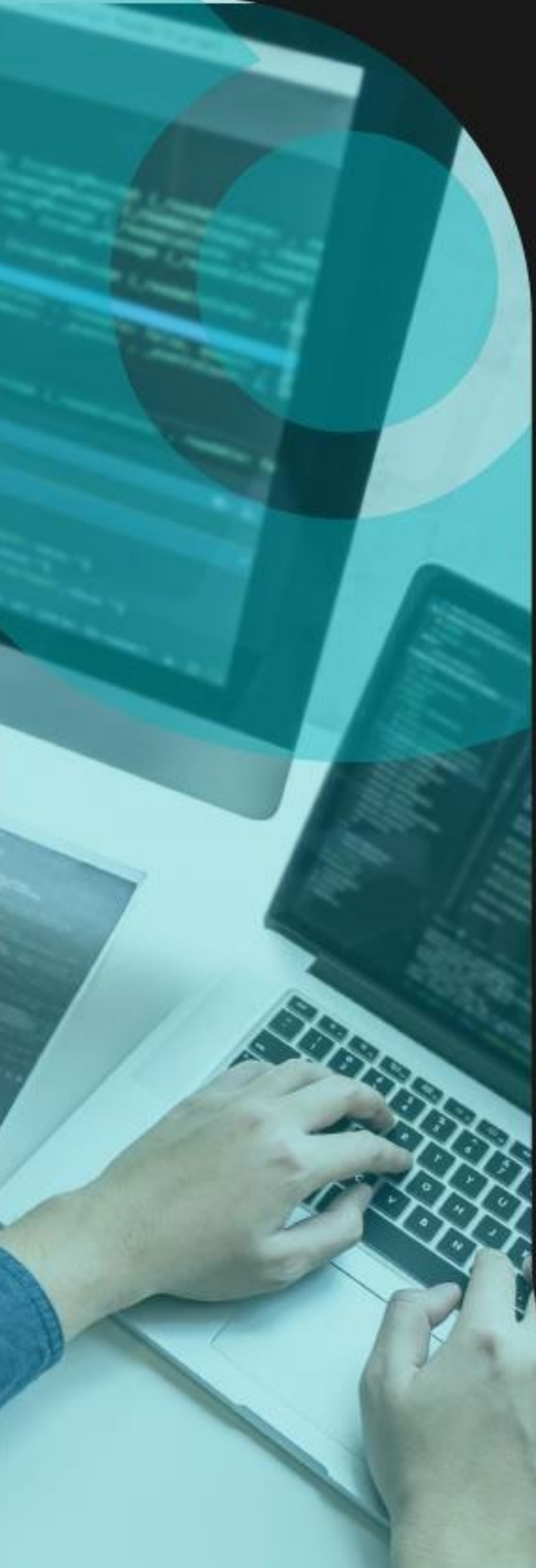


AWAKELAB

BASECAMP

Ciencia de Datos

Módulo: Aprendizaje de Máquina No Supervisado

Aprendizaje Esperado

---

1. Describir los principales conceptos y técnicas asociadas al aprendizaje de máquina no supervisado para resolver un problema.

Aprendizaje de máquina

Aprendizaje no supervisado

El aprendizaje automático no supervisado es el proceso de inferir patrones ocultos subyacentes a partir de datos históricos. Dentro de este enfoque, un modelo de aprendizaje automático intenta encontrar similitudes, diferencias, patrones y estructuras en los datos por sí mismo. No se necesita intervención humana previa.

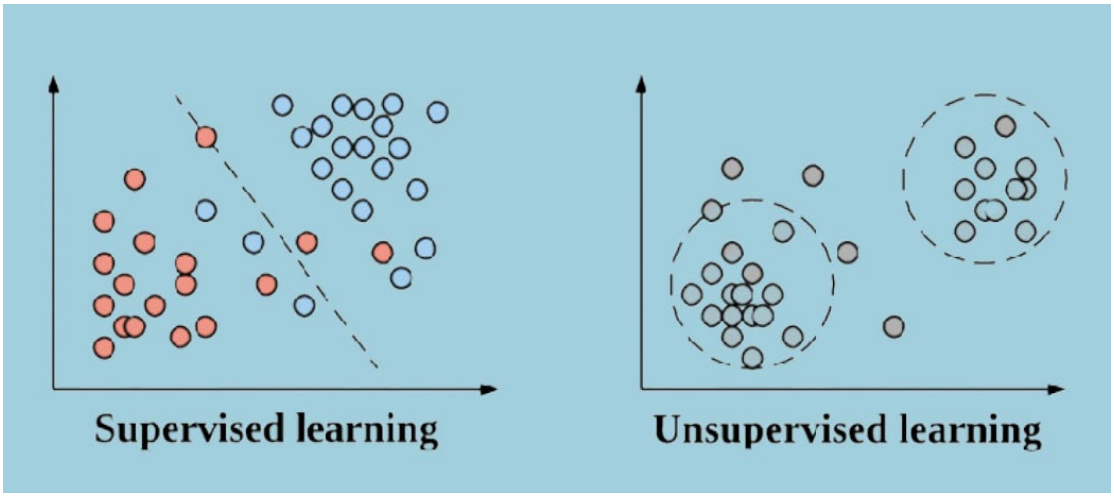
El algoritmo aprende a partir de la variable explicativa sin ninguna variable respuesta asociada, lo que le permite determinar los patrones de datos por sí mismo. Los algoritmos se dejan guiar por sus propios mecanismos para descubrir la estructura de datos.

Algunos algoritmos de Aprendizaje No Supervisado *:

- K - Medias
- K - Medians
- Sistemas de Recomendación

(*) Estudiaremos esto en mayor profundidad en las clases siguientes

Aprendizaje supervisado y aprendizaje no supervisado



La diferencia clave es que con *el aprendizaje supervisado* , un modelo aprende a predecir resultados en función del conjunto de datos etiquetado, lo que significa que ya contiene los ejemplos de respuestas correctas cuidadosamente trazados por supervisores humanos. *El aprendizaje no supervisado* , por otro lado, implica que un modelo nada en el océano de datos de entrada sin etiquetar, tratando de encontrarles sentido sin supervisión humana.

Clusterización

La agrupación en clústeres es el acto de organizar objetos similares en grupos dentro de un algoritmo de aprendizaje automático. La asignación de objetos relacionados en clústeres es beneficiosa para los modelos de IA. La agrupación en clústeres tiene muchos usos en la ciencia de datos, como el procesamiento de imágenes, el descubrimiento de conocimiento en datos, el aprendizaje no supervisado y varias otras aplicaciones. El análisis de conglomerados, o agrupamiento, se realiza escaneando los conjuntos de datos sin etiquetar en un modelo de aprendizaje automático y configurando medidas para características de puntos de datos específicos. El análisis de conglomerados luego clasificará y colocará los puntos de datos en un grupo con características coincidentes. Una vez que los datos se hayan agrupado, se les asignará un número de ID de grupo para ayudar a identificar las características del grupo. Desglosar conjuntos de datos grandes e intrincados en un modelo de aprendizaje automático utilizando la técnica de agrupación puede aliviar el estrés al descifrar datos complejos.

Ejemplos de agrupamiento

Instancias que se benefician del análisis de conglomerados de datos:

- Optimización de la planificación de la ciudad
- Personalización de conjuntos de entrenamiento para atletas profesionales
- Detección de amenazas de spam y actividad delictiva
- Identificando la desinformación
- Análisis de documentos
- Personalización de anuncios para los clientes
- Seguimiento del tráfico comercial en línea

Las capacidades de la IA que utiliza el análisis de conglomerados son amplias. Los grandes conjuntos de datos de aprendizaje automático se pueden compactar y numerar para simplificar el seguimiento de datos. Los ID de clúster pueden transformar puntos de datos minuciosos en herramientas de minería de datos que agilizan la predicción de tendencias de aprendizaje automático.

Data Compression

La compresión de datos (Data Compression) es una reducción en el número de bits (bit, abreviatura de dígito binario) necesarios para representar los datos. La compresión de datos puede ahorrar capacidad de almacenamiento, acelerar la transferencia de archivos y reducir los costos de hardware de almacenamiento y ancho de banda de la red.

Cómo funciona la compresión

La compresión la realiza un programa que utiliza una fórmula o algoritmo para determinar cómo reducir el tamaño de los datos. Por ejemplo, un algoritmo puede representar una cadena de bits, o 0 y 1, con una cadena más pequeña de 0 y 1 usando un diccionario para la conversión entre ellos, o la fórmula puede insertar una referencia o puntero a una cadena de 0s y 1s que el programa ya ha visto.

La compresión de texto puede ser tan simple como eliminar todos los caracteres innecesarios, insertar un solo carácter repetido para indicar una cadena de caracteres repetidos y sustituir una cadena de bits más pequeña por una cadena de bits que aparece con frecuencia. La compresión de datos puede reducir un archivo de texto al 50% o un porcentaje significativamente mayor de su tamaño original.

Para la transmisión de datos, la compresión se puede realizar en el contenido de los datos o en toda la unidad de transmisión, incluidos los datos de título. Cuando la información se envía o recibe a través de Internet, los archivos más grandes, ya sea individualmente o con otros como parte de un archivo, pueden transmitirse en formato ZIP, GZIP u otro formato comprimido.

¿Por qué es importante la compresión de datos?

La compresión de datos puede reducir drásticamente la cantidad de almacenamiento que ocupa un archivo. Por ejemplo, en una relación de compresión de 2:1, un archivo de 20 megabytes ocupa 10 MB de espacio. Como resultado de la compresión, los administradores gastan menos dinero y menos tiempo en almacenamiento.
La compresión será un método importante de reducción de datos a medida que los datos continúen creciendo exponencialmente.
Se puede comprimir prácticamente cualquier tipo de archivo, pero es importante seguir las mejores prácticas al elegir cuáles comprimir. Por ejemplo, es posible que algunos archivos ya vengan comprimidos, por lo que comprimir esos archivos no tendría un impacto significativo.

Métodos de compresión de datos: compresión sin pérdida y con pérdida

La compresión de datos puede ser un proceso con pérdida y sin pérdida de información. La compresión sin pérdida permite la restauración de un archivo a su estado original, sin la pérdida de un solo bit de datos, cuando el archivo está descomprimido. La compresión sin pérdida es el enfoque típico con ejecutables, así como archivos de texto y hojas de cálculo, donde la pérdida de palabras o números cambiaría la información.

La compresión con pérdida elimina permanentemente bits de datos que son redundantes, sin importancia o imperceptibles. La compresión con pérdida es útil con gráficos, audio, video e imágenes, donde la eliminación de algunos bits de datos tiene poco o ningún efecto perceptible en la representación del contenido.

Compresión frente a deduplicación de datos

La compresión a menudo se compara con la duplicación de datos, pero las dos técnicas funcionan de manera diferente. La deduplicación es un tipo de compresión que busca fragmentos de datos redundantes en un sistema de archivo o almacenamiento y luego reemplaza cada fragmento duplicado con un puntero al original. Los algoritmos de compresión de datos reducen el tamaño de las cadenas de bits en un flujo de datos que tiene un alcance mucho menor y generalmente no recuerda más que el último megabyte o menos de datos.

Ventajas y Desventajas de la compresión.

Las principales ventajas de la compresión son la reducción del hardware de almacenamiento, el tiempo de transmisión de datos y el ancho de banda de comunicación, y el consiguiente ahorro de costes. Un archivo comprimido requiere menos capacidad de almacenamiento que un archivo sin comprimir, y el uso de la compresión puede conducir a una disminución significativa en los gastos de disco. Un archivo comprimido también requiere menos tiempo para la transferencia y consume menos ancho de banda de red que un archivo sin comprimir.

La principal desventaja de la compresión de datos es el impacto en el rendimiento resultante del uso de recursos de CPU y memoria para comprimir los datos y realizar la descompresión. Muchos proveedores han diseñado sus sistemas para tratar de minimizar el impacto de los cálculos intensivos del procesador asociados con la compresión. Si la compresión se ejecuta en línea, antes de que los datos se escriban en el disco, el sistema puede descargar la compresión para preservar los recursos del sistema. Por ejemplo, IBM usa una tarjeta de aceleración de hardware separada para manejar la compresión con algunos de sus sistemas de almacenamiento empresarial.

Ejemplos en que se usa

- **Detección de anomalías.** Con el agrupamiento, es posible detectar cualquier tipo de valores atípicos en los datos. Por ejemplo, las empresas dedicadas al transporte y la logística pueden utilizar la detección de anomalías para identificar obstáculos logísticos o exponer piezas mecánicas defectuosas. Las organizaciones financieras pueden utilizar la técnica para detectar transacciones fraudulentas y reaccionar rápidamente, lo que en última instancia puede ahorrar mucho dinero.

- **Segmentación de clientes y mercados.** Los algoritmos de agrupamiento pueden ayudar a agrupar personas que tienen rasgos similares y crear personajes de clientes para campañas de marketing y focalización más eficientes.

- **Estudios clínicos de cáncer.** Los métodos de aprendizaje automático y agrupación se utilizan para estudiar los datos de expresión génica del cáncer y predecir el cáncer en etapas tempranas.

Cubriremos los casos de uso con más detalle un poco más adelante. Por ahora, comprendamos los aspectos esenciales del aprendizaje no supervisado comparándolo con su primo, el aprendizaje supervisado.

Referencias

[1] Aprendizaje de Máquina – Video

https://youtu.be/WyDGryVC8Iw

[2] Aprendizaje no supervisado

https://aprendeia.com/aprendizaje-no-supervisado-machine-learning/

Material Complementario

[1] ¿Qué es el Aprendizaje no supervisado?

https://www.youtube.com/watch?v=WyDGryVC8Iw

[2] ¿Qué es un cluster?

https://www.youtube.com/watch?v=yv2xwnvmbW4