1. Introducción a Machine Learning

1.1 Conceptos básicos y definiciones

1.1.1 ¿Qué es el machine learning?

El *machine learning* (aprendizaje automático) es una rama de la inteligencia artificial que se centra en el desarrollo de algoritmos y modelos que permiten a las computadoras aprender patrones y tomar decisiones sin intervención humana explícita. Se basa en la idea de que los sistemas pueden aprender y mejorar con la experiencia.

1.1.2 Multidisciplinar

El *machine learning* es multidisciplinario, involucrando conceptos de estadísticas, matemáticas, informática y domain knowledge específico del problema que se está abordando. Requiere una colaboración estrecha entre expertos en diferentes campos para lograr soluciones efectivas. Hoy en dia son muy pocos los ámbitos y campos del conocimiento donde no se pueda aplicar la inteligencia artificial y más específicamente el machine learning.

1.2 Tipos de Aprendizaje

1.2.1 Aprendizaje supervisado

En el aprendizaje supervisado, el modelo se entrena utilizando un conjunto de datos etiquetado, donde se conocen las respuestas correctas. El objetivo es que el modelo aprenda a mapear las entradas a las salidas correctas, y luego pueda hacer predicciones precisas sobre datos no etiquetados.

1.2.2 Aprendizaje no supervisado

Contrariamente, en el aprendizaje no supervisado, el modelo se enfrenta a datos no etiquetados y debe encontrar patrones y estructuras por sí mismo. Este tipo de aprendizaje es útil para descubrir relaciones inherentes en los datos y agrupar información de manera significativa.

1.2.3 Aprendizaje por refuerzo

En el aprendizaje por refuerzo, un agente aprende a tomar decisiones secuenciales al interactuar con un entorno. Recibe retroalimentación en forma de recompensas o penalizaciones, lo que guía al agente a mejorar su rendimiento a lo largo del tiempo.

1.3 Aplicaciones y ejemplos de machine learning

El *machine learning* se aplica en una variedad de campos, como reconocimiento facial, recomendación de productos, diagnóstico médico, conducción autónoma y mucho más. Estos ejemplos muestran cómo los algoritmos pueden adaptarse y mejorar su rendimiento con el tiempo. Cada día son innumerables los nuevos escenarios donde se intrduce machine learning, con aportes importantes.

1.4 Herramientas y lenguajes de programación para machine learning

Se utilizan diversas herramientas y lenguajes, como Python con bibliotecas como TensorFlow y scikit-learn, R, y plataformas especializadas como Azure Machine Learning y Google Colab, para implementar y entrenar modelos de *machine learning*.

1.5 Conjuntos de datos

Los conjuntos de datos son colecciones de datos que se utilizan para entrenar, validar y evaluar los modelos de machine learning. Los datos pueden provenir de diversas fuentes, como archivos, bases de datos, sensores, aplicaciones web, etc. Los datos pueden tener diferentes formatos, como texto, imágenes, audio, video, etc. Es importante conocer el origen, la estructura y el significado de los datos para poder aplicar las técnicas adecuadas de machine learning.

Antes de aplicar cualquier algoritmo de machine learning, es necesario explorar y analizar los datos para comprender sus características, distribuciones, relaciones, patrones y anomalías. Esto se puede hacer mediante técnicas estadísticas, gráficas y descriptivas, como calcular medidas de tendencia central, dispersión, correlación, visualizar histogramas, diagramas de caja, de dispersión, etc. La exploración y el análisis de los datos permiten identificar los problemas, las oportunidades y las limitaciones de los datos para el machine learning.

Los datos rara vez están listos para ser usados directamente en el machine learning, ya que suelen contener errores, inconsistencias, ruido, valores faltantes, duplicados, etc. Por eso, es necesario preprocesar y limpiar los datos para mejorar su calidad y adecuarlos a los requisitos de los algoritmos de machine learning.

Una vez que los datos están preprocesados y limpios, es necesario dividirlos en conjuntos de entrenamiento, validación y prueba, que se usan para diferentes propósitos en el machine learning. El conjunto de entrenamiento se usa para ajustar los parámetros del modelo de machine learning, el conjunto de validación se usa para seleccionar el mejor modelo entre varios candidatos, y el conjunto de prueba se usa para evaluar el rendimiento del modelo final en datos nuevos y no vistos. La división de los datos en estos conjuntos se puede hacer de forma aleatoria, estratificada o secuencial, dependiendo del tipo y la naturaleza de los datos. La proporción de los datos que se asignan a cada conjunto puede variar según el tamaño y la complejidad de los datos.

1.5.1 Fuentes y formatos de datos

Los conjuntos de datos pueden provenir de diversas fuentes, como bases de datos, sensores, archivos CSV, entre otros. Los formatos pueden variar, incluyendo datos tabulares, de texto, imágenes o secuencias temporales. Lo mas importante es conocer y estar seguro del tipo de archivo que contiene los diferentes ejemplos, para elegir la herramienta o librería adecuada, garantizando incluso la internacionalización.

1.5.2 Exploración y análisis de datos

Antes de construir modelos, es crucial explorar y analizar los datos para comprender su distribución, identificar patrones, verificar la integridad y valides de los mismo, pues no todos los datos llegan en formato numerico, por lo que es importante conocer todos los pormenores para asumir las acciones adecuadas.

1.5.3 Preprocesamiento y limpieza de datos

El preprocesamiento implica la limpieza de datos, tratamiento de valores atípicos y normalización para mejorar la calidad y relevancia de los datos utilizados en el entrenamiento.

1.5.4 Manejo de datos faltantes, datos inconsistentes o datos cualitativos

Los datos faltantes son aquellos que no tienen un valor asignado en el conjunto de datos. Pueden deberse a errores de medición, de transmisión, de almacenamiento, etc. Los datos faltantes pueden afectar al rendimiento y la precisión de los modelos de machine learning, por lo que es necesario tratarlos adecuadamente. Algunas de las estrategias para manejar los datos faltantes son: eliminar las filas o columnas que contienen datos faltantes, imputar los datos faltantes con valores estadísticos, como la media, la mediana o la moda, o con valores basados en otros atributos, como la regresión o el k-vecinos más cercanos, o ignorar los datos faltantes si el algoritmo de machine learning lo permite. El manejo de datos faltantes implica estrategias como la imputación, eliminación o sustitución para abordar la ausencia de información en ciertos puntos del conjunto de datos.

Ante la ausencia o inconsistencia de datos en los conjuntos de entrenamiento en machine learning, existen diversas estrategias para abordar este problema. Aquí hay algunas estrategias comunes:

1. Eliminación de datos faltantes: Eliminar filas o columnas: Puedes optar por eliminar las filas o columnas que contienen datos faltantes. Sin embargo, esta estrategia puede llevar a la pérdida de información valiosa.
2. Imputación de datos: Media o mediana: Puedes sustituir los valores faltantes con la media o la mediana de la variable correspondiente.
3. Valor constante: Reemplazar los valores faltantes con un valor constante.
4. Imputación avanzada: Utilizar técnicas más avanzadas como la imputación por regresión, K-Vecinos más Cercanos (K-NN), o métodos basados en modelos.
5. Manejo de variables categóricas: Eliminación de categorías, en el caso de variables categóricas, puedes eliminar categorías con datos faltantes o agruparlas en una categoría "desconocida".
6. Imputación específica para categorías: Aplicar técnicas específicas para imputar valores en variables categóricas.
7. Escalado de características: Normalización o estandarización: Asegurarte de que las variables tengan escalas comparables para evitar sesgos en los modelos.
8. Uso de modelos de machine learning para imputación: Modelos predictivos, puedes utilizar modelos predictivos para estimar los valores faltantes. Esto implica entrenar un modelo en el conjunto de datos sin datos faltantes y utilizarlo para predecir los valores que faltan.
9. Manejo especializado de series temporales: Interpolación temporal, en el caso de datos temporales, puedes utilizar técnicas de interpolación para estimar valores faltantes basándote en los valores circundantes en el tiempo.

Es importante elegir la estrategia de manejo de datos faltantes según el contexto específico de tu conjunto de datos y el problema que estás abordando. Además, es fundamental evaluar el impacto de estas decisiones en el rendimiento del modelo.

1.5.5 Codificación de variables categóricas

Las variables categóricas son aquellas que tienen un número limitado de valores posibles, que representan categorías o clases, como el género, el color, el país, etc. Los algoritmos de machine learning suelen requerir que los datos sean numéricos, por lo que es necesario codificar las variables categóricas en números. Algunas de las técnicas de codificación de variables categóricas son: asignar un número entero a cada categoría, como 1, 2, 3, etc., usar una codificación binaria, como 0 y 1, para

cada categoría, o usar una codificación de variables ficticias o dummy, que consiste en crear una columna por cada categoría, con valor 1 si la fila pertenece a esa categoría y 0 si no.

Las variables categóricas se codifican numéricamente para que los algoritmos puedan interpretarlas adecuadamente durante el entrenamiento del modelo. La transformación de variables cualitativas en variables cuantitativas, permite que los datos registrados como texto, deben ser categorizados como números que representen adecuadamente una característica respectiva que se encontraba como texto, en un valor numerico.

1.5.6 Escalado de características

El escalado de características consiste en transformar los valores de las características o atributos de los datos para que tengan un rango o una escala común, como [0, 1] o [-1, 1]. Esto se hace para evitar que las características con valores más altos o más variados dominen o sesguen los resultados de los algoritmos de machine learning, especialmente los que se basan en medidas de distancia, como el k-means o el k-vecinos más cercanos. Algunas de las técnicas de escalado de características son: normalización, que consiste en restar el valor mínimo y dividir por el rango de cada característica, estandarización, que consiste en restar la media y dividir por la desviación estándar de cada característica, o escalado robusto, que consiste en restar el percentil 25 y dividir por el rango intercuartílico de cada característica.

El escalado de características asegura que todas las variables tengan un impacto equitativo en el modelo, evitando que aquellas con magnitudes mayores dominen el proceso de entrenamiento. Lo que se busca es que todos los valores de las características se encuentren en una sola escala para que el modelo pueda converger con mayor rapidez.

1.5.7 División de datos en conjuntos de entrenamiento, validación y prueba

Los datos se dividen en conjuntos de entrenamiento, validación y prueba para entrenar, ajustar hiperparámetros y evaluar el rendimiento del modelo, respectivamente. Esto garantiza la generalización y evaluación adecuada del modelo. Existen dos posibilidades: Una que propone dividir el total de datos en entrenamiento y prueba y otra que considera la división en datos de entrenamiento, validación y prueba, en ambos casos se recomienda asumir una de las siguiente divisiones: 70% para entrenamiento y 30% para prueba, 80% para entrenamiento y 20% para prueba o 90% para entrenamiento y 10% para prueba o porcentaje cercanos a estos, esto dependerán de los datos, el modelo y sobre todo en los resultados que se generen adicionalmente de un análisis y prueba de combinación de varias alternativas para los hiperparametros.