Aca debe ir un titulo

Pichetti Augusto
Salse, Lucas Antonio

April 15, 2021

1 Introduction

Hoy en día como sabemos existe una importancia extremadamente alta sobre los datos que se recolectan en el día a día, ya sean datos obtenidos de carácter científico (mediciones climáticas, radiación, satelitales, médicos, etc.) o bien datos personales de cada individuo (gastos, ubicación, compras, pasatiempo, etc.). Todo esto generó que existiera la necesidad de poder emplear dichos datos y obtener un provecho de los mismos, con el objetivo de solucionar infinidades de problemas de carácter general (clima, enfermedades, etc.) como así también para beneficios económicos de las empresas (publicidad específica, segmentación, etc.). Es por esta razón para poder conseguir estos objetivos se emplea el concepto de inteligencia artificial.

1.1 Inteligencia Artificial

- La Inteligencia Artificial (Artificial Intelligence) se define como el estudio de los "agentes inteligentes", i.e. cualquier dispositivo que perciba su entorno y tome medidas que maximicen sus posibilidades de lograr con éxito sus objetivos. Poole et al. [1]

- La capacidad de un sistema para interpretar correctamente datos externos, para aprender de dichos datos y emplear esos conocimientos para lograr tareas y metas concretas a través de la adaptación flexible. Kaplan y Michael Haenlein. [2]

Esta definición nos da la idea de que la IA es un sistema reactivo, que reacciona a cambios externos y actúa en consecuencia.

1.2 Aprendizaje Automático

Una de las subramas de la inteligencia artificial es el aprendizaje automático que a su vez el subcampo de la ciencia de la computación, como se ve en la siguiente imagen:

El aprendizaje automático (Machine Learning) es el estudio científico de algoritmos y modelos estadísticos que los sistemas informáticos utilizan para realizar una tarea específica sin utilizar instrucciones explícitas, sino que se basan en patrones e inferencias. Es visto como un subcampo de inteligencia artificial. Los algoritmos de aprendizaje automático crean un modelo matemático basado en datos de muestra, conocidos como "datos de entrenamiento", para hacer predicciones o decisiones sin ser programado explícitamente para realizar la tarea. Bishop. [3]

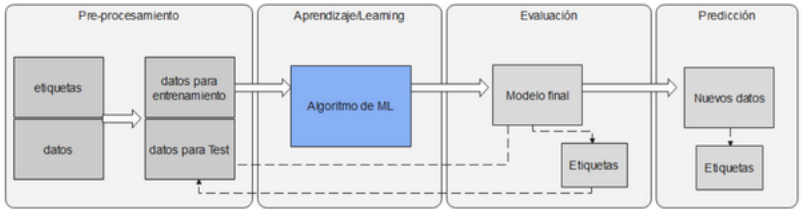

Figure 1: Diagrama de flujo de una aplicación de Machine Learning.

Por tanto el Aprendizaje Automático es la generación de un modelo de predicción de salida a partir de grandes cantidades de datos de entrada, realizando un tratamiento de los mismos a través de diferentes etapas bien definidas, como se pueden apreciar en la Fig. 1, las cuales iremos desarrollando en diferentes secciones.

Es importante destacar la independencia del aprendizaje automático al momento de tomar decisiones a partir de los datos proporcionados sin intervención externa, es decir que no hay una especificación de reglas que dictan cómo deben ser tomadas estas decisiones. A su vez, los modelos obtenidos a partir de los algoritmos de Machine Learning deben tener la capacidad de predecir a partir de nuevos datos, nunca antes procesados por el modelo, a esto se lo conoce como generalización.

1.3 Preprocesamiento

A menudo, los resultados van a depender más de la calidad de los datos en relación al problema, que de la parte de aprendizaje automático.

Como vemos la importancia de un buen manejo de datos es vital para poder conseguir los objetivos propuestos es por esto que debemos ser capaces de realizar un preprocesamiento de los mismos.

- Los datos del mundo real están "sucios": incompletos (datos perdidos, ausencia de datos de alguna tipología), ruidosos (con errores o outliers, variables redundantes), inconsistentes (diferencias en codificación o nombres).
- Datos de calidad → Resultados de calidad.

- Es necesario analizar las características de los datos para conocer mejor el problema, detectar posibles datos erróneos, dependencias, outliers, valores perdidos, etc.
- Deberemos ser capaces de utilizar ciertas técnicas que nos permitan analizar y visualizar que los datos estén dentro de todo correctos.

Tareas típicas de esta etapa:

- Selección de datos.
  - Extracción de características.
  - Selección de características descartables para reducir el número de variables.
- Limpieza de datos:
  - Recuperación de valores perdidos (imputación de datos).
  - Tratamiento de valores anómalos (outliers).
  - Suavizar el ruido.
  - Eliminar inconsistencias.
- Transformación de datos. Por ej, convertir una variable nominal en varias variables binarias, escalado de datos, fecha de nacimiento -¿ edad.
- Reducción de dimensionalidad. Consiste en emplear técnicas como PCA (análisis de componentes principales) para obtener combinaciones lineales de variables que reduzcan la dimensión de los datos.

Por lo tanto, una vez que se obtiene el dataset de entrada, es primordial realizar estas tareas de forma que presentemos un conjunto de datos que esté en condiciones de ser entrenado y luego el modelo resultante, al momento de ser probado con datos desconocidos, tenga un desempeño óptimo. Además, de ser necesario, se pueden utilizar algoritmos de Aprendizaje no supervisado los cuales resultan muy útiles cuando se cuentan con grandes cantidades de datos en contextos no conocidos.